



AthNLP 2024



**Barcelona  
Supercomputing  
Center**

Centro Nacional de Supercomputación

# Health NLP

Dr. Martin Krallinger  
Head of NLP for Biomedical Information Analysis (NLP4BIA),  
Barcelona Supercomputing Center (BSC)

[<mkralin@bsc.es>](mailto:<mkralin@bsc.es>)

AthNLP 2024, NCSR Demokritos Campus, Athens, Greece (19 - 25 September 2024)

# Talk outline

- A. Introduction & background
- B. Health language models
- C. Clinical NLP components, use cases and applications
- D. Shared tasks & evaluation of health NLP systems
- E. Example projects involving Health NLP
- F. Conclusions

# Introduction & background

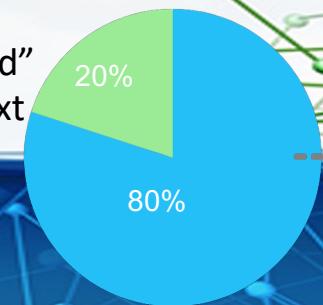


*Barcelona  
Supercomputing  
Center*  
Centro Nacional de Supercomputación

## Unstructured Clinical Data:

- **Images**
- **Natural language texts (EHRs)**

Around 80% of health data “locked” in unstructured text



### *Information source for:*

- Clinical decision support
- Patient stratification & selection
- Disease/adverse drug event surveillance
- Health management
- Predictive/modelling
- Many others,...

## Unstructured text: Clinical narrative

Transforming clinical text written by healthcare professionals into structured clinical data representations

## Structured Clinical data:

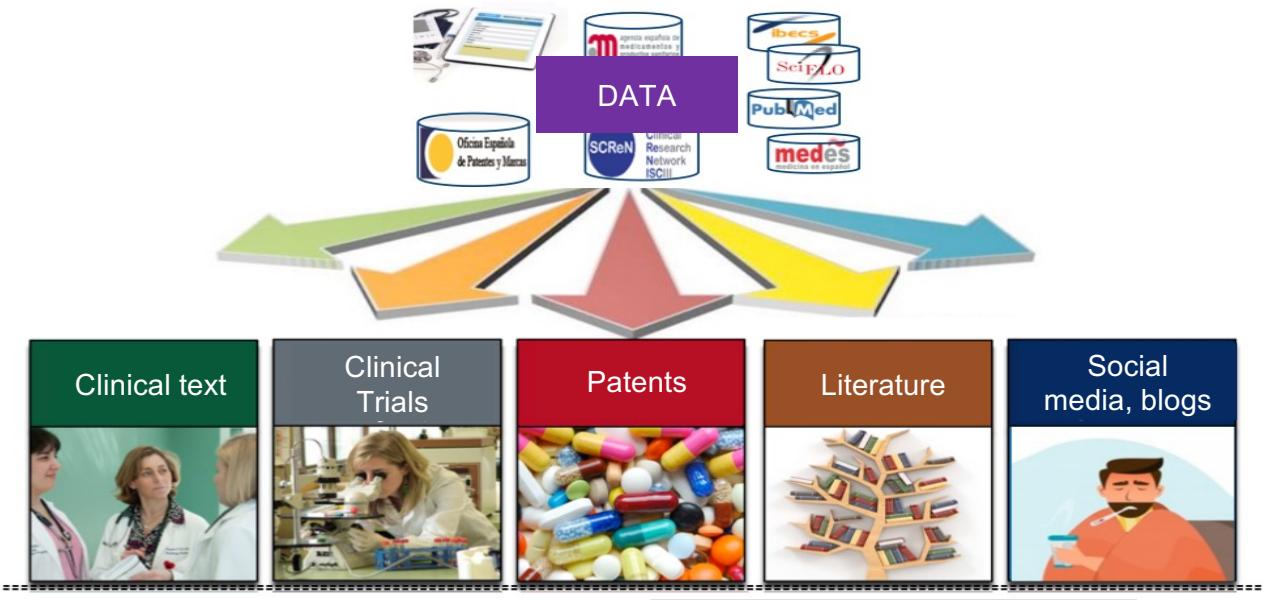
- **Clinical coding (ICD10)**
- **Lab tests/results**
- **Genomic/sequencing data**
- **Basic sociodemographic patient characteristics**

**Most of published studies that use EHR data still use ONLY structured data**

### Example - Galician health system:

- **200.000 clinical notes per day (population 2,7 million)**

# Diversity of health textual data content



Data sources include:

- ✓ Scientific/biomedical literature
- ✓ Clinical records
- ✓ Clinical trials
- ✓ Health web content
- ✓ Social media
- ✓ Patient forum
- ✓ *Patents*
- ✓ *Thesis & books*
- ✓ *Drug leaflets*
- ✓ Medical surveys & questionnaires,....

*NLP of diverse sources of data could provide a more **comprehensive** view of patient/population health*



# Challenges of clinical NLP approaches

- Language used to communicate & document results of observations & assessments during the care episode
- Difficulties & issues:
  - Proliferation of synonymy & polysemy
  - Use of **neologisms**
  - **Telegraphic** language, **abbreviations**, acronyms & apocopes (derma instead of dermatology)
  - **Localisms** & lexical language variants (nations, regions, areas,...)
  - **Errors:** grammatical, typographical or style, lack/errors in accentuation/spelling/ punctuation marks, sentences without verbs, etc.
  - Importance of: **negation**, speculation (e.g. does not show symptom)

**High variability depending on document type & specialities: No out of the box ?**

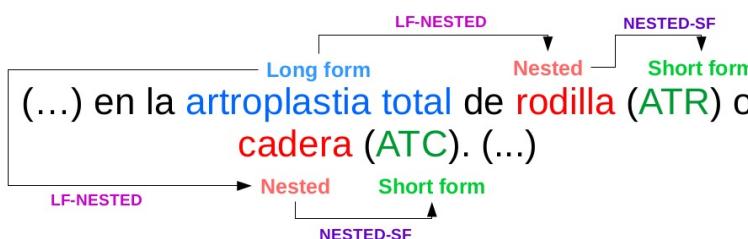
**NLP solution will usually work well enough -> need adaptation/fine tuning ?**

# Telegraphic language & abbreviations

Mujer de 84 años sin **ACM**. Niega hábitos tóxicos. Parcialmente dependiente **ABVD**. Vive en residencia. Antecedentes de **HTA**, **DLF** y **FA** antiagregada. Ictus **POCI ACP** izquierda en 2008, etiología cardioembólica.



Mujer de 84 años sin **alergias medicamentosas conocidas**. Niega hábitos tóxicos. Parcialmente dependiente **adriamycin bleomycin vinblastine and dacarbazine**. Vive en residencia. Antecedentes de **hipertensión arterial**, **depresión a largo plazo** y **fibrilación auricular** antiagregada. **Ictus circulación posterior, arteria cerebral posterior** izquierda en 2008, etiología cardioembólica.



## Barriers for medical NLP (beyond English)

- Lack of access to shared data
- Lack of annotated datasets for training and benchmarking
- Insufficient common conventions and standards for annotations
- The formidability of reproducibility
- Limited collaboration
- Lack of user-centered development and scalability

Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions 

Wendy W Chapman, Prakash M Nadkarni, Lynette Hirschman, Leonard W D'Avolio, Guergana K Savova, Ozlem Uzuner

*Journal of the American Medical Informatics Association*, Volume 18, Issue 5, September 2011, Pages 540–543, <https://doi.org/10.1136/amiajnl-2011-000465>

## Potential solutions

- **Shared tasks**—a partial solution for progress
- Thinking creatively to foster:
  - Reproducibility of results
  - Collaboration
  - User-centered design
  - Scalability and tackling real problems





# Biomedical NLP

## Current trends, needs & possibilities



### High quality annotated data sets

(MIMIC IV, GENIA, CHEMDNER, SPACCC, Merlot,...)

Multilingual, human in the loop, synthetic



### Software, platforms, open code, libraries

(HuggingFace, Spacy, CogStack, cTakes, GATE,...)

Model sharing, end-to-end solutions



### Shared tasks: benchmark, reproducible

(i2b2/n2c2, CLEF, TREC, BioCreative, BioNLP-ST, IberEval, EVALITA, SEMEVAL,...)

Increased diversity, multilingual



### Need of Computing infrastructure

Continuous access to GPUs for model training for:



### Language models

(LLMs, PLMs, Transformers, multilingual, domain-specific)



### Legal scenario

(Data & model licencing, privacy preservation & anonymization, legal data usage agreement, ethics)

Synthetic data, federated approach



### Annotation guidelines

(interpretation, extension, adaptation, quality, reusable, multilingual reproduceable data annotations)

# Barries and solutions to promote clinical NLPO resource development

## Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions FREE

Wendy W Chapman , Prakash M Nadkarni, Lynette Hirschman, Leonard W D'Avolio, Guergana K Savova, Ozlem Uzuner

*Journal of the American Medical Informatics Association*, Volume 18, Issue 5, September 2011, Pages 540–543, <https://doi.org/10.1136/amiajnl-2011-000465>

Published: 01 September 2011 Article history ▾

Névéol et al. *Journal of Biomedical Semantics* (2018) 9:12  
<https://doi.org/10.1186/s13326-018-0179-8>

Journal of  
Biomedical Semantics

REVIEW

Open Access



## Clinical Natural Language Processing in languages other than English: opportunities and challenges

Aurélie Névéol<sup>1</sup> , Hercules Dalianis<sup>2</sup>, Sumithra Velupillai<sup>3,4</sup>, Guergana Savova<sup>5</sup> and Pierre Zweigenbaum<sup>1</sup>

Shared tasks & evaluations

Sustainability of resources

Legal aspects

Interoperability

Research

Open Access

## Text mining for biology - the way forward: opinions from leading scientists

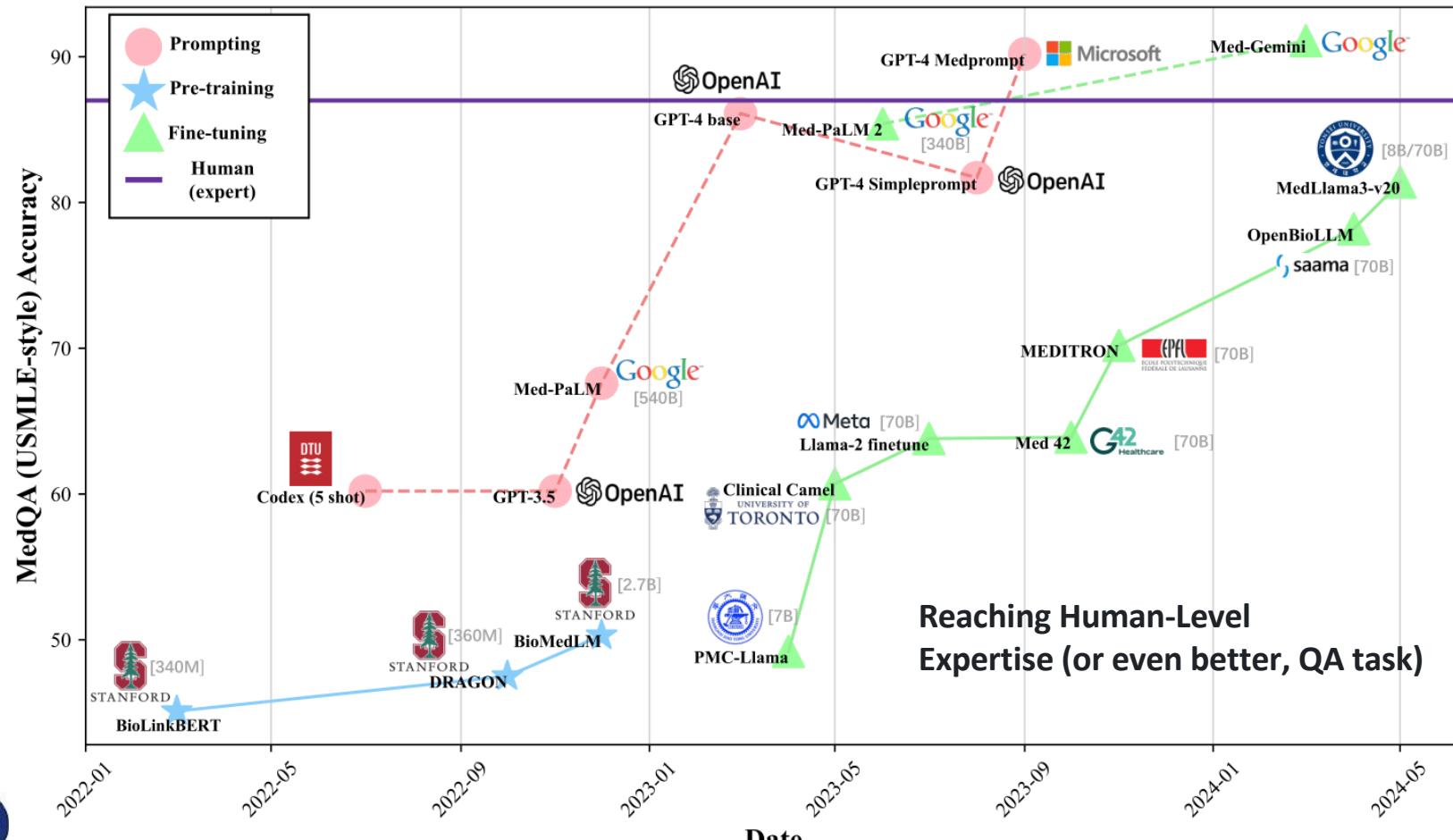
Russ B Altman<sup>1</sup>, Casey M Bergman<sup>2</sup>, Judith Blake<sup>3</sup>, Christian Blaschke<sup>4</sup>, Aaron Cohen<sup>5</sup>, Frank Gannon<sup>6</sup>, Les Grivell<sup>7</sup>, Udo Hahn<sup>8</sup>, William Hersh<sup>5</sup>, Lynette Hirschman<sup>9</sup>, Lars Juhl Jensen<sup>10,11</sup>, Martin Krallinger<sup>12</sup>, Barend Mons<sup>13</sup>, Seán I O'Donoghue<sup>10</sup>, Manuel C Peitsch<sup>14</sup>, Dietrich Rebholz-Schuhmann<sup>15</sup>, Hagit Shatkay<sup>16</sup> and Alfonso Valencia<sup>12</sup>

# Health Language Models

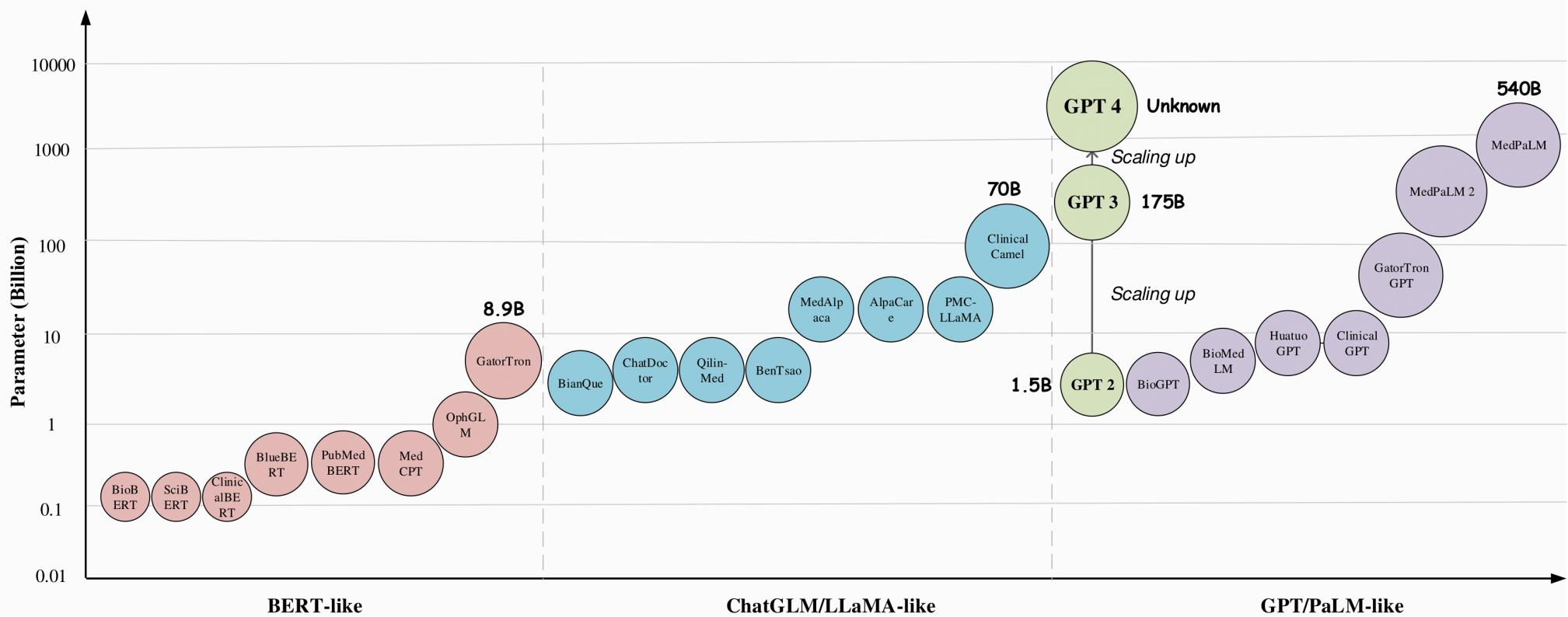


*Barcelona  
Supercomputing  
Center*  
Centro Nacional de Supercomputación

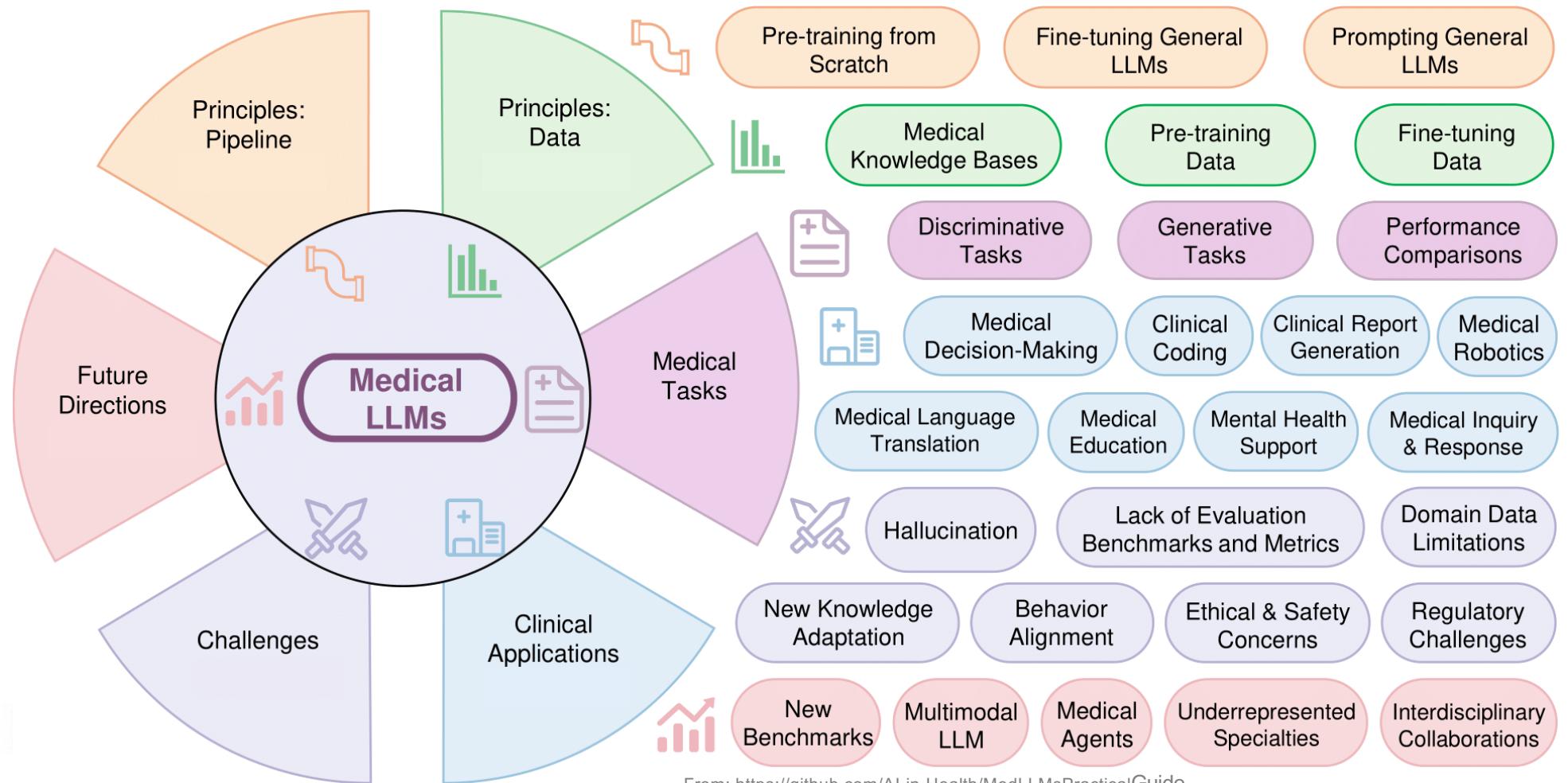
# Aims of medical Language Models



# Model size - Parameters

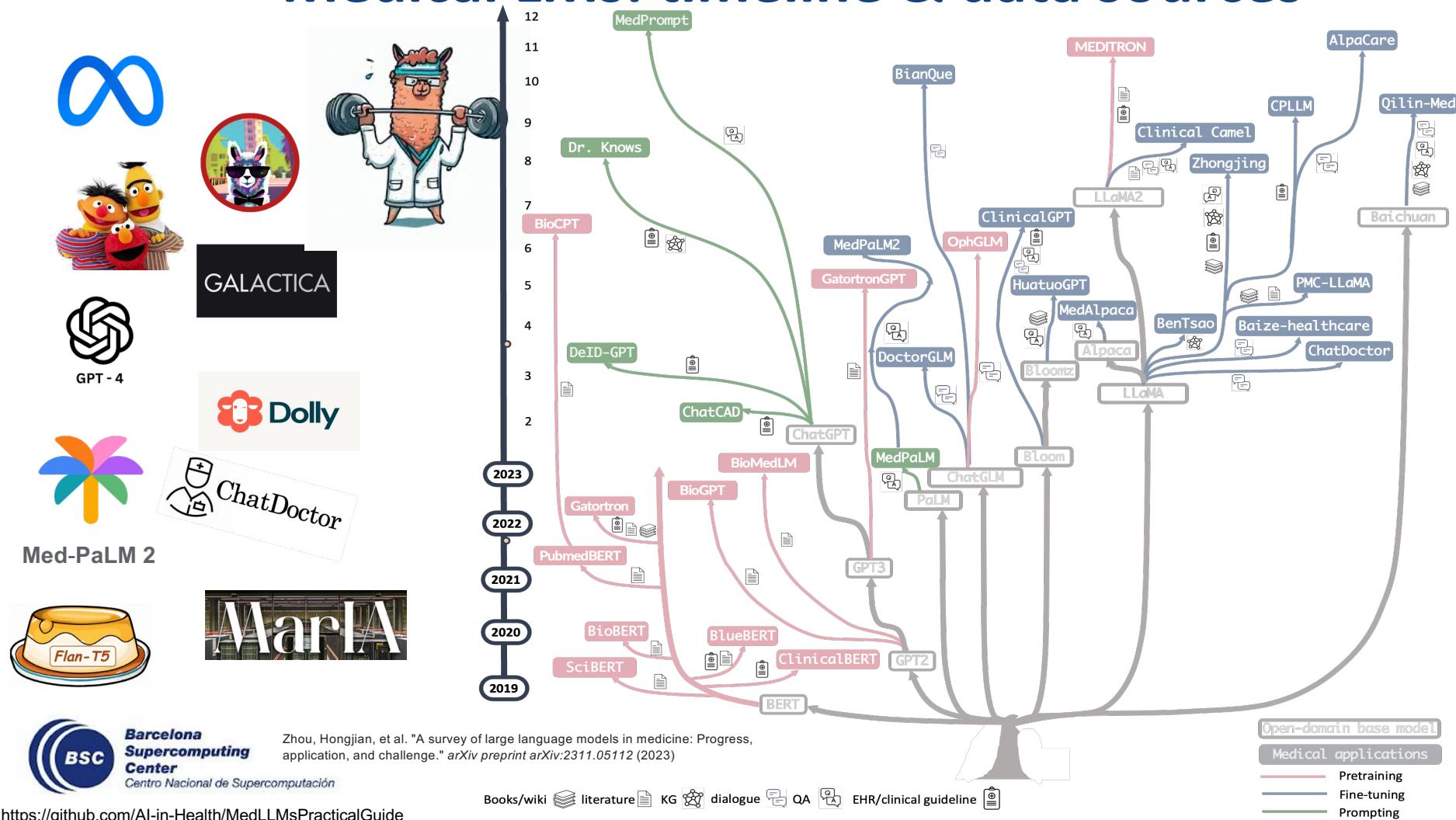


# Medical Language Models: main aspects

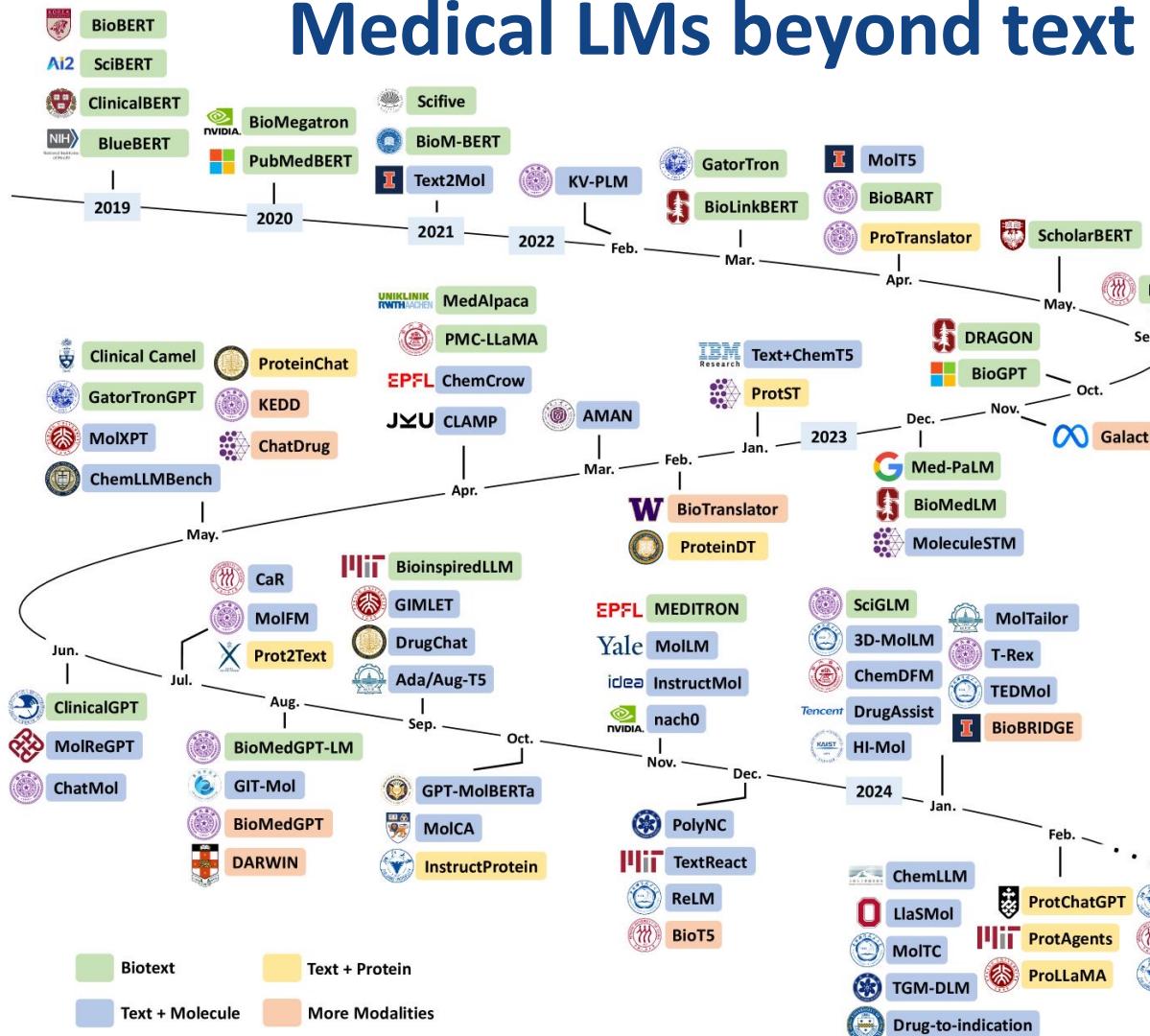


From: <https://github.com/AI-in-Health/MedLLMsPracticalGuide>

# Medical LMs: timeline & data sources



# Medical LMs beyond text

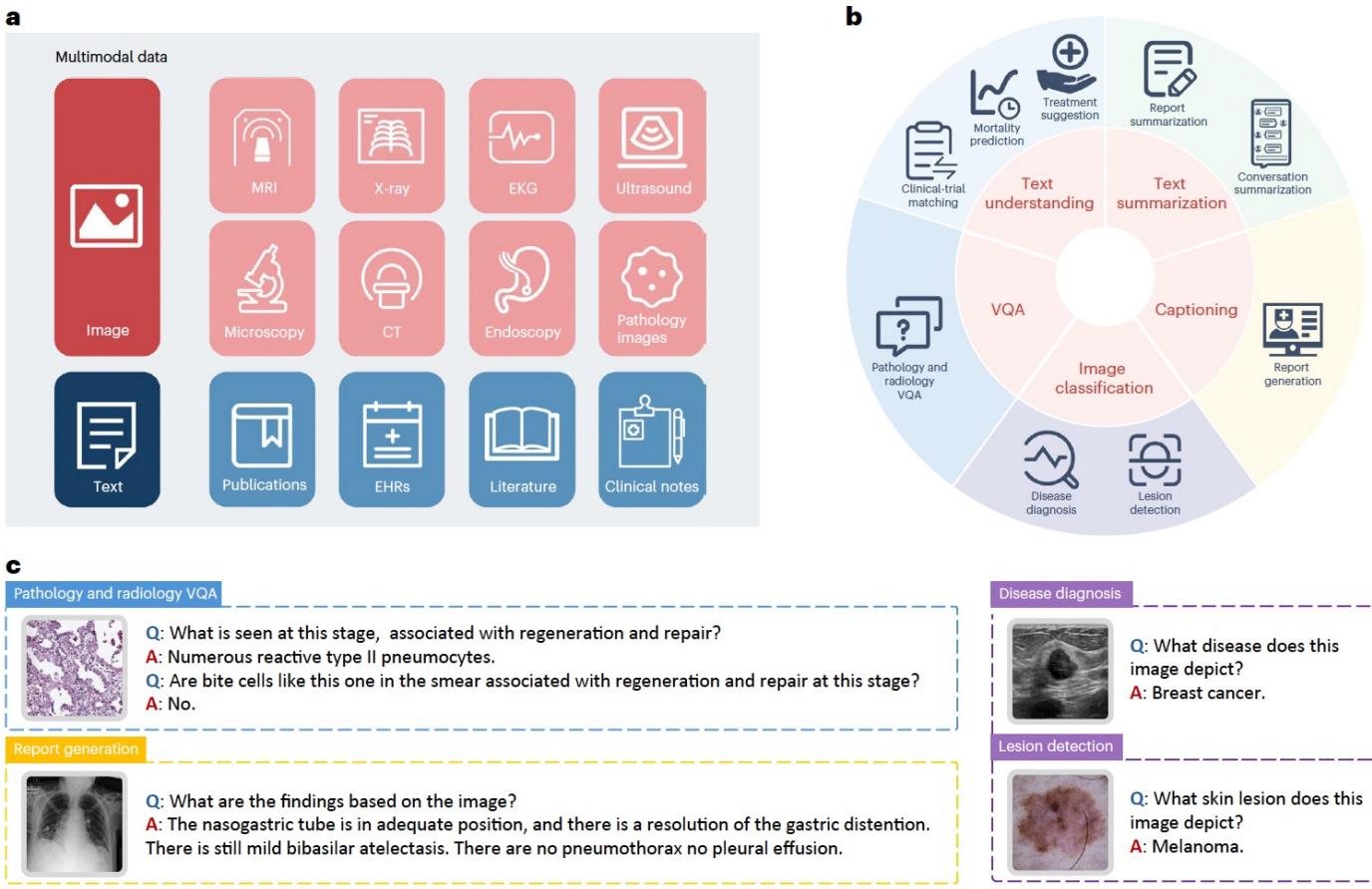


Pei, Qizhi, et al. "Leveraging Biomolecule and Natural Language through Multi-Modal Learning: A Survey." *arXiv preprint arXiv:2403.01528* (2024).

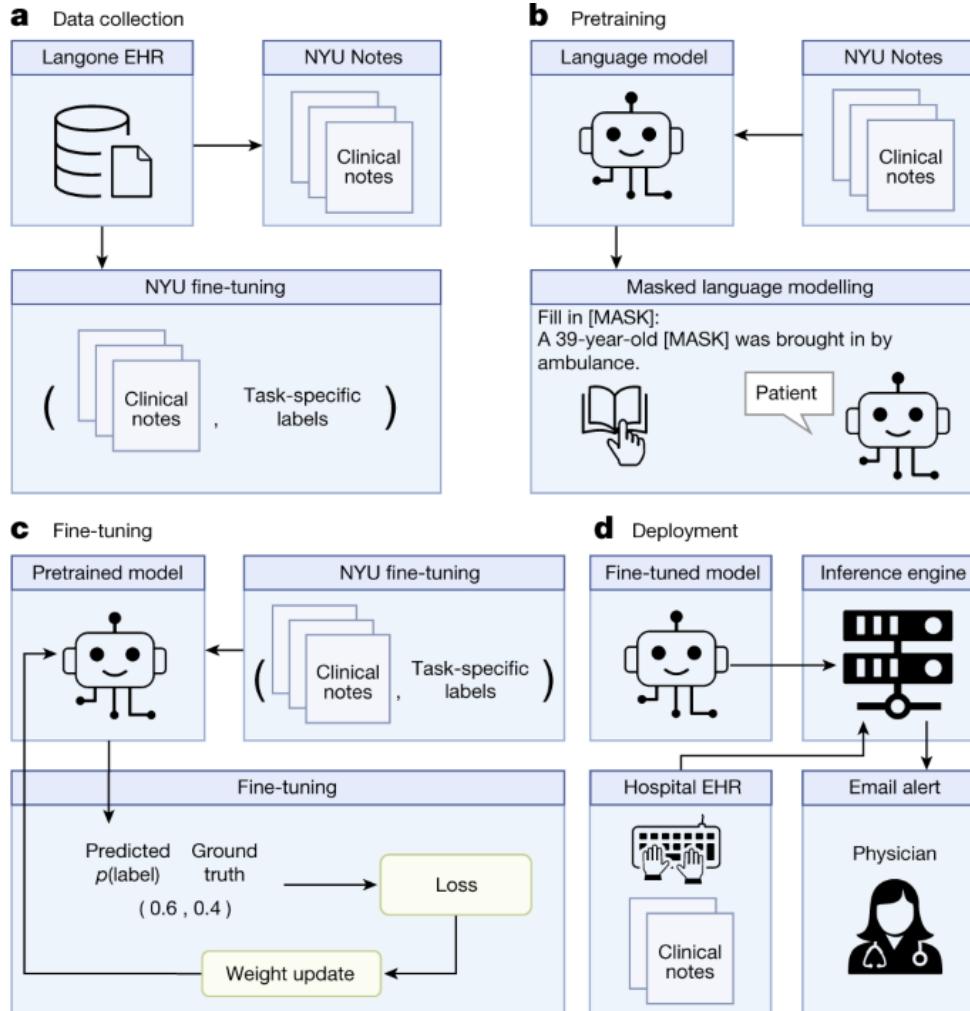
# Pre-training from Scratch

- **BiomedGPT**: A generalist vision–language foundation model, SOTA in 16 out of 25 tasks (Zhang et al. Nature Medicine, 2024)
- **NYUTron** Health system-scale language models are all-purpose prediction engines (Jiang et al. Nature, 2023)
- **GatorTronGPT**: A Study of Generative Large Language Model for Medical Research and Healthcare (Peng et al. Digital Medicine, 2023)
- **MedCPT**: Contrastive Pre-trained Transformers with Large-scale PubMed Search Logs for Zero-shot Biomedical Information Retrieval (Jin et al. Bioinformatics, 2023)
- **BioGPT**: Generative Pre-trained Transformer for Biomedical Text Generation and Mining (Luo et al. Bioinformatics, 2022)
- **DRAGON**: Deep Bidirectional Language-Knowledge Graph Pretraining (Yasunaga et al. NeurIPS, 2022)
- **BioLinkBERT/LinkBERT**: Pretraining Language Models with Document Links (Yasunaga et al. ACL, 2022)
- **GatorTron**: A Large Language Model for Electronic Health Records (Yang et al. Digital Medicine, 2022)
- **PubMedBERT**: Domain-specific Language Model Pretraining for Biomedical Natural Language Processing (Gu et al. ACM HEALTH 2021)
- **BioBERT**: A Pre-trained Biomedical Language Representation Model for Biomedical Text Mining (Lee et al. Bioinformatics, 2020)
- **SciBERT**: A Pretrained Language Model for Scientific Text (Beltagy et al. ENNLP, 2019)
- **ClinicalBERT**: Publicly Available Clinical BERT Embeddings (Alsentzer et al. NAACL Workshop, 2019)
- **BlueBERT**: Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets (Peng et al. BioNLP Workshop, 2019)

# BioMedGPT



# NYUTron



**a**

## In-hospital mortality prediction

How likely is the patient to die in the hospital before discharge?



## Binned comorbidity index imputation

Without structured ICDS, how sick/chronically ill is the patient?



## 30-day all-cause readmission prediction

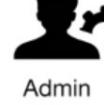
How likely is the patient to come back within 30 days of discharge?

**Operational task**



## Binned LOS prediction

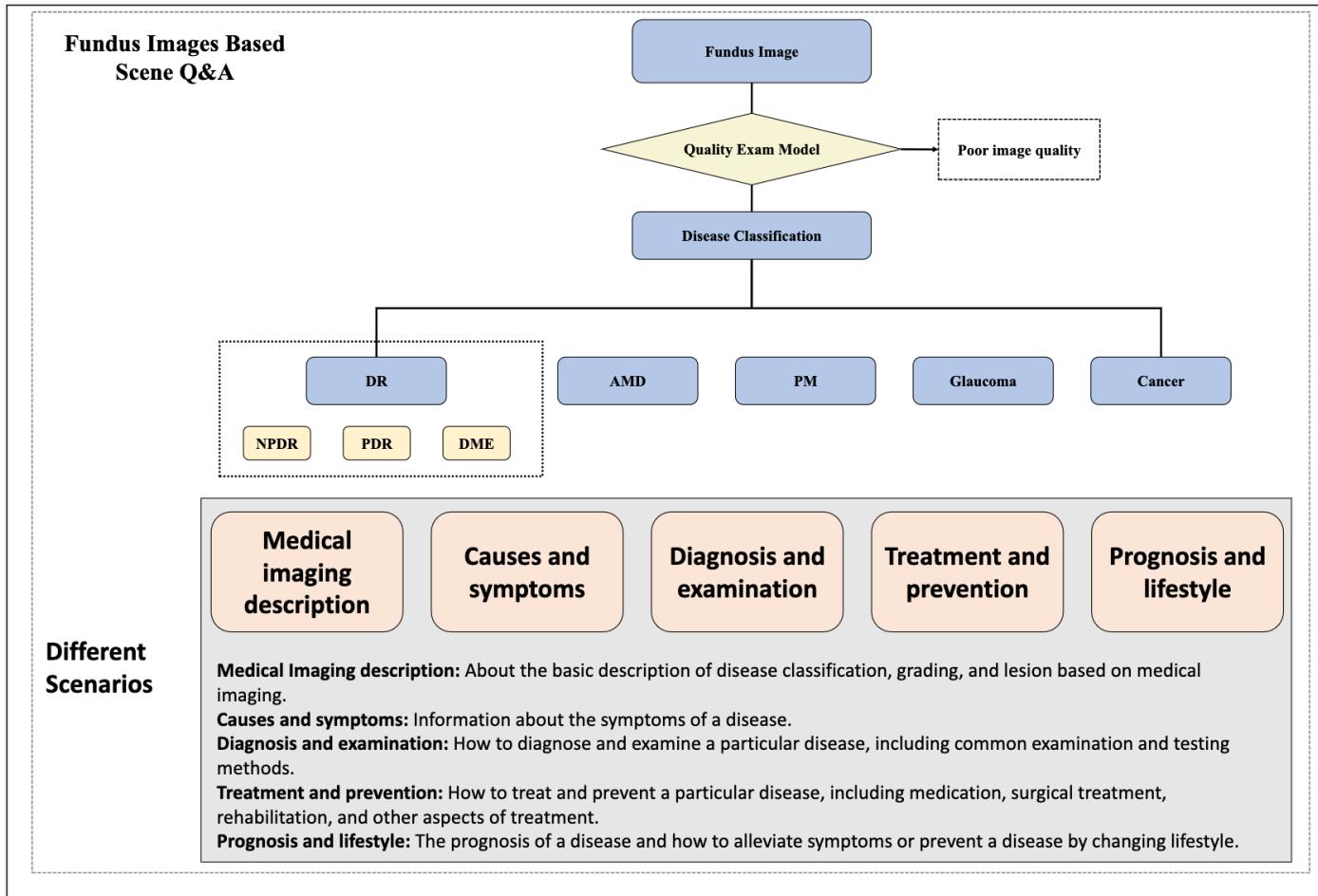
How long will the patient stay in the hospital?



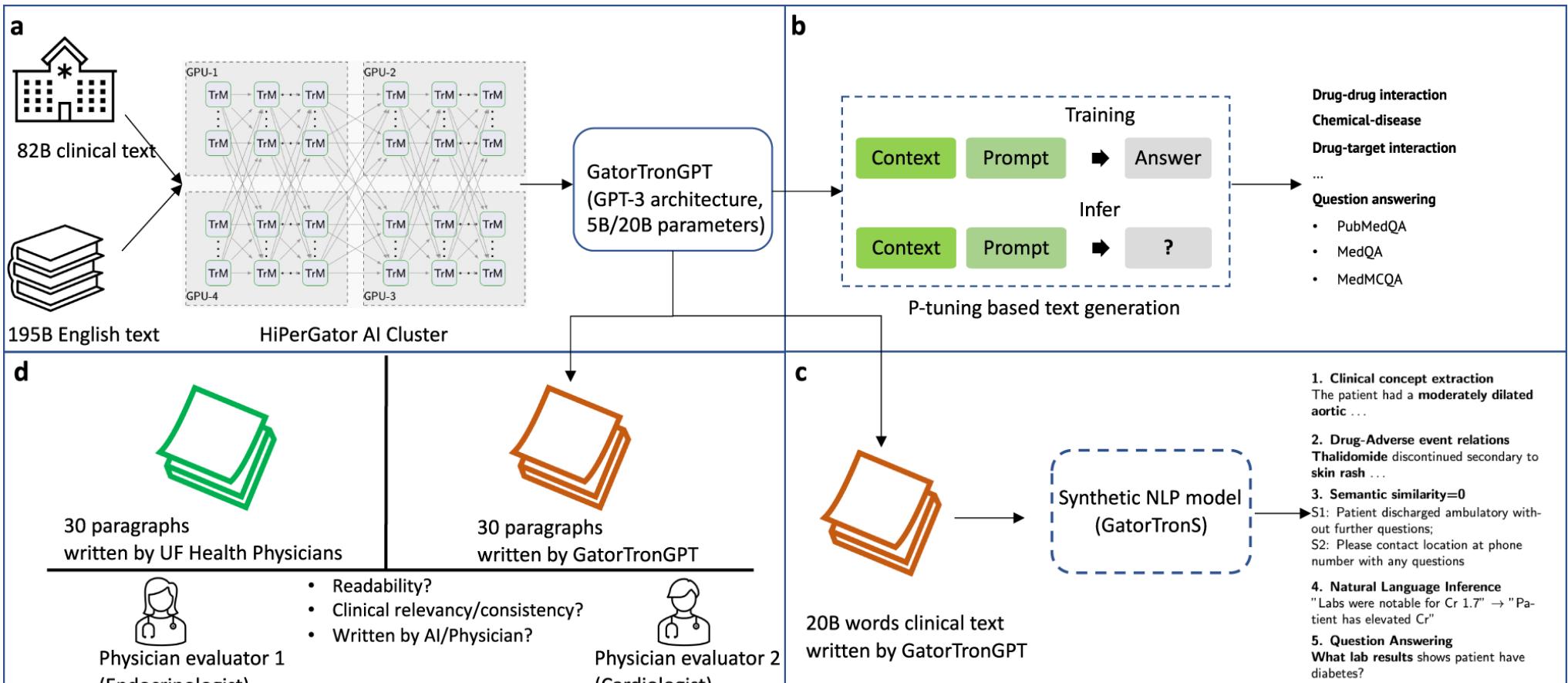
## Insurance denial prediction

How likely is the patient's insurance claim to be denied?

