Instituto Milenio
Fundamentos
de los datos

CMM
Centro de
Modelamiento
Matemático

Privacidad en la era de los datos

Jocelyn Dunstan, PhD MSc
jdunstan@uc.cl

¿Qué tan grande es un *Large Language Model*?

Ciclos de Charlas ReLeLa

https://relela.com/ciclos/

Si llevamos a texto todo lo que escuchamos para aprender a comunicarnos, ¿cómo se compara con GPT?

Charla de Jorge Ortiz Fuentes

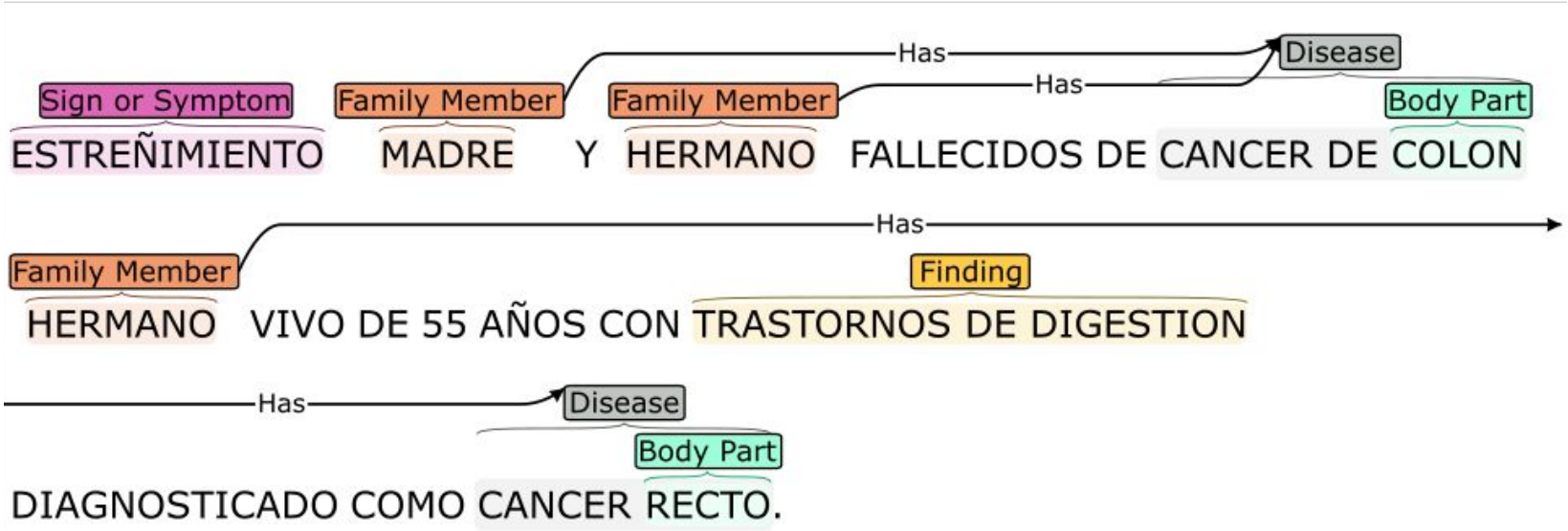
¿Cuánto pesa la transcripción de lo que escuchamos en 18 años?

- Una persona habla entre 150 y 160 palabras por minuto
- Supongamos que un niño pasa 15 horas diarias despierto
- En 15 horas el niño puede escuchar 15*60*155 = 139,500 palabras
- Multipliquemos esto por 365 días por 18 años = 916,5 millones de palabras
- En total esto corresponde a **75 mb** de texto

Modelo	GB de textos
GPT	4.8 gb
GPT2	40 gb
GPT3	570 gb
GPT3.5 (ChatGPT)	¿570 gb?
GPT4 (ChatGPT4)	?
Humano de 18 años	Entre 75 y 300 mb

¿Por qué esto es importante en el contexto de los modelos del lenguaje aplicados a medicina?