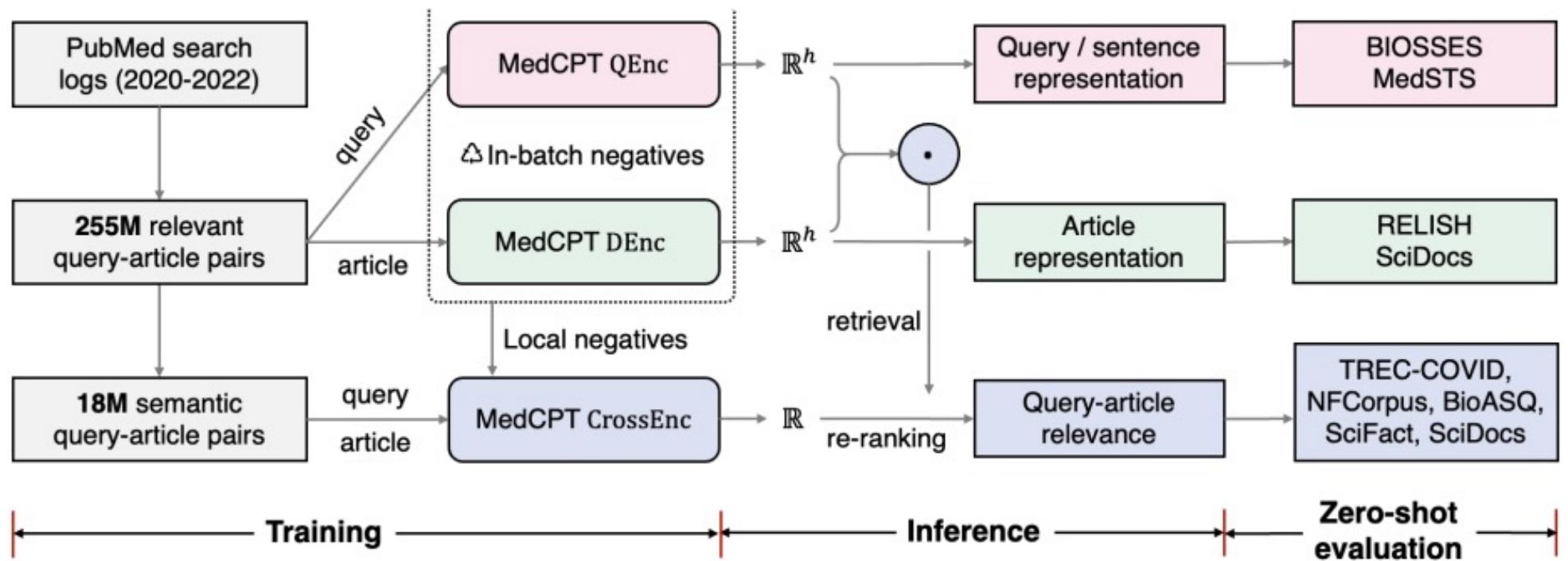
# OphGLM



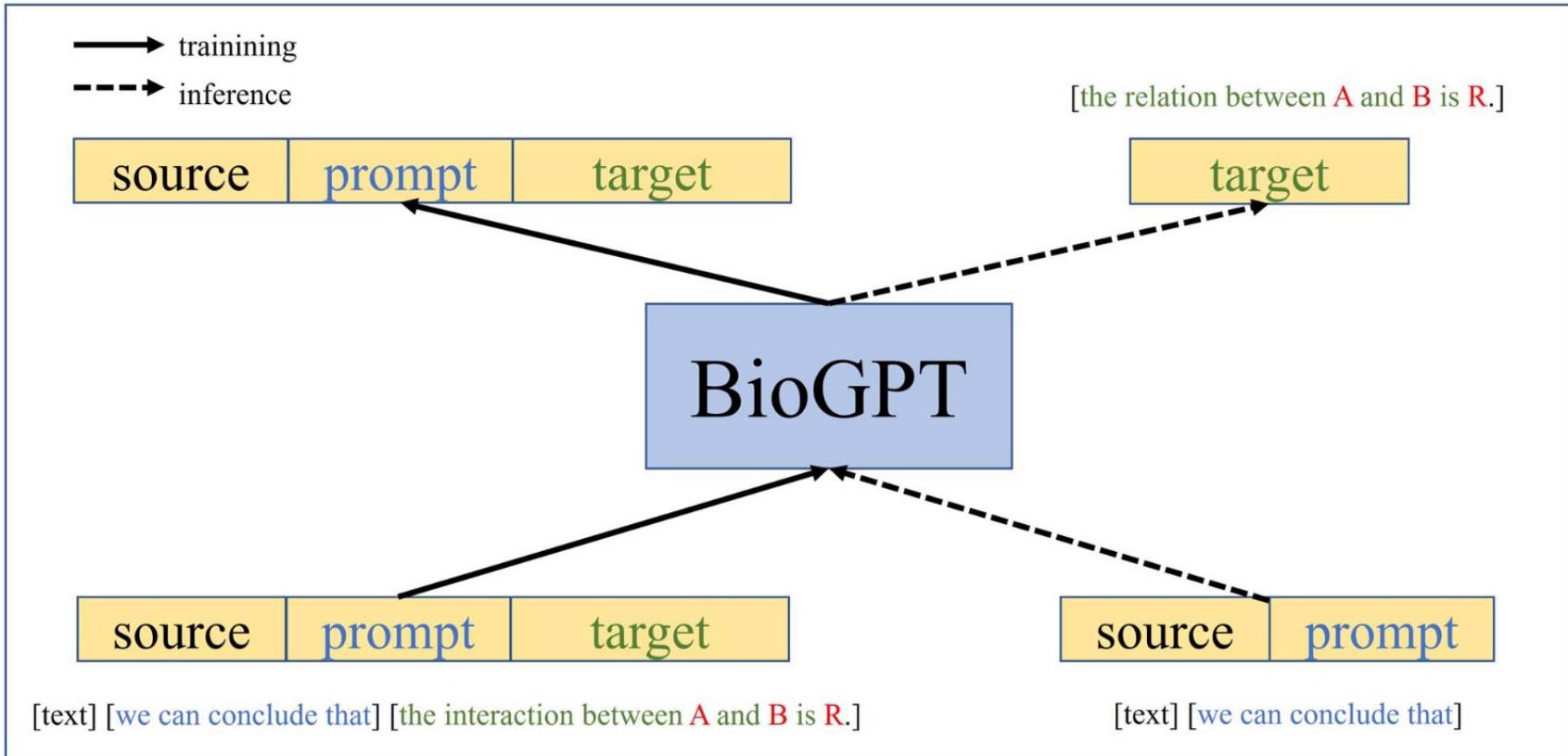
# GatorTronGPT



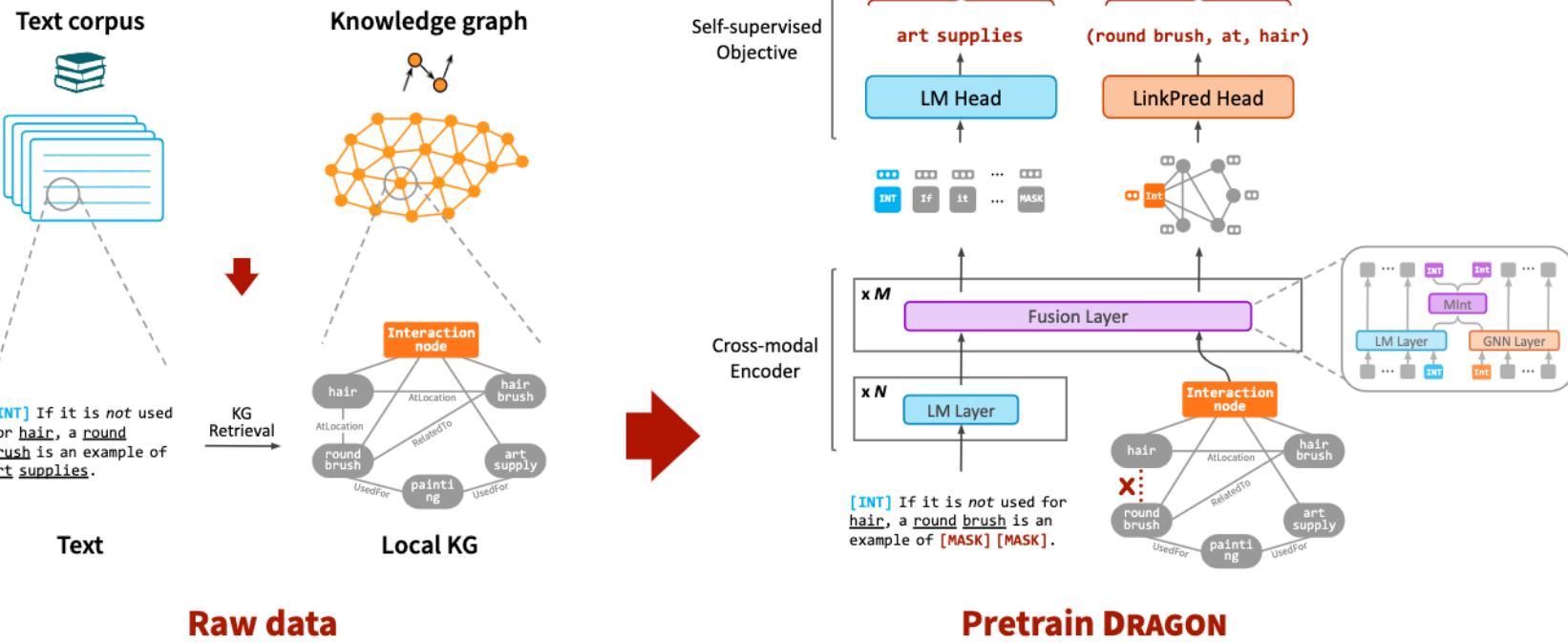
# MedCPT: Zero-shot Biomedical IR Model

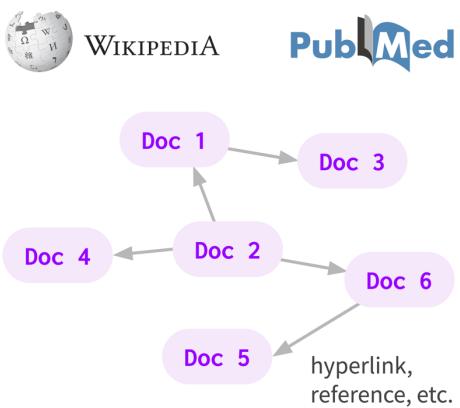


# BioGPT

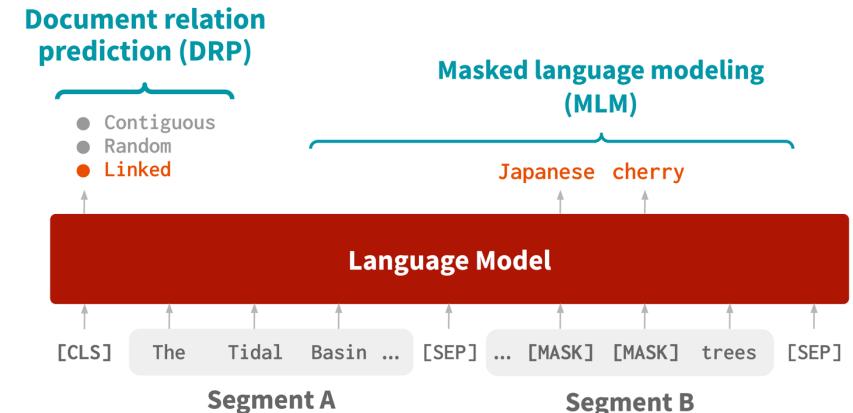


# DRAGON





# BioLinkBERT/LinkBERT



## Corpus of linked documents

Model	Size	Domain	Pretraining Corpus	Download Link (🤗 HuggingFace)
LinkBERT-base	110M parameters	General	Wikipedia with hyperlinks	<a href="#">michiyasunaga/LinkBERT-base</a>
LinkBERT-large	340M parameters	General	Wikipedia with hyperlinks	<a href="#">michiyasunaga/LinkBERT-large</a>
BioLinkBERT-base	110M parameters	Biomedicine	PubMed with citation links	<a href="#">michiyasunaga/BioLinkBERT-base</a>
BioLinkBERT-large	340M parameters	Biomedicine	PubMed with citation links	<a href="#">michiyasunaga/BioLinkBERT-large</a>

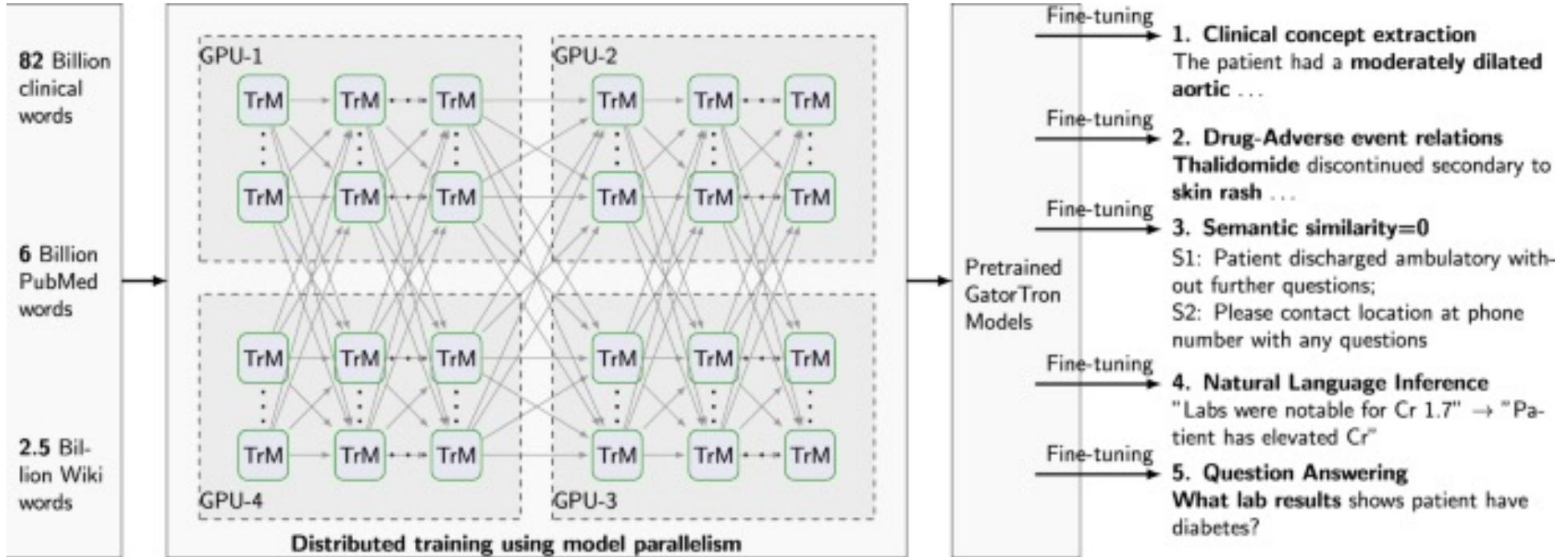
<https://github.com/michiyasunaga/LinkBERT>

## Create LM inputs

## Pretrain the LM

	PubMed-BERT <sub>base</sub>	BioLink-BERT <sub>base</sub>	BioLink-BERT <sub>large</sub>
<b>Named entity recognition</b>			
BC5-chem (Li et al., 2016)	93.33	<b>93.75</b>	<b>94.04</b>
BC5-disease (Li et al., 2016)	85.62	<b>86.10</b>	<b>86.39</b>
NCBI-disease (Dogan et al., 2014)	87.82	<b>88.18</b>	<b>88.76</b>
BC2GM (Smith et al., 2008)	84.52	<b>84.90</b>	<b>85.18</b>
JNLPBA (Kim et al., 2004)	<b>80.06</b>	79.03	<b>80.06</b>
<b>PICO extraction</b>			
EBM PICO (Nye et al., 2018)	73.38	<b>73.97</b>	<b>74.19</b>
<b>Relation extraction</b>			
ChemProt (Krallinger et al., 2017)	77.24	<b>77.57</b>	<b>79.98</b>
DDI (Herrero-Zazo et al., 2013)	82.36	<b>82.72</b>	<b>83.35</b>
GAD (Bravo et al., 2015)	82.34	<b>84.39</b>	<b>84.90</b>
<b>Sentence similarity</b>			
BIOSSES (Sogancioglu et al., 2017)	92.30	<b>93.25</b>	<b>93.63</b>
<b>Document classification</b>			
HoC (Baker et al., 2016)	82.32	<b>84.35</b>	<b>84.87</b>
<b>Question answering</b>			
PubMedQA (Jin et al., 2019)	55.84	<b>70.20</b>	<b>72.18</b>
BioASQ (Nentidis et al., 2019)	87.56	<b>91.43</b>	<b>94.82</b>
<b>BLURB score</b>	81.10	<b>83.39</b>	<b>84.30</b>

# GatorTron



# PubMedBERT

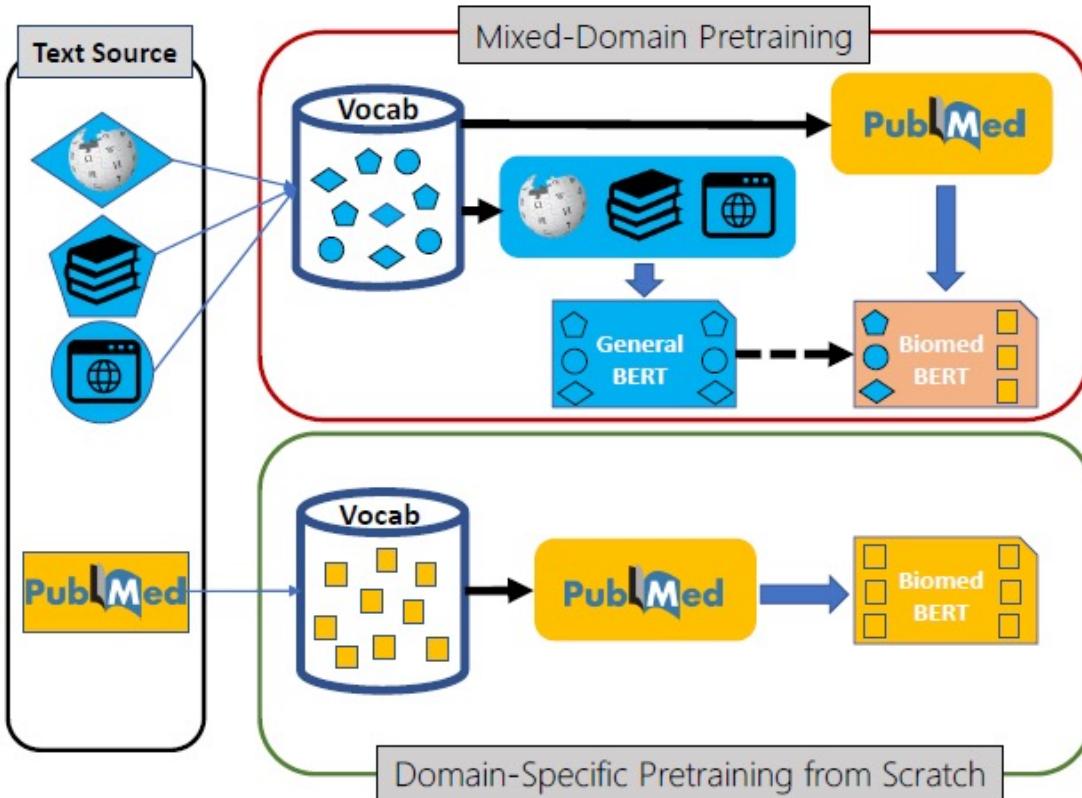
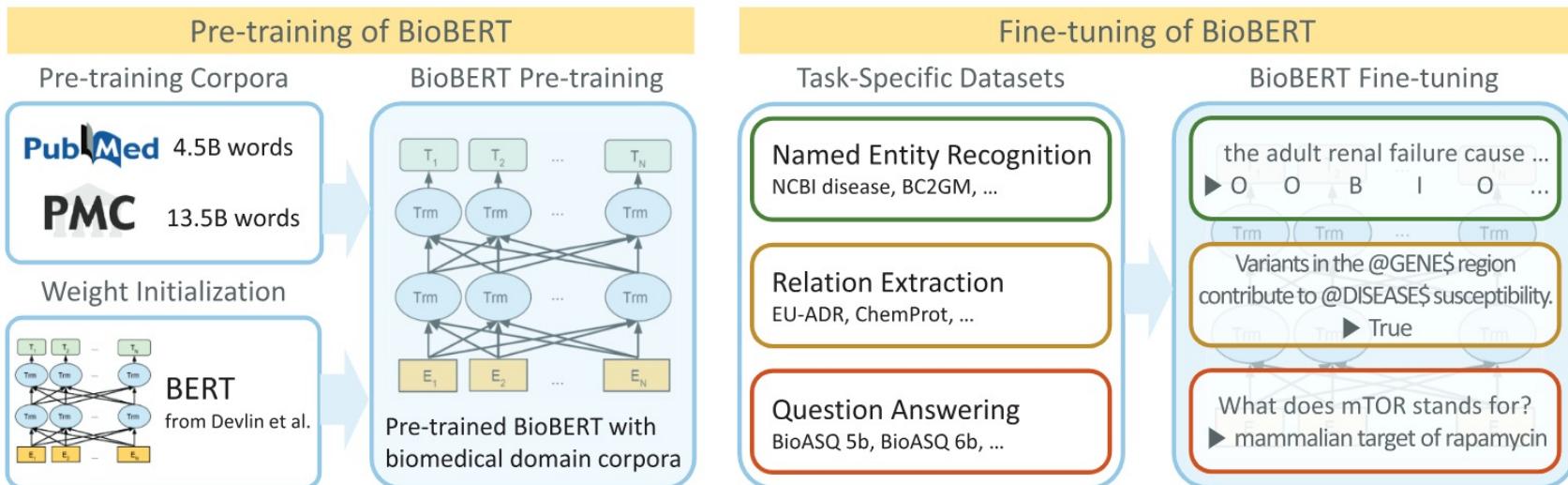


Fig. 1. Two paradigms for neural language model pretraining. Top: The prevailing mixed-domain paradigm assumes that out-domain text is still helpful and typically initializes domain-specific pretraining with a general-domain language model and inherits its vocabulary. Bottom: Domain-specific pretraining from scratch derives the vocabulary and conducts pretraining using solely in-domain text. In this paper, we show that for domains with abundant text such as biomedicine, domain-specific pretraining from scratch can substantially outperform the conventional mixed-domain approach.

Dataset	Task	Train	Dev	Test	Evaluation Metrics
BC5-chem	NER	5203	5347	5385	F1 entity-level
BC5-disease	NER	4182	4244	4424	F1 entity-level
NCBI-disease	NER	5134	787	960	F1 entity-level
BC2GM	NER	15197	3061	6325	F1 entity-level
JNLPBA	NER	46750	4551	8662	F1 entity-level
EBM PICO	PICO	339167	85321	16364	Macro F1 word-level
ChemProt	Relation Extraction	18035	11268	15745	Micro F1
DDI	Relation Extraction	25296	2496	5716	Micro F1
GAD	Relation Extraction	4261	535	534	Micro F1
BIOSSES	Sentence Similarity	64	16	20	Pearson
HoC	Document Classification	1295	186	371	Micro F1
PubMedQA	Question Answering	450	50	500	Accuracy
BioASQ	Question Answering	670	75	140	Accuracy

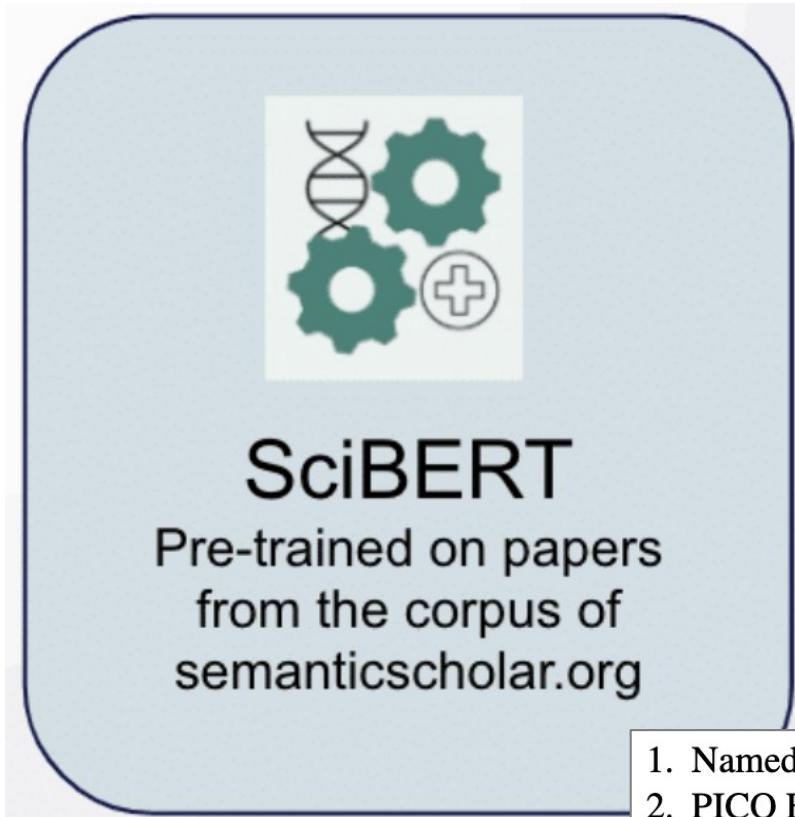
Table 3. Datasets used in the BLURB biomedical NLP benchmark. We list the numbers of instances in train, dev, and test (e.g., entity mentions in NER and PICO elements in evidence-based medical information extraction).

# BioBERT



Corpus	Number of words	Domain
English Wikipedia	2.5B	General
BooksCorpus	0.8B	General
PubMed Abstracts	4.5B	Biomedical
PMC Full-text articles	13.5B	Biomedical

Model	Corpus combination
BERT (Devlin et al., 2019)	Wiki + Books
BioBERT (+PubMed)	Wiki + Books + PubMed
BioBERT (+PMC)	Wiki + Books + PMC
BioBERT (+PubMed + PMC)	Wiki + Books + PubMed + PMC



# SciBERT

**Background:** based on a multilayer bidirectional Transformer (Vaswani et al., 2017). Trained on two tasks: predicting randomly masked tokens & predicting whether two sentences follow each other.

**Architecture:** follows the same architecture as BERT but is instead pretrained on scientific text

**Vocabulary:** BERT uses WordPiece (Wu et al., 2016) for unsupervised tokenization of the input text.

**Corpus:** Trained on a random sample of 1.14M papers from Semantic Scholar (Ammar et al., 2018). Corpus consists of 18% papers from computer science & 82% from broad biomedical domain

## NLP tasks

1. Named Entity Recognition (NER)
2. PICO Extraction (PICO)
3. Text Classification (CLS)
4. Relation Classification (REL)
5. Dependency Parsing (DEP)

<https://github.com/allenai/scibert/>

# ClinicalBERT

## Publicly Available Clinical BERT Embeddings

<b>Emily Alsentzer</b> Harvard-MIT Cambridge, MA emilya@mit.edu	<b>John R. Murphy</b> MIT CSAIL Cambridge, MA jr murphy@mit.edu	<b>Willie Boag</b> MIT CSAIL Cambridge, MA wboag@mit.edu	<b>Wei-Hung Weng</b> MIT CSAIL Cambridge, MA ckbjimmy@mit.edu
--	--	---	--

<b>Di Jin</b> MIT CSAIL Cambridge, MA jindi15@mit.edu	<b>Tristan Naumann</b> Microsoft Research Redmond, WA tristan@microsoft.com	<b>Matthew B. A. McDermott</b> MIT CSAIL Cambridge, MA mmd@mit.edu
--	--	---

**Data:** clinical text from the approximately 2 million notes in the MIMIC-III v1.4 database

Trained two varieties of BERT on MIMIC notes:

- **Clinical BERT:** used text from all note types, initialized from BERT-Base
- **Discharge Summary BERT:** used only discharge summaries to tailor the corpus to downstream tasks, initialized from BioBERT.

Model	MedNLI	i2b2 2006	i2b2 2010	i2b2 2012	i2b2 2014
BERT	77.6%	93.9	83.5	75.9	92.8
BioBERT	80.8%	<b>94.8</b>	86.5	78.9	<b>93.0</b>
Clinical BERT	80.8%	91.5	86.4	78.5	92.6
Discharge Summary BERT	80.6%	91.9	86.4	78.4	92.8
Bio+Clinical BERT	<b>82.7%</b>	94.7	87.2	<b>78.9</b>	92.5
Bio+Discharge Summary BERT	<b>82.7%</b>	94.8	<b>87.8</b>	78.9	92.7

# BlueBERT

## Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets

Yifan Peng   Shankai Yan   Zhiyong Lu

National Center for Biotechnology Information

National Library of Medicine, National Institutes of Health

Bethesda, MD, USA

{yifan.peng, shankai.yan, zhiyong.lu}@nih.gov

### 4.1.1 Pre-training BERT

BERT ([Devlin et al., 2019](#)) is a contextualized word representation model that is pre-trained based on a masked language model, using bidirectional Transformers ([Vaswani et al., 2017](#)).

In this paper, we pre-trained our own model BERT on PubMed abstracts and clinical notes ([MIMIC-III](#)). The statistics of the text corpora on which BERT was pre-trained are shown in Table 2.

Corpus	Words	Domain
PubMed abstract	> 4,000M	Biomedical
MIMIC-III	> 500M	Clinical

Table 2: Corpora

Task	Metrics	SOTA*	ELMo	BioBERT	Our BERT			
					Base (P)	Base (P+M)	Large (P)	Large (P+M)
MedSTS	Pearson	83.6	68.6	84.5	84.5	<b>84.8</b>	84.6	83.2
BIOSSES	Pearson	84.8	60.2	82.7	89.3	<b>91.6</b>	86.3	75.1
BC5CDR-disease	F	84.1	83.9	85.9	<b>86.6</b>	85.4	82.9	83.8
BC5CDR-chemical	F	93.3	91.5	93.0	<b>93.5</b>	92.4	91.7	91.1
ShARe/CLEFE	F	70.0	75.6	72.8	75.4	<b>77.1</b>	72.7	74.4
DDI	F	72.9	78.9	78.8	78.1	79.4	<b>79.9</b>	76.3
ChemProt	F	64.1	66.6	71.3	72.5	69.2	<b>74.4</b>	65.1
i2b2	F	73.7	71.2	72.2	74.4	<b>76.4</b>	73.3	73.9
HoC	F	81.5	80.0	82.9	85.3	83.1	<b>87.3</b>	85.3
MedNLI	acc	73.5	71.4	80.5	82.2	<b>84.0</b>	81.5	83.8
Total			78.8	80.5	82.2	<b>82.3</b>	81.5	79.2

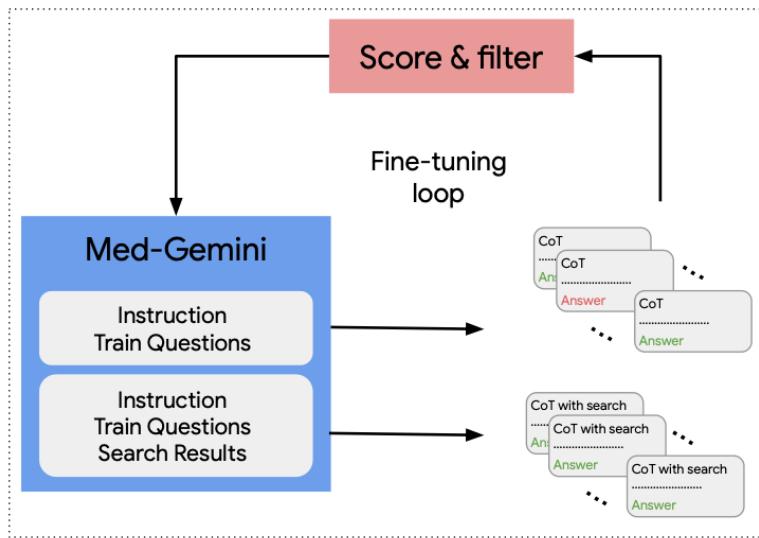
\* SOTA, state-of-the-art as of April 2019, to the best of our knowledge: MedSTS, BIOSSES ([Chen et al., 2019](#)); BC5CDR-disease, BC5CDR-chem ([Yoon et al., 2018](#)); ShARe/CLEFE ([Leaman et al., 2015](#)); DDI ([Zhang et al., 2018](#)). Chem-Prot ([Peng et al., 2018](#)); i2b2 ([Rink et al., 2011](#)); HoC ([Du et al., 2019](#)); MedNLI ([Romanov and Shivade, 2018](#)). P: PubMed, P+M: PubMed + MIMIC-III

# Fine Tuning General LLMs (Selected subset)

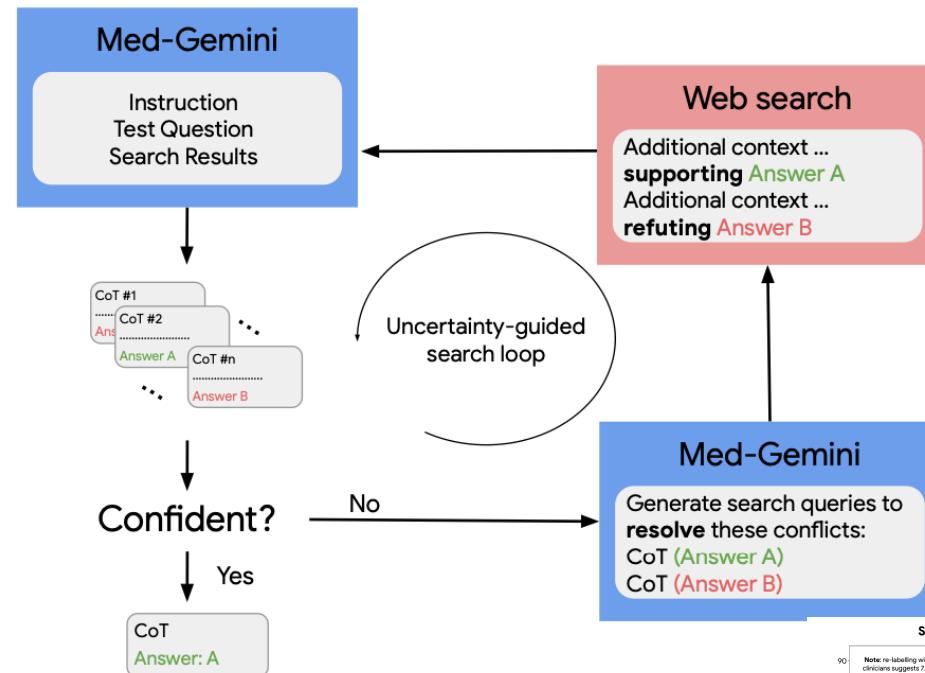
- **Med-Gemini** Capabilities of Gemini Models in Medicine (Saab et al, 2024.4)
- **BioMistral** A Collection of Open-Source Pretrained Large Language Models for Medical Domains (Labrak et al. Arxiv, 2024.2)
- **Taiyi**: A Bilingual (English& Chinese) Fine-Tuned Large Language Model for Diverse Biomedical Tasks (Luo et al. , 2023.11)
- **AlpaCare**: Instruction-tuned Large Language Models for Medical Application (Zhang et al. Arxiv, 2023.10)
- **MEDITRON-70B**: Scaling Medical Pretraining for Large Language Models (Chen et al. Arxiv, 2023.10)
- **BioMedGPT/OpenBioMed** Open Multimodal Generative Pre-trained Transformer for BioMedicine (Luo et al. Arxiv, 2023.8)
- **ClinicalGPT**: Large Language Models Finetuned with Diverse Medical Data and Comprehensive Evaluation. 2023 (Wang et al. Arxiv, 2023.6)
- **MedPaLM 2**: Towards expert-level medical question answering with large language models (Singhal et al. Arxiv, 2023.6)

# MedGemini

Self-training with search



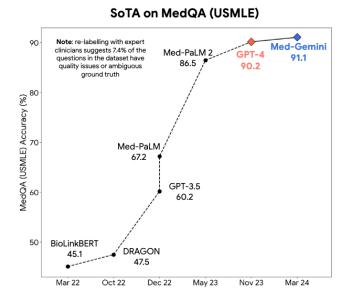
Uncertainty-guided search at inference



Building on Gemini 1.0 and Gemini 1.5

Multimodal models specialized in medicine

Evaluate on 14 medical benchmarks



# BioMistral

BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains

Model Name	Base Model	Model Type	Sequence Length	Download
BioMistral-7B	<a href="#">Mistral-7B-Instruct-v0.1</a>	Further Pre-trained	2048	<a href="#">HuggingFace</a>
BioMistral-7B-DARE	<a href="#">Mistral-7B-Instruct-v0.1</a>	Merge DARE	2048	<a href="#">HuggingFace</a>
BioMistral-7B-TIES	<a href="#">Mistral-7B-Instruct-v0.1</a>	Merge TIES	2048	<a href="#">HuggingFace</a>
BioMistral-7B-SLERP	<a href="#">Mistral-7B-Instruct-v0.1</a>	Merge SLERP	2048	<a href="#">HuggingFace</a>



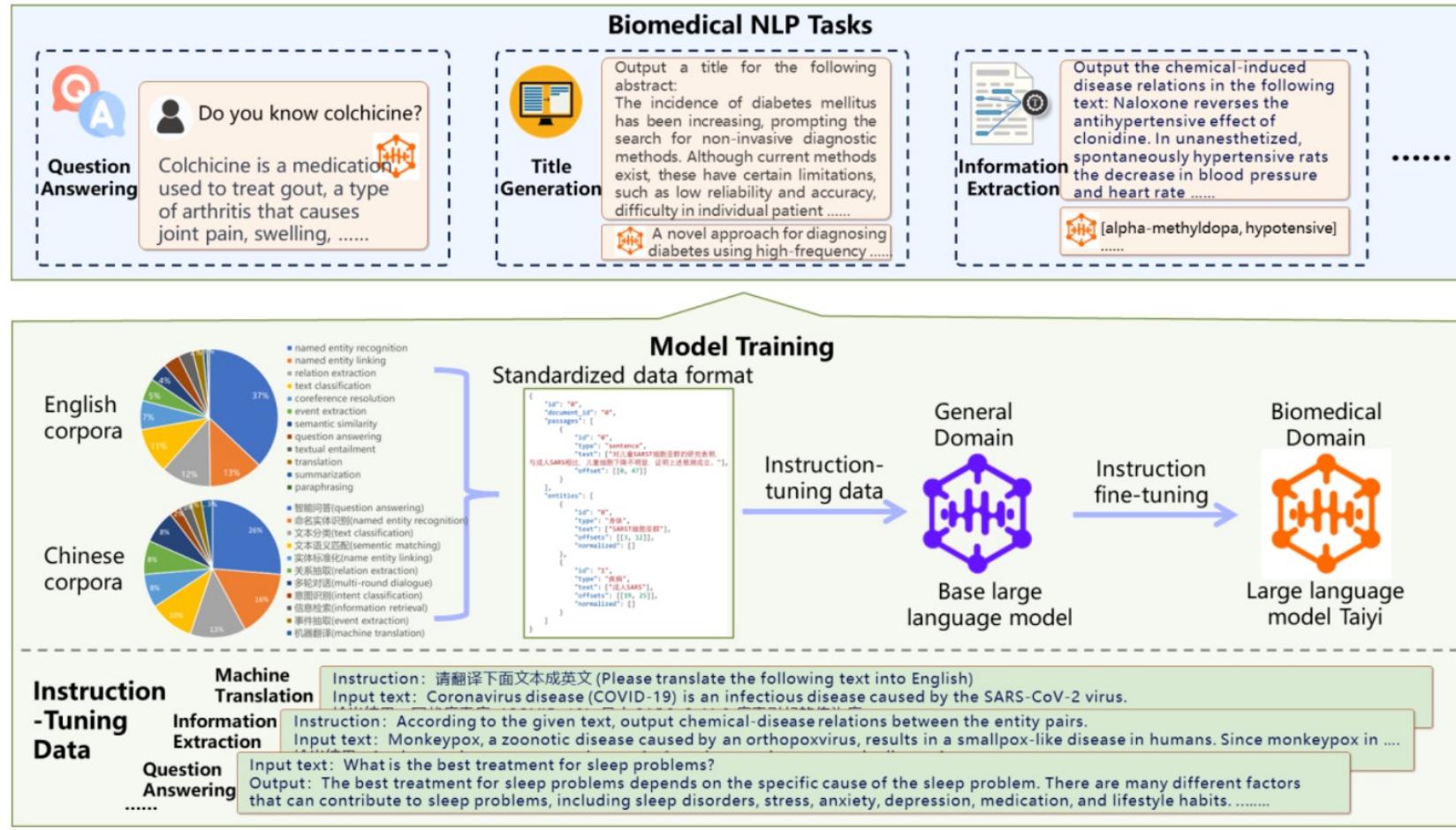
**Barcelona  
Supercomputing  
Center**  
Centro Nacional de Supercomputación

# BioMistral

- BioMistral, open-source LLM tailored for biomedical domain, utilizing Mistral as its foundation model
- Pre-trained on PubMed Central (full text)
- Evaluation on a benchmark comprising 10 medical question-answering (QA) tasks in English
- For Multilingual generalization of medical LLMs they automatically translated & evaluated benchmarks into 7 languages

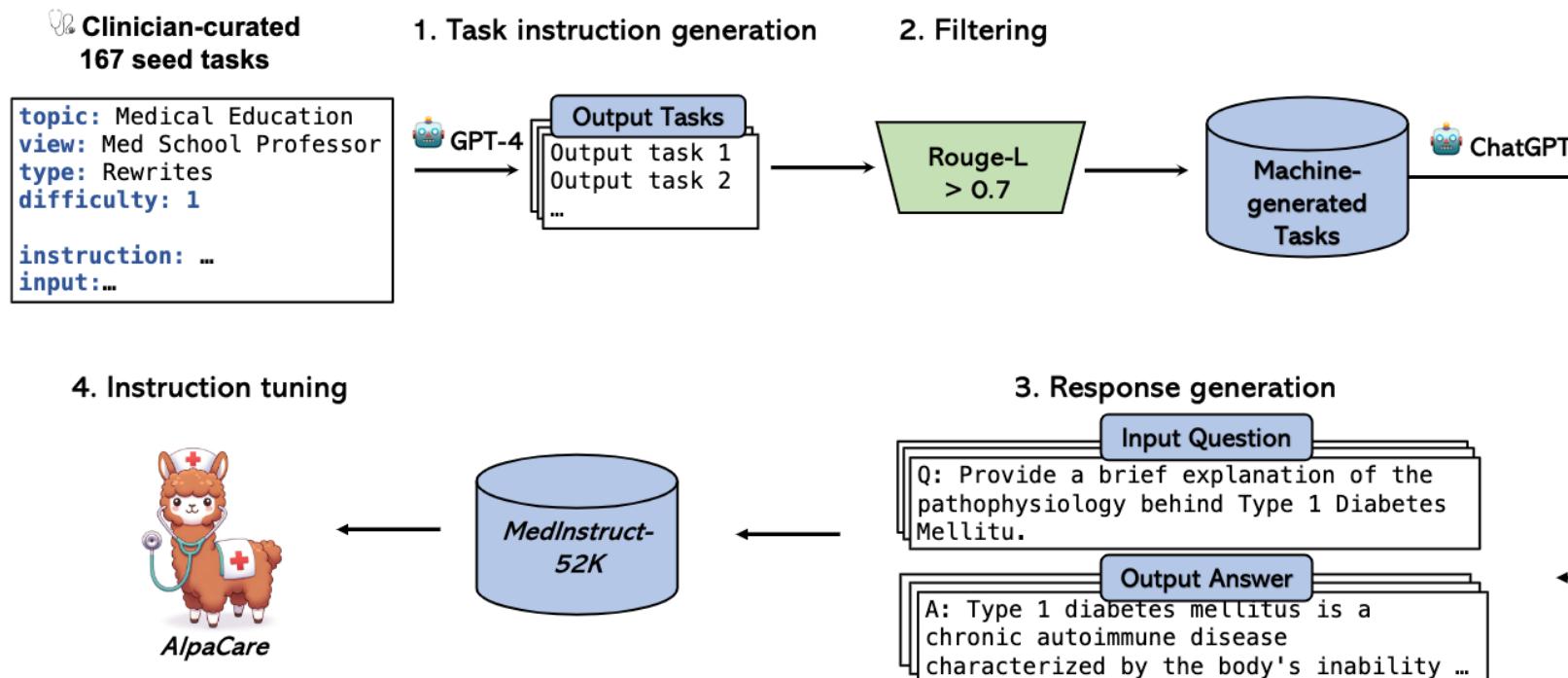
<https://huggingface.co/BioMistral/BioMistral-7B>

# Taiyi



Qwen-7B as our pre-trained model

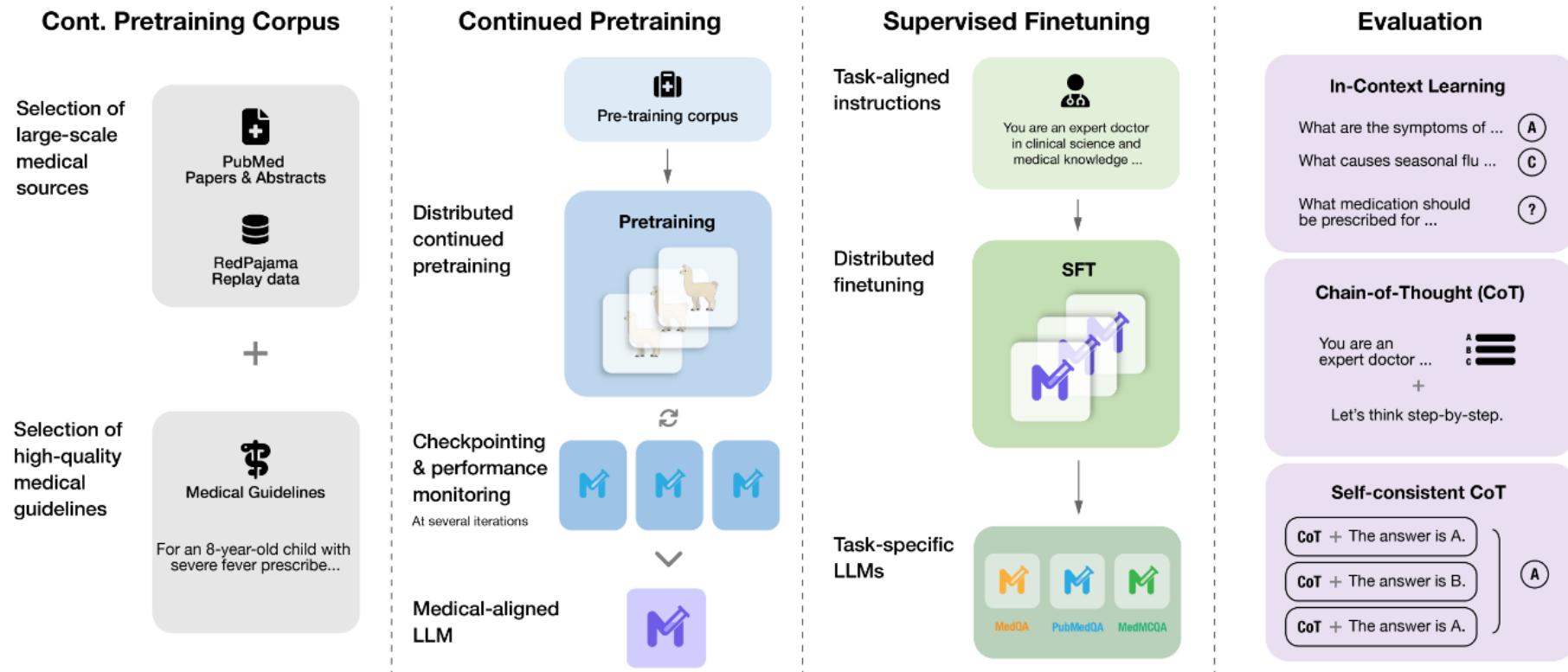
# AlpcaCARE



- Machine-generated medical instruction-fine tuning (IFT) dataset (MedInstruct-52k) using GPT-4 and Chat-GPT with high-quality expert-curated seed set
- Then fine-tune LLaMA-series models on it to develop AlpCare

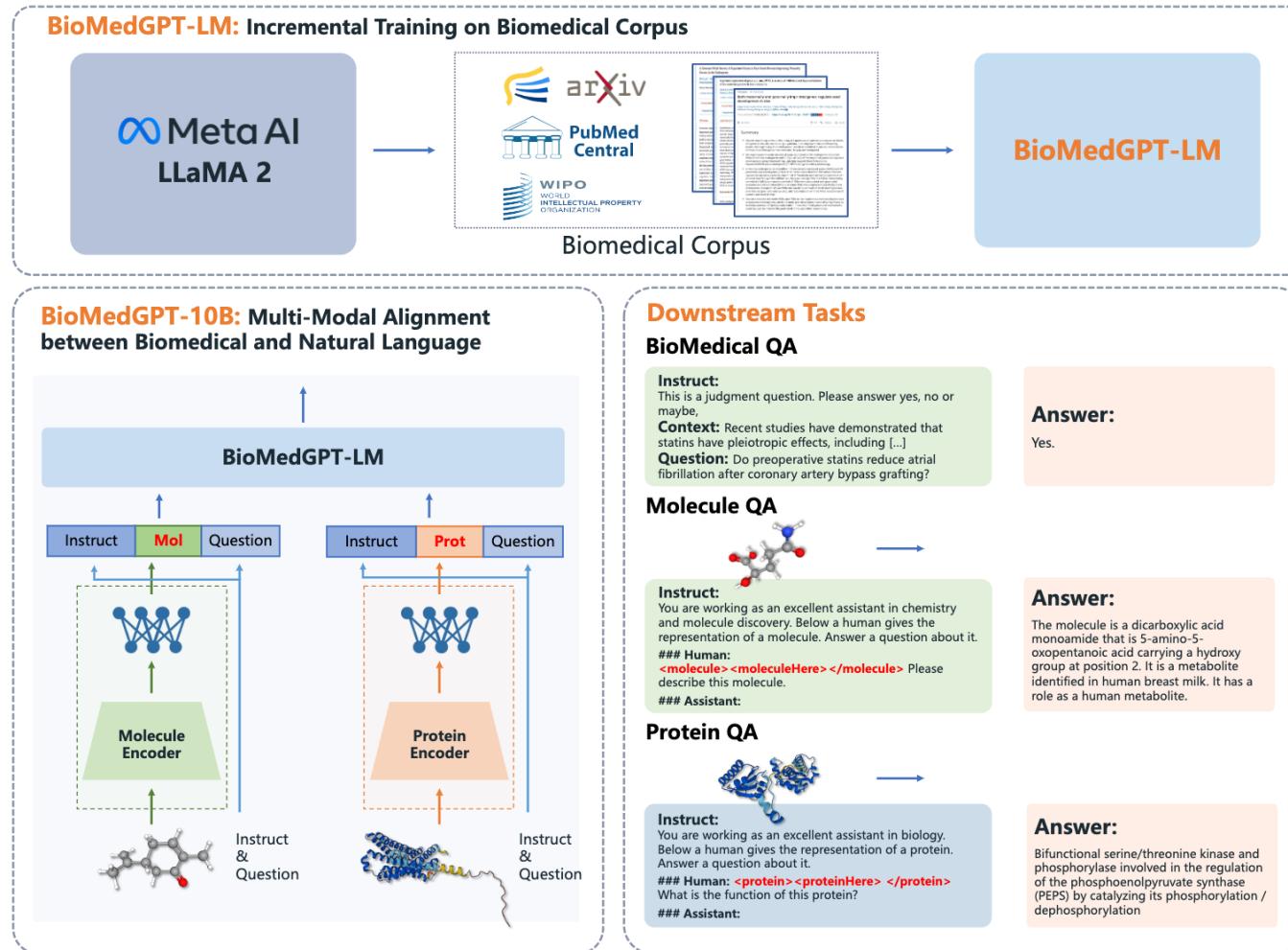
<https://github.com/XZhang97666/AlpCare>

# MediTron-70B

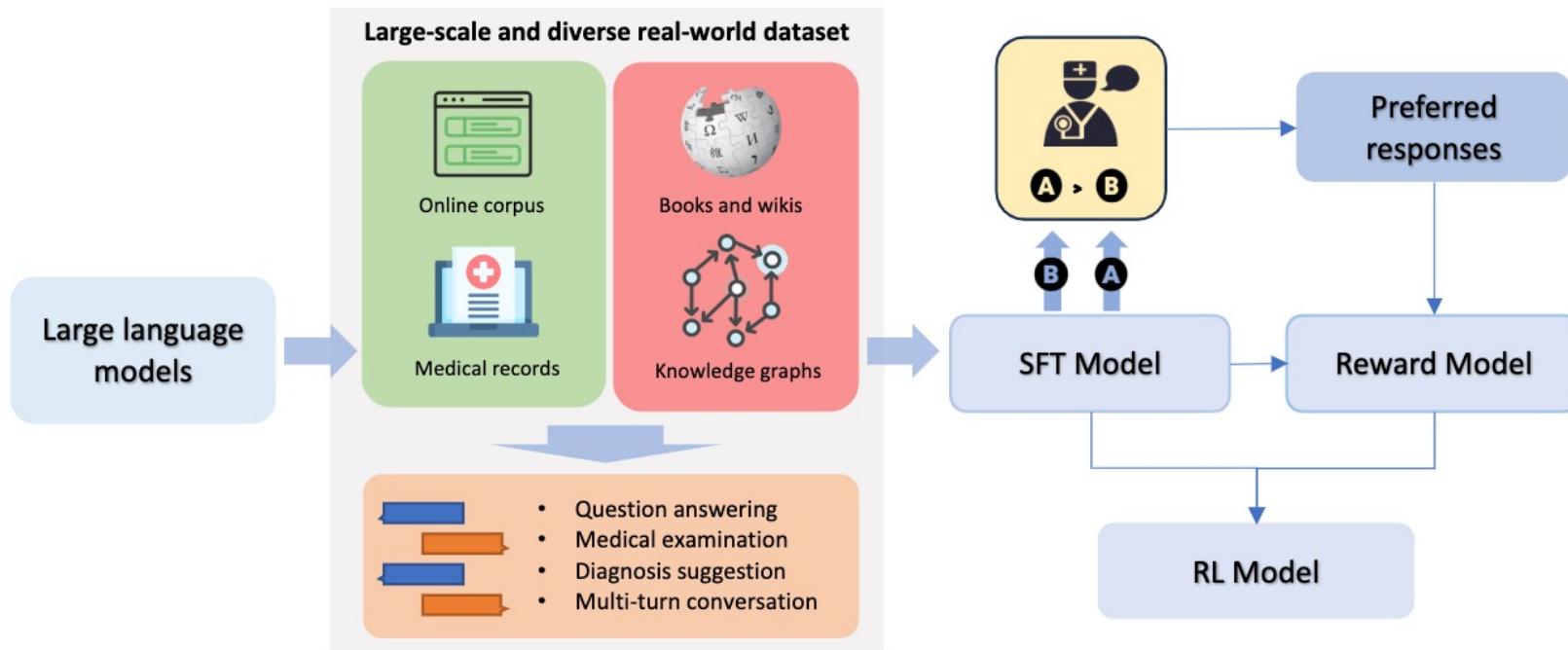


- MEDITRON builds on Llama-2
- Pretraining on PubMed articles, abstracts, medical guidelines
- Evaluations using four major medical benchmark

# BioMedGPT

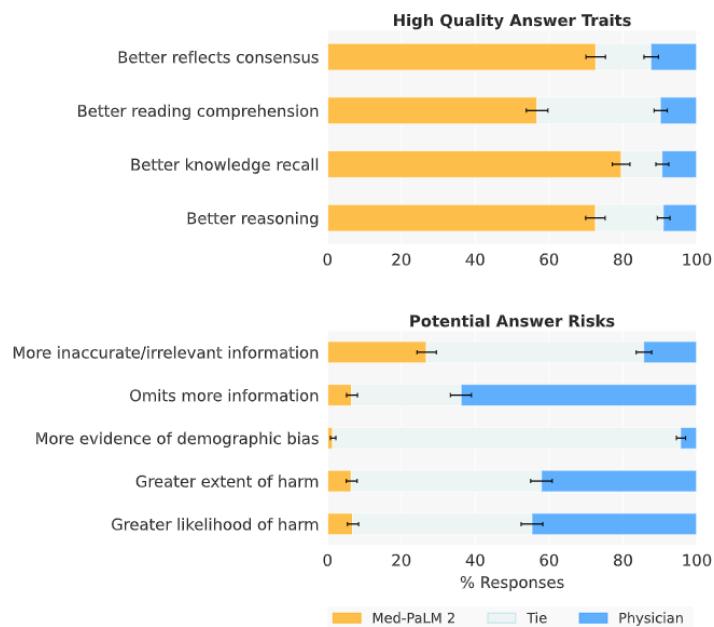
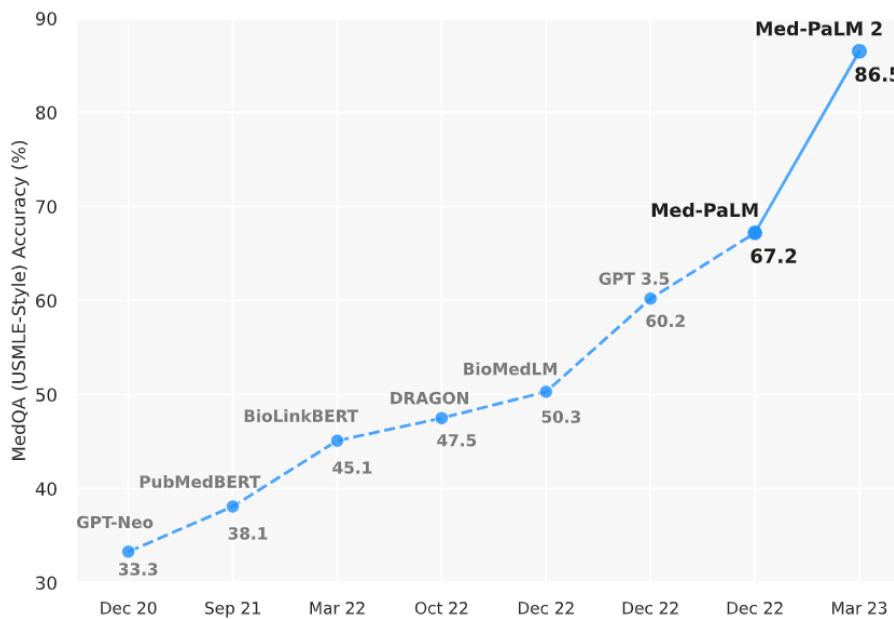


# ClinicalGPT



- Medical datasets including cMedQA2 (chinese medical question-and-answer dataset), cMedQA-KG, MD-EHR, MEDQA-MCMLE, and MedDialog, for the training and evaluation
- BLOOM-7B as base model (open-source nature, multilingual support)

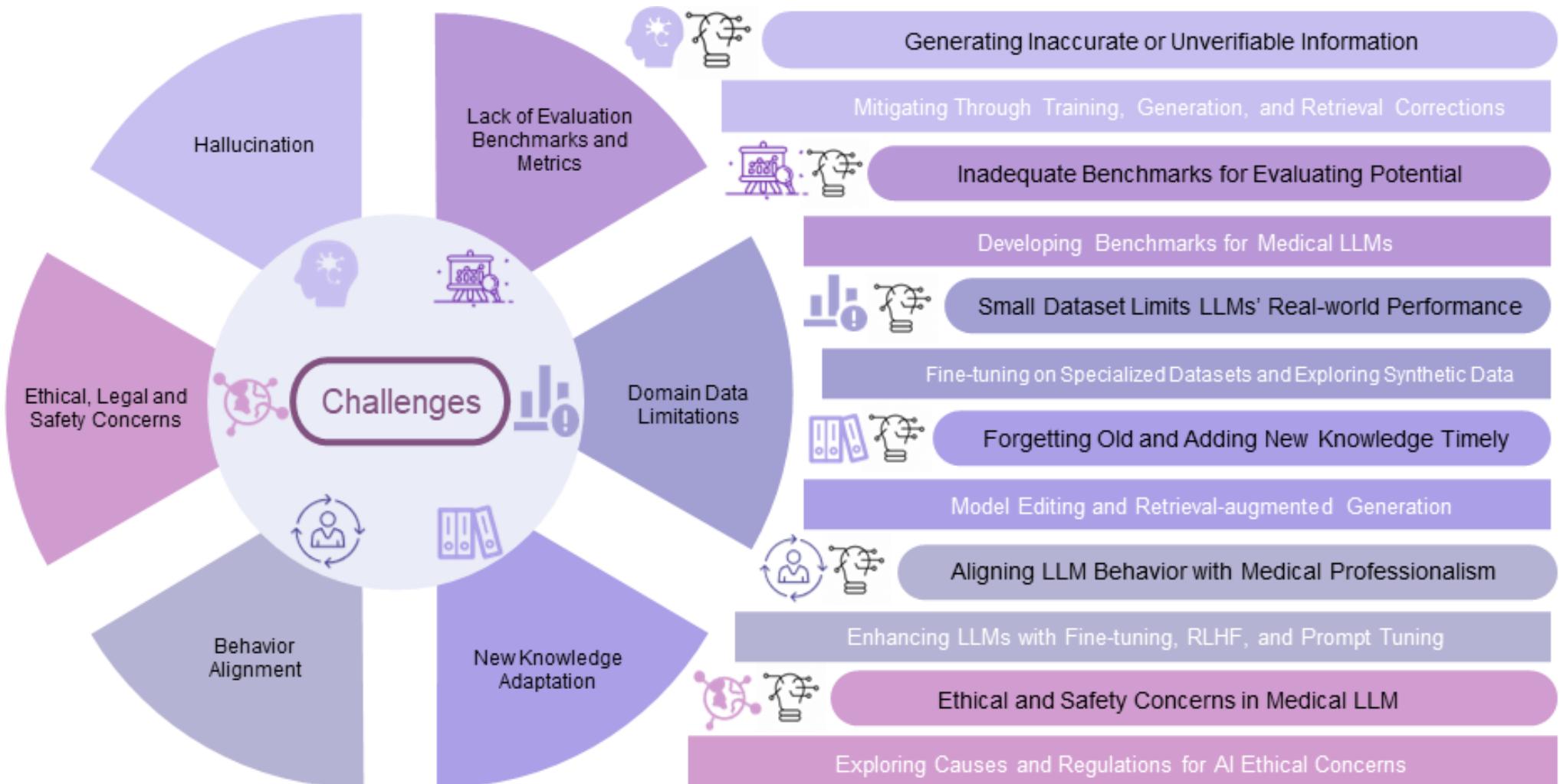
# MedPalm-2

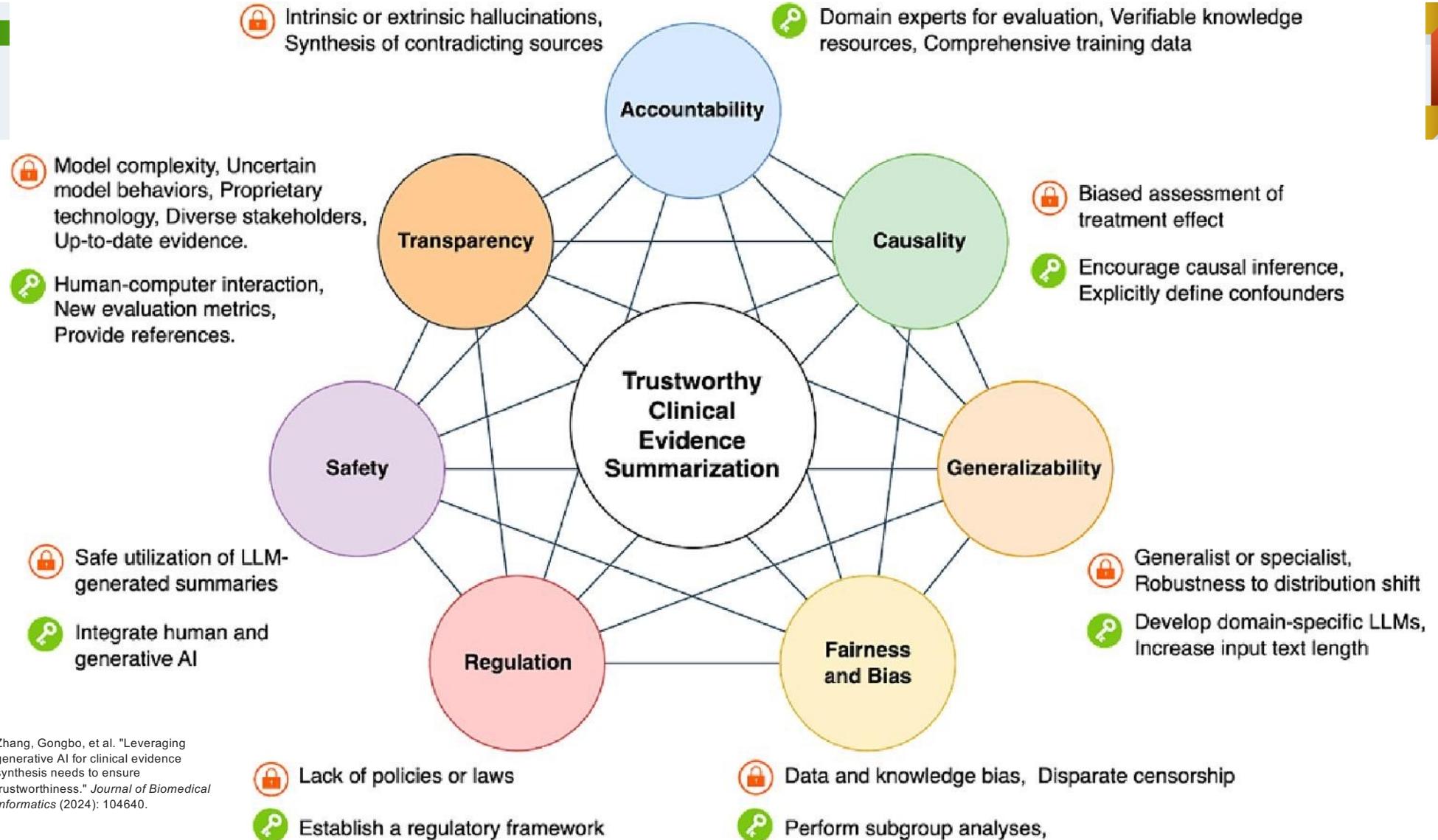
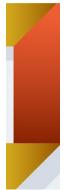
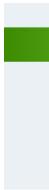


- Medical LLM trained using as base model PaLM 2
- Applied instruction finetuning to the base LLM
- Datasets: MultiMedQA—namely MedQA, MedMCQA, HealthSearchQA, LiveQA and MedicationQA

# Medical data sources for Language Models

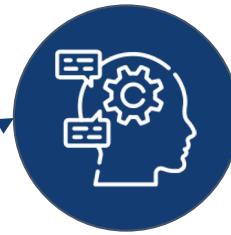
- Scientific publication abstracts: PubMed (English), SCIELO (Spanish, Portuguese, English)
- Scientific publication full text papers: PubMed Central (and Semantic scholar)
- Clinical case reports from PMC Patient
- Clinical practice guidelines (CPGs) found in PubMed and PubMed Central
- Clinical Records: mainly in house EHRs as well as accessible datasets like MIMIC-III/IV
- Medical dialogue datasets (e.g. Meddialog)
- Medical web content / crawler
- Health-related / medical Wikipedia content
- Clinical vocabularies and terminologies UMLS (for SapBERT)
- Biomedical Databases/knowledgebases





# Large corpora for Language Models: Spanish

Spanish Biomedical Crawled Corpus	745M tokens
Clinical cases of many specialities	102M tokens
Spanish scientific publications	100M tokens
Patents	135 M tokens
Spanish clinical trials	4.1M tokens



Biomedical MarIA

Spanish spoken by **> 572 million** people (with 477 native speakers) and **> 900 of romance languages worldwide**

We need to go beyond English: all languages, and also multilingual resources!

Hugging Face

PlanTL-GOB-ES bsc-bio-ehr-es

Fill-Mask Transformers PyTorch Spanish roberta biomedical clinical ehr spanish

Model card Files and versions Community

<https://huggingface.co/PlanTL-GOB-ES/bioc-ehr-es>

Biomedical-clinical language model for Spanish

BSC Barcelona Supercomputing Center Centro Nacional de Supercomputación

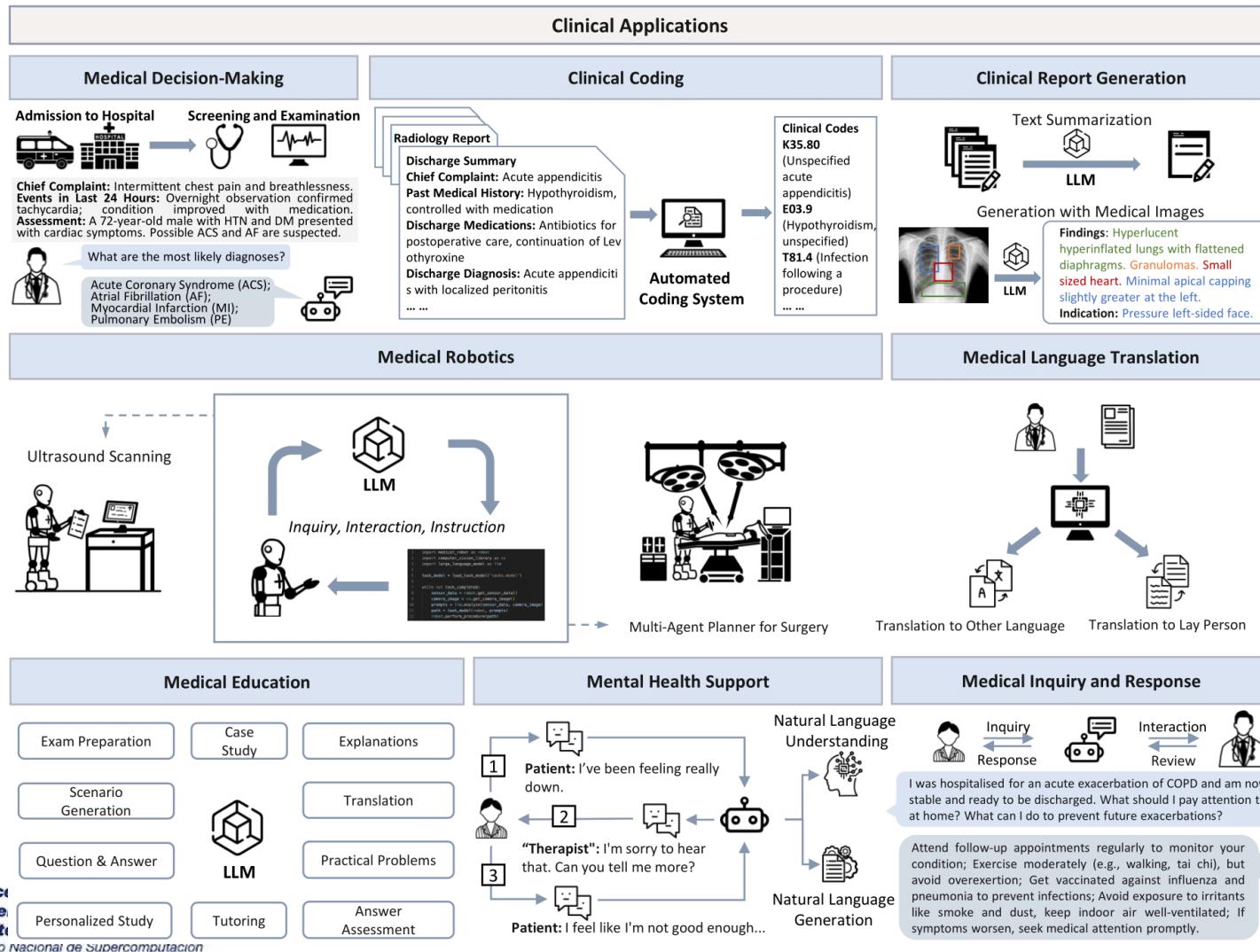
Plan TL Plan de Impulso de las Tecnologías del Lenguaje

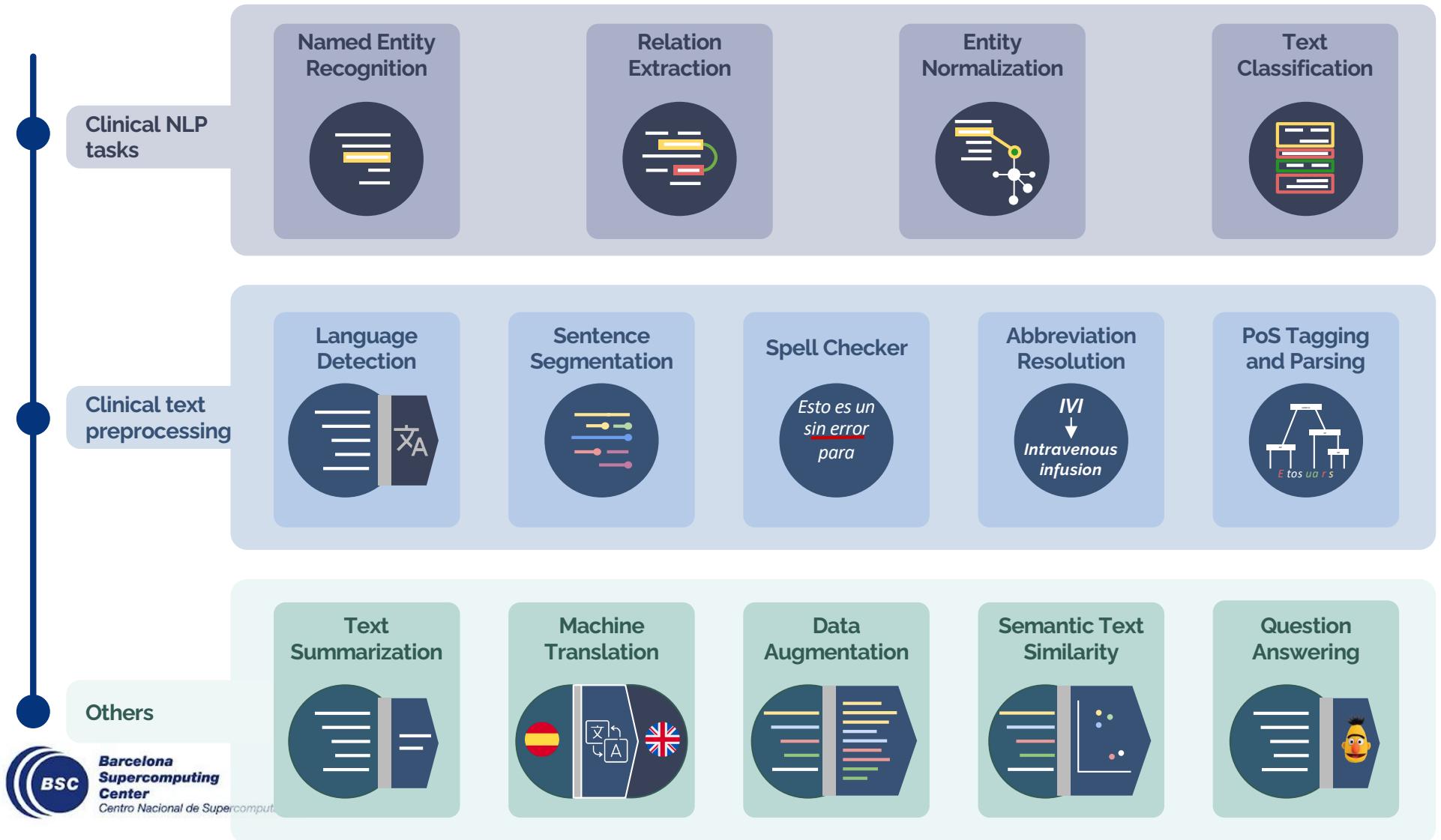
BIBLIOTECA NACIONAL DE ESPAÑA BNE

# Clinical NLP components, use cases and applications (beyond English)



*Barcelona  
Supercomputing  
Center*  
Centro Nacional de Supercomputación





# Data augmentation using Generative AI

A 54-year-old male who had a medical history of membranous nephropathy II with nephrotic syndrome was administered with long-term oral glucocorticoids and immunosuppressants. The patient had a 20 pack-year history of smoking, and denied a family history of hereditary diseases. Chest x-ray demonstrated normal findings at one month before admission. On August 8, 2016, the patient was hospitalized for fever accompanied by progressive dyspnea, cough, and expectoration for 5 days. On admission, the BMI of the patient was 24.5 kg/m<sup>2</sup>, and his body temperature was 39.0°C. Furthermore, the patient had symptoms of tachypnea (35 bpm) and severe hypoxemia (SaO<sub>2</sub> 86%). On auscultation, the patient had good air entrance bilaterally with scattered diffuse crackles and rhonchi. Furthermore, the chest CT scan revealed multiple ground-glass opacities, and laboratory tests revealed normal white blood cell (WBC) count, but with elevated neutrophil count, C-reactive protein (CRP), erythrocyte sedimentation rate (ESR), and (1→3)-β-D-glucan. The patient was diagnosed as RSV infection on the fourth day of hospitalization when positive RSV-Ab was detected.

On admission, the patient was immediately given respiratory monitoring and supplemental oxygen to improve the low oxygen saturation, as well as antibiotics (moxifloxacin for 4 days, followed by cefminoxine for 8 days), and antifungal therapy (voriconazole for 10 days). The dose of the glucocorticoids and immunosuppressants remained largely unchanged. After 10 days of treatment, the patient's condition became worse. Chest CT revealed the progression of the disease, and oxygen partial pressure was further decreased. The patient was transferred to the Emergency Intensive Care Unit, where the patient was intensively treated, including noninvasive mechanical ventilation, broad-spectrum antibiotics (i.v. meropenem, oral moxifloxacin, and cotrimoxazole), antifungal therapy (micafungin), corticosteroids (methylprednisolone 40 mg bid iv) to relieve the inflammation, and other supportive treatment. Ganciclovir was also prescribed due to a possibility of viral infection, such as cytomegalovirus. Five days later, the patient's condition was further aggravated based on the chest x-ray evaluation. Despite receiving another round of treatments, including invasive ventilator-assisted ventilation therapy, methylprednisolone (80 mg bid), antibacterial agents (cefoperazone sulbactam, tigecycline, and cotrimoxazole) and antifungal (micafungin) therapy, the patient eventually died after 2 days.

## Version of case report as Discharge Summary (1)

**\*\*Patient Name:\*\*** [Insert Patient Name]

**\*\*Medical Record Number:\*\*** [Insert MRN]

**\*\*Date of Admission:\*\*** August 8, 2016

**\*\*Date of Discharge:\*\*** [Insert Date]

**\*\*Attending Physician:\*\*** [Insert Physician Name]

**\*\*Final Diagnosis:\*\*** Respiratory syncytial virus (RSV) infection with progression to respiratory failure and subsequent death.

### **\*\*History of Present Illness:\*\***

A 54-year-old male with a known history of membranous nephropathy II associated with nephrotic syndrome, maintained on long-term oral glucocorticoids and immunosuppressants, was admitted with a 5-day history of fever, progressive dyspnea, cough, and expectoration. The patient had a 20 pack-year history of smoking and no significant family history of hereditary diseases. A chest x-ray one month prior to admission showed normal findings.

### **\*\*Physical Examination on Admission:\*\***

- **BMI:** 24.5 kg/m<sup>2</sup>
- **Temperature:** 39.0°C
- **Respiratory Rate:** 35 bpm (tachypnea)
- **Oxygen Saturation (SaO<sub>2</sub>):** 86% (severe hypoxemia)
- **Auscultation Findings:** Bilateral air entry with scattered diffuse crackles and rhonchi.

### **\*\*Laboratory and Imaging Findings:\*\***

- **Chest CT Scan:** Multiple ground-glass opacities.
- **Laboratory Tests:**
  - Normal white blood cell (WBC) count.
  - Elevated neutrophil count.
  - Elevated C-reactive protein (CRP), erythrocyte sedimentation rate (ESR), and (1→3)-β-D-glucan.
- **RSV Diagnosis:** Positive RSV-Ab detected on the fourth day of hospitalization.

## Spanish Version of case report as Discharge Summary (1)

**\*\*Nombre del Paciente:\*\*** [Insertar Nombre del Paciente]

**\*\*Número de Historia Clínica:\*\*** [Insertar Número de Historia]

**\*\*Fecha de Ingreso:\*\*** 8 de agosto de 2016

**\*\*Fecha de Alta:\*\*** [Insertar Fecha]

**\*\*Médico Tratante:\*\*** [Insertar Nombre del Médico]

**\*\*Diagnóstico Final:\*\*** Infección por virus respiratorio sincitial (VRS) con progresión a insuficiencia respiratoria y fallecimiento.

### **\*\*Historia de la Enfermedad Actual:\*\***

Paciente masculino de 54 años con antecedentes de nefropatía membranosa II asociada a síndrome nefrótico, bajo tratamiento con glucocorticoides orales e inmunosupresores a largo plazo. El paciente fue ingresado con un cuadro de 5 días de evolución de fiebre, disnea progresiva, tos y expectoración. Antecedente de 20 paquetes-año de tabaquismo y sin historia familiar significativa de enfermedades hereditarias. Una radiografía de tórax realizada un mes antes del ingreso mostró hallazgos normales.

### **\*\*Examen Físico al Ingreso:\*\***

- **\*\*IMC:\*\*** 24.5 kg/m<sup>2</sup>
- **\*\*Temperatura:\*\*** 39.0°C
- **\*\*Frecuencia Respiratoria:\*\*** 35 respiraciones por minuto (taquipnea)
- **\*\*Saturación de Oxígeno (SaO<sub>2</sub>):\*\*** 86% (hipoxemia severa)
- **\*\*Auscultación:\*\*** Entrada de aire bilateral con crepitantes y roncus dispersos.

### **\*\*Hallazgos de Laboratorio e Imágenes:\*\***

- **\*\*TAC de Tórax:\*\*** Múltiples opacidades en vidrio esmerilado.
- **\*\*Pruebas de Laboratorio:\*\***
  - Recuento de leucocitos (WBC) normal.
  - Neutrofilia.
  - Elevación de la proteína C reactiva (CRP), velocidad de sedimentación globular (VSG) y (1→3)-β-D-glucano.
- **\*\*Diagnóstico de VRS:\*\*** Detección de anticuerpos contra VRS (VRS-Ab) en el cuarto día de hospitalización.

# Data Augmentation: synthetic texts

## Original clinical case report

**PROC**  
Anamnesis  
**HUMANO** **SEX** **EDAD**  
Paciente mujer de 57 años que acude a Urgencias por presentar un cuadro clínico de desorientación tómporo-espacial, tendencia al sueño, lenguaje incoherente, descoordinación motora y pérdida de control de esfínteres de horas de evolución. Niega pérdida de conocimiento, disminución de fuerza o sensibilidad, alteraciones en la marcha (salvo enlentecimiento de la misma), ingesta de alcohol u otros tóxicos. La paciente no tiene alergias conocidas ni antecedentes personales de interés, salvo dislipemia y osteoporosis tratada con suplementos de calcio, vitamina D y ácido ibandrónico. Fumadora de 10 cigarrillos/ día y bebedora muy ocasional de una cerveza. Entre sus antecedentes familiares, su padre falleció por causa tumoral, aunque la paciente no recuerda el primario, y un hermano falleció en la infancia por leucemia.



**SINTOMA** **SINTOMA** **SINTOMA** **SINTOMA**  
**HUMANO** **ENFERMEDAD** **HUMANO** **ENFERMEDAD** **ENFERMEDAD** **FARMACO** **FARMACO** **FARMACO** **ENFERMEDAD**  
**ENFERMEDAD** **HUMANO** **FAM** **SINTOMA** **ENFERMEDAD** **HUMANO** **FAM** **SINTOMA** **HUMANO**  
**ENFERMEDAD** **HUMANO** **ENFERMEDAD** **ENFERMEDAD** **ENFERMEDAD** **HUMANO**

## Re-written clinical case report (oncology)

Antecedentes  
**SEX** **HUMANO** **EDAD**  
Una mujer de 57 años se presenta en Urgencias con síntomas de desorientación temporal y espacial, somnolencia, discurso incoherente, falta de coordinación motora y pérdida de control de los esfínteres durante varias horas. Niega pérdida de conocimiento, debilidad, cambios en la marcha (excepto por lentitud), consumo de alcohol u otras drogas. No tiene alergias conocidas excepto dislipemia y osteoporosis tratada con calcio, vitamina D y ácido ibandrónico. Fuma 10 cigarrillos al día y bebe cerveza ocasionalmente. En sus antecedentes familiares, su padre falleció de cáncer y un hermano de leucemia en la infancia.



**SINTOMA** **SINTOMA** **SINTOMA** **SINTOMA** **ENFERMEDAD** **ENFERMEDAD** **ENFERMEDAD** **HUMANO**  
**FARMACO** **FARMACO** **FARMACO** **HUMANO** **FAM** **SINTOMA** **ENF** **HUMANO** **FAM** **ENFERMEDAD** **HUMANO**  
**ENFERMEDAD** **HUMANO** **ENFERMEDAD** **ENFERMEDAD** **HUMANO**

# Data Augmentation: synthetic noisy texts

Version of clinical case report (oncology) with grammar, typography and orthographic errors

PROCEDIMIENTO	HUMANO	SINTOMA	SINTOMA
4 Exploración física			
5 A su llegada a Urgencias, la paciente presenta regular estado general. Afebril, con buenas constantes. Bien hidratada, nutrita y perfundida. Exploración cardiopulmonar anodina. El abdomen es blando, depresible, doloroso a la palpación en el epigastrio. Se palpa hepatomegalia a expensas del lóbulo izquierdo a dos traveses de dedo, con percusión mate en esta zona.	SINTOMA	PROCEDIMIENTO	SINTOMA
	SINTOMA	PROCEDIMIENTO	ENFERMEDAD
	SINTOMA	PROCEDIMIENTO	ENFERMEDAD
	SINTOMA	PROC	
	SINTOMA	SINTOMA	SINTOMA
	Ruidos hidroáreos conservados. Sin defensa abdominal ni signos de irritación peritoneal.	SINTOMA	
	SINTOMA	ENFERMEDAD	
	Miembros inferiores sin edemas, sin signos de insuficiencia venosa crónica ni de trombosis venosa profunda. Pulsos periféricos presentes y simétricos. Neurológicamente, consciente y orientada en persona, no así en espacio ni tiempo. Bradipsíquica, con lenguaje entrecortado y discurso incoherente. Nomina y repite. Sin apraxias, heminegligencia ni extinción.	ENFERMEDAD	SINTOMA
	HUMANO	SINTOMA	SINTOMA
	SINTOMA	SINTOMA	SINTOMA
	SINTOMA	SINTOMA	SINTOMA
	PROCEDIMIENTO	SINTOMA	ENFERMEDAD
	CAMPIMETRÍA POR CONFRONTACIÓN NORMAL: Pupilas isocóricas y normorreactivas; ausencia de nistagmo.	SINTOMA	
	SINTOMA	SINTOMA	
	Pares craneales normales. Sin signos de irritación meníngea.	SINTOMA	
	SINTOMA	SINTOMA	SINTOMA
	Fuerza y sensibilidad conservadas y simétricas. No dismetría ni disdiadocinesia.	SINTOMA	
PROCEDIMIENTO	SINTOMA	SINTOMA	
4 Exploración física			
5 A su llegada a Urgncs, la pacnt presnta reglar estdo gral. Afebril, cn buens constantes. Ben hidrata, nutrida y prfundida. Exploración cardpulmonar arnodna. El abdmen es blndo, depresble, doloroso a la palpación en el epigastro. Se plpa hepatomegalia a expnsas dl lóbulo izquierdo a dos traveses d dedo, cn prcusón mat en esta zona. Ruidos hidroáreos conservados. Sin defnsa abdominal ni signos d irrtón peritoneal. Miembrs inferiors sin edems, sin signos d insuficiencia venosa crónica ni d trombosis venosa profunda. Pulsos periférics presnts y simétricos. Neurológicmnt, conscient y orientda en persona, no así en espacio ni tiempo. Bradipsíquica, cn lenguaj enlntecido y discursin incohrrnte. Nomina y rept. Sin apraxias, heminegligencia ni extincón.	SINTOMA	PROCEDIMIENTO	SINTOMA
	SINTOMA	ENFERMEDAD	ENFERMEDAD
	SINTOMA	SINTOMA	SINTOMA
	SINTOMA	ENF	ENFERMEDAD
	SINTOMA	ENFERMEDAD	
	PROCEDIMIENTO	SINTOMA	ENFERMEDAD
	CAMPIMETRÍA POR CONTRRTACIÓN NORMAL: Pupilas isocrócas y normorreactivas; ausnca d nistagmo.	SINTOMA	
	SINTOMA	SINTOMA	SINTOMA
	Pares craneales normales. Sin signos d irrtón menígea. Fuerza y sensibilidad consrvaas y simétricas. No dismetría ni disdiadocinesia. Reflejos osteotendinosos normles, cn refijo cutáneo-plantar flexor bilatr.	SINTOMA	
	PROC	SINTOMA	SINTOMA
	SINTOMA	SINTOMA	
	Destaca en la explracón la presnci d asterixis en ls miembrs supiores, cn alguna mioclonía espontánea aislada.	SINTOMA	



# Data Augmentation: synthetic cross-language strategy

## Translated versions of clinical case report

Hombre de 72 años con alergia a penicilinas, hipertensión arterial, ex fumador desde hace 20 años y con un stent coronario desde 1986, acude a Urgencias en agosto de 2009 por disartria y debilidad en extremidades izquierdas de inicio repentino. Refiere dolor dorsal izquierdo de dos meses que no mejora con carbamazepina. Es diagnosticado de accidente isquémico transitorio y hospitalizado para estudio.

### Exploración física

Consciente, orientado, afebril, sin soplos carotídeos, EVA: 9. No adenopatías palpables. Normocardia y normoventilación. Abdomen blando, sin masas ni megalias. Extremidades inferiores sin edemas, sensibilidad y fuerza conservadas.

### Pruebas y Diagnóstico

Hemograma con anemia leve. ECG con onda T negativa. Eco-doppler carotídeo muestra ateromatosis sin estenosis significativa. TC revela diversas alteraciones, incluyendo masa adrenal, lesiones óseas y atrofia cerebral. Biopsia adrenal confirma carcinoma.

### Tratamiento y Evolución

Inicia quimioterapia. Se controla dolor con analgésicos y radioterapia paliativa. Presenta episodios de accidente isquémico transitorio. Desarrolla deterioro general y melena, sugiriendo hemorragia digestiva. Se opta por cuidados paliativos y suspenden anticoagulantes. Fallece por fallo multiorgánico debido a sepsis y enfermedad oncológica avanzada.

### Medical History

A 72-year-old man with penicillin allergy, hypertension, a former smoker for 20 years, and a coronary stent since 1986 presents to the Emergency Department in August 2009 with sudden onset dysarthria and weakness in the left extremities. He reports left dorsal pain for two months not relieved by carbamazepine. He is diagnosed with a transient ischemic attack and hospitalized for further evaluation.

### Anamnese

Ein 72-jähriger Mann mit Penicillinallergie, Hypertonie, ehemaliger Raucher seit 20 Jahren und einem Koronarstent seit 1986 kommt im August 2009 mit plötzlicher Dysarthrie und Schwäche in den linken Extremitäten in die Notaufnahme. Er berichtet über seit zwei Monaten bestehende Schmerzen im linken Rücken, die nicht durch Carbamazepin gelindert werden. Er wird mit einer vorübergehenden ischämischen Attacke diagnostiziert und zur weiteren Abklärung ins Krankenhaus eingewiesen.

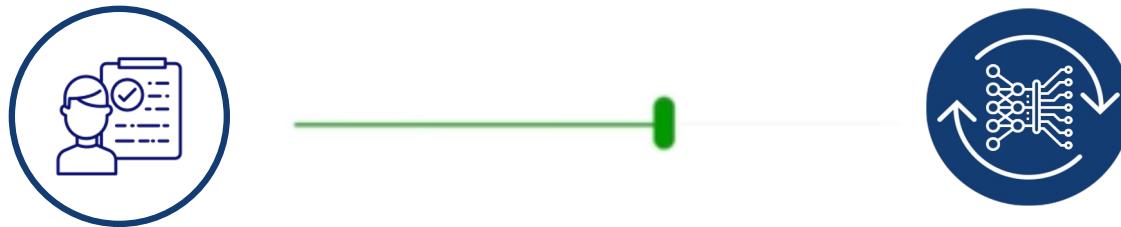
### Antécédents médicaux

Un homme de 72 ans, allergique à la pénicilline, hypertendu, ancien fumeur depuis 20 ans et porteur d'un stent coronarien depuis 1986, se présente aux urgences en août 2009 avec une dysarthrie soudaine et une faiblesse des membres gauches. Il signale des douleurs dorsales gauches depuis deux mois, non soulagées par la carbamazépine. Il est diagnostiqué d'un accident ischémique transitoire et hospitalisé pour des examens complémentaires.

### 病史

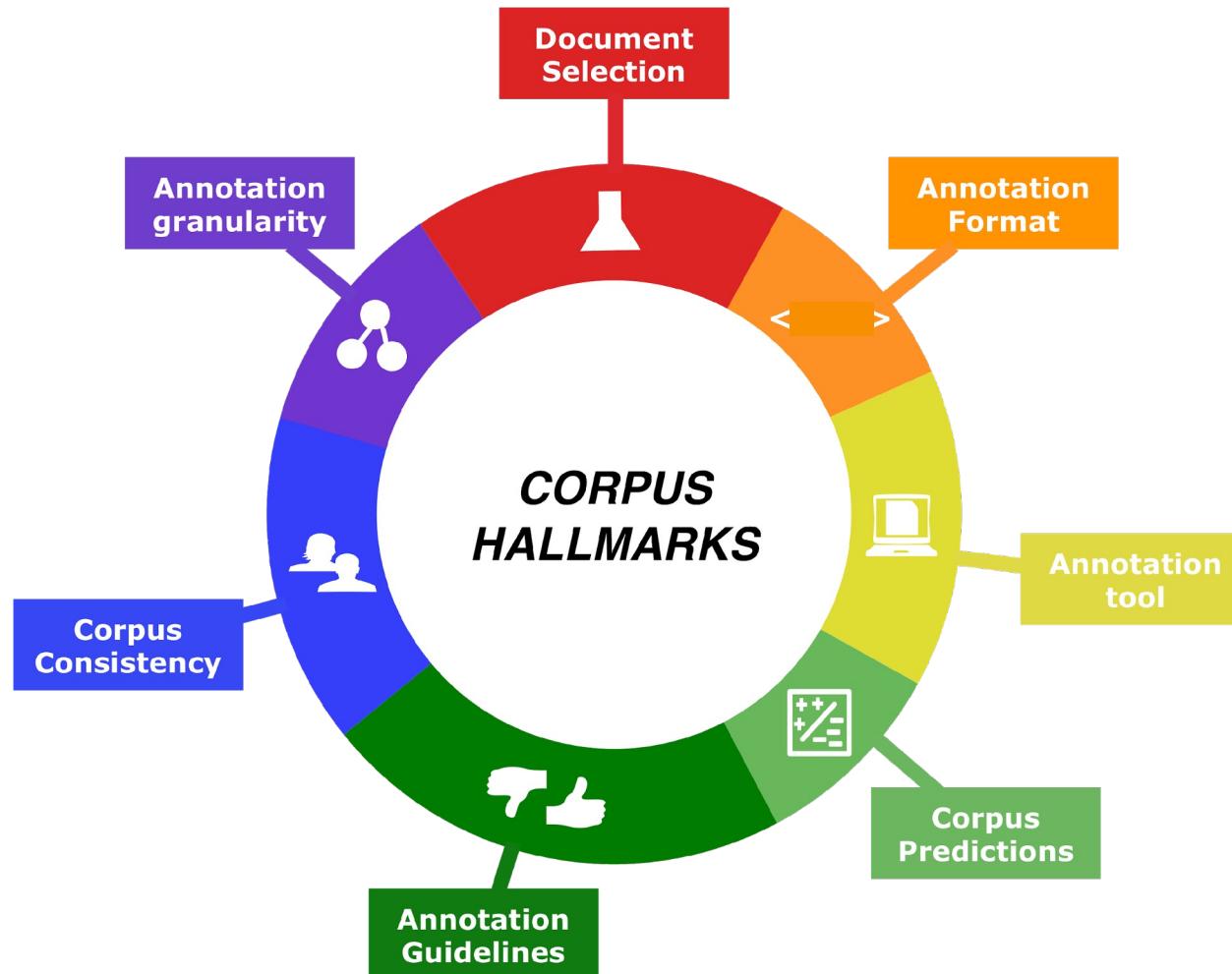
一位72岁的男性患者，对青霉素过敏，患有高血压，自1986年以来已经戒烟20年，并植入了冠状动脉支架，在2009年8月因突发性的言语困难和左侧肢体无力而前往急诊科。他反映左背痛已有两个月之久，卡马西平无效。诊断为短暂性脑缺血发作，并住院进一步检查。

# Labeling is a human-machine collaboration

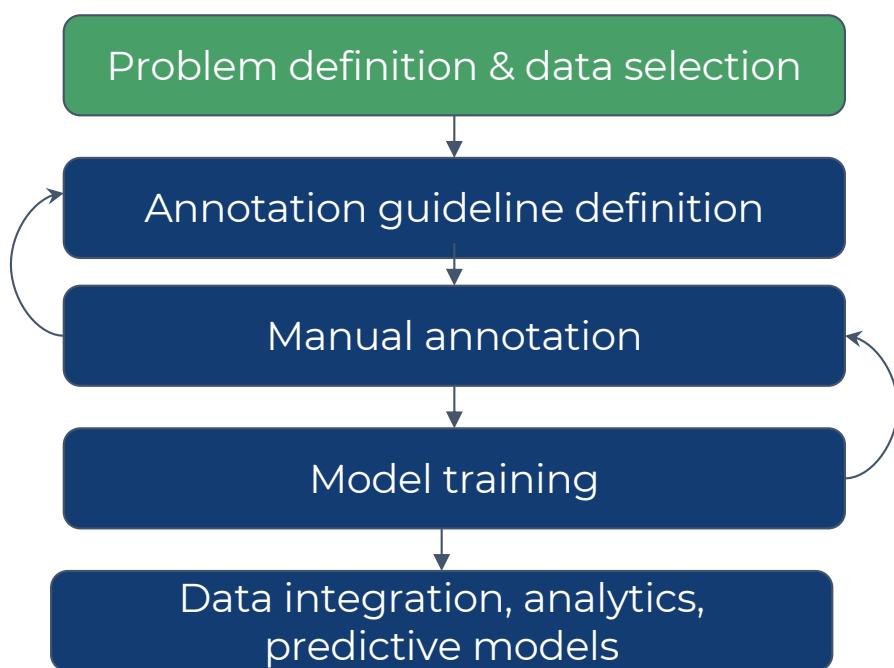


LLMs should be able to reference and follow the labelling instructions like humans

## Critical aspects for Corpus construction



# Data annotation pipeline

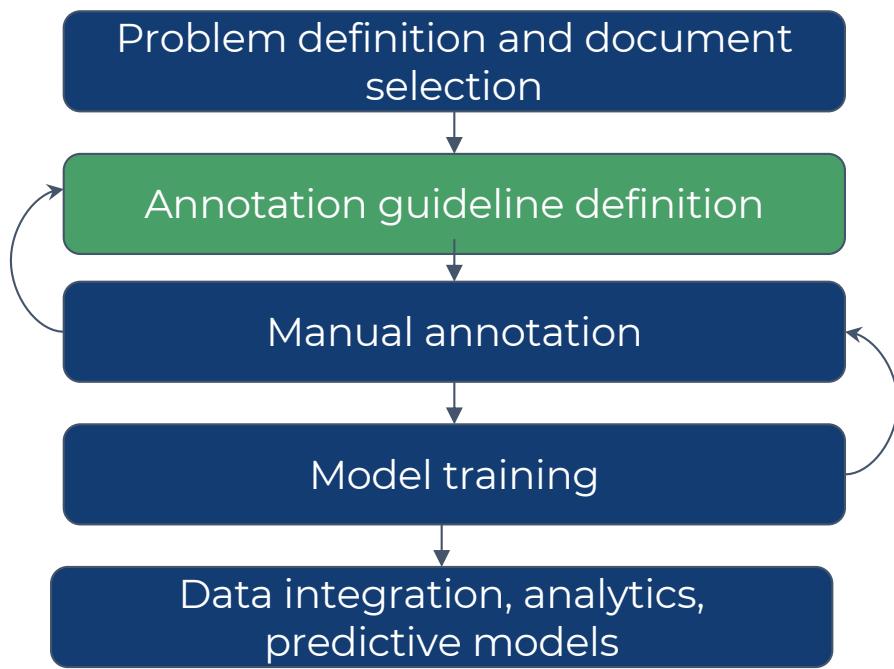


- Search of a “knowledge gap”
- Definition of use cases
- Discussion with potential end users
- Selection of relevant reports (often semi-manual or meta-based)

## Two typical scenarios:

- A. **Exhaustive extraction of a predefined list of clinical variables**
- B. **Large scale data structuring using NLP**

# Data annotation pipeline



- Written protocol on how to label data
- Input of experts (e.g, doctors, linguists,...)
- Continuously refined during data annotation
- Important for quality control, reproducibility, consistency and interpretability
- Also mapping to controlled vocabularies for data harmonization/normalization & interoperability
- Controlled vocabularies: SNOMED CT, MeSH, ICD10, Human Phenotype ontology, ESCO,...

## Diseases



**Guías DISTEMIST:**  
Anotación y normalización de enfermedades en textos clínicos

V1 [Abril 2022]

AUTORES

Eulàlia Farré-Maduell (Barcelona Supercomputing Center)  
Luis Gascó Sánchez (Barcelona Supercomputing Center)  
Salvador Lima López (Barcelona Supercomputing Center)  
Antonio Miranda-Escalada (Barcelona Supercomputing Center)  
Martin Krallinger (Barcelona Supercomputing Center)



Este estudio ha sido realizado dentro del ámbito del Plan de Impulso de las Tecnologías del Lenguaje con financiación de la Secretaría de Estado para el Avance Digital. Dicho contenido es de uso exclusivo de sus autores. Reservados todos los derechos. Se permite su copia y distribución por cualquier medio siempre que se mantenga el reconocimiento de sus autores, no se haga uso comercial de las obras y no se realice ninguna modificación de las mismas.

## Places



**Guías MEDDOPLACE:**  
Anotación, normalización y clasificación de lugares e información relacionada en textos clínicos

V1 [Marzo 2023]

AUTORES

Salvador Lima López (Barcelona Supercomputing Center)  
Eulàlia Farré-Maduell (Barcelona Supercomputing Center)  
Vicent Briva-Iglesias (Dublin City University)  
Martin Krallinger (Barcelona Supercomputing Center)



Este estudio ha sido realizado dentro del ámbito del Plan de Impulso de las Tecnologías del Lenguaje con financiación de la Secretaría de Estado para el Avance Digital, que no comparte necesariamente los contenidos expresados en el mismo. Dichos contenidos son responsabilidad exclusiva de sus autores. Reservados todos los derechos. Se permite su copia y distribución por cualquier medio siempre que se mantenga el reconocimiento de los autores, no se realice ninguna modificación de las mismas.

## Species



**Guías LivingNER:**  
anotación, normalización y clasificación de especies, seres vivos y patógenos/enfermedades infecciosas

(LivingNER Guidelines: annotation, mapping and classification of pathogens, living organisms and infectious diseases)

V3 [Abril 2022]

AUTORES

Eulàlia Farré-Maduell (BSC)  
Salvador Lima López (BSC)  
Gloria González (BIFI-CSIC)  
Antonio Miranda-Escalada (BSC)  
Martin Krallinger (BSC)



Este estudio ha sido realizado dentro del ámbito del Plan de Impulso de las Tecnologías del Lenguaje con financiación de la Secretaría de Estado para el Avance Digital, que no comparte necesariamente los contenidos expresados en el mismo. Dichos contenidos son responsabilidad exclusiva de sus autores. Reservados todos los derechos. Se permite su copia y distribución por cualquier medio siempre que se mantenga el reconocimiento de los autores, no se realice ninguna modificación de las mismas.

## Clinical procedures



**Guías MedProcNER/ProcTEMIST:**  
Anotación y normalización de procedimientos en textos clínicos

V1 [Marzo 2023]

AUTORES

Eulàlia Farré-Maduell (Barcelona Supercomputing Center)  
Luis Gascó Sánchez (Barcelona Supercomputing Center)  
Salvador Lima López (Barcelona Supercomputing Center)  
Antonio Miranda-Escalada (Barcelona Supercomputing Center)  
Martin Krallinger (Barcelona Supercomputing Center)



Este estudio ha sido realizado dentro del ámbito del Plan de Impulso de las Tecnologías del Lenguaje con financiación de la Secretaría de Estado para el Avance Digital, que no comparte necesariamente los contenidos expresados en el mismo. Dichos contenidos son responsabilidad exclusiva de sus autores. Reservados todos los derechos. Se permite su copia y distribución por cualquier medio siempre que se mantenga el reconocimiento de los autores, no se realice ninguna modificación de las mismas.



**MEDDOPROF:** Identificación de Profesiones y Ocupaciones en Casos Clínicos

Guías de anotación

Plan de impulso de las Tecnologías del Lenguaje

[E. Farré-Maduell, M. Krallinger, A. Miranda, S. Lima, M. Agüero] y [V. Briva-Iglesias]

**Profession/occupation**



**Guías de anotación de información de salud protegida**

Plan de impulso de las Tecnologías del Lenguaje

Enrique Mota<sup>1</sup>, Nelson Martín<sup>2</sup>, Ángel Moreno<sup>3</sup>, Elvira Ferrete<sup>4</sup>, Jesús Santamaría<sup>4</sup>, Montserrat Marimon<sup>4</sup>, Ander Intxaurreondo<sup>5</sup>, Aitor González-Agiro<sup>6</sup>, Marta Villegas<sup>6</sup>, Martin Krallinger<sup>2,4</sup>

1 Indien Technologies

2 Hospital Universitario "12 de Octubre"

3 Centro Nacional de Investigaciones Oncológicas

4 Centro Nacional de Supercomputación

10 - 2018

**Anonymization**



**GUÍA DE ANOTACIÓN Y NORMALIZACIÓN DE COMPUESTOS QUÍMICOS**

Plan de impulso de las Tecnologías del Lenguaje

Obdulia Rabal

Ander Intxaurreondo

Martin Krallinger

Julio 2018

**Chemicals, drugs, genes, proteins**



**GUIA DE ANOTACIÓN DE MORFOLOGÍAS NEOPLÁSICAS**

Junio 2020

Versión 1.3

Autores

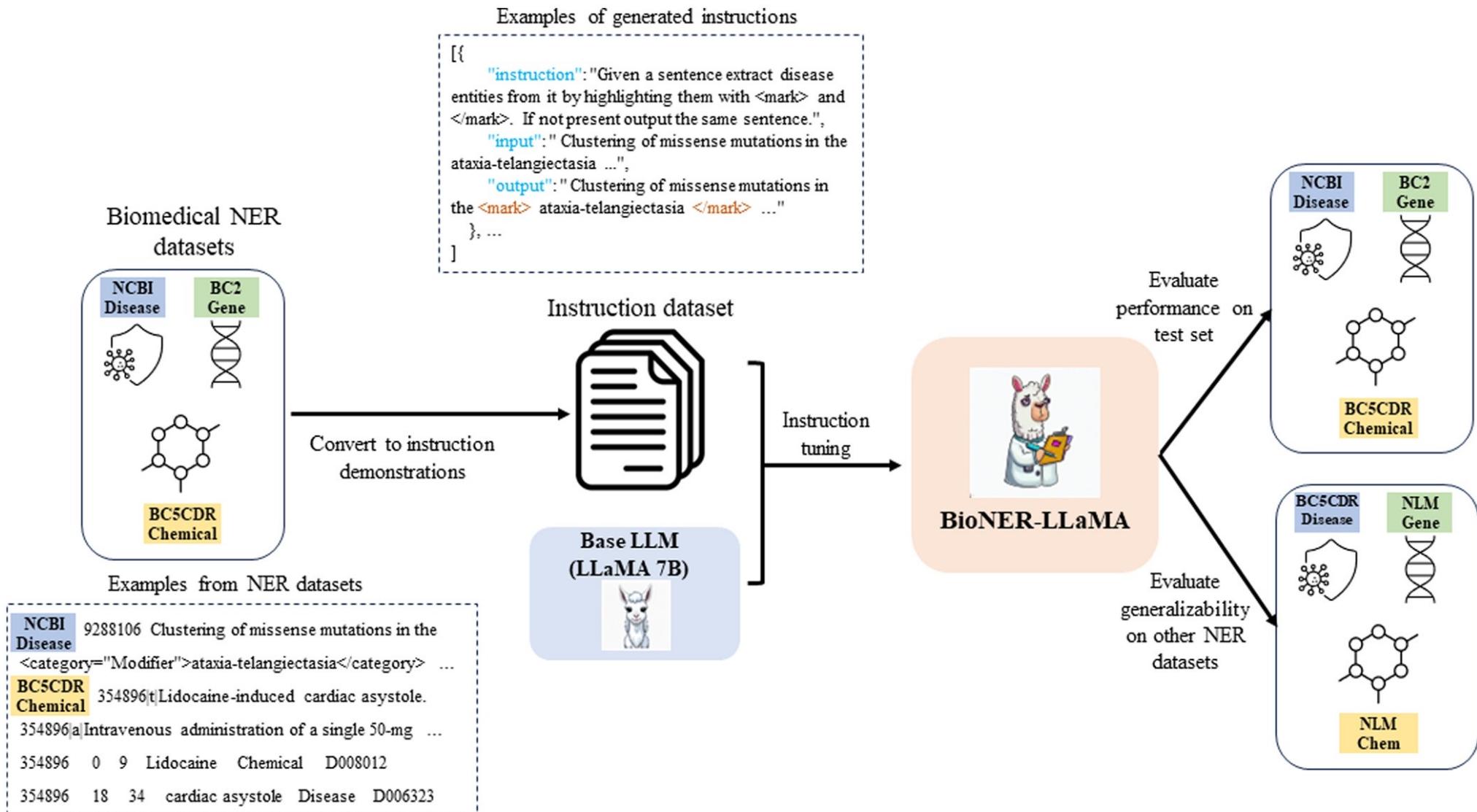
Eulàlia Farré  
Gloria González  
Martin Krallinger  
Toni Mas  
Antonio Miranda

**Tumor morphology**

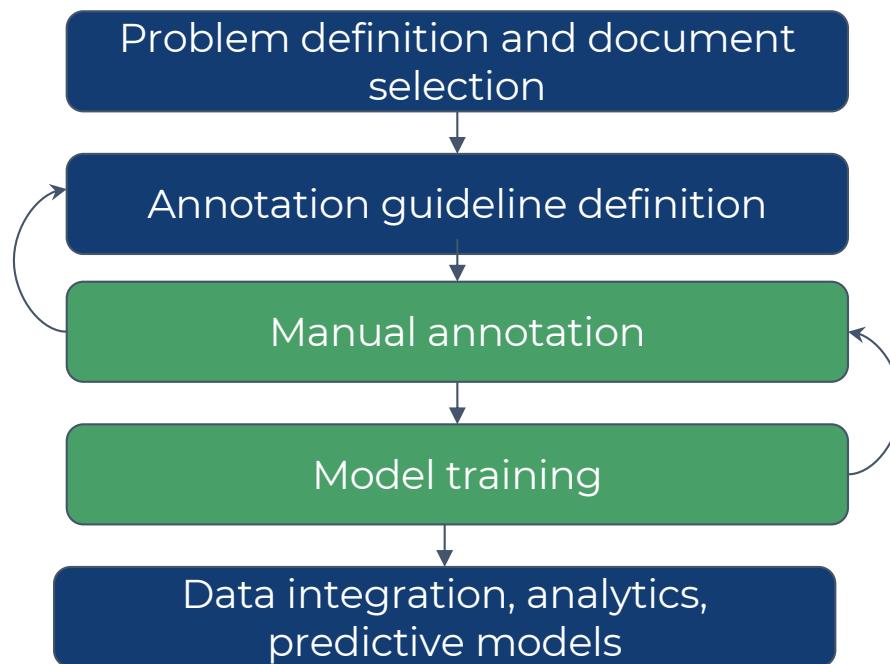
# Data annotation rules/criteria

[EN-N3] [severidad]	No se deben incluir dentro de la mención modificadores relacionados con severidad: "fatal", "severo", "leve", ...  [ejemplos]  88. Lo trasladaron con el diagnóstico de <b>neumonía tuberculosa</b> severa  Lo trasladaron con el diagnóstico de <b>neumonía tuberculosa severa</b>	[G5] [errores-ortográficos]	Las menciones que incluyan algún tipo de error ortográfico (por ejemplo: letras de más o de menos, espacios incorrectos) también deben anotarse.  [ejemplos]  4. <b>Heptitis B</b> ( <i>Hepatitis B</i> ) 5. <b>Poliomelitis vacunal</b> ( <i>Poliomielitis vacunal</i> ) 6. <b>Osteo mielitis</b> ( <i>Osteomielitis</i> )
[EN-N4] [patología-onco]	No anotaremos patologías oncológicas de causa infecciosa.  [ejemplos]  89. Por positividad a <b>HHV8</b> se estableció el diagnóstico de <b>Sarcoma de Kaposi</b> clásico.  90. ... <b>linfomas</b> con la estimulación antigenica crónica por otras <b>infecciones</b> como <b>H. pylori, VEB o VHH8</b> .	[G6] [palabras-completas]	Todas las menciones deben estar compuestas por palabras enteras. No se pueden anotar palabras a medias.  [ejemplos]  7. <b>Abceso laterocervical</b>  <b>Abceso laterocervical</b>

- Annotation rule type; General, Positive, Negative, linguistic, normalization
- Each rule: unique Id, short name, description/definition, examples
- Iterative guideline refinement, versioning, introduction, required expertise or annotators
- Translation to other languages: English, Italian, Dutch, Swedish, Romanian, Czech

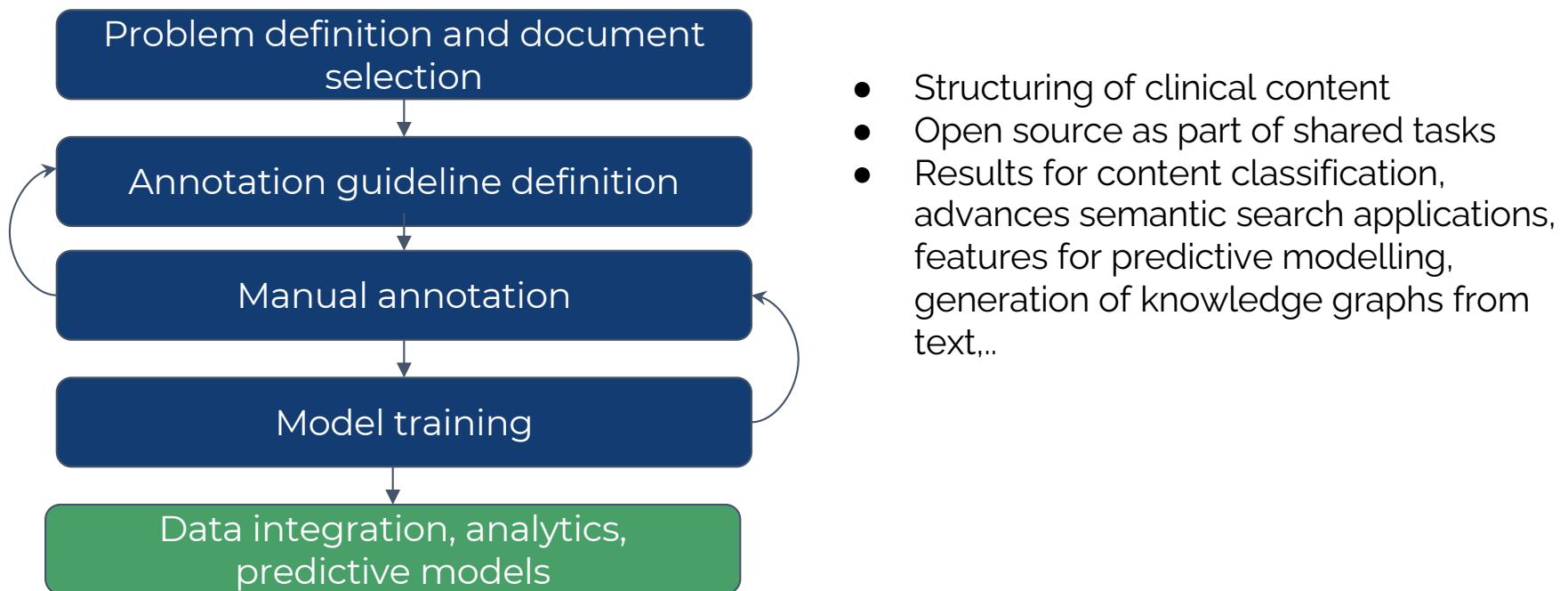


# Data annotation pipeline



- Parallel annotations and inter-annotator agreement
- Use of pre-annotations to bootstrap manual annotation
- Intermediate model training for better pre-annotations and validation
- Most expensive in terms of time and effort

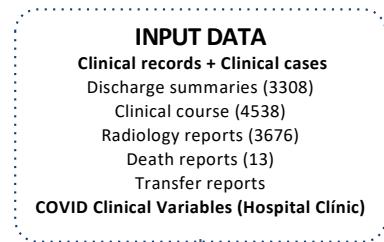
# Data annotation pipeline



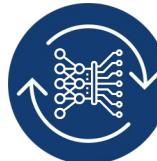
# Clinical NLP processing of patient records

GOAL: Structuring of written clinical reports

## 1. Baseline



## 2. Iterative AI refinement



*"tag a little, train a little"*

### Improve baseline model by

#### Annotating more data

- Improve data generation protocol for AI refinement
- Adaptation of Guidelines
- Better document selection
- Annotation of samples together with hospital prof.

#### System modifications

- NER improvement (post-processing, transformers, embeddings, attention layers)
- Mapping improvement (dep learning-based entity grounding techniques)

## 3. Results



**Structured Health Records with clinical variables from Hospital**

"El paciente tiene fiebre, no disnea, sospecha de COVID-19"

"El paciente tiene fiebre, no **disnea**, sospecha de **COVID-19**"

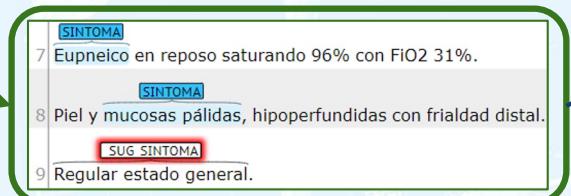
Record #01		
Affirmed	Negated	Uncertain
fever	dyspnea	COVID-19

# Clinical expert/linguist correction



## Text Selection and Preparation (BSC)

Semi-manual document selection by BSC using documents with different properties (pre-annotation frequency, doc. length) and special focus on concepts requested by HC



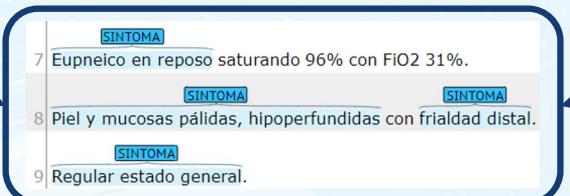
Pre-annotated text



## Annotation Sessions (clinical expert)

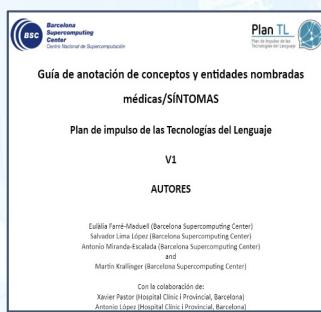
## Correction (Clínica-BSC)

The annotated documents are used to improve the baseline model and generate new pre-annotations for next stage



Pre-annotated text

- + manually annotated and reviewed
- + additional classes classification



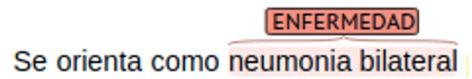
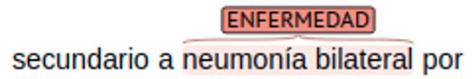
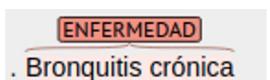
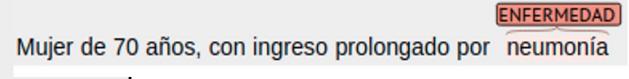
Annotation tool training

Annotation Guidelines Refinement +  
Addition of Extra Attributes

Individual Annotations for IAA

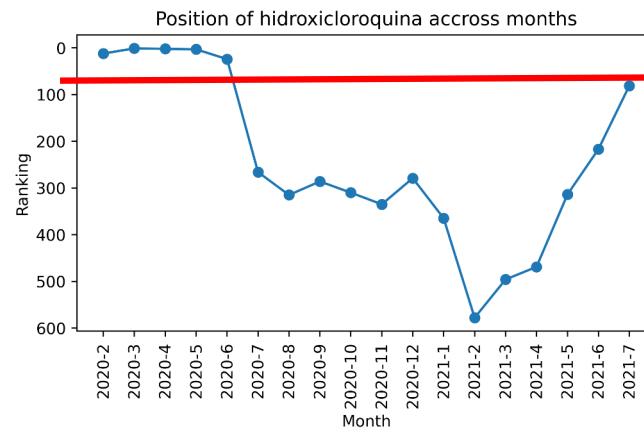
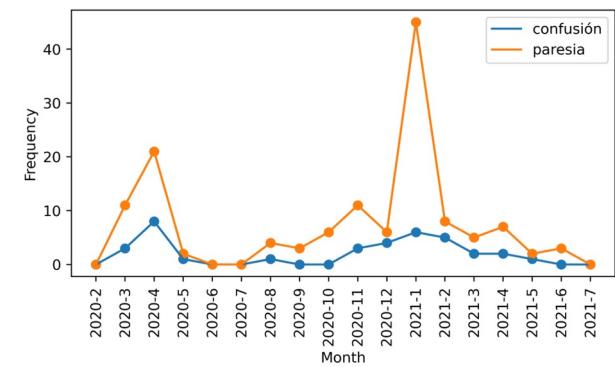
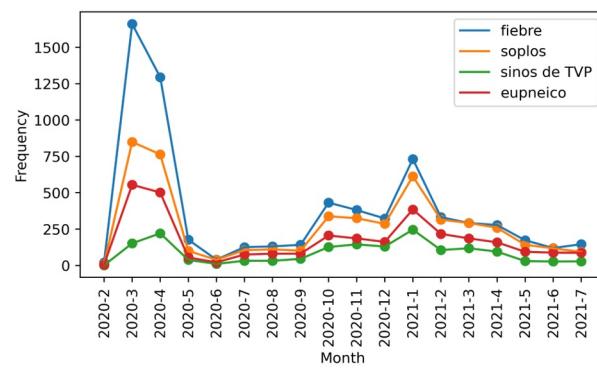
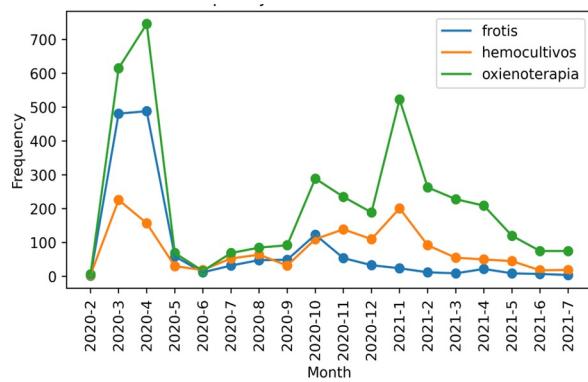


## Extracted disease mentions: Hospital Clinic of Barcelona reports

Type of report	# of doc.	# of tokens	Extracted diseases mentions		Unique diseases names		Example
			model initial	retrained	model initial	retrained	
Discharge summary	5333	5064727	64764 neg: 20% spe: 11.6%	73505 <u>neg: 18.5%</u> <u>spe: 10.4%</u>	15651	21164	 Snomed CT ID: 407671000
Exitus reports	23	4946	213 neg: 6.6% spe: 6.6%	224 <u>neg: 5.8%</u> <u>spe: 6.7%</u>	167	181	 Snomed CT ID: 407671000
Clinical course	6085	46614881	241232 neg: 8.2% spe: 5.8%	345933 <u>neg: 6.5%</u> <u>spe: 4.4%</u>	38755	44255	 Snomed CT ID: 63480004
Radiology reports	5637	659618	29595 neg: 18.9% spe: 14.6%	32292 <u>neg: 19.3%</u> <u>spe: 13.3%</u>	7519	7553	 Snomed CT ID: 233604007
Transfer reports	1021	277184	4877	6208	1678	2039	

# Issues: variability across time

Ranking of entities (e.g, hidroxicloroquina) and frequency over time



# Automatic clinical entity detection

> 40 entity classes

ORG VIVO	PACIENTE-PROFESION	ENFERMEDAD	ENFERMEDAD	FARMACO	NEG
Varón	de 50 años, transportista de alimentos empaquetados.	Exfumador, hipertenso	en tratamiento con Enalapril sin		
NSCO	ENFERMEDAD	SINTOMA	SINTOMA	SINTOMA	SINTOMA
otros antecedentes de interés o hábitos tóxicos. Ingrera por cuadro de 4 días de evolución de	astenia, hiporexia, dispepsia, cefalea, mialgias,				
SINTOMA	SINTOMA	SINTOMA	NEG	NSCO	NEG
SINTOMA	SINTOMA	SINTOMA	Niega	fiebre.	Niega consumo de drogas, productos de parafarmacia, setas silvestres o nuevos fármacos. Niega
ORG VIVO	NEG	ORG VIVO	ORG VIVO	NSCO	ORG VIVO
Presentamos el caso de una	mujer	de 38 años, sin	antecedentes personales ni familiares de interés y de profesión	pescadera	desde
SINTOMA	ENFERMEDAD				
los 17 años. Acude a urgencias por un cuadro de prurito generalizado y lesiones habonosas confluentes por todo el cuerpo, precisando dosis altas					
PROCEDIMIENTO	PROCEDIMIENTO				
de corticoides	y antihistamínicos para su cese. Un mes más tarde vuelve a reproducirse idéntica sintomatología, acompañada además en esta				
SINTOMA	SINTOMA	FARMACO	PROCEDIMIENTO	ORG VIVO	
ocasión de poliartralgias y rigidez, sobre todo en rodillas y tobillos, con buena respuesta a indometacina. En la anamnesis, la paciente refería					
HUMAN	AGE	ENFERMEDAD			
AF: Hermano fallecido a los 35 años por	IAM.				
SEXO-SUJETO-ASISTENCIA	EDAD-SUJETO-ASISTENCIA	SINTOMA	SINTOMA	DURATION	
DATE	HUMAN	DATE	AGE	DATE	
Enfermedad actual:	Varón	de	42 años	que presenta cuadro de dolor epigástrico y precordial opresivo asociado a cortejo vegetativo de	3 horas de duración.
ENFERMEDAD	PROCEDIMIENTO				
El mes previo había presentado un cuadro de	sinusitis	tratado con	antibiótico.		
FAC	PROCEDIMIENTO				
Ante la clínica que presenta, acude a su centro de salud, donde realizan un	ECG.				

## Clinical information:

- Symptoms
- Diseases
- Procedures
- Drugs
- Organisms
- Tumor morphology
- Chemicals& proteins
- Observable entities....

## Linguistic modifiers:

- Negation
- Speculation
- Temporality
- ...

## Sociodemographic information:

- Locations
- Occupations
- Toxic habits
- Sensitive data
- ...

# From unstructured...

## ANTECEDENTES

Recién nacido pre-término (RNPT) de 32 semanas de gestación (SG), peso al nacimiento 1.740 g (P50-P75), ingresada en UCI neonatal. Desde el nacimiento presenta crisis mioclónicas que llegan a status epiléptico, asociadas a taquicardia, hipertensión, hipertermia y quejido sin distrés, acompañadas de un patrón de brote/supresión en el electroencefalograma integrado por amplitud (EEGa).

**ANTECEDENTES OBSTÉTRICOS:** Padres cosanguíneos (primos hermanos). 3 embarazos previos, de los cuales 1 resultó en aborto. Antecedente de hermano prematuro (26+6 SG) fallecido a los 22 días por sepsis nosocomial, presentando en primeros días de vida anemización gradual con necesidad de transfusión.

Hermana sana. Ingreso a las 29 SG por oligoamnios grave y febrícula, descartándose rotura prematura de membranas. Maduración pulmonar completa. Remitida al Hospital San Cecilio de Granada con 29+5 SG por sospecha ecográfica de anemia fetal severa con alteración del doppler, realizándose transfusión fetal (Hb pre = 9,5 g/dL; Hb post = 15 g/dL) y tomándose muestra fetal para cariotipo, fenotipo eritrocitario y PCR de CMV y Parvovirus. Revalorada con 31+4 SG en dicho centro ante nueva sospecha de anemia fetal grave con sospecha de hemorragia feto-materna, sin indicación de nueva transfusión. Cesárea electiva a las 32 SG por riesgo de pérdida de bienestar fetal. Nace con escaso esfuerzo respiratorio precisando CPAP para traslado a la UCI neonatal.

# ... to structured clinical content



Each mention is linked to an ontology entry (normalized)  
> 45 semantic classes/entity types (incl. negation, temporal expressions, symptoms,..)

## ANTECEDENTES

Recién nacido pre-término (RNPT) de 32 semanas de gestación (SG), peso al nacimiento 1.740 g (P50-P75), ingresada en UCI neonatal. Desde el nacimiento presenta crisis mioclónicas que llegan a status epiléptico, asociadas a taquicardia, hipertensión, hipertermia y quejido sin distrés, acompañadas de un patrón de brote/supresión en el electroencefalograma integrado por amplitud (EEGa).

ANTECEDENTES OBSTÉTRICOS: Padres consanguíneos (primos hermanos). 3 embarazos previos, de los cuales 1 resultó en aborto. Antecedente de hermano prematuro (26+6 SG) fallecido a los 22 días por sepsis nosocomial, presentando en primeros días de vida anemización gradual con necesidad de transfusión.

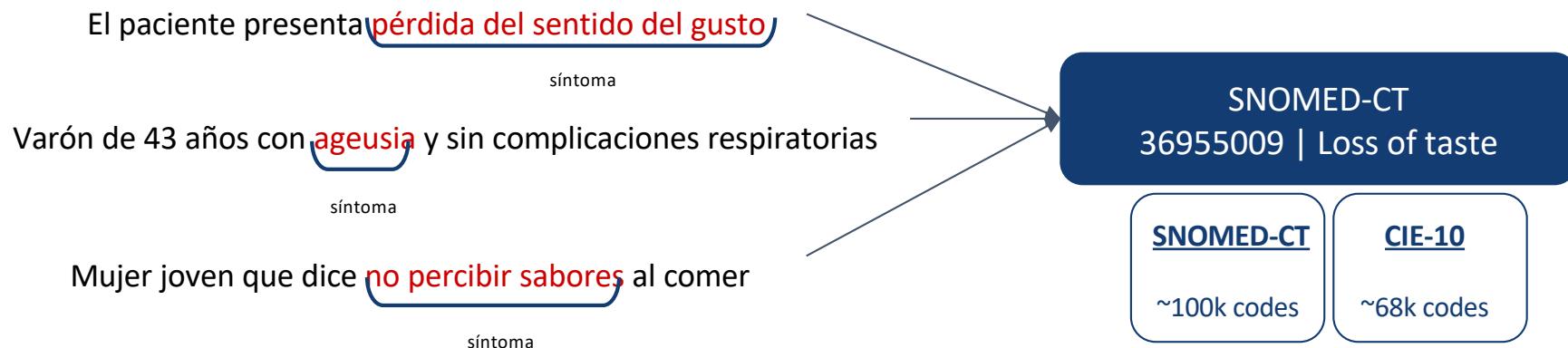
Hermana sana. Ingreso a las 29 SG por oligoamnios grave y febrícula, descartándose rotura prematura de membranas. Maduración pulmonar completa. Remitida al Hospital San Cecilio de Granada con

29+5 SG por sospecha ecográfica de anemia fetal severa con alteración del doppler, realizándose transfusión fetal (Hb pre = 9,5 g/dL; Hb post = 15 g/dL) y tomándose muestra fetal para cariotipo, fenotipo

eritrocítario y PCR de CMV y Parvovirus. Revalorada con 31+4 SG en dicho centro ante nueva sospecha de anemia fetal grave con sospecha de hemorragia feto-materna, sin

indicación de nueva transfusión. Cesárea electiva a las 32 SG por riesgo de pérdida de bienestar fetal. Nace con escaso esfuerzo respiratorio precisando CPAP para traslado a la UCI neonatal.

# Medical Entity Linking or mapping (normalization) intro



Harmonization



Interoperability



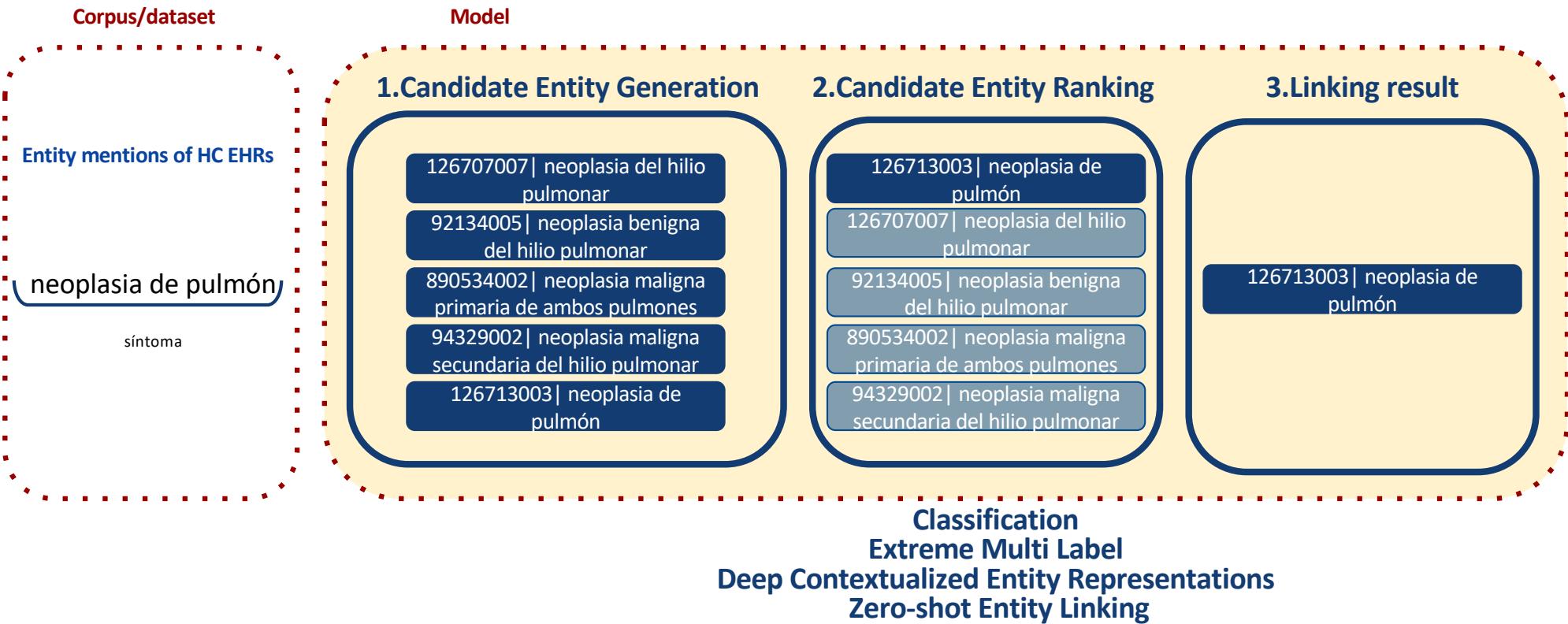
Data integration



Data Mining

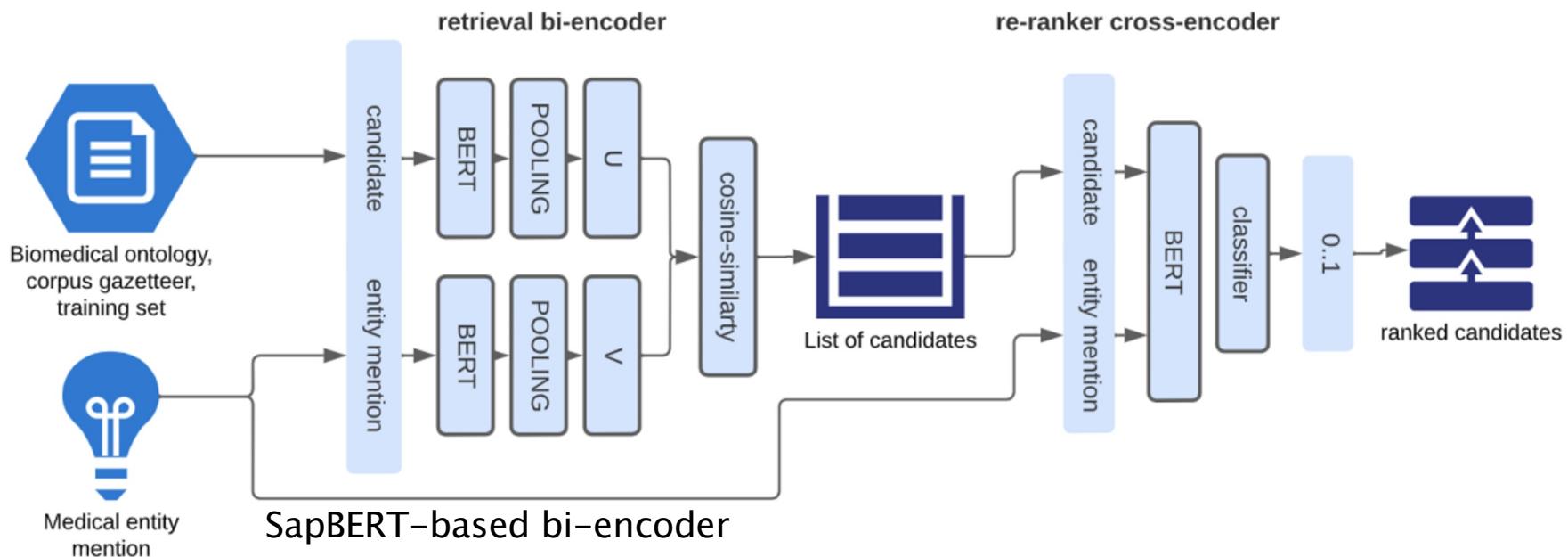
**Medical Entity Linking** (or Entity Normalization): linking or mapping mentions of clinical concepts found in text to standardized controlled vocabularies/terminologies

# Clinical Entity Linking System Design



# ClinLinker clinical concept normalization & mapping

Schema of the implemented ClinLinker pipeline



# Medical Entity Linking

## Background

Many structured vocabularies: multilingual resources for Entity Linking in Health

### SNOMED-CT

Collection of medical terms with codes, synonyms and definitions used in clinical documentation

~100k codes

### CIE-10

Medical classification structured vocabulary

~68k codes

### MeSH

Scientific documents indexing

~27k entries

### HPO

Medical genetics and genomics structured vocabulary

~13k terms  
~156k annotations

### IPC

Patents indexing

~76k groups

# Medical Entity Linking: Lack of Manually annotated Entity Linking corpora!

Limited number of Gold Standard datasets and not sufficiently addresses in benchmarking leaderboards:

- [n2c2/UMass Track on Clinical Concept Normalization Task 3](#): Track on Clinical Concept Normalization
- [Bacteria Biotope at BioNLP-OST 2019 Task](#): Biological Entity Linking



Most top-performing systems<sup>1,2</sup> used Neural Language Models

- BERT<sup>3</sup>
- XLNet<sup>4</sup>

- [1]. IHan, J. C., & Tsai, R. T. H. (2020). NCU-IISR: Pre-trained Language Model for CANTEMIST Named Entity Recognition. In Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR Workshop Proceedings.
- [2]. Wang, Y., Fu, S., Shen, F., Henry, S., Uzuner, O., & Liu, H. (2020). The 2019 n2c2/OHNLP Track on Clinical Semantic Textual Similarity: Overview. *JMIR Medical Informatics*, 8(11), e23375.
- [3]. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- [4]. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems* (pp. 5753-5763).

# Common data Model (CDM) of clinical NLP output: FHIR NLP profile

## DT4H FHIR NLP-profile

Fields suffixed with a “\*” are mandatory values

Represents the individual text analysis results.

- Clinical site identifier\*:
  - Values:
- Patient ID\*
- Record ID\*
- Admission or contact ID
- Record type\*
  - Values: progress report, discharge summary, discharge letter...
- Record format\*
  - Values: txt, PDF, XML, Json, docx
- Date record created\*
- Date record last revised
- Record character encoding
  - Values: ASCII, UTF-8, UTF-16, UTF-32, No encoding, Unknown, ...
- Extraction date
- NLP-processing date\*
- NLP-processing date update
- NLP-processing pipeline name\*
- NLP-processing pipeline version\*
- Report section
- Report language\*
  - Values: en, nl, es, it, cs, ro, sv, ca
- De-identified text\*
  - Value: yes/no
- De-identification pipeline name
- De-identification pipeline version
- Text\*
- Annotations

## Information contained within the annotation:

- Concept class
  - Values: symptom, disorder/disease, procedure, medication, cardiology entity, other
  - Concept in pre-defined clinical variable list (based on defined data-dictionary) Y/N
- Start offset (type ‘int’)
- End offset (type ‘int’)
- Concept mention string (type ‘str’)
- Concept confidence/likelihood
- NER component type
  - dictionary lookup, transformer, other
- NER component version (type ‘str’)
- Value Negation
  - Values: yes, no
- Negation confidence
- Qualifier Negation
- Qualifier Temporal
- DT4H concept identifier (type ‘str’)
- NEL component type
  - lexical similarity, transformer, other
- NEL component version (type ‘str’)
- Controlled vocabulary namespace (type ‘str’)
  - Values: UMLS, SNOMED CT, ICD10, MedDRA, ICD9, DT4H, HPO, LOINC, ISO, GeoNames, MeSH, ESCO, ATC, ICPC, other, none.
- Controlled vocabulary version (type ‘str’)
- Controlled vocabulary concept identifier (type str)
  - Value: ranked list
- Controlled vocabulary concept official term (type ‘str’)
- Controlled vocabulary source (type ‘str’)
  - original, machine translation, manual translation

[Plot network](#) [Refresh session](#)

## Network visualization panel

Filter by co-occurrence count:



Select labels:

- PROCEDIMIENTO
- SINTOMA
- FARMACO
- EDAD-SUJETO-ASISTENCIA
- ENFERMEDAD
- ENTIDAD-OBSERVABLE
- FECHAS
- PROTEINAS
- SPECIES
- TERRITORIO
- PAIS
- SEXO-SUJETO-ASISTENCIA
- ID-SUJETO-ASISTENCIA
- CALLE
- HOSPITAL

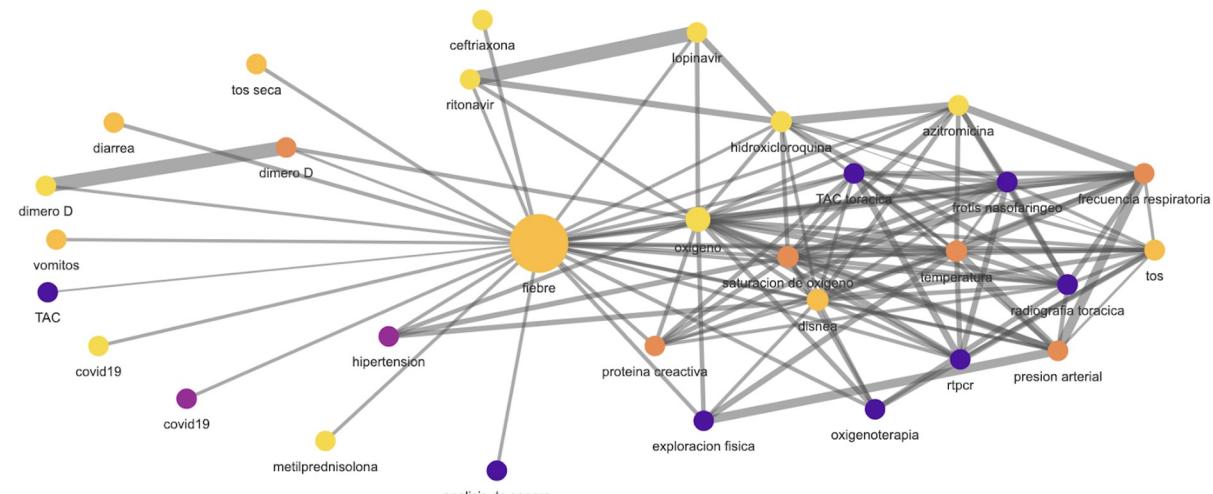
### Network summary

num of edges    num of nodes

121    31

Highlight label ▼

## Clinical knowledge graph (KD, hypothesis generation)



[Co-occurrence](#) [Node ranking](#) [Communities](#)

word1	word2	count word1	count word 2	count of co-occurrences	PMI
lopinavir	ritonavir	38	40	38	2.32
dimero D (entidad-observable)	dimero D (farmaco)	38	29	27	2.29
frecuencia respiratoria	presion arterial	50	52	33	1.34
exploracion fisica	presion arterial	43	52	25	1.16
azitromicina	hidroxicloroquina (farmaco)	47	69	35	1.11
frecuencia respiratoria	temperatura	50	74	36	0.96

# Co-occurrence network example: occupational health



77



**Barcelona  
Supercomputing  
Center**  
Centro Nacional de Supercomputación

Luis Gasco, Darryl Estrada-Zavala, Eulàlia Farré- Maduell, Salvador Lima-López, Antonio Miranda- Escalada, and Martin Krallinger. 2022. Overview of the SocialDisNER shared task on detection of diseases mentions from healthcare related and patient generated social media content: methods, evaluation and corpora. In Proceedings of the Seventh Social Media Mining for Health (#SMM4H) Workshop and Shared Task



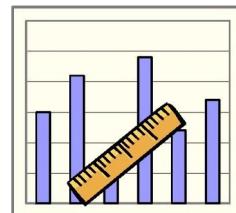
# Shared tasks & evaluation of biomedical NLP systems



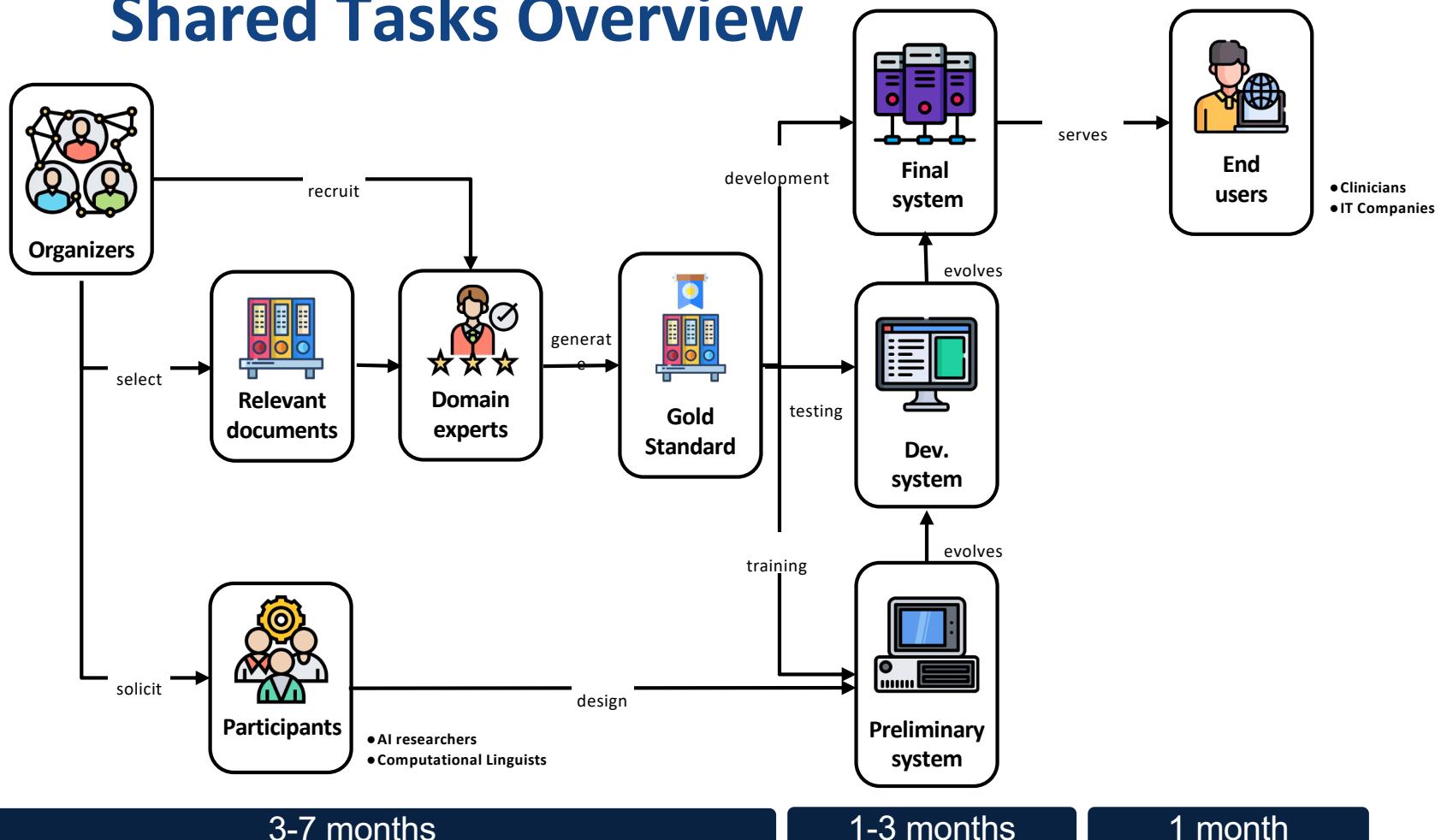
*Barcelona  
Supercomputing  
Center*  
Centro Nacional de Supercomputación

# Importance of shared tasks

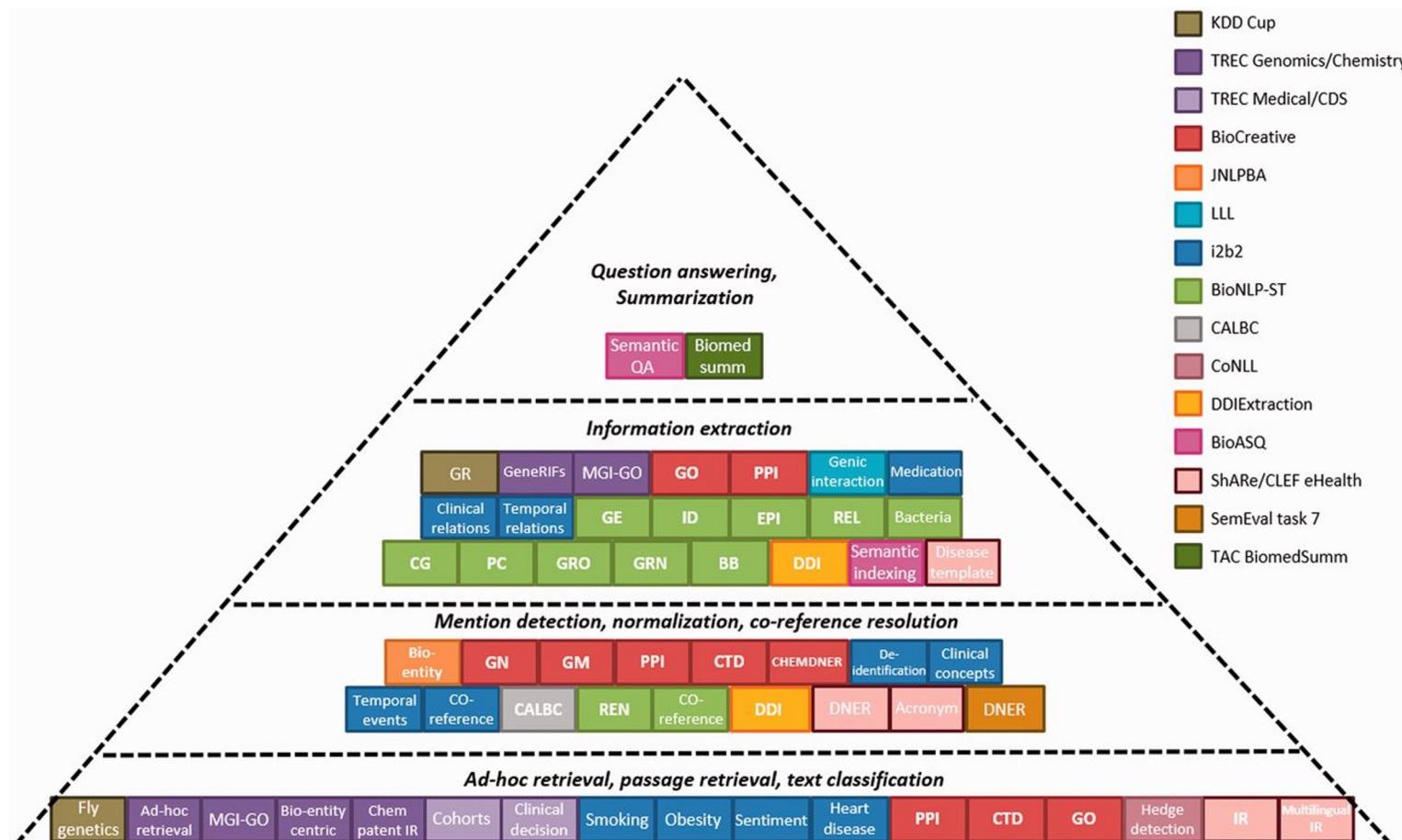
- Objective and independent benchmarking
- Interoperability, harmonization of resources and combined systems
- Generation of new resources and technologies
- Building trust and quality for solutions (especially for health domain!)
- Promote the development of both commercial & academic solutions
- Design decision support (what works & what doesn't)
- Reproducibility, replicability, interoperability, scalability, sustainability
- Generalizability & adaptability of methodologies and systems



# Shared Tasks Overview



# NLP shared task in biomedicine



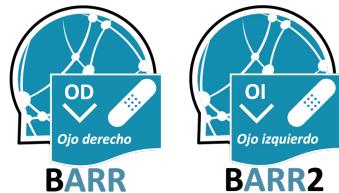
Barcelona  
Supercomputing  
Center  
Centro Nacional de Supercomputación

Chung-Chi Huang, and Zhiyong Lu Brief Bioinform  
2015;bib.bbv024

Briefings in  
Bioinformatics

# Overview Medical NLP Shared tasks

Medical abbreviation detection and resolution



Semantic indexing:  
literature, patents,  
clinical trials, projects



Medical document  
anonymization



MEDDOCAN

Detection of drugs,  
chemicals, genes



PharmaCoNER

Detection & clinical  
coding of tumor  
morphology



CANTEMIST

Clinical coding: ICD-10: diagnosis & procedures



CODIESP

Diseases in health  
social media



SocialDisNER

Profession, occupation  
detection in health  
social media



ProfNER

## Some of our past Shared Tasks (II)



MEDDOPROF

Detection of occupations and laboral statuses + normalization to SNOMED CT and ESCO

<https://temu.bsc.es/meddoprof/>



ClinSpEn

English <-> Spanish translation of clinical case reports, terminology and ontologies  
(biomedical WMT subtrack)

<https://temu.bsc.es/clinspen/>



MEDDOPLACE

Detection of places, clinical departments and related content + normalization to SNOMED CT, GeoNames and PlusCodes

<https://temu.bsc.es/meddoplace/>

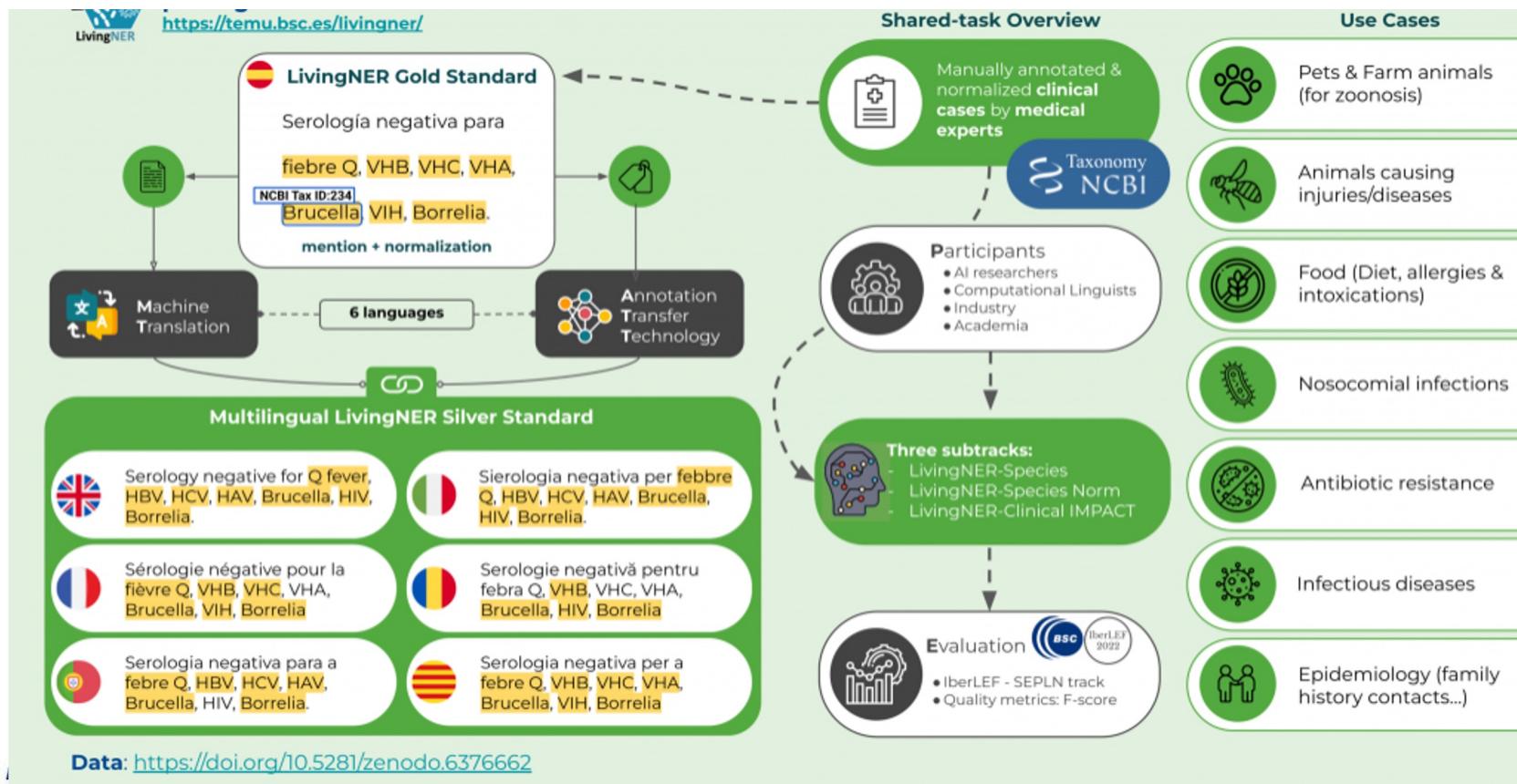


LivingNER

Detection of pathogens and living beings + normalization to NCBI-Taxonomy

<https://temu.bsc.es/livingner/>

# LivingNER Multilingual Silver Standard



**Supercomputing Center**  
Centro Nacional de Supercomputación

Miranda-Escalada, Antonio, et al. "Mention detection, normalization & classification of species, pathogens, humans and food in clinical documents: Overview of the LivingNER shared task and resources." *Procesamiento del Lenguaje Natural* 69 (2022): 241-253.

## LivingNER participant results

- MiF: micro-averaged F-score (main metric)
- MiP: micro-avg. Precision
- MiR: micro-avg. Recall

Top team: 0.951 f-score for NER,  
0.93 for Normalization

Team Name	SPECIES NER			SPECIES Norm		
	MiP	MiR	MiF	MiP	MiR	MiF
Vicomtech NLP	.9583	<u>.9438</u>	<b>.951</b>	0.9376	<b>0.9234</b>	<b>0.9304</b>
racai	.9569	<b>.9439</b>	<u>.9503</u>	-	-	-
READ-Biomed	.954	.9411	.9475	-	-	-
SINAI	.9571	.9346	.9457	.8733	.8527	.8629
plncmm	.9455	.9373	.9414	.9139	<u>.906</u>	.9099
Sumam Francis	.9443	.9307	.9375	-	-	-
Clac	.9385	.9256	.932	.9495	.891	<u>.9193</u>
john_snow_labs	.916	.9327	.9243	-	-	-
avacaondata	.9228	.908	.9153	.512	.4799	.4954
Pumas	.9284	.8899	.9087	.9389	.8075	.8682
IAM	.9209	<b>.8733</b>	<u>.8965</u>	-	-	-
IGES	.9112	<u>.8638</u>	.8869	.8979	.8512	.874
NLP-CIC-WFU	.8303	.8704	.8499	.7768	.8143	.7951
Vitor	.9492	<u>.5634</u>	<b>.7071</b>	-	-	-
zzz	.8012	.6138	.6951	-	-	-
Kformer-OEG	.7306	.6057	.6623	-	-	-
Mark *pw	.8214	.6145	.703	-	-	-
Han *pw	.5399	.1965	.2881	-	-	-
Sapphire	.6875	.0149	.0291	-	-	-
Boun-ner	0.126	0.078	0.0963	-	-	-
PathoTagIt-Base	0.9461	0.8507	0.8958	-	-	-

Miranda-Escalada, Antonio, et al. "Mention detection, normalization & classification of species, pathogens, humans and food in clinical documents: Overview of the LivingNER shared task and resources." *Procesamiento del Lenguaje Natural* 69 (2022): 241-253.



## MEDDOPLACE

Website: <https://temu.bsc.es/meddoplace/>  
Data: <https://doi.org/10.5281/zenodo.7707567>



IberLEF  
2023

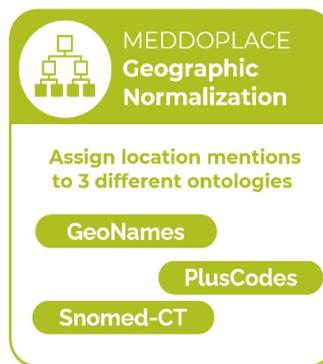
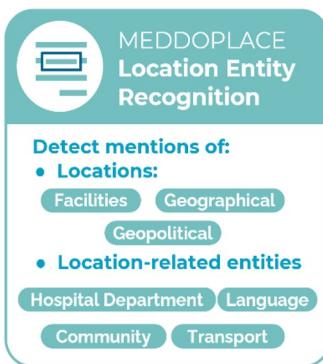
### Medical Document Place-related Content Extraction Shared-Task



#### Gold Standard

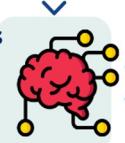
- 1,000 clinical cases
- Clinical and linguistic experts
- +9,000 annotations
- Guidelines with +50 pages

He was referred to the **Neurology Dept.** of the **Hospital 12O (Madrid)** and the **Movement Disorders Unit** of the **Hospital Clinic Barcelona**, detecting continuous muscular hyperexcitability without dystonic or myopathic criteria.



#### Participants

MEDDOPLACE Training Data



Evaluation metrics  
micro-average precision, recall and F-score



Setting a new state-of-the-art in location detection in Spanish clinical documents

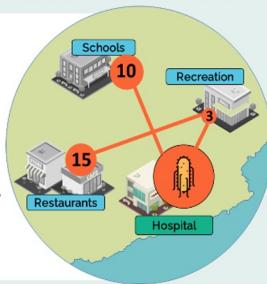


Lima-López, S., Farré-Maduell, E., Brivá-Escalada, V., Gascó, L., & Krallinger, M. (2023). MEDDOPLACE Shared Task overview: recognition, normalization and classification of locations and patient movement in clinical texts. Procesamiento del Lenguaje Natural, 71.

### Example Use Cases

#### Geographic location and health risk factors

He had travelled through Lombardy and Tuscany, travelling in a rented car and staying in private rented houses, before returning to Scotland from Milan on day -2. He was not aware of any contact with COVID-19 cases.



#### Emerging diseases

In the last three months, he traveled through Central America from Colombia to the USA, staying a few weeks in New York where he visited rural and forest areas, returning later to Spain.



#### Patient route through health services

From the doctor's office, she was sent to the primary care dentistry department, from where she was referred to her referral hospital. There, she consults the Haematology Dept., where they cannot solve the anaesthesia problem and refer her to the hospital's Surgery Dept.



# MEDDOPLACE (Dataset for locations)



**MEDDOPLACE**

Medical Document Place-related  
Content Extraction Shared Task  
[temu.bsc.es/meddoplace](http://temu.bsc.es/meddoplace)

## Data annotation protocol (in Spanish & English)

1,000 clinical case reports



**MEDDOPLACE Guidelines:**  
Annotation, Normalization and Classification  
of Locations and Place-Related Information  
in Clinical Texts

V1 [March 2023]

AUTHORS

Salvador Lima López (Barcelona Supercomputing Center)  
Eulàlia Farré-Maduell (Barcelona Supercomputing Center)  
Vicent Briva-Iglesias (Dublin City University)  
Martin Krallinger (Barcelona Supercomputing Center)

Data normalization &  
Semantic interoperability

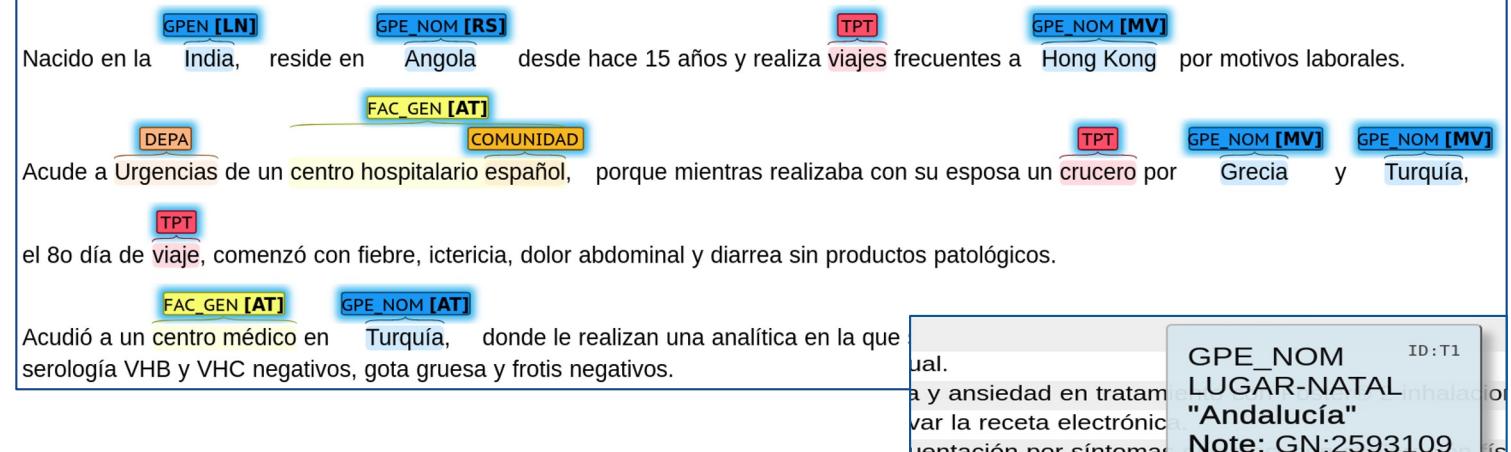


**SNOMED CT**  
• Plus Codes  
**GeNames**



**Barcelona  
Supercomputing  
Center**  
Centro Nacional de Supercomputación

Lima-López, S., Farré-Maduell, E., Brivá-Escalada, V., Gascó, L., & Krallinger, M. (2023). MEDDOPLACE Shared Task overview: recognition, normalization and classification of locations and patient movement in clinical texts. Procesamiento del Lenguaje Natural, 71.



- Around 10,000 annotations distributed in 10 different labels
- Almost all are normalized (Snomed, Geonames,...)
- Further classified in five clinically-relevant classes

Available at: <https://zenodo.org/records/8403498>

Spanish

# MEDDOPLACE example case

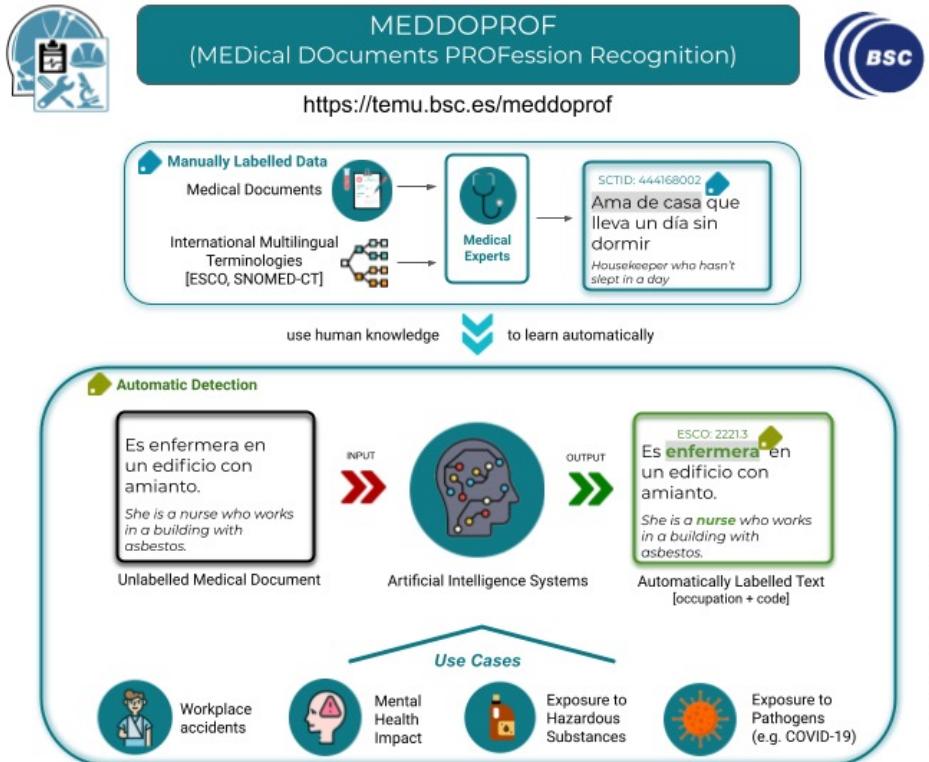
English

1	Antecedentes personales
2	Varón de treinta años, <b>COMUNIDAD</b> , gaditano, fumador de veinte cigarrillos al día, sin criterios de bronquitis crónica y bebedor social.
3	Realizó un viaje turístico tipo "mochilero" con su pareja durante cuarenta días por <b>TRANSPORTE</b>
	Africa noroccidental, visitando tanto zonas rurales como urbanas: Marruecos, <b>GPE_NOM [MV]</b> , <b>GEO_GEN [MV]</b> , <b>GPE_NOM [MV]</b> , Sahara Occidental, Mauritania, <b>GPE_NOM [MV]</b> , <b>GPE_NOM [MV]</b> , <b>GPEN [MV]</b> , <b>GPE_NOM [MV]</b> , Mali y Guinea Ecuatorial.
4	Recibió consejos al viajero en el <b>FAC_GEN [AT]</b> Centro de Salud Internacional, con quimioprofilaxis bien cumplimentada con cloroquina.
5	Recibió vacunación tetravalente meningocócica, fiebre amarilla y tifoidea oral.
6	No refería conductas de riesgo específicas durante el <b>TPT</b> viaje salvo baños en agua dulce en <b>GPE_NOM [MV]</b> , Bamako.
8	Enfermedad actual
9	A las 24 horas tras el regreso, acudió al <b>DEPARTAMENTO</b> Servicio de Urgencias, aquejado de fiebre elevada, con escalofríos, cuadro catarral y cefalea holocraneal.
10	Tras evaluación básica inicial y sin foco aparente, es dado de alta con tratamiento con amoxicilina/clavulánico.

1	Personal history
2	Thirty-year-old man from Cádiz, <b>GPEN</b> smoker of twenty cigarettes a day, without criteria of chronic bronchitis and social drinker.
3	He went on a 40-day backpacking <b>TPT</b> trip with his partner in north-west Africa, visiting both rural and urban areas: Morocco, <b>GEO_GEN</b> , <b>GPE_NOM</b> , <b>GPE_NOM</b> , <b>GPE_NOM</b> , <b>GPEN</b> , <b>GPE_NOM</b> , <b>FAC_GEN</b> and Equatorial Guinea.
4	He received travel advice at the International Health Centre, with well complimented chemoprophylaxis with chloroquine.
5	She received tetravalent meningococcal, yellow fever and oral typhoid vaccinations.
6	She reported no specific risk behaviours during the <b>TPT</b> trip except for bathing in fresh water in <b>GPE_NOM</b> , <b>GPEN</b> .
8	Present illness
9	Twenty-four hours after his return, he presented to the <b>DEPARTAMENTO</b> Emergency Department with high fever, chills, catarrhal symptoms and holocranial headache.
10	After initial basic assessment and with no apparent outbreak, he was discharged with treatment with amoxicillin/clavulanic acid.
11	He consulted again 24 hours later due to persistent febrile symptoms, holocranial headache, profuse sweating, rhinoconjunctivitis and non-productive cough.
12	Physical examination The patient had blood pressure 100/60; heart rate 110 bpm; temperature 37.8°C; baseline O2 saturation 98%; Glasgow score 15/15.

- Multilingual silver standard in 8 languages: Catalan, English, French, Italian, Dutch, Portuguese, Romanian and Swedish

# MEDDOPROF (Dataset for professions, occupations)



- Original motivation: Detect healthcare professionals with COVID
- Almost 2K clinical case reports** in Spanish from variety of specialties
- Manually labelled by clinicians and linguists with mentions of **professions**, **activities** and **working status** and classified according to their holder.
- Inter-annotator Agreement (Quality and consistency): 0.9.

Available at: <https://zenodo.org/records/7116201>

Desde la aplicación de las directivas de confinamiento, refirió que había podido proseguir su trabajo como

PACIENTE-PROFESION

oficinista; su

esposa, una maestra de escuela, y su hijo de 21 años, desarrollador web, también trabajaban desde el hogar.

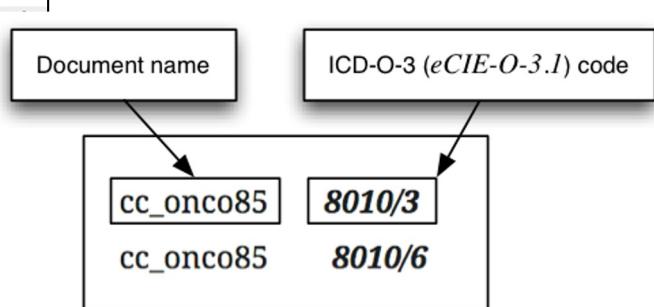
FAMILIAR-PROFESION

FAMILIAR-PROFESION

FAMILIAR-SITUACION LABORAL

# CANTEMIST: oncology corpus (tumor morphology, ICD-O-3)

Tratamiento  
En espera de la realización de una nueva biopsia, se inició tratamiento de primera línea con cisplatino y gencitabina, considerando que el tratamiento con gencitabina podría ser efectivo en tumores de estirpe sarcomatoide.  
Presentó tolerancia regular al tratamiento durante el primer ciclo, con mucositis grado 1, un episodio de fiebre sin focalidad y neutropenia grado 3, que se pospuso el inicio del segundo ciclo.  
Esta clínica obligó también a retrasar la biopsia quirúrgica que estaba prevista.  
En la TC toracoabdominopélvica de revaloración tras dos ciclos, se observó disminución del tamaño de la masa en LSD (de 80 mm a 56 mm) y un aumento de tamaño de la lesión suprarrenal derecha (de 31 mm a 54 mm).  
En octubre de 2016, el Equipo de Cirugía Plástica realizó nueva biopsia con exéresis de músculo piramidal, sin evidencia de malignidad, lo que se realizó una TC/PET para planificar la biopsia con mayor rentabilidad.  
Finalmente se realizó una biopsia guiada por TC de la lesión paramediastínica derecha, con el diagnóstico de carcinoma no célula pequeña (CPNCP) sugestivo de adenocarcinoma (TTF1 y p63 negativos), con estudio molecular negativo para ROS1, MET, ALK, KRAS y EGFR.

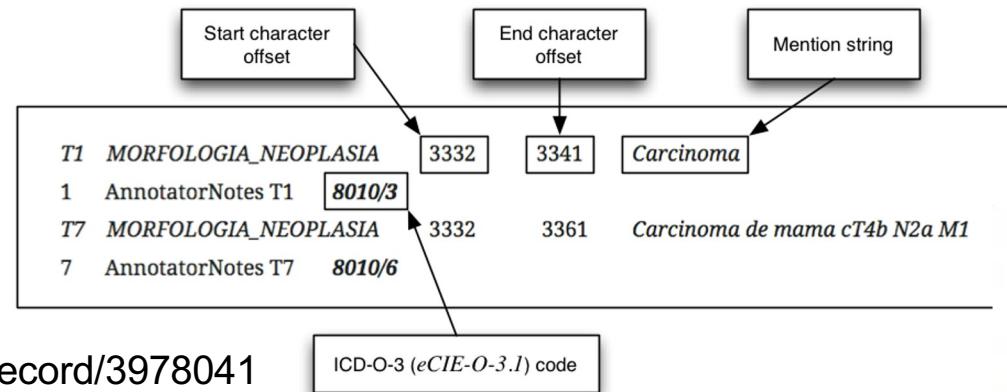


## TSV: ICD-O coding subtask

Miranda-Escalada, Antonio, Eulàlia Farré, and Martin Krallinger. "Named Entity Recognition, Concept Normalization and Clinical Coding: Overview of the Cantemist Track for Cancer Text Mining in Spanish, Corpus, Guidelines, Methods and Results." *IberLEF@SEPLN* (2020): 303-323.

## Brat:

### NER subtask & Normalization subtask



## CANTEMIST results

- **NER subtask:** 11 teams with F1 > 0.80
- **Norm subtask:** 6 teams with F1 > 0.75
- **ICD-O coding subtask:** highly competitive results

Miranda-Escalada, Antonio, Eulàlia Farré, and Martin Krallinger. "Named Entity Recognition, Concept Normalization and Clinical Coding: Overview of the Cantemist Track for Cancer Text Mining in Spanish, Corpus, Guidelines, Methods and Results." *IberLEF@SEPLN* (2020): 303-323.

Team Name	NER			Norm			Coding			
	P	R	F1	P	R	F1	P	R	F1	MAP
HITSZ-ICRC	.871	.868	.87	.824	.826	.825	-	-	-	-
Vicomtech	.868	<b>0.871</b>	.869	.822	.821	.821	.875	.836	<b>.855</b>	.847
SINAI	.859	.851	.855	.763	.755	.759	-	-	-	-
NLNDE	.854	.852	.853	.767	.766	.767	.77	.771	.77	.749
NCU-IISR	.849	.851	.85	-	-	-	-	-	-	-
Recognai	.85	.84	.845	-	-	-	-	-	-	-
mhjabreel	.837	.84	.839	.775	.779	.777	.797	.812	.805	.737
HULAT-UC3M	.826	.843	.834	-	-	-	-	-	-	-
Fadi	.844	.818	.831	.798	.774	.786	.826	.838	.832	.797
rrz-uc3m	.823	.824	.823	.202	.14	.165	-	-	-	-
baciero-fdez	.808	.802	.805	-	-	-	-	-	-	-
HULATUC3M-GI	.828	.769	.797	-	-	-	-	-	-	-
IBS_Software	.765	.764	.764	-	-	-	-	-	-	-
lasigeBioTM	.787	.714	.749	.064	.058	.061	.211	.601	.312	.506
Tong Wang	.757	.736	.746	-	-	-	-	-	-	-
DTIMAI	.727	.741	.734	-	-	-	-	-	-	-
episource	.691	.758	.723	.557	.61	.582	.68	.681	.681	.575
XIntao	.716	.721	.719	-	-	-	-	-	-	-
UAB	.688	.744	.715	-	-	-	-	-	-	-
Bigbyte	.649	.469	.545	.645	.467	.542	.794	.73	.761	.68
PaccanaroLab	.159	.595	.251	-	-	-	-	-	-	-
fernandez	0	0	0	-	-	-	-	-	-	-
ICB-UMA	-	-	-	-	-	-	.007	<b>.928</b>	.013	.847
kathrync	-	-	-	-	-	-	.182	.51	.268	.394

# MEDDOCAN : clinical document anonymization

Corpus annotated by PlanTL for anonymization and de-identification task:  
MEDDOCAN evaluation campaign (IberLEF)

## MEDDOCAN Corpus

- Annotation of protected health information.
- Guide/scheme for annotation and quality analysis (consistency).

**Best f-score:** 0.98530

NOMBRE PERSONAL SANITARIO	ID_TITULACION PERSONAL SANITARIO				
Médico: Luis Moyano Calvo	NºCol: 28 31 23567.				
EDAD SUJETO ASISTENCIA	SEXO SUJETO ASISTENCIA	EDAD SUJETO ASISTENCIA			
informe clínico del paciente: Adolescentes	Varón	de diecisiete años			
En la analítica de orina se aprecian 30-50 hematíes por campo. Urocultivo negativo.					
Se practica ecografía abdominal observándose pequeña lesión de medio centímetro de diámetro, sólida con refuerzo hiperecogénico anterior.					
Realizamos cistoscopia observándose en cara lateral derecha, por fuera de orificio ureteral dos pequeñas lesiones sobre elevadas, con muco:					
Sospechándose lesión inflamatoria se prescribe tratamiento con A.I.N.E. durante diez días sin que desparezcan las lesiones, decidiéndose in					
Se realiza RTU de ambas lesiones vesicales, siendo el informe anatopatológico el de leiomioma vesical, describiendo la lesión como "pro					
eosinófilo sin atipia, necrosis ni actividad mitótica significativa. Con el estudio inmunohistoquímico se demostró intensa positividad citoplasmá					
NOMBRE PERSONAL SANITARIO	CALLE	TERRITORIO	TERRITORIO	PAÍS	CORREO ELECTRONICO
Remitido por: Dr. Luis Moyano Calvo	C/ Eduardo Rivas, 3	28018	Madrid.	España.	e-mail: joseluismoyano@ya.com



<http://temu.bsc.es/maddocan/>



<https://zenodo.org/record/4279323>



Marimon, Montserrat, et al. "Automatic De-identification of Medical Texts in Spanish: the MEDDOCAN Track, Corpus, Guidelines, Methods and Evaluation of Results." *IberLEF@SEPLN*. 2019.



# CARMEN I : clinical corpus with symptoms at PhysioNET

Anonymized Spanish clinical corpus with improved guidelines, both direct and indirect re-identification attributes, corpus resynthesis (substitution of equivalent mentions, e.g. name by another name)

The screenshot shows the PhysioNet CARMEN-I homepage. At the top, there's a navigation bar with links for Database, Find, Share, About, News, Projects, Martin Krallinger, and Search. Below the navigation is a section titled "CARMEN-I: A resource of anonymized electronic health records in Spanish and Catalan for training and testing NLP tools". It lists the authors: Balalia Faria Maduell, Salvador Llinares-López, Santiago Andrés Frías, Artur Conesa, Elisa Ausín, Antonio López-Rodríguez, Helena Arino, Elena Calvo, María-Jesús Bertran, María Ángeles Marcos, Montserrat Nofre Muñoz, Laura Tañá Vela, Antonia Martí, Ricardo Ferrero, Xavier Pastor, Xavier Borrell Frigola, and Martin Krallinger. It also includes publication information: Published: Nov. 2, 2023; Version: 1.0.

**When using this resource, please cite:** (show more options)  
 Farre Maduell, E., Llinares-López, S., Frías, S. A., Conesa, A., Ausín, E., López-Rodríguez, A., Arino, H., Calvo, E., Bertran, M. J., Marcos, M. A., Marcos, M., Vela, L., Martí, A., Ferrero, R., Pastor, X., Borrell, X., & Krallinger, M. (2023). CARMEN-I: A resource of anonymized electronic health records in Spanish and Catalan for training and testing NLP tools (version 1.0). PhysioNet. <https://doi.org/10.13026/bxwv-3y44>

**Please include the standard citation for PhysioNet:** (show more options)  
 Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P. C., Mark, R., ... & Stanley, H. E. (2000). Physiobank, Physiobank, and Physioedit: Components of a new research resource for complex physiologic signals. Circulation (Orlando), 101(23), pp. e479-e479.



Se inició el tratamiento de la **miocarditis** con **inmunoglobulinas** (80 mg/día) durante 4 días y  
**tratamiento antiviral:** **IFN B** (0,25 mg/48 h) y (**ritonavir** 400 mg/**lopinavir** 100 mg/12 h).

Desarrollo y adaptación de las tecnologías de lenguaje a historia clínica

<https://physionet.org/content/carmen-i/1.0/>

Datos clínicos



DIABETES MELLITUS tipo 2 en tratamiento antidiabético oral

NER

Extracción de entidades

DIABETES MELLITUS tipo 2 en tratamiento antidiabético oral con metformina

Normalización de entidades

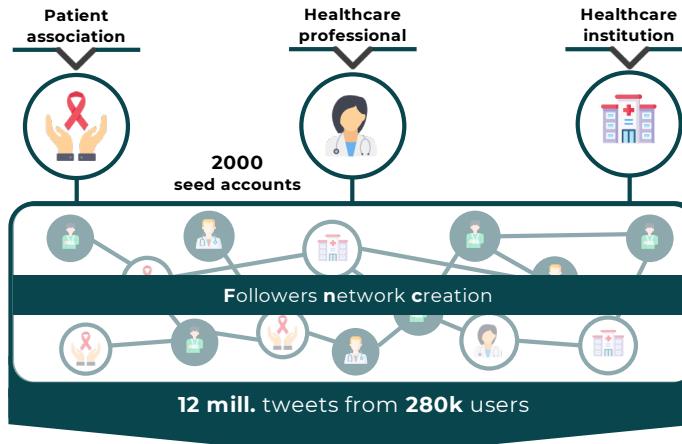
Datos clínicos estructurados

enfermedad	0   22	SCTID:73211009
procedimiento	26   52	SCTID: 415708007
fármaco	56   65	SCTID: 85188009

# Health NLP & Social media

SocialDisNER website: <https://temu.bsc.es/socialdisner/>  
Data: <https://zenodo.org/record/6406706#.YmlbMtpByUk>

COLING 2022  Barcelona Supercomputing Center  
Centro Nacional de Supercomputación



## Selection criteria



First person patient statements



Patient family members



Trusted healthcare professional content

## Gold Standard



Manual labelling of 10,000 tweets by clinical expert.



Cuando recibí el diagnóstico de esclerosis múltiple, lo más duro fue asimilar que la tenía. #esclerosis

When I received the diagnosis of multiple sclerosis, the hardest thing was to assimilate that I had it. #sclerosis

## Participants



## Corpus

## Evaluation

- Workshop: SMM4H track
- Conference: COLING 22
- Evaluation platform: CodaLab
- Quality metric: F-score

## Applications

Public opinion mining & sentiment analysis of diseases

Real-time disease outbreak surveillance & monitoring

Post-market drug safety

Prevalence of work-associated diseases

Characterization of patient-reported symptoms

Detection of hate speech or exclusion of sick people

Epidemiology and population health



Barcelona Supercomputing Center  
Centro Nacional de Supercomputación

# DisTEMIST: Disease Corpus & SNOMED CT normalization

  
DisTEMIST

**Guías DISTEMIST:**  
**Anotación y normalización de enfermedades en textos clínicos**

VI [Abril 2022]

**AUTORES**

Eulàlia Farré-Maduell (Barcelona Supercomputing Center)  
Luis Gascó Sánchez (Barcelona Supercomputing Center)  
Salvador Lima López (Barcelona Supercomputing Center)  
Antonio Miranda-Escalada (Barcelona Supercomputing Center)  
Martin Krallinger (Barcelona Supercomputing Center)

Niña de 3 años y 8 meses de edad.  
Diabetes gestacional y madre Rh - grado A, gestación de 38 semanas y parto normal.  
Cariotipo femenino normal 46 XX.

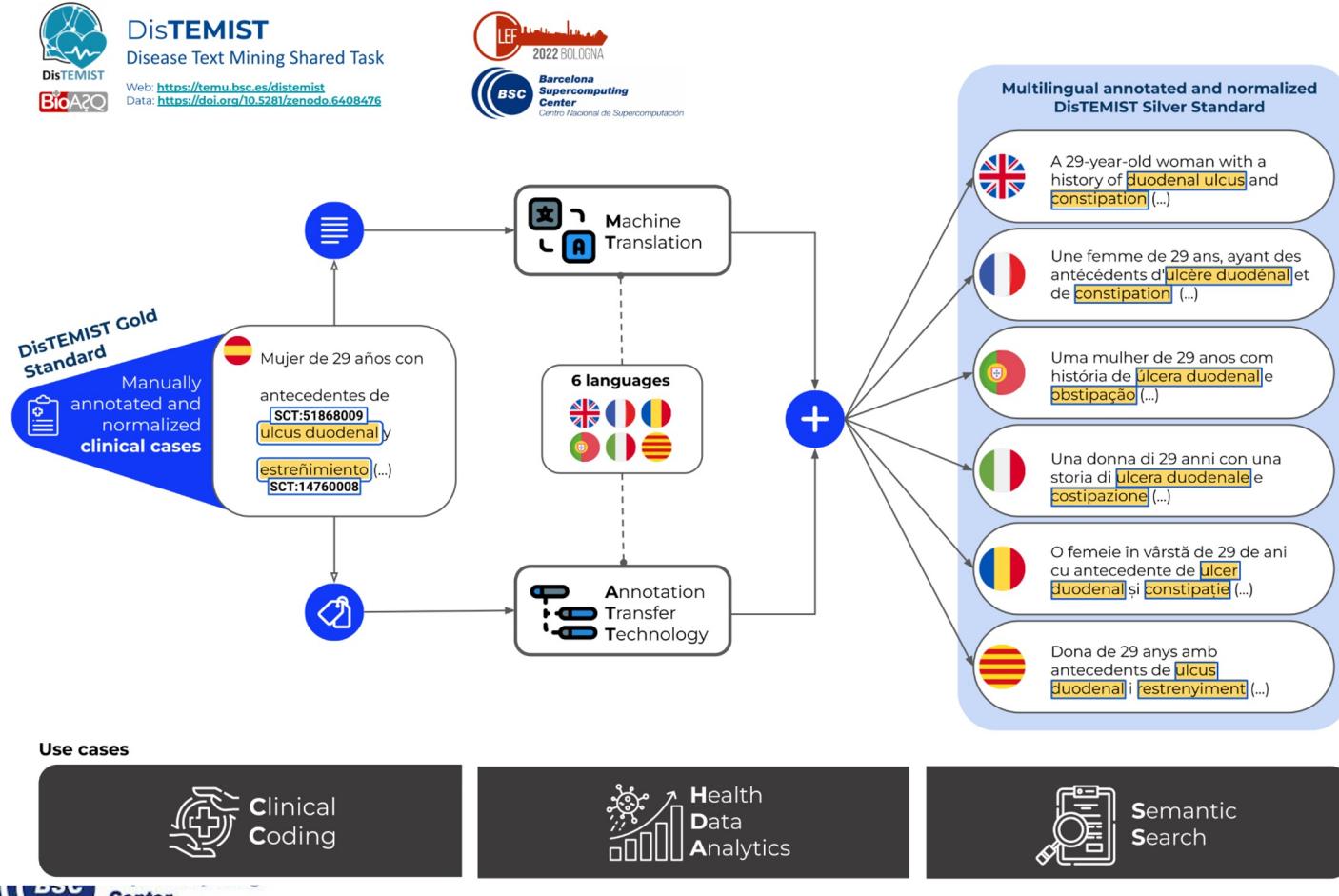
El diagnóstico por ecografía prenatal de esta niña muestra agenesia del cuerpo calloso. **ENFERMEDAD**  
riñón derecho displásico multiquístico. **ENFERMEDAD**  
En el sistema nervioso central presenta un quiste de plexos coroideos y un quiste en la región interhemisférica profunda. **ENFERMEDAD**  
En el examen de resonancia magnética nuclear, no se detectan anomalías de la migración neuronal. **ENFERMEDAD**  
Los registros electroencefalográficos (EEG) fueron normales.  
Los estudios cognitivos muestran un cociente de desarrollo psicomotor normal.  
La exploración del fondo de ojo muestra una apariencia prácticamente normal del fondo de ojo derecho, con una papila óptica normal, dos pequeñas lagunas coriorretinianas hipopigmentadas y ausencia de afectación macular. **ENFERMEDAD**  
Sin embargo, observamos en el fondo de ojo izquierdo una malformación colobomatosa del nervio óptico y

	Documents	Annotations	Unique codes	Tokens
<b>Training</b>	750	8,066	4,819	305,166
<b>Test</b>	250	2,599	2,484	101,152
<b>Total</b>	1,000	10,665	7,303	406,318

- Inter-Annotator Agreement (IAA): 82.3
- Best team: 0.77 F-score

doc1	T1	ENFERMEDAD	209	236	agenesia del cuerpo calloso	5102002	EXACT	
doc1	T2	ENFERMEDAD	239	277	riñón derecho displásico multiquístico	82525005		NARROW
doc1	T3	ENFERMEDAD	322	348	quiste de plexos coroideos	230790004	EXACT	
doc1	T4	ENFERMEDAD	354	399	quiste en la región interhemisférica profunda	40720005		NARROW
doc1	T5	ENFERMEDAD	914	957	malformación colobomatosa del nervio óptico	77157004		NARROW
doc1	T6	ENFERMEDAD	960	984	lagunas coriorretinianas	302893000	NARROW	
doc1	T7	ENFERMEDAD	462	496	anomalías de la migración neuronal	253146009	EXACT	
doc1	T8	ENFERMEDAD	837	855	afectación macular	312999006	EXACT	
doc1	T9	ENFERMEDAD	1005	1023	afectación macular	312999006	EXACT	

# DisTEMIST Multilingual Silver Standard



## Phase 0: Spanish

## Phase 1: 7 languages:

- ✓ English
- ✓ French
- ✓ Portuguese
- ✓ Italian
- ✓ Romanian
- ✓ Catalan
- ✓ Galician

## Phase 2: +3 languages:

- Dutch
- Swedish
- Czech

## Phase 3: +3 languages:

- German
- Danish
- Norwegian

Miranda-Escalada, Antonio, et al. "Overview of distemist at bioasq: Automatic detection and normalization of diseases from clinical texts: results, methods, evaluation and multilingual resources." *Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings*. 2022.

# DisTEMIST Multilingual Silver Standard

## Spanish Gold Standard

Snomed ID:  
416098002

ENFERMEDAD

1 Paciente de 70 años de edad, minero jubilado, sin alergias medicamentosas conocidas, que presenta como antecedentes personales: accidente laboral antiguo con fracturas vertebrales y costales; intervenido de enfermedad de Dupuytren en mano derecha y by-pass iliofemoral izquierdo; Diabetes Mellitus tipo II, hipercolesterolemia e hiperuricemia; enolismo activo, fumador de 20 cigarrillos / Snomed ID:  
13644009

ENFERMEDAD ENFERMEDAD  
ENFERMEDAD ENFERMEDAD ENFERMEDAD ENFERMEDAD

## English Silver Standard

Snomed ID:  
416098002

ENFERMEDAD

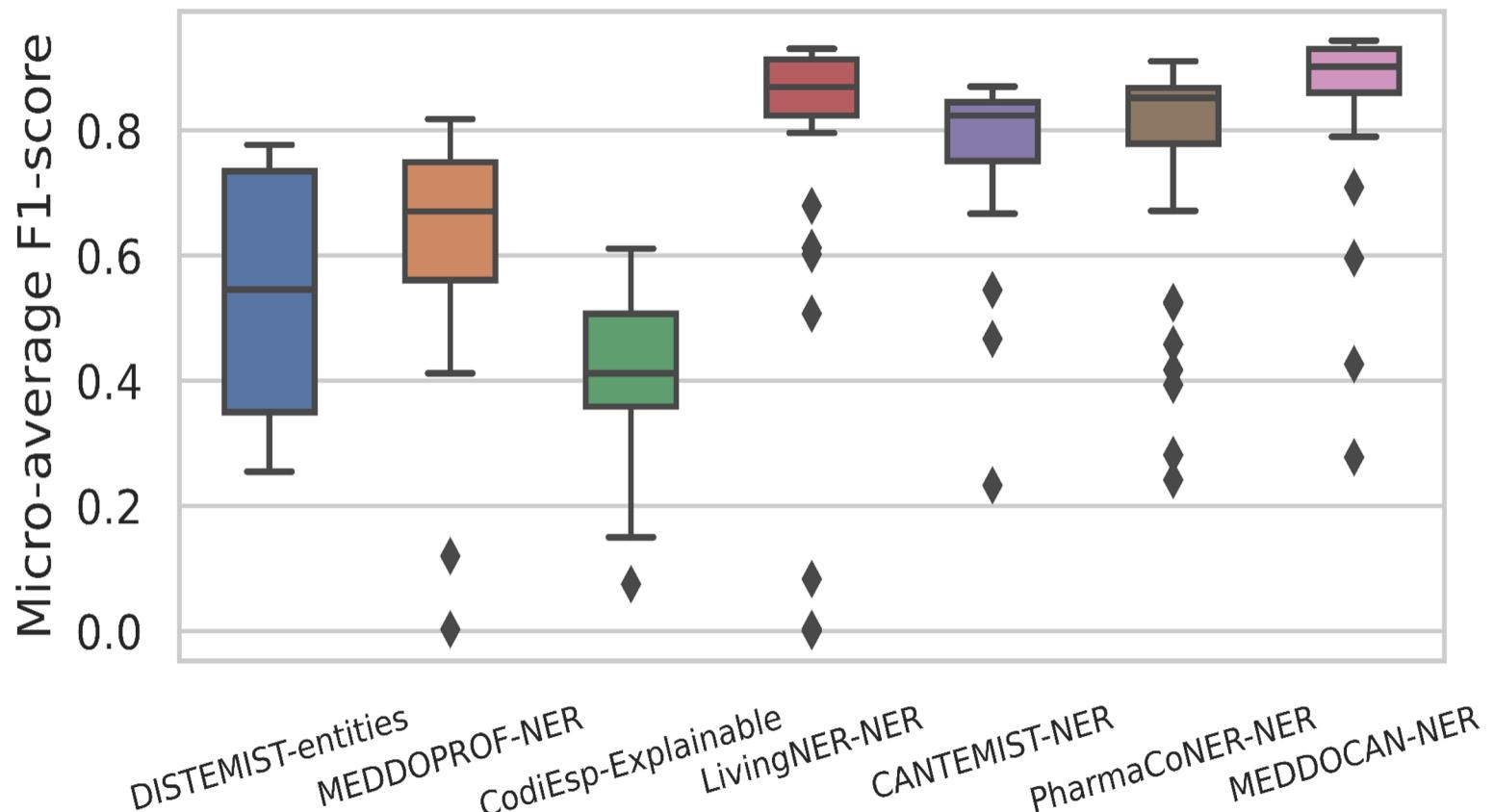
1 Patient aged 70 years, retired miner, with no known drug allergies, with a personal history: old work accident with vertebral and rib fractures; operated on for Dupuytren's disease in the right hand and coronary by-pass; Diabetes Mellitus type II, hypercholesterolemia and hyperuricaemia; active alcoholism, smoker of 20 cigarettes/day.

ENFERMEDAD ENFERMEDAD ENFERMEDAD  
ENFERMEDAD ENFERMEDAD ENFERMEDAD



Online side-by side browser (BRAT): temu.bsc.es/mDistemist/diff.xhtml#/translations/en/train/S0004-06142005000500011-1?diff=/gold-standard/train/

## Variability in performance across different clinical entity types



# Impact of shared tasks

- International participation of both academy and industry: 17 tracks (13 Spanish, 2 English, 2 MT en<>es)





## Overview of the MEDIQA-M3G 2024 Shared Task on Multilingual Multimodal Medical Answer Generation

Wen-wai Yim, Asma Ben Abacha, Yujuan Fu, Zhaoyi Sun, Fei Xia, Meliha Yetisgen, Martin Krall

### Abstract

Remote patient care provides opportunities for expanding medical access, saving healthcare costs, and off-demand convenient services. In the MEDIQA-M3G 2024 Shared Task, researchers explored solutions for dermatological consumer health visual question answering, where user-generated queries and images are input and a free-text answer response is generated as output. In this novel challenge, eight teams with 10 submissions were evaluated across three language test sets. In this work, we provide a summary of the data results and approaches. We hope that the insights learned here will inspire future research directions that technology can deburden clinical workload and improve care.

Query	Responses
  <p>帮忙诊断一下:三个月前出现如下图,自己用达克宁喷雾两个月无明显效果,之后去乡村诊所,医生指导用鸡眼膏,之后出现变红变多,请帮忙诊断下</p> <p>Please help with the diagnosis: Three months ago, the condition shown in the picture below appeared. The patient used Daknening spray for two months without any noticeable effect. Afterwards, they went to a rural clinic, where the doctor advised them to use corn ointment. Subsequently, the condition turned red and worsened.</p> <p>Please help with the diagnosis.</p> <p>Por favor, ayude con el diagnóstico : Hace tres meses, apareció la condición mostrada en la imagen de abajo. El paciente utilizó el spray Daknening durante dos meses sin ningún efecto notable. Posteriormente, acudió a una clínica rural, donde el médico le aconsejó que utilizara pomada de maíz. Posteriormente, la condición se volvió roja y empeoró. Por favor, ayude con el diagnóstico.</p>	<p>RESPONSE1: 是鸡眼。 It's a corn. Es un callo.</p> <p>RESPONSE2: 考虑: 跖疣 Consideration: Plantar wart Consideración: ¿Verruga plantar</p> <p>RESPONSE3: 是跖疣, 不 是鸡眼, 激光 治疗。 It's a plantar wart, not a corn. Laser treatment is recommended. Es una verruga plantar, no un callo. Se recomienda el tratamiento con láser.</p>



# Example projects involving Health NLP



*Barcelona  
Supercomputing  
Center*  
Centro Nacional de Supercomputación

# How to deal with the multilingual reality by taking advantage of existing resources?



Generating high quality language resources is expensive



Analysis and modelling of the problem



Creation and validation of annotation guidelines



Training of professionals and annotation



Leveraging existing linguistic resources in biomedical field



Obtaining corpora more efficiently



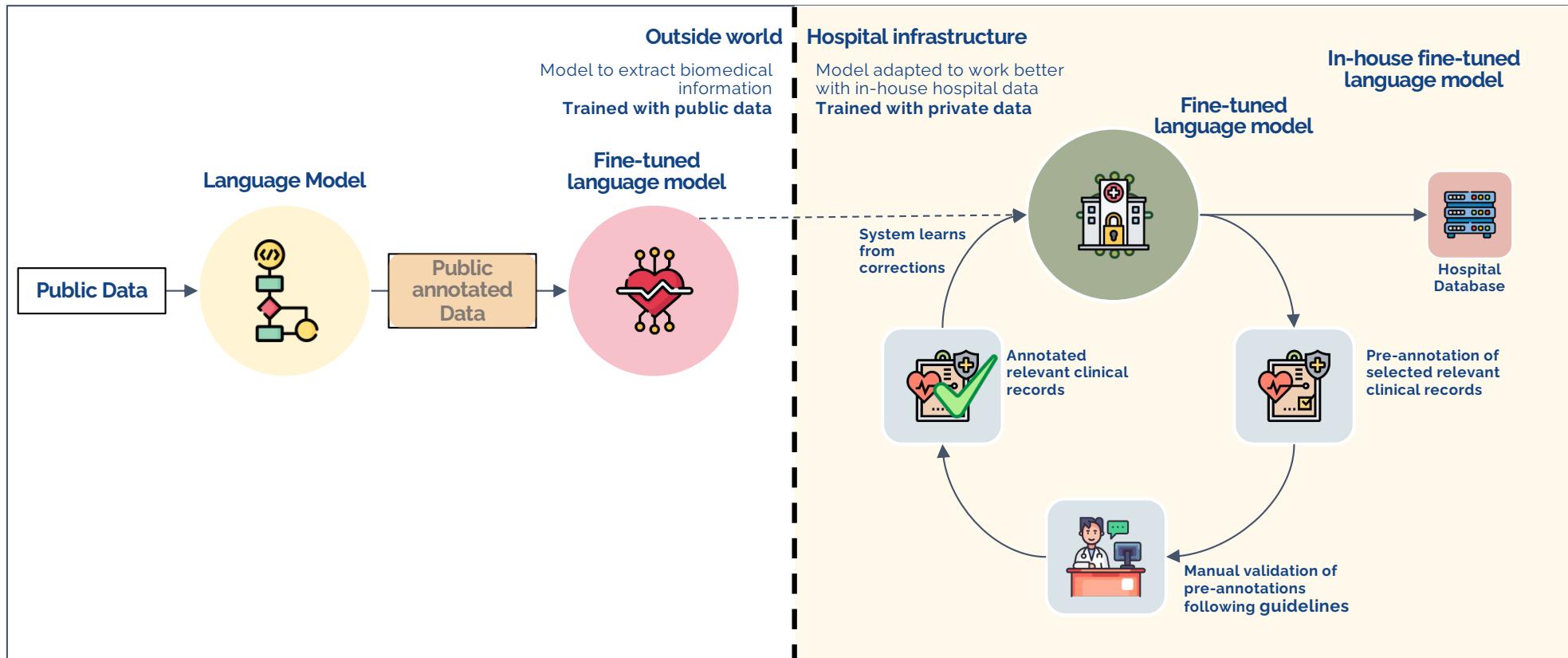
Robust data to train consistent models



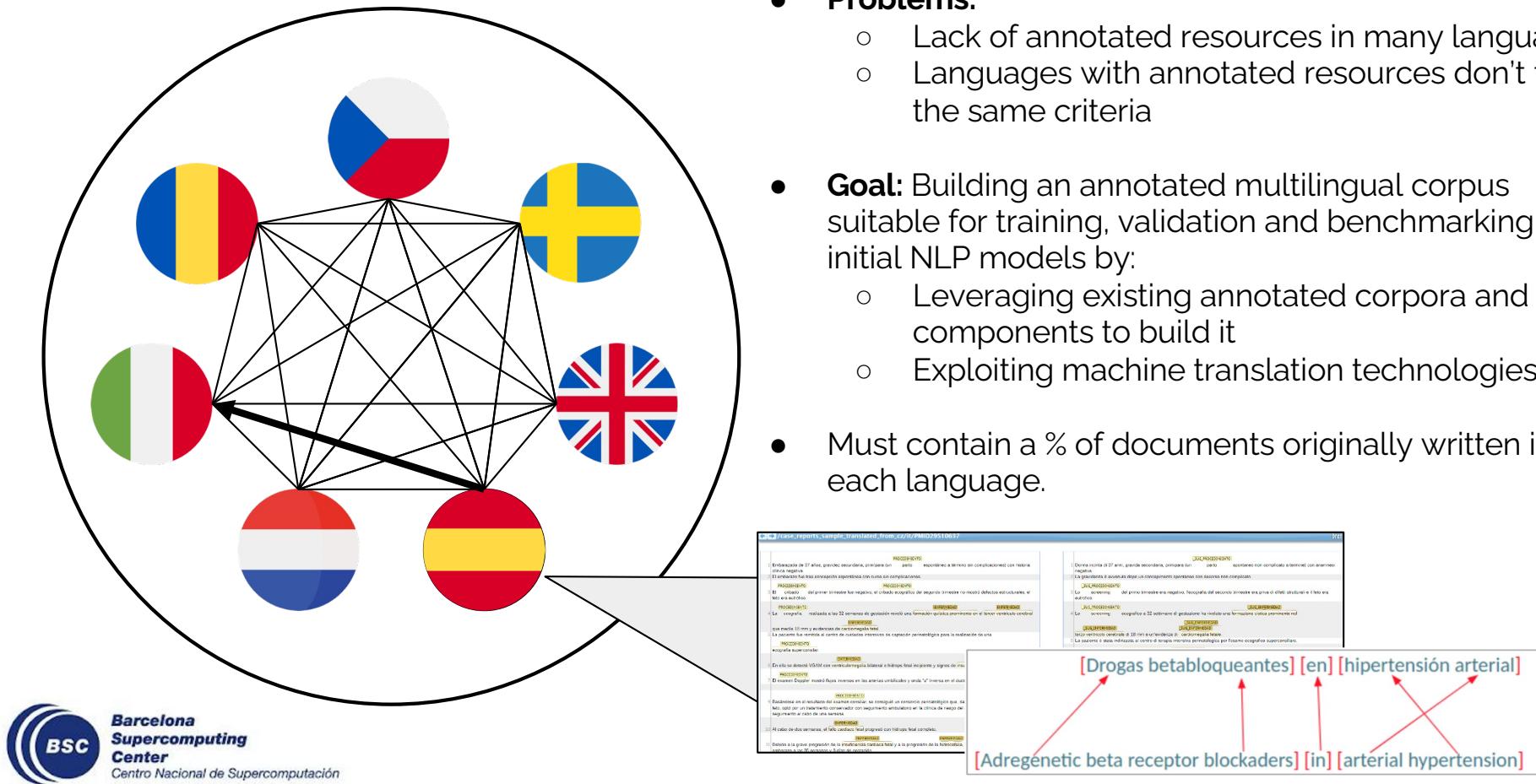
Data access restrictions in federated scenarios



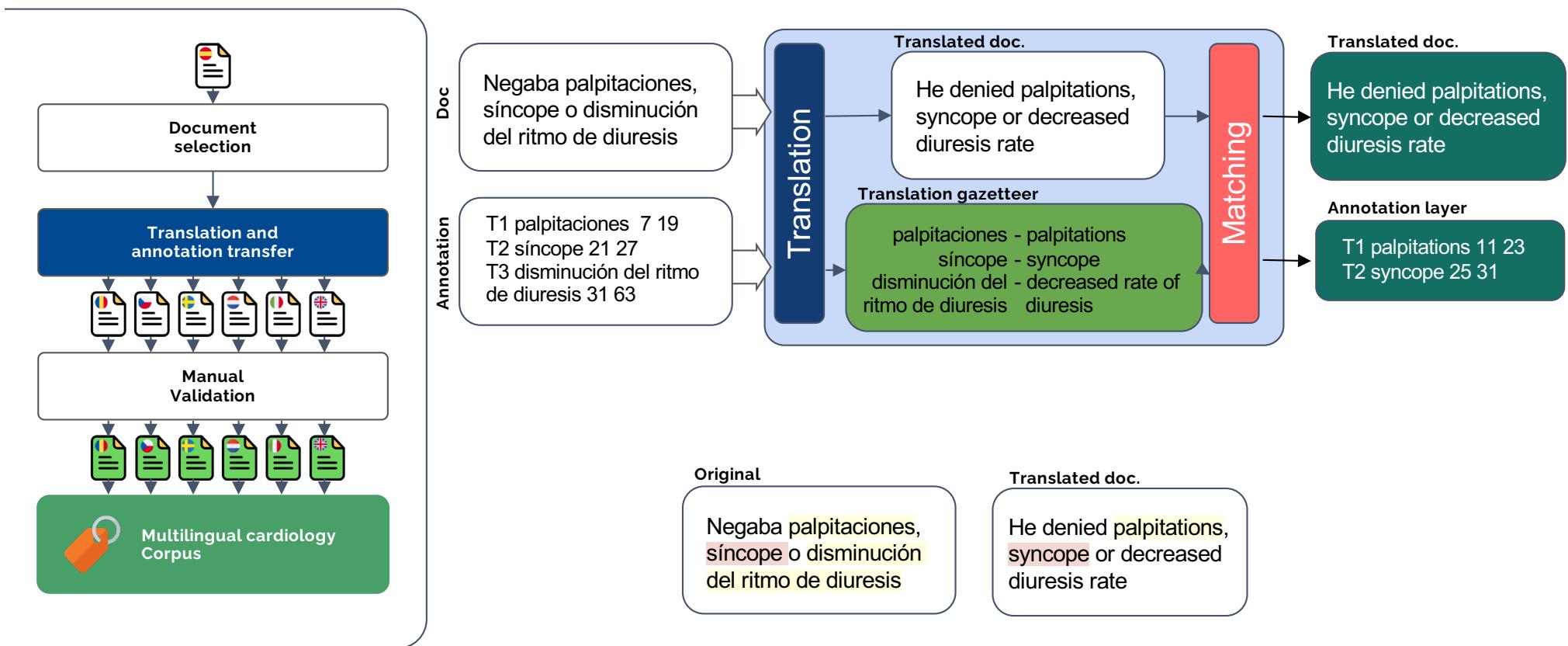
# Training IE models process in federated scenarios with clinician-in-the-loop



# Multilingual Public Annotated Data

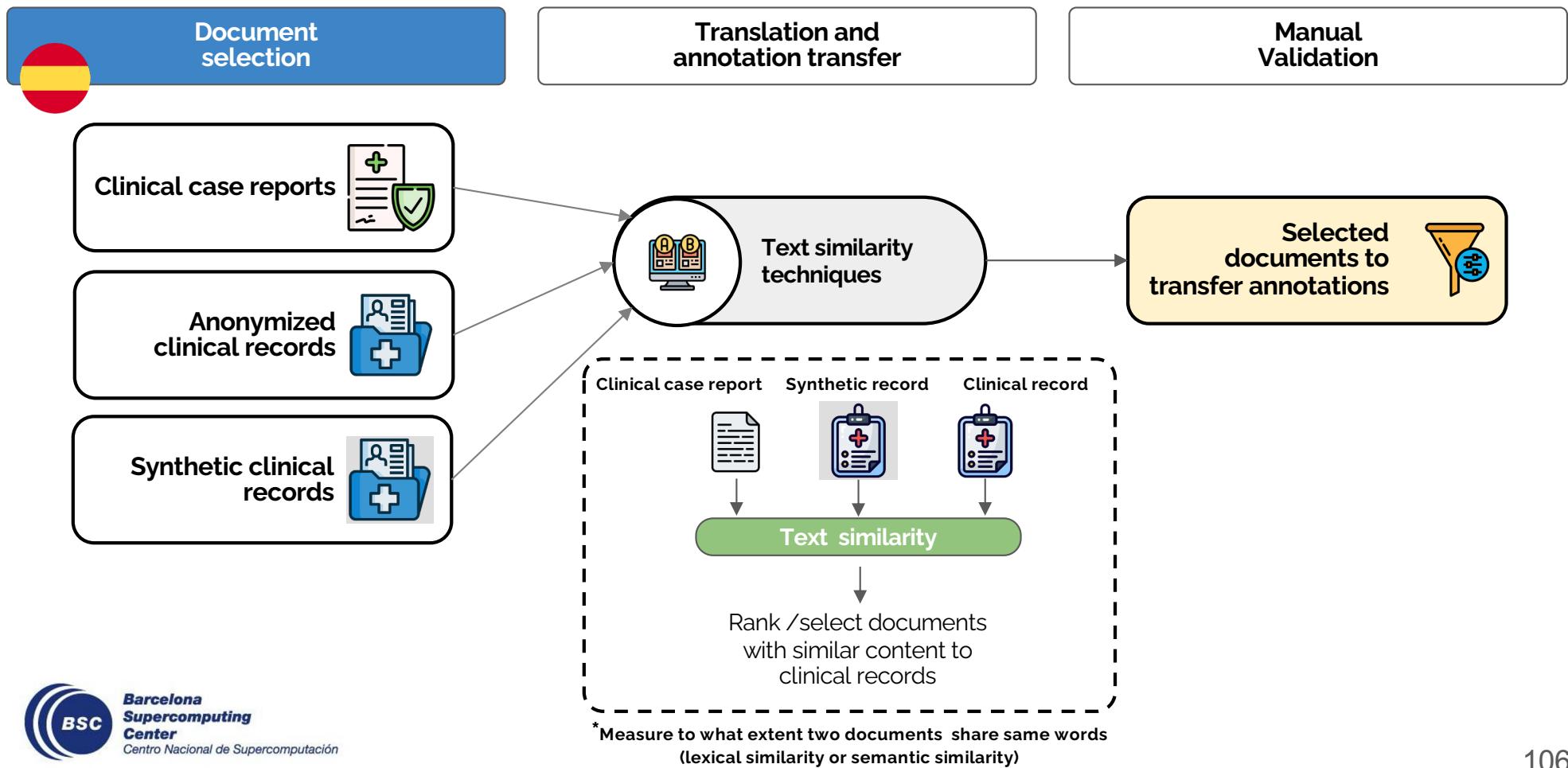


# Multilingual resources and annotation projection approaches

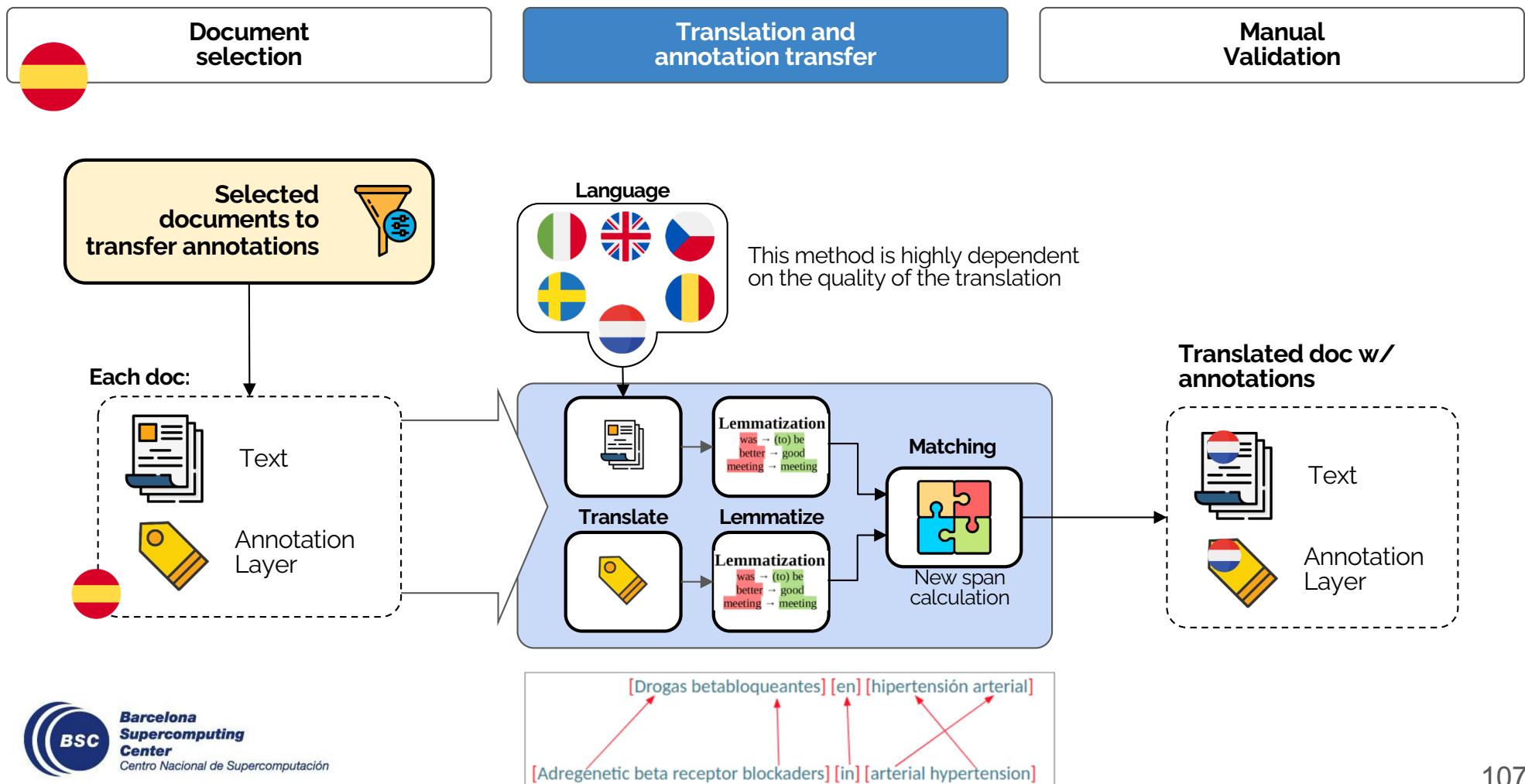


There might be false positives and false negatives!

# Multilingual Public Annotated Data - Document selection



## Multilingual Public Annotated Data - Annotation transfer



# Multilingual Public Annotated Data - Validation

[/cardioccc/nl/casos\\_clinicos\\_cardiologia168](#)

brat

ANTECEDENTES PERSONALES	
2	Sin alergias medicamentosas conocidas.
3	Hábitos tóxicos: exfumador desde hace 5 años, fumó 1 paquete/día durante > 20 años.
4	Factores de riesgo cardiovascular: hipertensión arterial (HTA) (con mal control en domicilio, cifra en torno a 180 mmHg), dislipemia (DL), diabetes mellitus tipo II (DM).
5	Antecedentes cardiológicos: cardiopatía isquémica crónica desde 2006, oclusión de descendente anterior (DA) con relleno por colaterales, ateromatosis difusa.
SÍNTOMA	
6	Ergometría positiva a altas cargas, por lo que se decide manejo médico.
PROCEDIMIENTO	
7	FA paroxística anticoagulada con sintrom.
ENFERMEDAD	
8	Médico-quirúrgicos: artropatía psoriásica, recibió tratamiento con metotrexato.
ENFERMEDAD	
9	Ingresó en 2016 por infección por gripe A e insuficiencia respiratoria.
FARMACO	
10	Tratamiento domiciliario: atenolol 50 mg/24h, Uniket retard 50 mg/24h, telmisartán 80 mg/24h, amlodipino 5 mg/24h, atorvastatina 80 mg/24h, ezetimiba 10 mg/24h, metformina 1000 mg + sitagliptina 50 mg/12h, deflazacort 6 mg/24h, omeprazol 20 mg/24h, sintrom 4 mg, empagliflozina 10 mg/24h, ácido fólico 5 mg/24h, metotrexato 15 mg sc cada 7 días, vesomni 6/0,4 mg/24h.

 **Center**  
Centro Nacional de Supercomputación

SOONLIJKE GESCHIEDENIS	
2	Geen bekende geneesmiddelenallergieën.
3	Gifgewoonten: ex-roker sinds 5 jaar, rookte 1 pakje/dag gedurende > 20 jaar.
4	Cardiovasculaire risicofactoren: hypertensie (HT) (met slechte controle thuis, SBP rond 180), dyslipidemie (LD), diabetes mellitus type II (DM).
5	Cardiologische voorgeschiedenis: chronische ischemische hartziekte sinds 2006, anterieve descenderende (AD) occlusie met collaterale vulling, diffuse atheromatose.
SÍNTOMA	
6	Positieve ergometrie bij hoge belasting, dus werd besloten tot medisch beheer.
PROCEDIMIENTO	
7	Paroxysmaal AF geanticoaguleerd met synthrom.
ENFERMEDAD	
8	Medisch-chirurgisch: psoriatische artropathie, behandeld met methotrexaat.
ENFERMEDAD	
9	In 2016 opgenomen voor influenza A infectie en respiratoire falen.
FARMACO	
10	Thuisbehandeling: atenolol 50 mg/24h, Uniket retard 50 mg/24h, telmisartan 80 mg/24h, amlodipine 5 mg/24h, atorvastatine 80 mg/24h, ezetimibe 10 mg/24h, metformine 1000 mg + sitagliptine 50 mg/12h, deflazacort 6 mg/24h, omeprazol 20 mg/24h, sintrom 4 mg, empagliflozin 10 mg/24h, foliumzuur 5 mg/24h, methotrexaat 15 mg sc om de 7 dagen, vesomni 6/0,4 mg/24h.

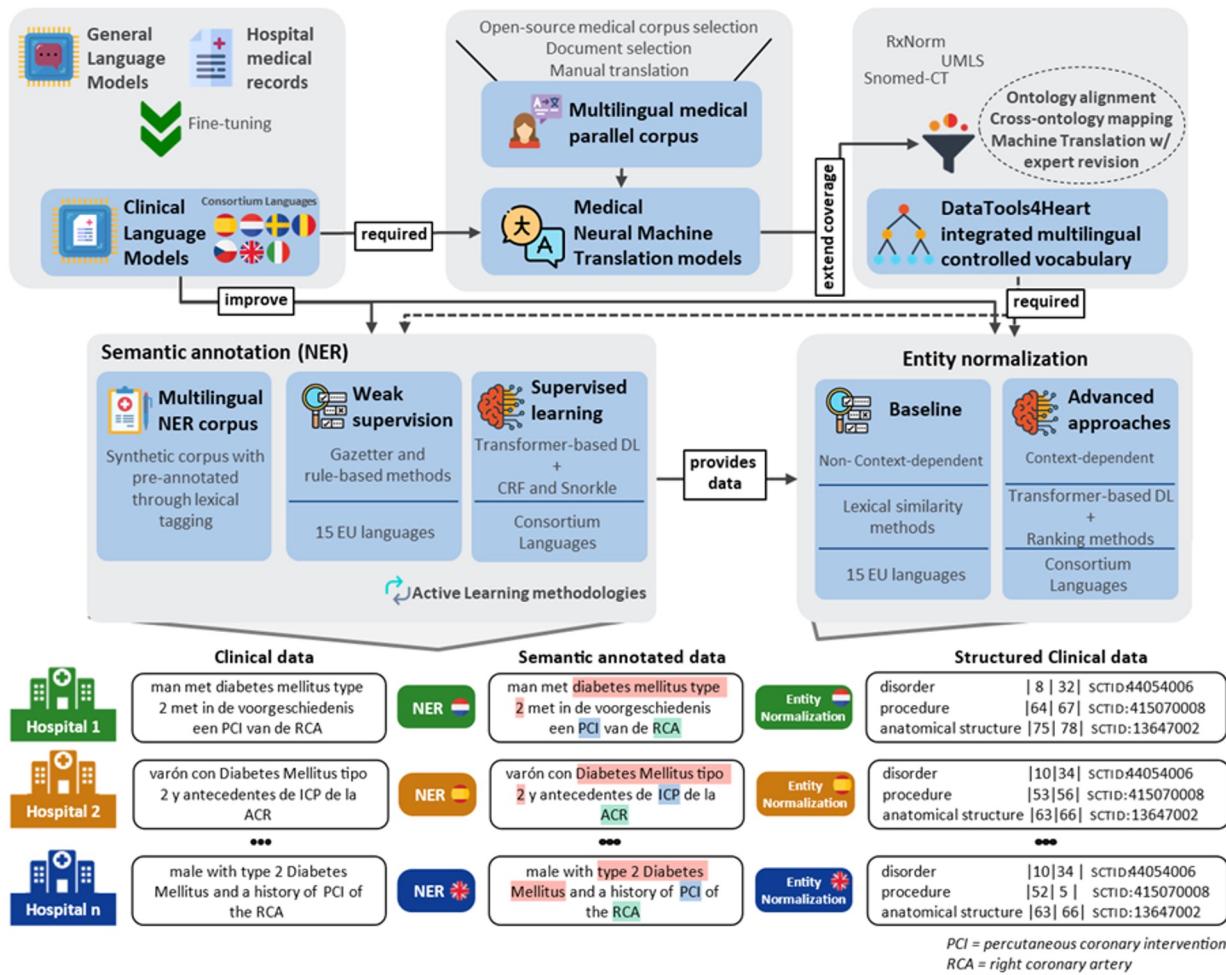
Translation and detection problems with acronyms?

Entities lost in the annotation transfer

Bad translations?



# European Health Data Toolbox for Enhancing Cardiology Data Interoperability, Reusability and Privacy



7 languages

Health care institutions



University College London Hospitals  
NHS Foundation Trust



ST. ANNE'S UNIVERSITY HOSPITAL BRNO  
INTERNATIONAL CLINICAL RESEARCH CENTER



Fondazione Policlinico Universitario Agostino Gemelli IRCCS  
Università Cattolica del Sacro Cuore



Funded by  
the European Union



# Trustworthy Artificial Intelligence for Personalised Risk Assessment in Chronic Heart Failure (AI4HF)

5 languages

Health care institutions

Development of: Natural language processing (NLP) pipeline to extract information from the clinical reports, incl. **risk factors, symptoms, family history & lifestyle, professions, locations**

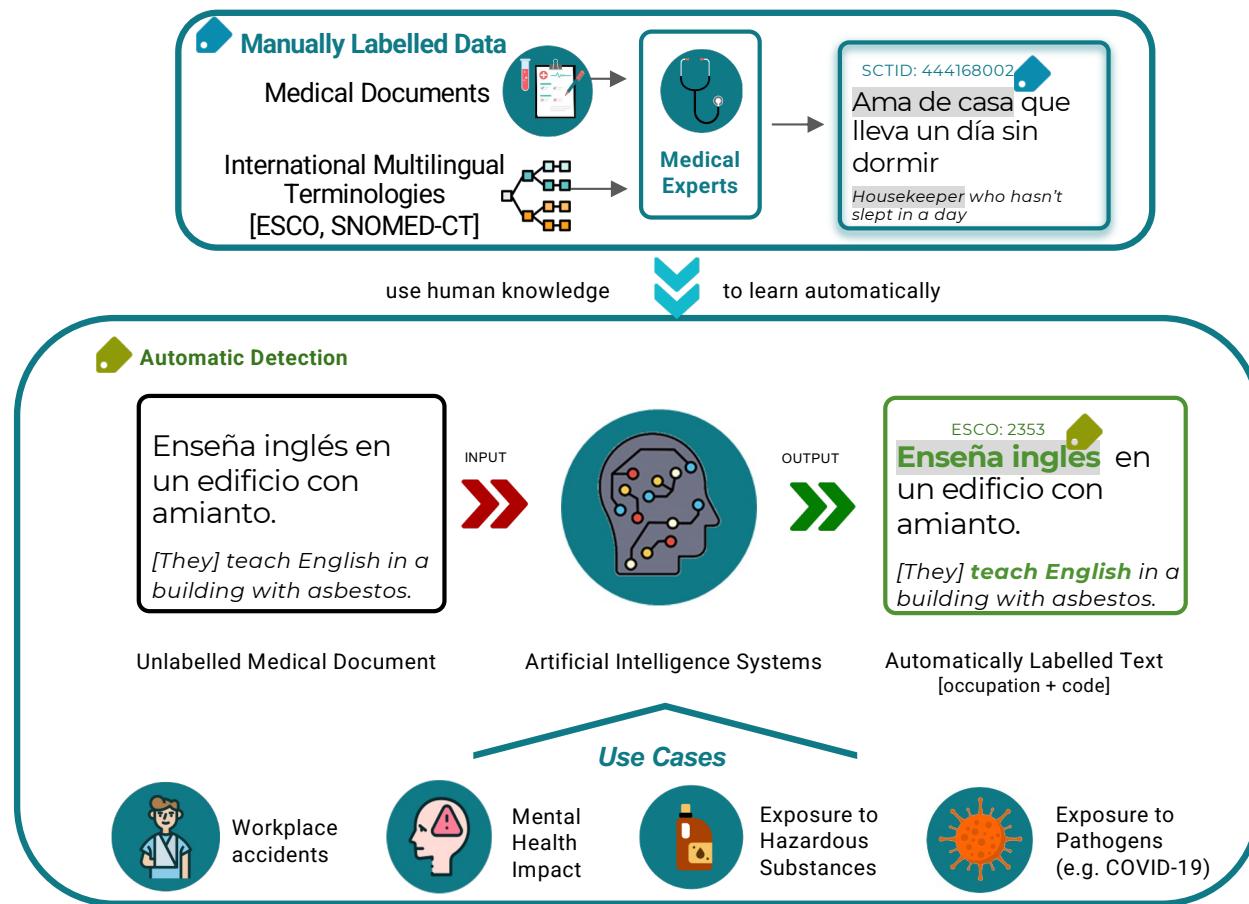
Construction of: multi-lingual corpus to achieve semantic interoperability between the data for **English, Spanish, Dutch, Catalan and Czech**

Use of seed languages (English & Spanish), followed by the application of NLP pipeline that exploits **multi-lingual controlled vocabularies & machine translation**

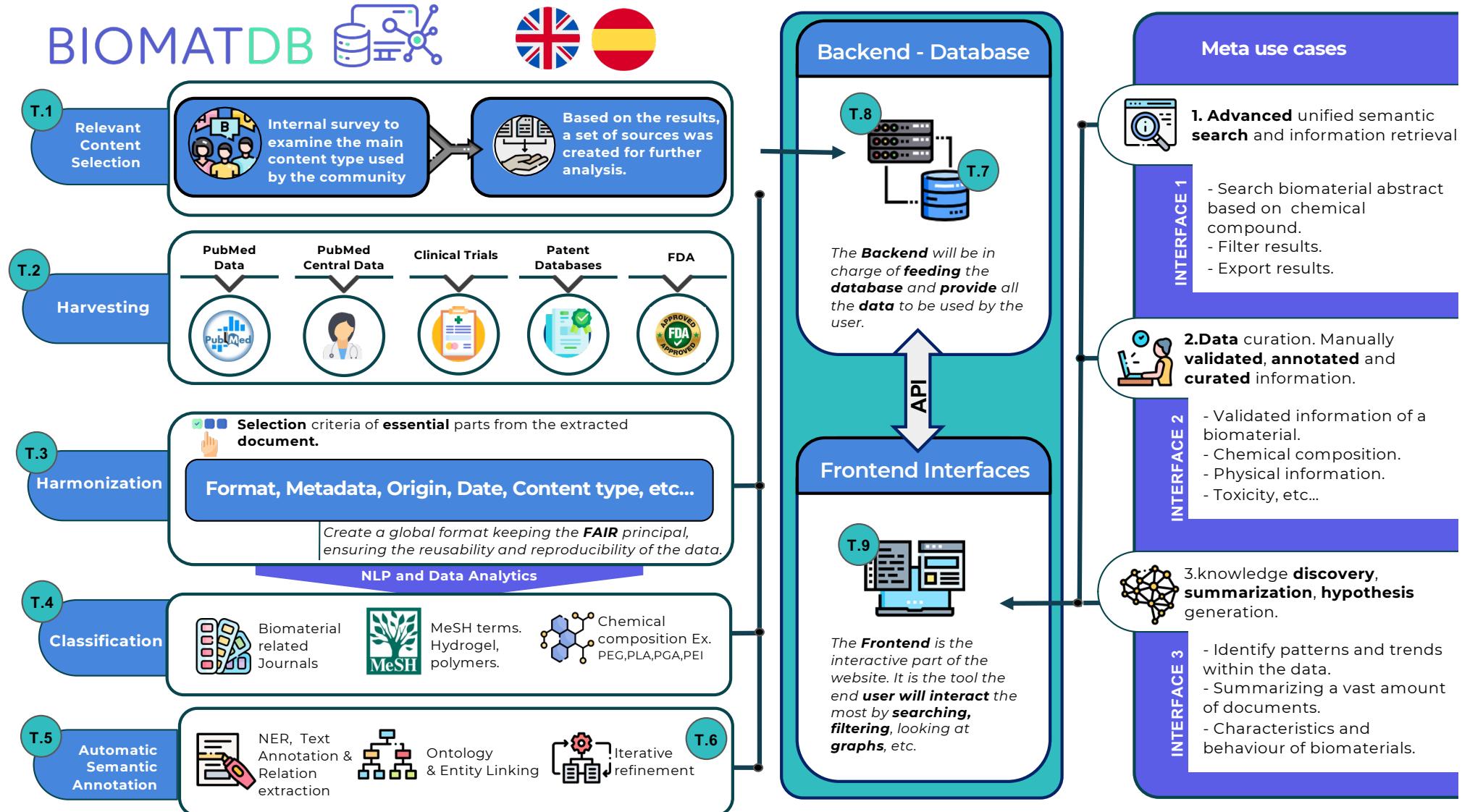


Funded by  
the European Union

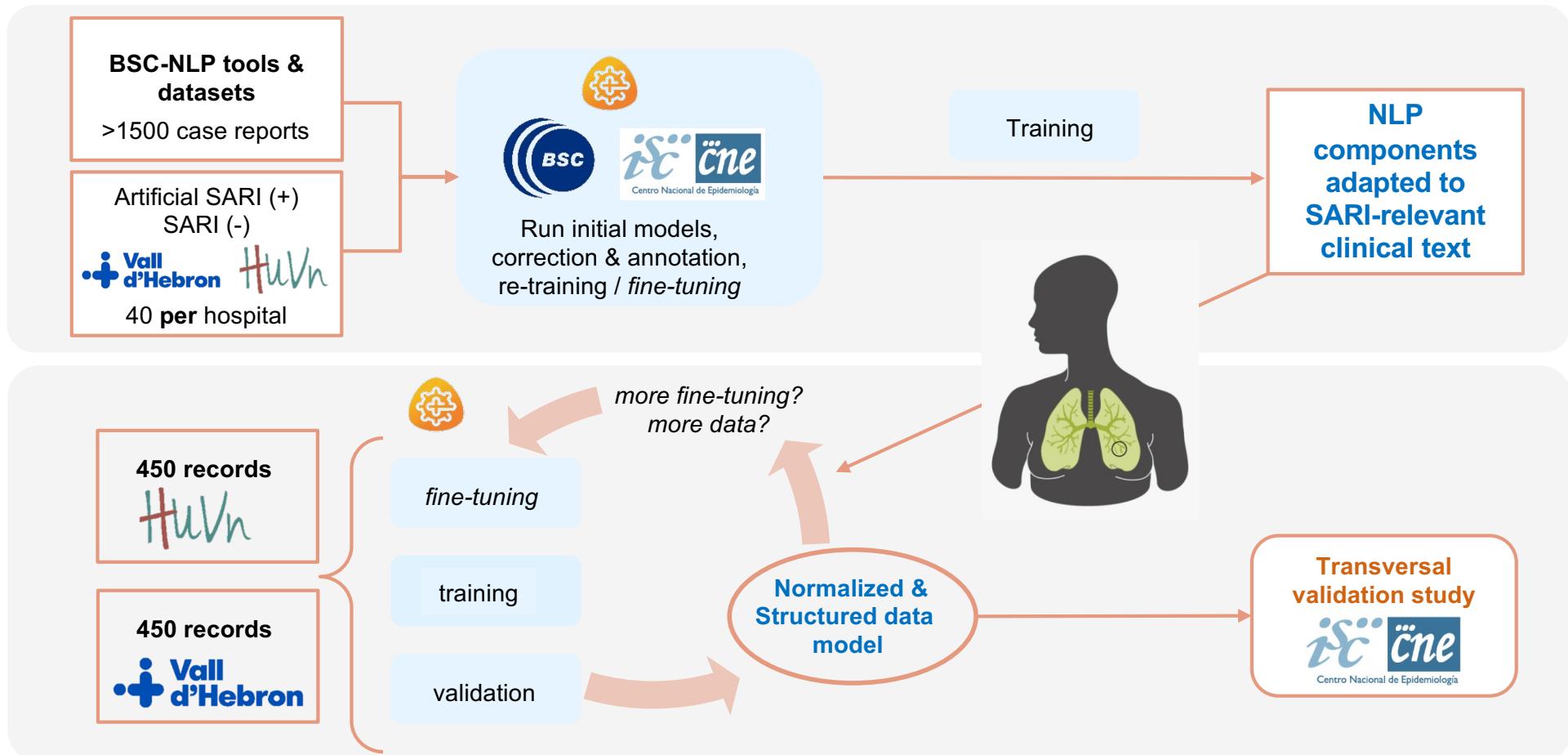
# AI4ProfHealth: NLP for occupational



# BIOMATDB



# Epidemiological vigilance: acute respiratory infections



Severe Acute Respiratory Infections (SARI): COVID, influenza, RSV

# Rare diseases

## BARITONE

2 languages



Subproject 2

Subproject 1



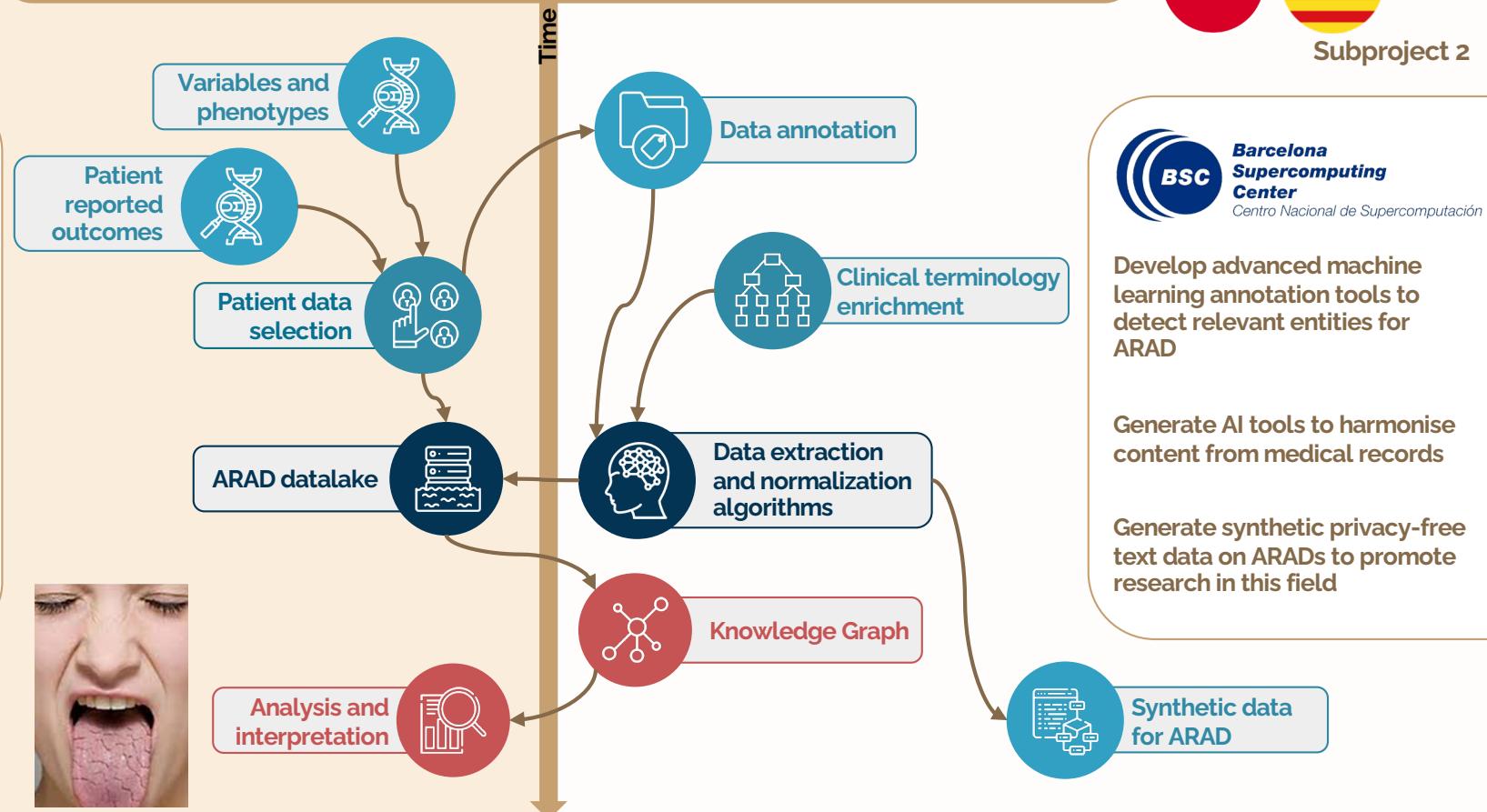
Identify clinical longitudinal phenotypic patterns of ARADs

Define omics studies to progress towards precision medicine

Incorporate patient reported outcomes to get their status and health condition



Boosting digital transition in healthcare: integration of clinical data and AI technologies for high accuracy phenotyping of complex diseases



Develop advanced machine learning annotation tools to detect relevant entities for ARAD

Generate AI tools to harmonise content from medical records

Generate synthetic privacy-free text data on ARADs to promote research in this field

# Rare diseases

## UVEITIS

2 languages



Development and validation of machine learning prediction models of health-related outcomes in uveitis: a multicentric project using electronic health record free text with state-of-the-art deep learning methods

Complejo Asistencial  
Universitario de León

H Clínico San Carlos  
HU 12 de Octubre  
HU Infanta Leonor  
HU Infanta Sofía  
HU La Paz

AUVEA

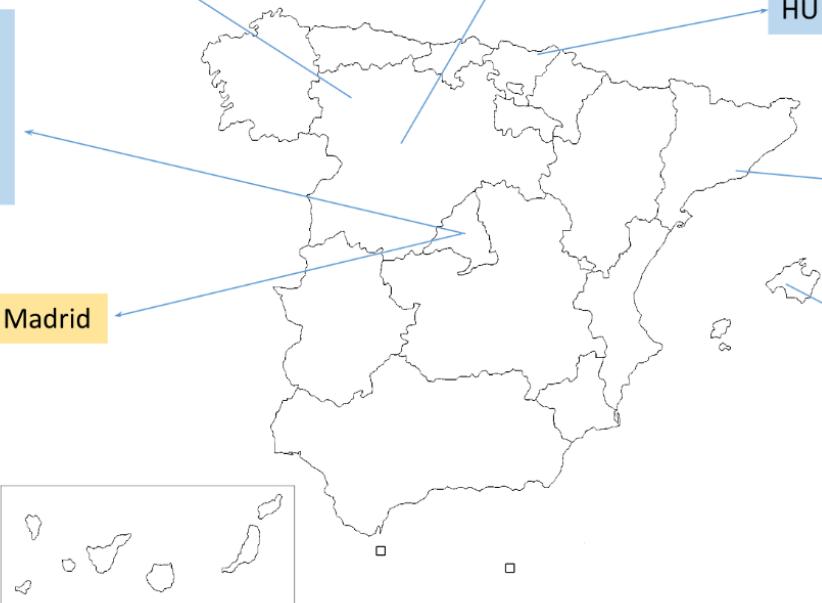
Hospital Clínico Universitario  
de Valladolid - IOBA

HU Donostia

Grupos Clínicos: 10

Ingenieros: 1

Asociación Pacientes: 1



# Conclusions



*Barcelona  
Supercomputing  
Center*  
Centro Nacional de Supercomputación

# Needs & Conclusions

- Disease
- Specialty
- Hospital
- Record type
- Time period

Data structure variation



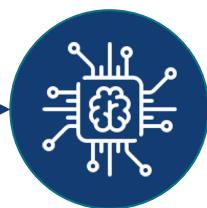
Privacy issues

- Synthetic data
- Anonymization/de-identification
- Federated scenarios
- Processing in situ



High quality  
clinical data

Help to  
train/evaluate



High quality  
clinical NLP  
models

- Exploitation of results
- Validation of usability
- Validation of quality & impact

Supporting  
medical needs



Interoperability

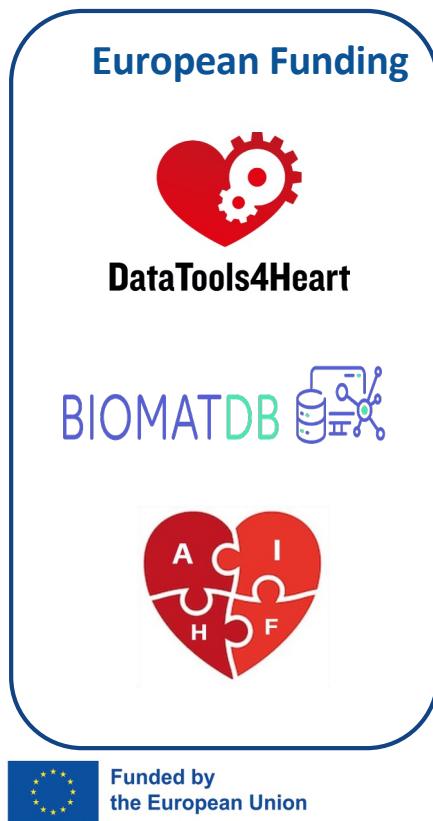
- Data harmonization
- Integration into data model
- Standardization of results

Training and benchmarking of  
language models

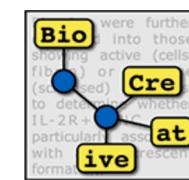
# Acknowledgements



**Barcelona  
Supercomputing  
Center**  
Centro Nacional de Supercomputación



## Terminologies



## Shared tasks



## EHR Processing

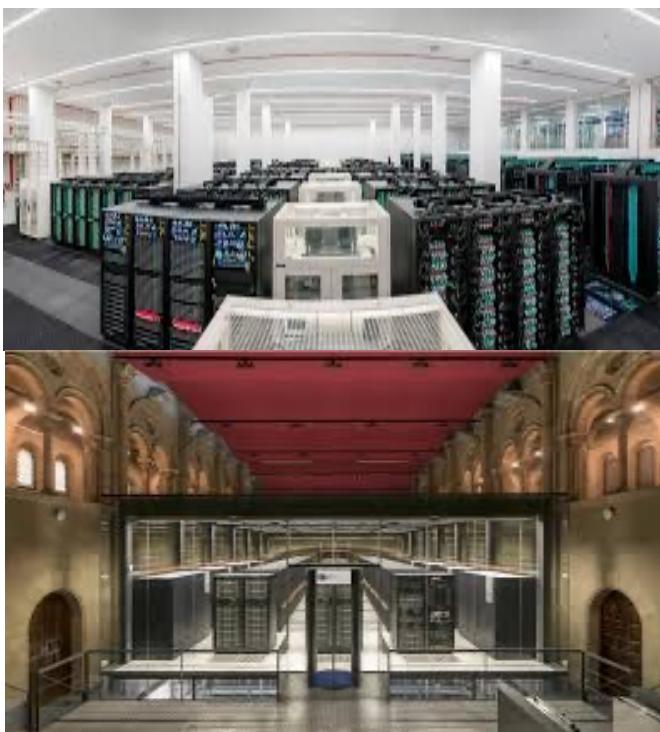




## MARTIN KRALLINGER

### LEADING RESEARCHER

Life Sciences - NLP for Biomedical Information Analysis



**Barcelona Supercomputing Center**  
Centro Nacional de Supercomputación



## JAN RODRIGUEZ MIRET

### JUNIOR RESEARCH ENGINEER

Clinical NER, semantic annotation, language models, multilingual annotations



## SERGI MARSOL TORRENT

Automatic terminology generation, Ontology enrichment/cross, relation extraction



## SALVADOR LIMA LOPEZ

### RESEARCH ENGINEER

Linguistic NER, linguistic & sociodemographic annotation process, shared task & evaluation campaigns



## MIGUEL RODRIGUEZ ORTEGA

### RESEARCH ENGINEER

Text similarity, text clustering, text classification, unsupervised NLP techniques



## ALBERTO BECERRA TOME

Entity linking/normalization, clinical language models, LLMs



## EULALIA FARRE

### SENIOR RESEARCH ENGINEER

Clinical corpus construction, multilingual annotation resources, clinical use cases & applications, ontologies



## LAURA VIGIL GIMENEZ

### RESEARCH ENGINEER

Clinical NLP applications, rheumatology & pneumology, corpus development, annotation guidelines, NER human-in the loop, entity linking



## JUDITH ROSELL ROSELL

### JUNIOR RESEARCH ENGINEER

NLP and NER approaches applied to biomaterials domain



## PABLO IGNACIO JESUS ARANCIBIA

### BARAHONA

NLP platform/ datalake full stack, integration & development language models



## LIDIA SALAS ESPEJO

### JUNIOR RESEARCH ENGINEER

Data management, clinical terminology data annotation, data harvesting, preprocessing



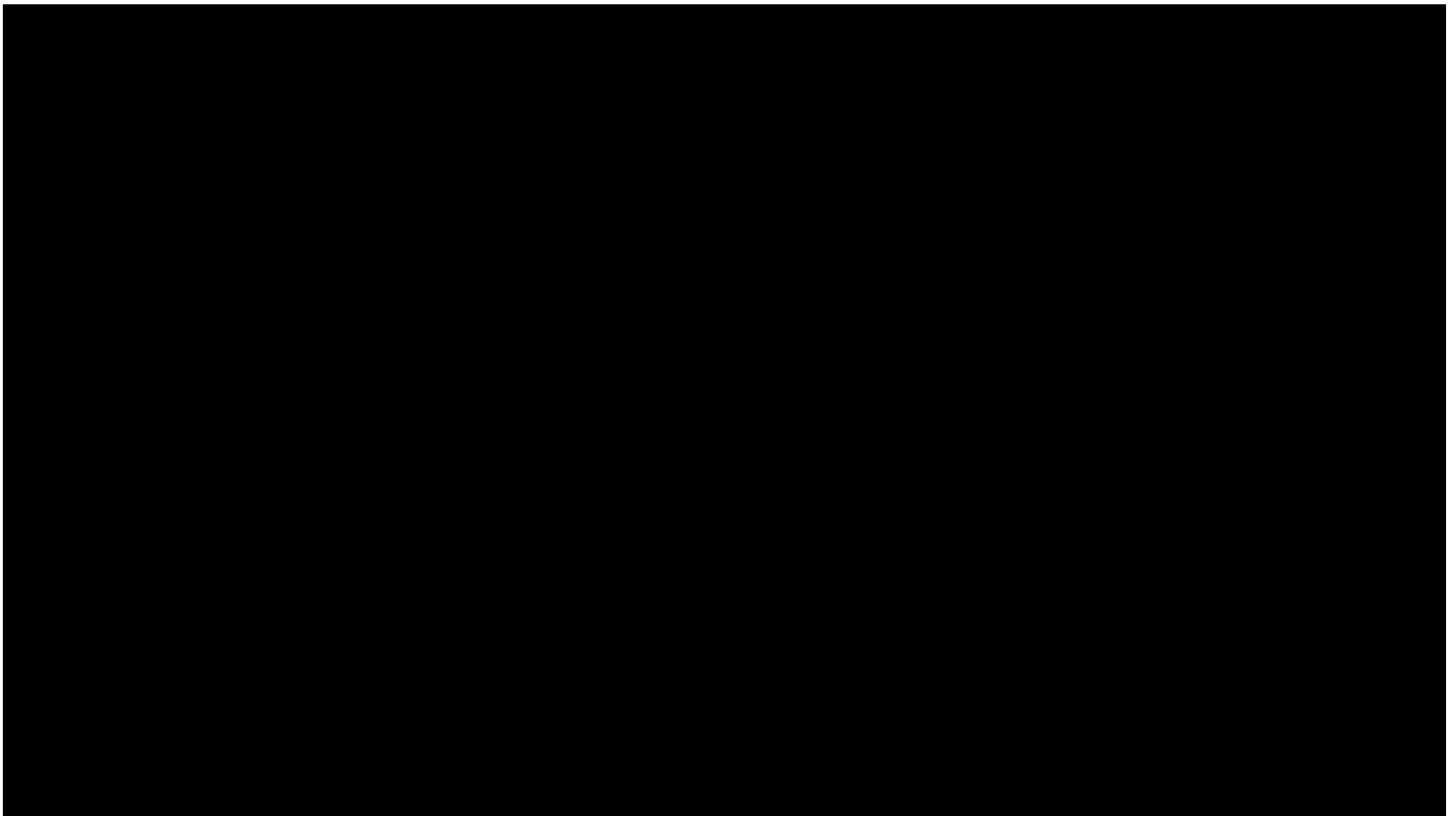
Two open positions !

[mkrallin@bsc.es](mailto:mkrallin@bsc.es)

**WE ARE HIRING**

# Questions ?

<b>SEXO-SUJETO-ASISTENCIA</b>	<b>EDAD-SUJETO-ASISTENCIA</b>	<b>DURATION</b>	<b>HUMAN</b>	<b>HUMAN</b>	<b>ENFERMEDAD</b>	<b>ENFERMEDAD</b>	<b>HUMAN</b>	<b>ENFERMEDAD</b>	<b>HUMAN</b>	<b>ENFERMEDAD</b>	<b>DATE</b>	<b>DURATION</b>	<b>EDAD-SUJETO-ASISTENCIA</b>	<b>PROCEDIMIENTO</b>	<b>DATE</b>	<b>NEG</b>	<b>PROCEDIMIENTO</b>	<b>EDAD-SUJETO-ASISTENCIA</b>	<b>DURATION</b>	<b>SINTOMA</b>	
Mujer	de	23 años	con antecedentes familiares de madre con Lupus Eritematoso Sistémico (LES) y hermana con enfermedad celíaca, y antecedentes personales de Síndrome de Kinsbourne diagnosticado en el primer año de vida, tratado hasta los 8 años con esteroides, actualmente asintomática, una laparoscopia exploradora a los 10 años por dolor abdominal siendo																		
<b>PROCEDIMIENTO</b>	<b>ENFERMEDAD</b>	<b>NEG</b>	<b>ENFERMEDAD</b>	<b>NEG</b>	<b>NSCO</b>	<b>PROFESION</b>	<b>DATE</b>	<b>FECHAS</b>	<b>DATE</b>	<b>SINTOMA</b>											
dicha laparotomía blanca y un aborto espontáneo en la 80 semana de gestación. No alergias medicamentosas conocidas. No tratamiento (tto) médico actualmente. Acude al servicio de urgencias en Abril de 2015, por cuadro de edematisación generalizada intermitente incluyendo cara, párpados y miembros inferiores durante los dos últimos meses. Asocia intensa astenia, más acentuada en los últimos días, cuando comienza con																					
<b>SINTOMA</b>	<b>NEG</b>	<b>NSCO</b>	<b>SINTOMA</b>	<b>SINTOMA</b>	<b>DURATION</b>	<b>SINTOMA</b>	<b>NEG</b>	<b>NSCO</b>	<b>PROFESION</b>	<b>HUMAN</b>	<b>ENFERMEDAD</b>	<b>DATE</b>	<b>DURATION</b>	<b>EDAD-SUJETO-ASISTENCIA</b>	<b>PROCEDIMIENTO</b>	<b>DATE</b>	<b>NEG</b>	<b>PROCEDIMIENTO</b>	<b>EDAD-SUJETO-ASISTENCIA</b>	<b>DURATION</b>	<b>SINTOMA</b>
sensación de opresión torácica no irradiada y disnea de moderados esfuerzos. Refiere cuadro de odinofagia 10 días antes acompañada de fiebre, termometrada de 38°C, siendo diagnosticada por su médico de familia de amigdalitis y tratada con ciprofloxacin sin mejoría. En urgencias, se realizó radiografía de tórax apreciándose un derrame pleural derecho y un aumento evidente de la silueta cardiaca confirmándose la																					
<b>ENFERMEDAD</b>	<b>PROCEDIMIENTO</b>																			<b>SINTOMA</b>	
presencia de derrame pericárdico mediante ecocardiografía. Se decidió derivar a nuestra unidad para estudio. Completando la anamnesis, la paciente negaba la presencia de otros síntomas cardiorrespiratorios, genitourinarios o síntomas sugerentes de patología sistémica: sequedad, artritis, lesiones dérmicas, uveítis, lesiones aftosas, etc... Si presentaba "sensación de retortijón" diariamente post-ingesta acompañado de																			<b>SEI</b>		
4-5 deposiciones de consistencia normal y sin productos patológicos. Pérdida ponderal de unos 10 kg durante los últimos años de forma intencionada con medidas higiénico-diéticas.																					
<b>PROCEDIMIENTO</b>	<b>ENTIDAD-OBSERVABLE</b>	<b>ENTIDAD-OBSERVABLE</b>	<b>SINTOMA</b>	<b>NEG</b>	<b>ENTIDAD-OBSERVABLE</b>	<b>ENTIDAD-OBSERVABLE</b>	<b>SINTOMA</b>	<b>NEG</b>	<b>ENTIDAD-OBSERVABLE</b>	<b>SINTOMA</b>	<b>SINTOMA</b>	<b>SINTOMA</b>	<b>SINTOMA</b>	<b>NEG</b>	<b>ENTIDAD-OBSERVABLE</b>	<b>SINTOMA</b>	<b>NEG</b>	<b>ENTIDAD-OBSERVABLE</b>	<b>SINTOMA</b>	<b>NEG</b>	<b>MORFOLOGIA-NEOPLASIA</b>
En la exploración física presentaba una TA 97/72 mmHg, FC 67 lpm, afebril, SpO2 95% basal con FR normal; buen aspecto general, consciente, orientada y colaboradora, eupneica en reposo, sin signos de bajo gasto, no ingurgitación jugular. No presentaba edema palpebral ni facial. Ausencia de adenopatías laterocervicales, supraclaviculares, axilares ni inguinales; no																					
<b>NSCO</b>	<b>ENFERMEDAD</b>	<b>SINTOMA</b>	<b>PROCEDIMIENTO</b>	<b>NEG</b>	<b>NSCO</b>	<b>SINTOMA</b>														<b>ENTIDAD-OBSERVABLE</b>	
bocio ni masas cervicales. Resto de la exploración cardiorrespiratoria, abdominal y de miembros dentro de la normalidad sin apreciarse en dicho momento signos de edema.																			<b>NEG</b>		
Los resultados de las pruebas complementarias solicitadas inicialmente fueron los siguientes:																					
<b>PROCEDIMIENTO</b>	<b>ENTIDAD-OBSERVABLE</b>	<b>SINTOMA</b>	<b>ENTIDAD-OBSERVABLE</b>	<b>PROCEDIMIENTO</b>	<b>NEG</b>	<b>NSCO</b>	<b>SINTOMA</b>														
— Hemograma: series blanca, roja y plaquetaria normales, coagulación y VSG normal.																					
<b>ENTIDAD-OBSERVABLE</b>	<b>ENTIDAD-OBSERVABLE</b>	<b>ENTIDAD-OBSERVABLE</b>	<b>ENTIDAD-OBSERVABLE</b>	<b>ENTIDAD-OBSERVABLE</b>	<b>ENTIDAD-OBSERVABLE</b>	<b>ENTIDAD-OBSERVABLE</b>	<b>ENTIDAD-OBSERVABLE</b>	<b>ENTIDAD-OBSERVABLE</b>	<b>ENTIDAD-OBSERVABLE</b>	<b>ENTIDAD-OBSERVABLE</b>	<b>ENTIDAD-OBSERVABLE</b>	<b>ENTIDAD-OBSERVABLE</b>	<b>ENTIDAD-OBSERVABLE</b>	<b>ENTIDAD-OBSERVABLE</b>	<b>ENTIDAD-OBSERVABLE</b>	<b>ENTIDAD-OBSERVABLE</b>	<b>ENTIDAD-OBSERVABLE</b>	<b>ENTIDAD-OBSERVABLE</b>	<b>ENTIDAD-OBSERVABLE</b>	<b>ENTIDAD-OBSERVABLE</b>	
— Bioquímica con función renal, ácido úrico, función hepática, amilasa, LDH, iones, calcio, fósforo, factor reumatoide, hormonas tiroideas, perfil ferrocínético, cortisol basal, lípidos, albúmina, proteínas totales y PCR normales.																					
<b>ENTIDAD-OBSERVABLE</b>	<b>ENTIDAD-OBSERVABLE</b>																				
— NTproBNP 288.4 pg/ml, TTu 6 ng/l;																					
<b>SINTOMA</b>	<b>PROCEDIMIENTO</b>																				
— Sedimento y bioquímica de orina: normales.																					
— Inmunoglobulinas y complemento normales.																					
<b>ENTIDAD-OBSERVABLE</b>	<b>ENTIDAD-OBSERVABLE</b>	<b>SPECIES</b>	<b>SPECIES</b>	<b>ENFERMEDAD</b>	<b>SPECIES</b>	<b>PROCEDIMIENTO</b>	<b>PROCEDIMIENTO</b>	<b>NEG</b>	<b>PROCEDIMIENTO</b>	<b>NEG</b>	<b>SINTOMA</b>										
— Autoinmunidad: ANA positivo a título 1/320 con patrón moteado; ENAS y Ac antifosfolípidos negativos. Test de coombs directo negativo.																					
<b>PROCEDIMIENTO</b>	<b>ENFERMEDAD</b>	<b>SINTOMA</b>	<b>ENFERMEDAD</b>	<b>ENFERMEDAD</b>	<b>ENFERMEDAD</b>	<b>ENFERMEDAD</b>	<b>ENFERMEDAD</b>	<b>ENFERMEDAD</b>	<b>ENFERMEDAD</b>	<b>ENFERMEDAD</b>	<b>ENFERMEDAD</b>	<b>ENFERMEDAD</b>	<b>ENFERMEDAD</b>	<b>ENFERMEDAD</b>	<b>ENFERMEDAD</b>	<b>ENFERMEDAD</b>	<b>ENFERMEDAD</b>	<b>ENFERMEDAD</b>	<b>ENFERMEDAD</b>	<b>ENFERMEDAD</b>	
— Microbiología: Serologías para VHB, VHC, Sífilis, VIH, hemocultivos, Mantoux y test de IgM negativos.																					
<b>PROCEDIMIENTO</b>	<b>ENFERMEDAD</b>	<b>MORFOLOGIA-NEOPLASIA</b>	<b>ENFERMEDAD</b>	<b>ENFERMEDAD</b>	<b>ENFERMEDAD</b>	<b>ENFERMEDAD</b>	<b>ENFERMEDAD</b>	<b>ENFERMEDAD</b>	<b>ENFERMEDAD</b>	<b>ENFERMEDAD</b>	<b>ENFERMEDAD</b>	<b>ENFERMEDAD</b>	<b>ENFERMEDAD</b>	<b>ENFERMEDAD</b>	<b>ENFERMEDAD</b>	<b>ENFERMEDAD</b>	<b>ENFERMEDAD</b>	<b>ENFERMEDAD</b>	<b>ENFERMEDAD</b>	<b>ENFERMEDAD</b>	
— TAC toraco-abdominal: poliserositis con afectación pleural, pericárdica y presencia de líquido libre peritoneal asociada a adenopatías axiales, inguinales y retroperitoneales a valorar enfermedad infecciosa/inflamatoria.																					
<b>PROCEDIMIENTO</b>	<b>ENFERMEDAD</b>	<b>ENFERMEDAD</b>	<b>NEG</b>	<b>NSCO</b>	<b>ENFERMEDAD</b>																
— Ecocardiograma: derrame pericárdico de predominio posterior moderado-severo sin signos de compromiso hemodinámico; válvula aórtica trivalva.																					
<b>DIAGNÓSTICO</b>																					
<b>NEG</b>	<b>NSCO</b>																				
Se determinaron los niveles de C1q cuyos valores se encontraban dentro de la normalidad con una actividad del 110% descartándose el Angioedema Hereditario. Asimismo, se solicitó un estudio genético para FMF en el que se informa de la presencia de la mutación del gen MEFV (variante p. Glu148Gln), lo que confirma el diagnóstico de esta entidad. Ante los antecedentes de celiacos y alteración en el hábito intestinal, se determinaron Ac antigliadina y trasglutamínasa que fueron negativos y estudio genético que fue positivo para el HLA-DQ2.																					
<b>TRATAMIENTO Y EVOLUCIÓN</b>																					
<b>NSCO</b>	<b>ENFERMEDAD</b>																				
Desde el ingreso, se inició terapia con omeprazol, ibuprofeno y colchicina asociándose hidroxicloroquina 200 mg cada 24 horas ante la sospecha de LES. La paciente presentó buena evolución clínica con desaparición de los síntomas cardiorrespiratorios y permaneciendo afebril sin presentar complicaciones durante el ingreso. Finalmente, se suspendió todo el tto dejando únicamente colchicina tras recibir los resultados del																					
test genético para FMF en el que se observó la citada mutación en el gen MEFV; en las sucesivas revisiones, la paciente refiere encontrarse asintomática.																					

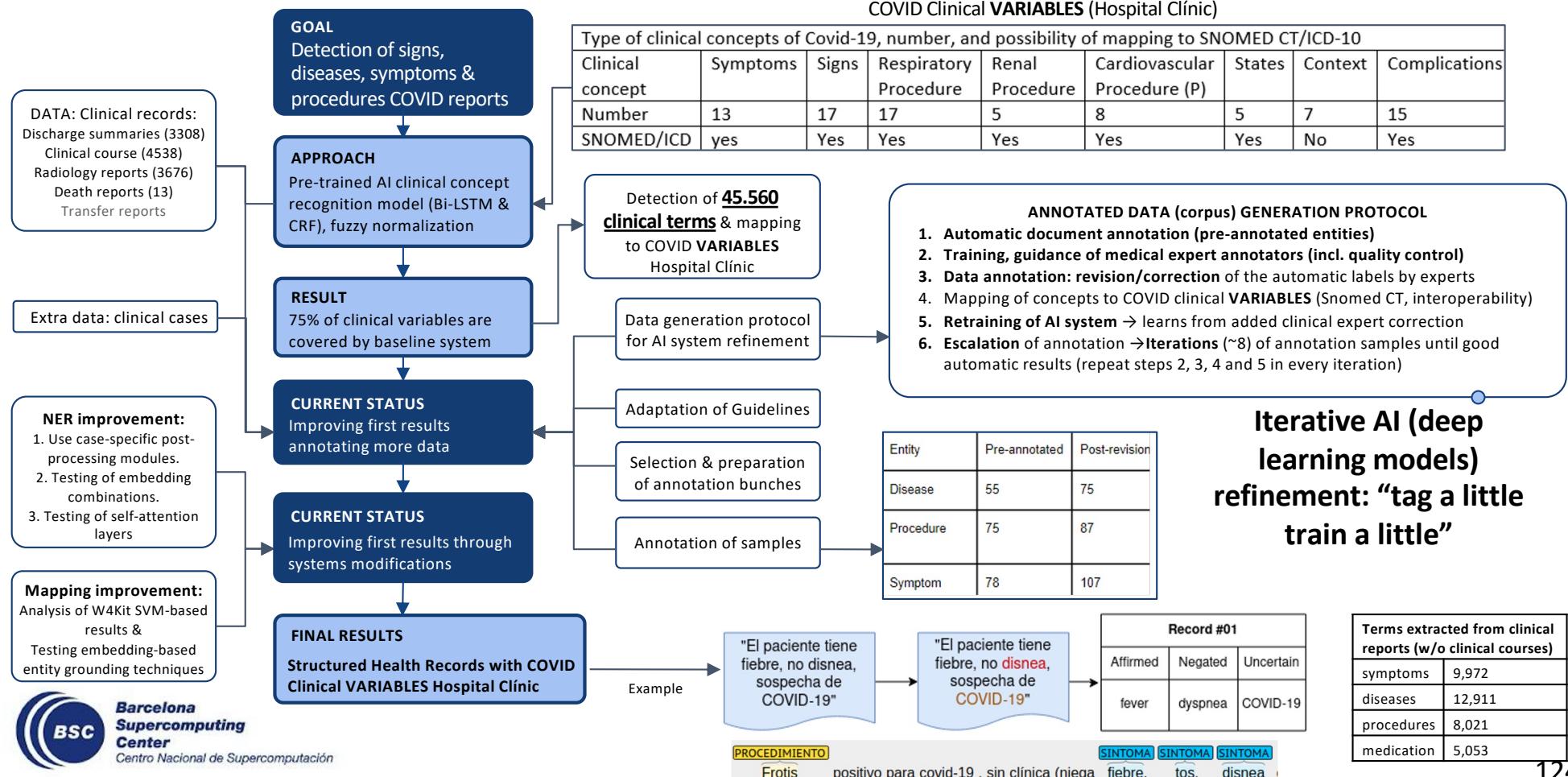


# Additional Slides



*Barcelona  
Supercomputing  
Center*  
Centro Nacional de Supercomputación

# Clinical NLP processing of COVID patient records



# Negation and Uncertainty

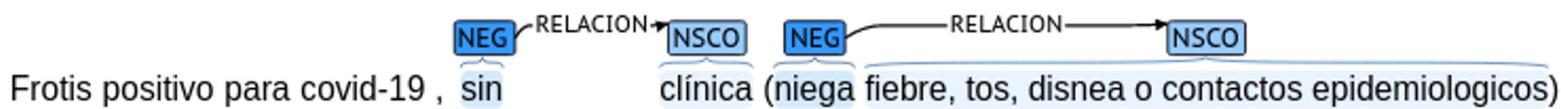
## Approach

- Trained AI models with existing corpora in Spanish (NUBES, Lima López et al., 2020) and used them to detect negations and speculations in HC's EHRs.
- Depending on file and specialty, up to 25 negations and 10 speculations in a single file.
- Next step: manual refinement of results to tune models (new Gold Standard).

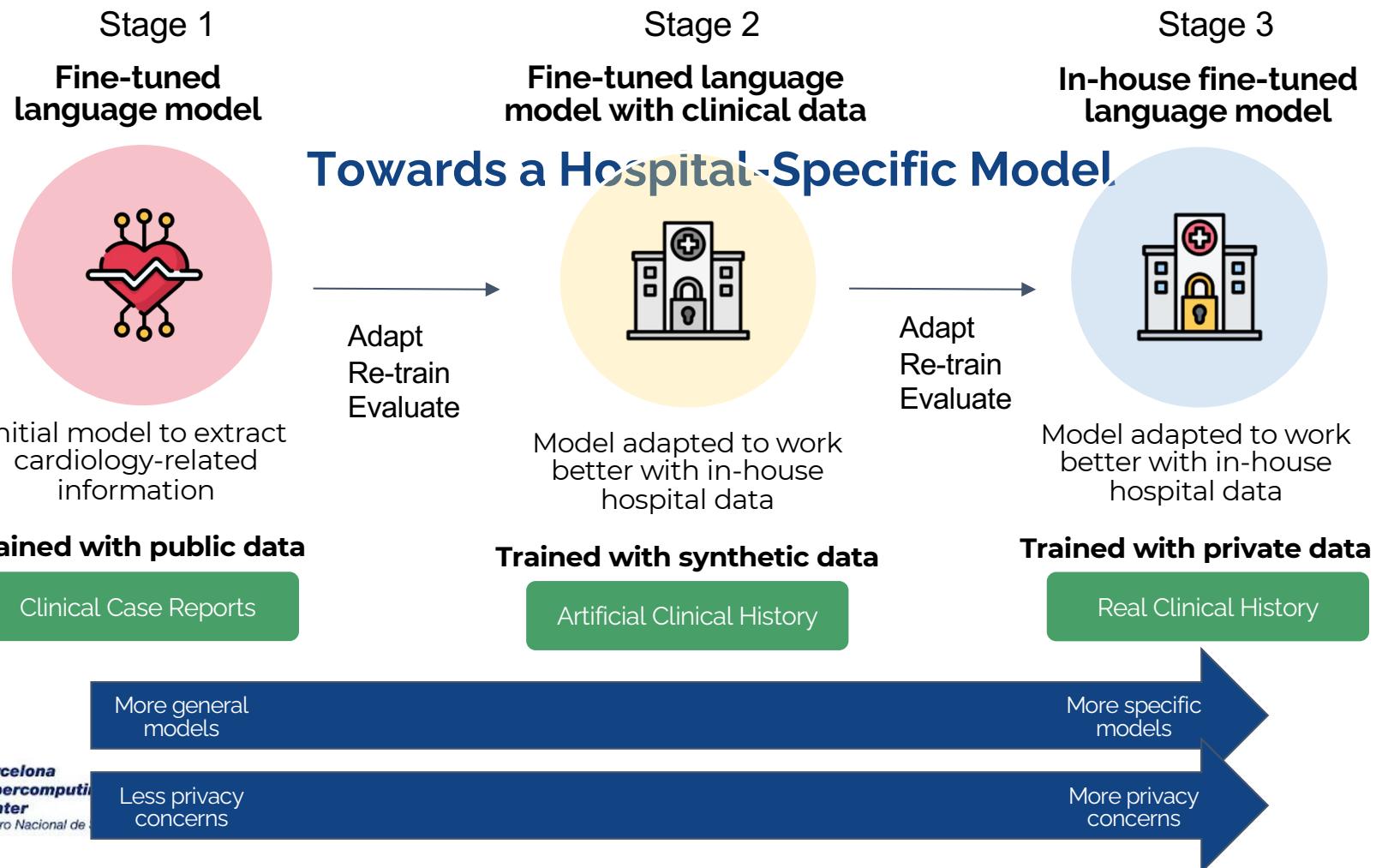


Annotation of data to improve NER models in collaboration with linguists from W4Kit (in process)

Iterative improvement of the models/creation of new ones.



## NLP pipeline development overview: stages



## DisTEMIST: ENFERMEDAD label

**ENFERMEDAD** (*disease*) includes clinical conditions that affect the normal functioning of the body and that have a certain extension in time.

- 1 Paciente de 70 años de edad, minero jubilado, sin alergias medicamentosas conocidas, que presenta como antecedentes personales: accidente laboral antiguo con fracturas vertebrales y costales; interventido de enfermedad de Dupuytren en mano derecha y by-pass iliofemoral izquierdo; Diabetes Mellitus tipo II, hipercolesterolemia e hiperuricemia; enolismo activo, fumador de 20 cigarrillos / día.
- 2 Es derivado desde Atención Primaria por presentar hematuria macroscópica postmiccional en una ocasión y microhematuria persistente posteriormente, con micciones normales.

## SympTEMIST: SINTOMA label

**SINTOMA** (*symptom*) includes subjective and objective manifestations of clinical conditions (symptoms and signs), as well as qualitative descriptions of procedures' results.

2 La exploración física revela: T<sup>a</sup> 40,2 C; T.A: 109/68 mmHg; Fc: 105 lpm. Se encuentra consciente, orientado, sudoroso, eupneico, con  
buen estado de nutrición e hidratación. En cabeza y cuello no se palpan adenopatías, ni bocio ni ingurgitación de vena yugular, con  
pulsos carotídeos simétricos. Auscultación cardíaca rítmica, sin soplos, roces ni extratonos. Auscultación pulmonar con  
conservación del murmullo vesicular. Abdomen blando, depresible, sin masas ni megalias. En la exploración neurológica no se detectan  
signos meníngeos ni datos de focalidad. Extremidades sin varices ni edemas. Pulsos periféricos presentes y simétricos. En la exploración urológica  
se aprecia el teste derecho aumentado de tamaño, no adherido a piel, con zonas de fluctuación e intensamente doloroso a la palpación, con  
pérdida del límite epidídimo-testicular y transiluminación positiva.

# ProcTEMIST: PROCEDIMIENTO label

**PROCEDIMIENTO** (*procedure*) includes actions or interventions performed by health professionals in order to diagnose or treat diseases and health problems.

- 2 En la exploración abdominal se palpó una tumoración dura en hipocondrio izquierdo que llegaba hasta el mesogastrio y fosa iliaca izquierda. Los estudios imagenológicos realizados fueron: ultrasonido abdominal, radiografía de tórax, urograma descendente y tomografía axial computarizada.
- 3 En la radiografía de tórax no se detectaron metástasis. Los hallazgos obtenidos en los demás estudios imagenológicos indicaron la presencia de una extensa masa bien delimitada, de contornos regulares y con señales heterogéneas sugerentes de áreas hemorrágicas y necróticas. Había desplazamiento de las estructuras vecinas con infiltración. No se demostró infiltración del hígado ni la vena cava inferior. El tratamiento consistió en la cirugía de exéresis por vía toraco-abdominal con clampaje de la aorta torácica para el control de la hemorragia transoperatoria. Se realizó además de la resección del tumor nefrectomía izquierda y esplenectomía en bloque con paquetes de adenopatías hilarias.

# DrugTEMIST: FARMACO label

**FARMACO** (*drug*) includes specific products, compounds or substances with a defined molecular composition and a therapeutic purpose.

1 \* MEDICACIÓN EN SALA

FARMACO

2 - Amoxicilina/clavulánico 875/125 mg/8h hasta 29/09 incluído

FARMACO

3 - Hidroxicloroquina 200mg/12h hasta 1/10 incluído

FARMACO

4 - Prednisona en pauta descendiente (26/09 a 28/09 30mg/24h, 29/09 a 1/10 20mg/24h, 2/10 a 4/10 10mg/24h y 5/10 STOP.

FARMACO

5 - Enoxaparina 100mg/24h

FARMACO

6 - AAS 100mg/24h

FARMACO

7 - Atorvastatina 40mg/24h

FARMACO

8 - Salbutamol 2INH/6h si precisa



Ba  
Su  
Ce

Centro Nacional de Supercomputación

## SPACCC: ENTIDAD\_OBSERVABLE label

**ENTIDAD\_OBSERVABLE** (*observable entity*) includes mentions of specific aspects assessed as part of exploratory and laboratory procedures.

- 3 Al ingreso al servicio de urgencias se encuentra con presión arterial (TA): 130/80, frecuencia cardíaca (FC): 80
- pulsaciones por minuto, frecuencia respiratoria (FR): 18 respiraciones por minuto, temperatura (T): 38.7 grados centígrados
- (D-dímero) mayor a 20.000ng/ml, trombocitopenia máxima de 47.000/l y tiempo de protrombina máximo de 1,58 ratio), fracaso renal agudo con cifras de creatinina de hasta 2,33mg/dl y elevación de la lactato deshidrogenasa (LDH) con valores máximos de

# LivingNER: SPECIES label

**SPECIES** (*species*) includes living beings with the exception of humans. It covers mostly pathogens, as well as animals and food.

si bien, las infecciones fúngicas producidas por Aspergillus, Histoplasma, Criptococo y Pneumocystis jirovecii, entre otras, y algunas bacterianas como Legionella y Neumococo, u otras menos comunes, como Listeria o Nocardia también deberían formar parte del diagnóstico diferencial. De entre las infecciones víricas que se manifiestan con frecuencia en forma de neumonía en este tipo de pacientes, habría que considerar adenovirus, influenza (A y B) y parainfluenza, virus respiratorio sincitial (VRS) y citomegalovirus (CMV). La tuberculosis pulmonar es probablemente la infección respiratoria cuya asociación más se ha estudiado en pacientes tratados con antiTNF- $\alpha$ . Estudios recientes mostraron que España y

## LivingNER: HUMAN label

**HUMAN** (*human*) includes mentions of human beings, including family members, occupations and more.

7 A la entrevista dirigida, la paciente nos refiere historia familiar materna de diabetes (bisabuela, abuela y hermana de madre) todas ellas tratadas con antidiabéticos orales. Existencia de varones en dicha rama familiar de múltiples quistes, sin haber sido sometidos a estudio, localizados en cabeza y cuello y síndrome de ovario poliquístico en hermana materna. No antecedente de enfermedad cardiovascular por rama paterna.

# CANTEMIST: MORFOLOGIA\_NEOPLASIA label

**MORFOLOGIA\_NEOPLASIA** (*neoplasms morphology*) includes mentions of neoplasms, tumours and similar growths.

- 13 - Carcinoma indiferenciado de pulmón T4N0M1b (única lesión cerebral de 2.5cm) proponiéndose tratamiento RT holocrebral, seguido de QT de inducción, posteriormente QT/RT con intención radical y finalmente radiocirugía
- 14 - Carcinoma indiferenciado de pulmón T4N0M1b (lesión cerebral, metástasis suprarrenal) proponiéndose RT holocraneal, resección por laparoscopia en las localizaciones que capta el PET-TAC y finalmente QT paliativa Dadas las dos posibilidades se solicita PET-TAC donde se describe nódulo hipermetabólico de 31mm ( SUV 4.3) en fosa renal izquierda, adyacente a cambios postquirúrgicos y próximo a la glándula suprarrenal homolateral, igualmente con infiltración maligna (probable recidiva local vs adenopatía), así como implante en grasa perihepática sugestivo de diseminación oligometástasica.
- 15 Ante la sospecha de 2 tumores sincrónicos: Carcinoma indiferenciado de pulmón y recidiva local de neoplasia renal, se

# PharmaCoNER: NORMALIZABLE/NO\_NORMALIZABLE label

**NORMALIZABLE/NO\_NORMALIZABLE** includes mentions of chemical substances, divided by whether they could be normalized to SNOMED CT or not.

3 Con la sospecha diagnóstica de rotura no traumática de un feocromocitoma pre-existente, se determinaron metanefrinas plasmáticas, que fueron

normales, y catecolaminas y metanefrinas urinarias. En la orina de 24 horas del día siguiente al ingreso se obtuvieron los siguientes resultados:

adrenalina: 65,1 mcg (valores normales -VN: 1,7-22,5), noradrenalina: 151,1 mcg (VN: 12,1-85,5), metanefrina: 853,5 mcg (VN: 74-297) y

normetanefrina: 1396,6 mcg (VN: 105-354). A los 10 días, todavía ingresado el paciente, las cifras urinarias se habían normalizado por completo de modo espontáneo.

1 Se trata de un paciente masculino de 70 años, quien ingresó en el servicio de urgencias del Hospital Pablo Tobón Uribe, con cuadro de aproximadamente una hora de evolución consistente en opresión torácica, malestar general, astenia y diaforesis; que iniciaron después de haber

ingerido 100 mg de sildenafil, niega ingesta de otro estimulante sexual o cocaína y sin relación sexual después de su consumo. El paciente como único antecedente clínico sufría de hipertensión arterial, controlada farmacológicamente y niega episodios previos de angina o

consumo de nitratos. El examen clínico y sus signos vitales eran normales; sin embargo, después de la valoración inicial presenta paro



# PharmaCoNER: PROTEINAS label

**PROTEINAS** (*proteins*) includes mentions of proteins and genes.

(positividad para vicentina, focalmente para CK22 y AEI-AE3 y S-100, siendo negativo para PLAP, CD30, CD117,

CD45, CD20 y cromogramina). Pleomorfismo nuclear, actividad mitótica, focos de necrosis, sin observarse embolización vascular. La

- 5 El análisis sanguíneo muestra valores dentro de la normalidad para alfa-feto proteína, HCG y LDH (HCG: 1,74 U/l; Alfa-feto proteína: 2.07 ng/ml; LDH: 299 mU/ml).

# Socio-demographic and World Knowledge Entities



*Barcelona  
Supercomputing  
Center*  
Centro Nacional de Supercomputación

## MEDDOPLACE: GPE\_GEN/GPE\_NOM label

**GPE** includes mentions of named (*NOM*) and generic (*GEN*) geo-political entities such as countries, cities or continents.

- 4 El paciente había viajado a Italia el día -9 para asistir a un partido de rugby en Roma; luego, había viajado con su pareja y dos amigos por la Lombardía, el Véneto y la Toscana, desplazándose en un coche de alquiler y haciendo estancia en casas privadas de alquiler, antes de regresar a Escocia desde Milán el día -2.
- 6 La madre no tenía fiebre ni tos, y no convivían con otros miembros de la familia.
- 7 El barrio de la familia del paciente se halla a unos 15 km del mercado mayorista de marisco de Huanan, en Wuhan, y no constan antecedentes familiares de contactos con dicho mercado.
- 8 Sin embargo, a 5 personas del vecindario en que reside el paciente se les había diagnosticado COVID-19.
- 9 El niño nació a término y el primer día se le administró la vacuna antituberculosa (BCG).

## MEDDOPLACE: GEO\_GEN/GEO\_NOM label

**GEO** includes mentions of named (*NOM*) and generic (*GEN*) geographical accidents and habitats such as mountains, seas and areas with specific characteristics.

1 Se trata de un joven de 20 años de edad, que se encontraba en las aguas de la playa Red Frog de Isla Bastimentos.

GEO

2 Los testigos observaron que el joven se encontraba en el mar y "pidió ayuda", se sumergió y volvió a salir a unos 15 metros de donde se había sumergido y no lo volvieron a ver.

GEO\_GEN

24 Sin embargo, se consideró poco probable que el caso correspondiera a rabia dado el antecedente de la zona epidemiológica de la exposición

GEO\_GEN

(zona urbana de Floridablanca) y el tiempo transcurrido desde la mordedura por el murciélagos.

GEO\_GEN [RSI]

42 Teniendo en cuenta que el paciente procedía de un área endémica de enfermedad de Chagas, se realizaron pruebas serológicas para detectar anticuerpos circulantes contra Trypanosoma cruzi: ambas resultaron positivas (ELISA 8 e inmunofluorescencia >1/160), lo que confirmó el diagnóstico de infección crónica por dicho parásito.

10 También se realizó un tránsito horitodo esofagogastrico y radiografía de abdomen que descartaron afectación gástrico-intestinal.

## MEDDOPLACE: FAC\_GEN/FAC\_NOM label

**FAC** includes mentions of named (*NOM*) and generic (*GEN*) human-made facilities such as hospitals, schools, airports, supermarkets and more.

72 En diciembre de 2014 firma el consentimiento informado para preselección dentro del ensayo clínico EDI1001 del hospital Virgen del Rocío, al que finalmente no es candidato. FAC\_NOM [AT]

1 Niña de tres años de edad, que acude a la FAC\_NOM [AT] Facultad de Odontología de la Universidad de Sevilla remitida por el FAC\_NOM [AT] Hospital General Juan Ramón Jiménez de Huelva.

3 Se le informa de la no disponibilidad de la vacuna en el FAC\_GEN [AT] centro de salud, ya que no es una vacunación presente en el calendario vigente de FAC\_GEN [AT] inmunizaciones de la Comunidad de Madrid, así como de la posibilidad de asistir al FAC\_GEN [AT] Centro de Vacunación Internacional.

4 En la anamnesis cabe destacar el antecedente de que ambas habían acudido el fin de semana anterior a un FAC\_GEN balneario y se bañaron también en una FAC\_GEN piscina de pequeño tamaño en el FAC\_GEN jardín de la vivienda. FAC\_GEN

## MEDDOPLACE: DEPARTAMENTO label

**DEPARTAMENTO** (*department*) includes mentions of clinical departments and locations within a hospital or healthcare facility.

12 El lactante se hospitalizó en la **DEPARTAMENTO** unidad de cuidados intensivos neonatales por microcefalia y frente amplia.

13 Posteriormente, el paciente se remitió a la **DEPARTAMENTO** consulta externa de dismorfología, a la cual llegó con 2 meses de edad y los siguientes hallazgos

12 Obesidad mórbida (110 kg) en seguimiento por **DEPARTAMENTO** Endocrinología Pediátrica.

13 Síndrome de Apnea Obstructiva de Sueño en seguimiento por **DEPARTAMENTO** Neumología Pediátrica.

14 Seguimiento rutinario por **DEPARTAMENTO** Neurología Pediátrica.

3 En planta previo a quirófano, el paciente no presenta antecedentes personales de interés salvo apendicectomía, no fumador, no toma ninguna medicación.

## MEDDOPLACE: COMUNIDAD label

**COMUNIDAD** (*community*) includes mentions of nationalities, religions and ethnicities.

1 Mujer Colombiana de 65 años que presenta sintomatología de 1 año y 9 meses de dolor episódico tipo punzada localizado en hipocondrio y flanco

58 Afecta a poblaciones de origen **COMUNIDAD** italiano, **COMUNIDAD** griego, **COMUNIDAD** español, **COMUNIDAD** árabe y **COMUNIDAD** judío.

## MEDDOPLACE: IDIOMA label

**IDIOMA** (*language*) includes mentions of languages and language barriers.

- 10 Las dificultades de manejo y los problemas de expresión del paciente en lengua española motivaron la intervención del trabajador social
- 1 Mujer haitiana de 29 años, de raza negra, con el antecedente de una ACF, pero de conocimiento en Chile solo desde el 2017 por limitaciones de idioma.

## MEDDOPLACE: TRANSPORTE label

**TRANSPORTE** (*transportation*) includes mentions of patient movement and methods of transportation.

4 **TPT**  
Viajes frecuentes a EEUU.

5 **TRANSPORTE**  
Vacaciones sur california 3 semanas e Israel.

10 Enfermedad actual: refiere dolor pleurítico y expectoración hemoptoica dese hace 2 días, tras viaje transoceánico.

1 Paciente 23 años que sufre accidente de tráfico al chocar su **motocicleta** contra una **furgoneta**.

9 Tías maternas en tratamiento psiquiátrico, presentando temblor en extremidades, una de ellas en **silla de ruedas**.

## MEDDOPROF: PROFESION label

**PROFESSION** (*profession*) includes mentions of occupations, including job titles and job descriptions.

- 15 Informe clínico del paciente: Varón de 26 años de edad **deportista profesional** (*jugador de fútbol*) sin otros antecedentes personales de interés para el caso que nos ocupa. Sufre una caída durante la práctica deportiva desde su propia altura, apoyando la muñeca izquierda y produciéndose una hiperextensión de ésta. A consecuencia del trauma presenta dolor, deformidad e impotencia funcional, con exploración neurovascular distal normal.
- 1 Una mujer de 27 años sin antecedentes, **residente del hospital**, presentó odinofagia seguida de artralgia difusa y una erupción de placas eritematosas pruriginosas extendidas, con una afectación básicamente facial y acra. El diagnóstico de urticaria fue confirmado por un **dermatólogo**.
- 1 Paciente de 27 años, sexo masculino, **operario de la industria química**, que consulta por dermatitis de un año de evolución. Refiere que sus síntomas cutáneos comenzaron a los cuatro meses de su **ingreso a una fábrica de sales de cromo**, a los que agrega episodios de rinitis serosa y lagrimeo



## De-identified health records in Spanish and Catalan for medical entity recognition and anonymization

Anonymized clinical reports corpus

Anonymization protocol

Guidelines for manual anonymization

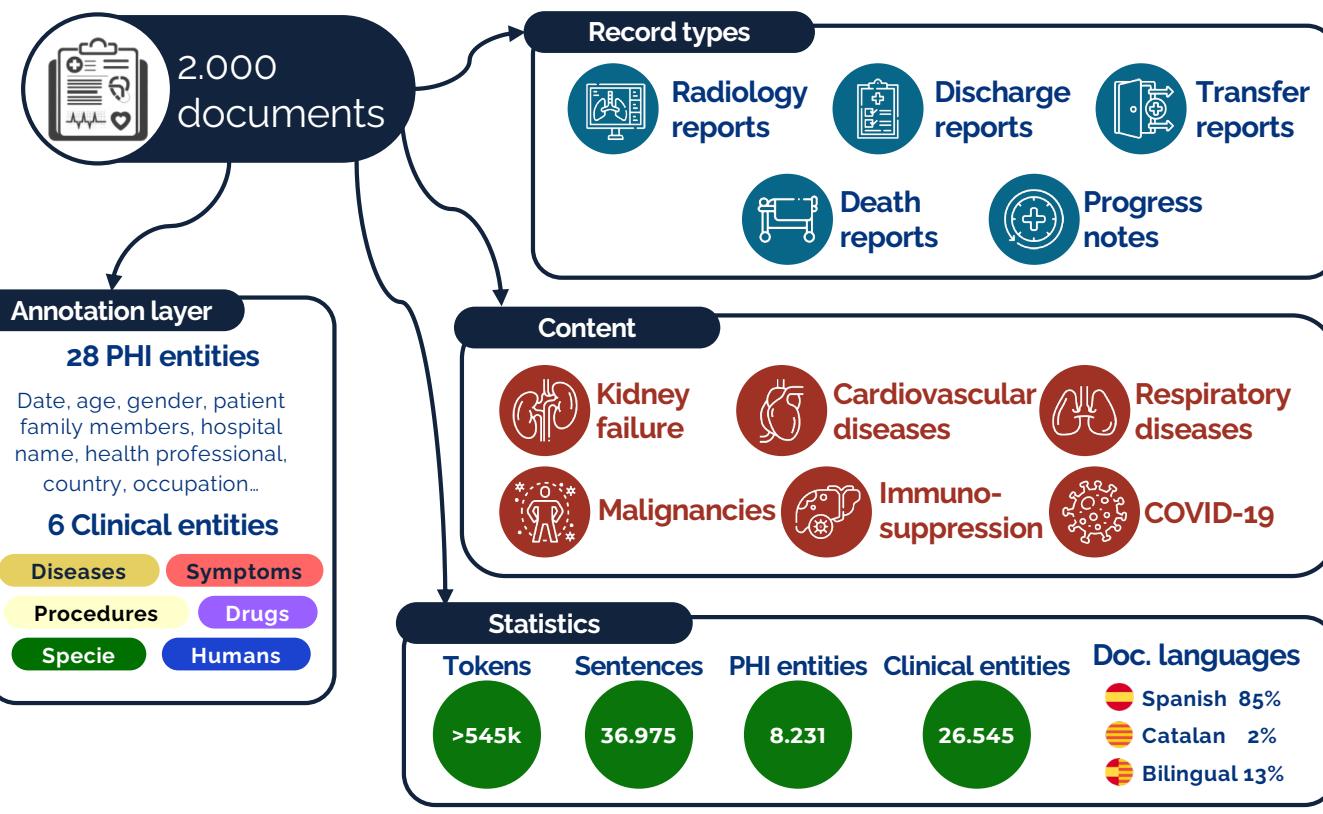
Guidelines for clinical entities annotation

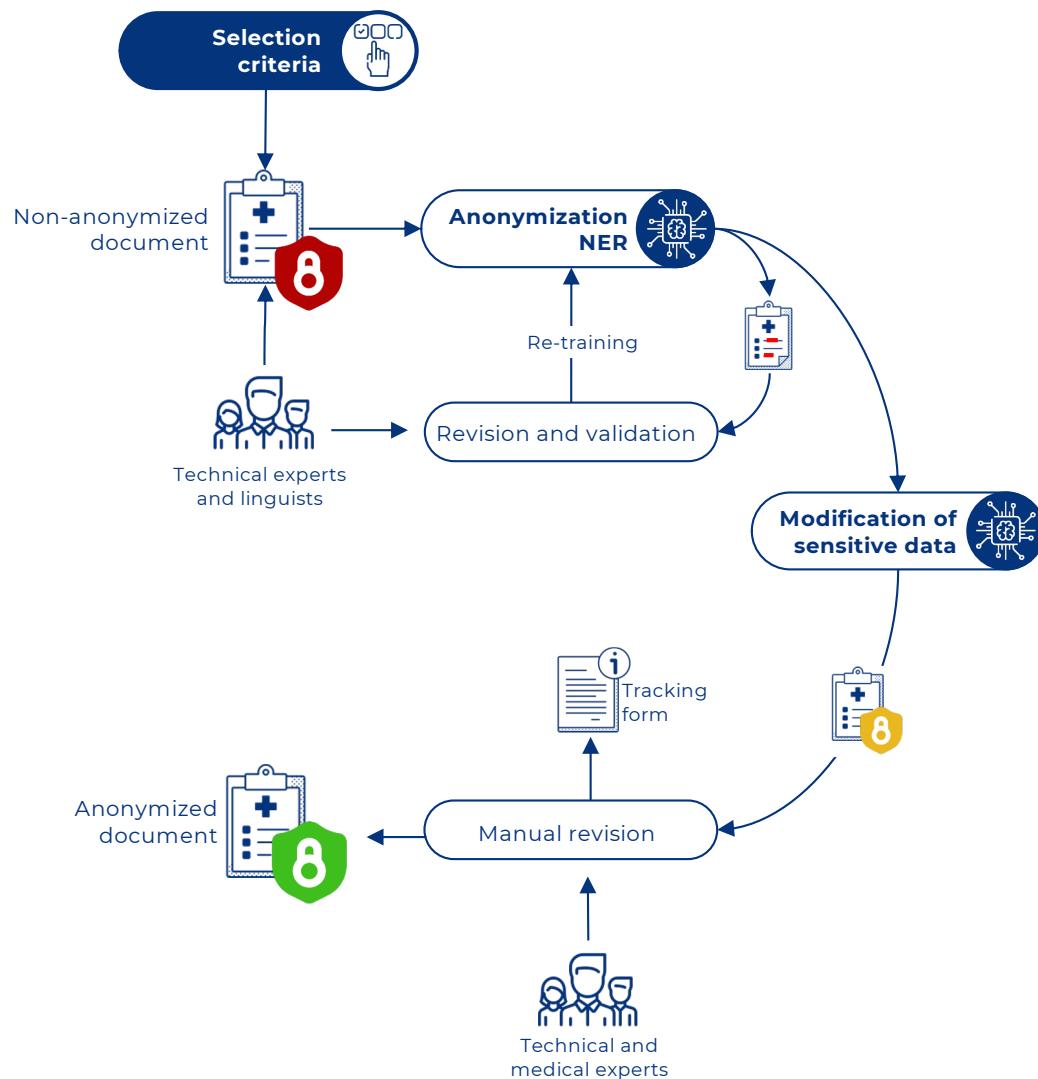
Pre-trained anonymization model

Pre-trained clinical annotation models



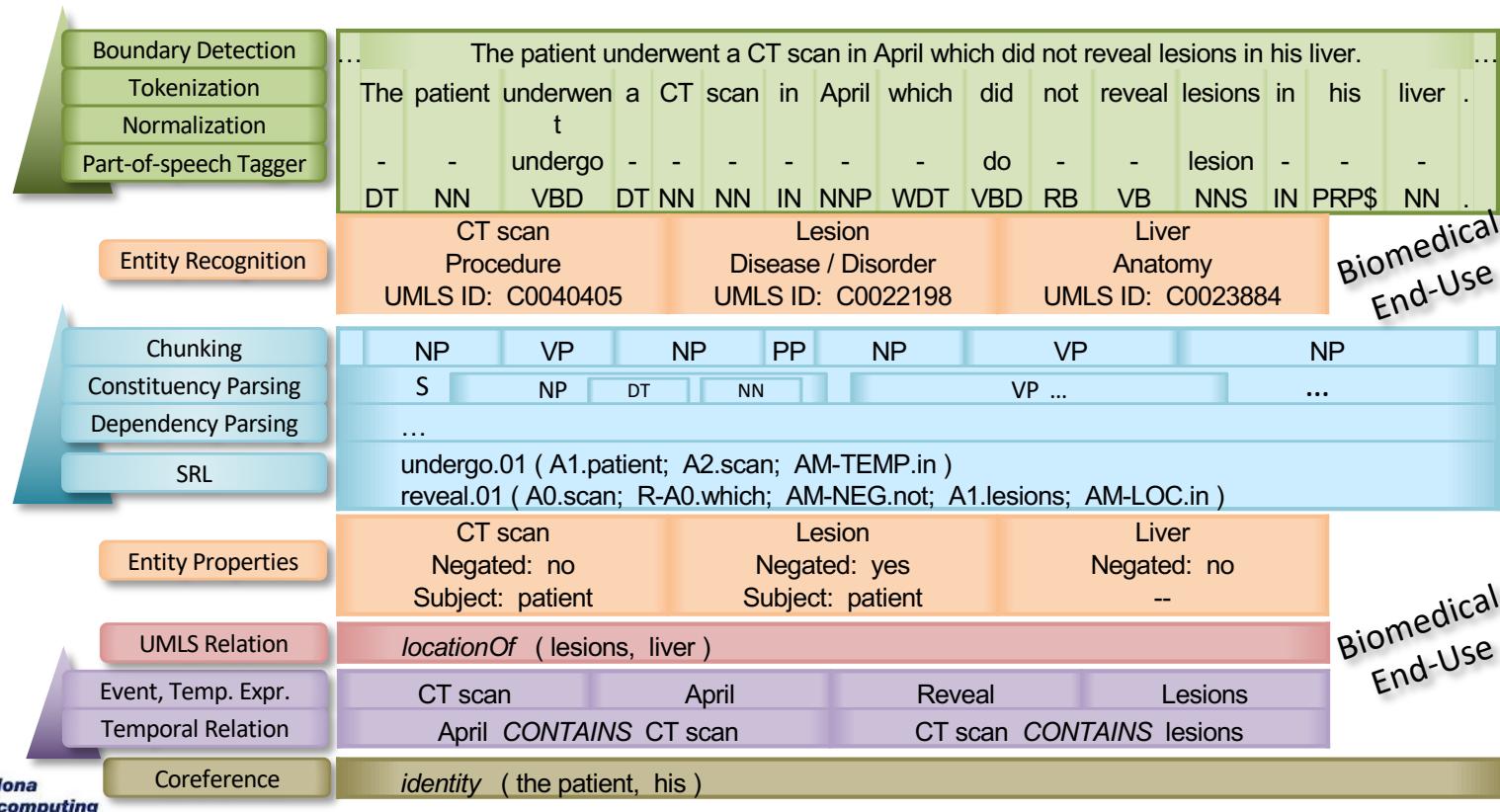
Barcelona Supercomputing Center  
Centro Nacional de Supercomputación





# Anotaciones relevantes para PLN clínico: cTakes

The patient underwent a CT scan in April which did not reveal lesions in his liver.



# Clinical text pre-processing: sectionizer

DIAGNÓSTICOS
1. ICTUS PACI EN TERRITORIO DE ARTERIA CEREBRAL MEDIA DERECHA.
2. ESTENOSIS CAROTÍDEA DERECHA NO SIGNIFICATIVA. 3. ICTUS LACUNAR CRÓNICO. ENFERMEDAD DE PEQUEÑO VASO CEREBRAL
4. HIPERTENSIÓN ARTERIAL
5. DIABETES MELLITUS TIPO 2
6. VASCULOPATÍA PERIFÉRICA
7. EPOC
8. OMALGIA IZQUIERDA
Paciente de 74 años, rankin 0, sin AMC, ex-tabaquismo (>60paq/año), ex-enolismo.]
ANTECEDENTES
- DM tipo 2 en tratamiento con antidiabéticos orales. Hb A1c 6,1 en enero 2016
- HTA en tratamiento médico.
- EPOC en tratamiento broncodilatador.
- Hiperuricemia.
- Ingreso en MI en 2013 por monoartritis de rodilla inflamatoria con cultivos negativos y síndrome diarreico con aislamiento i
- Hemicolectomía D por adenoma tubular con displasia de alto grado en 2011. Ulcus péptico diagnosticado por fibrogastrosc
epigastralgia
- Hernia discal
- Aneurisma de aorta infrarenal (TAC control 2015. Estabilidad del aneurisma aórtico infrarenal de 31 mm (diámetro AP - co
- Vasculopatía periférica: claudicación intermitente.
- Ateromatosis Carotídea Bilateral : Eco TSA (marzo 2016): eje derecho: placa homogénea y regular en bulbo e inicio de CI 30-
heterogénea, irregular en bulbo e inicio de CI y CE 50- 69%.
- Vertigo periférico de larga evolución.
Medicación: Adiro 100mg, Alopurinol 300 mg/24 h, Rilast Forte 1/24 h, Singulair 10 mg /24 h, Atrovent, Tramadol 50 mg, 1-1-0
40/10(olmesartan/antagonista del calcio) 1 comp/24 h, Hidroclorotiazida 50 mg/24 h, Diamicron 30 mg 2 -0-0, Metformina 1-0-1, Orfidal 0-0-
Tamsulosina 0.4.

Type	Set	Start	End	Id	
Diagnósticos		18	23	7612	{nombre=ictus}
Diagnósticos		132	137	7613	{nombre=ictus}
Antecedentes		406	408	7616	{nombre=diabetes}
Antecedentes		486	489	7617	{nombre=hipertension}



Barcelo  
Superco  
Center

Centro Nacional de Supercomputación