Interconsulta real obtenida por Ley de Transparencia



Hercules Dalianis

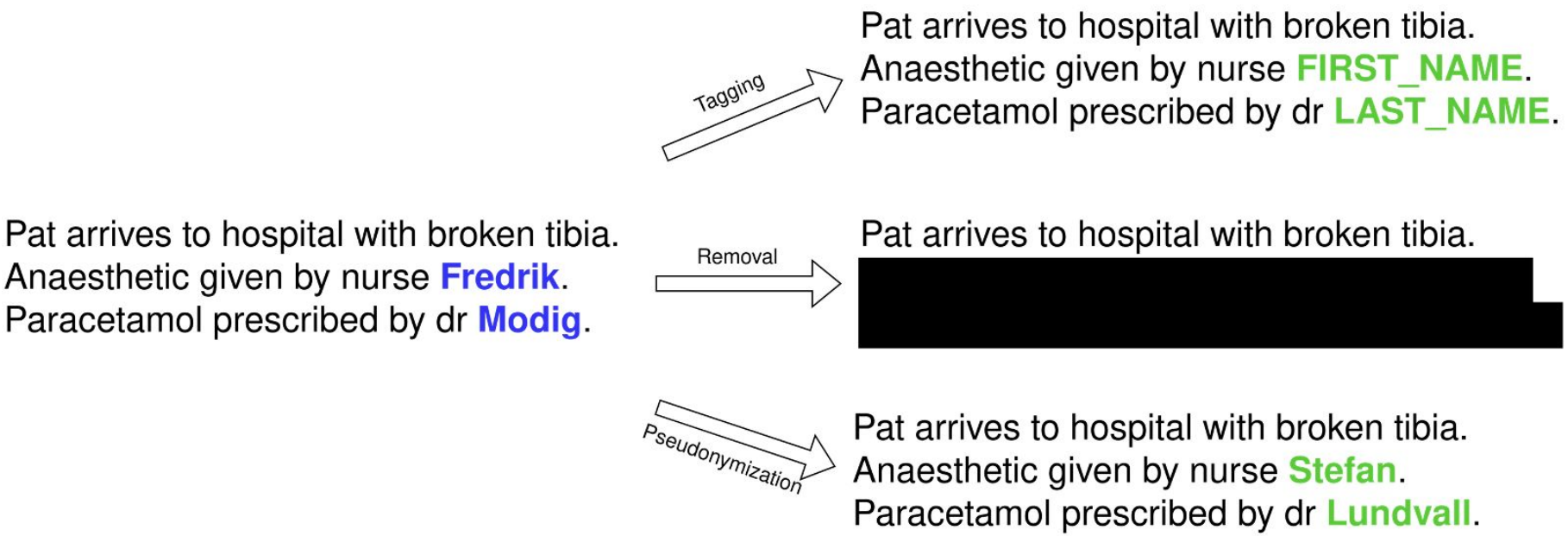
Thomas Vakili

Cortesía de Thomas Vakili

Memorización involuntaria de modelos del lenguaje

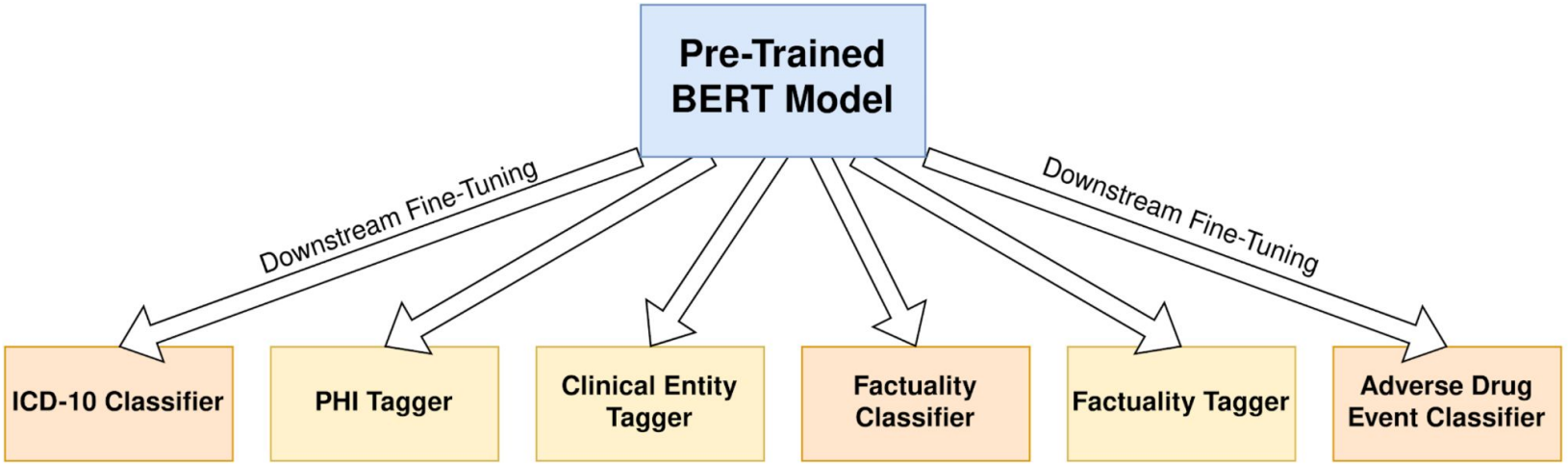
- El tamaño de los modelos conduce a una memorización involuntaria
- Carlini et al. (2020) extraen pasajes memorizados largos del GPT-2 entrenado con datos de dominio general
- Queremos adaptar los modelos para su uso clínico utilizando historiales médicos electrónicos
- La privacidad en medicina es crucial y además es un derecho humano

Cortesía de Hércules Dalianis



Utility Preservation of Clinical Text After De-Identification

Thomas Vakili and Hercules Dalianis



https://aclanthology.org/2022.bionlp-1.38.pdf

Cortesía de Hércules Dalianis

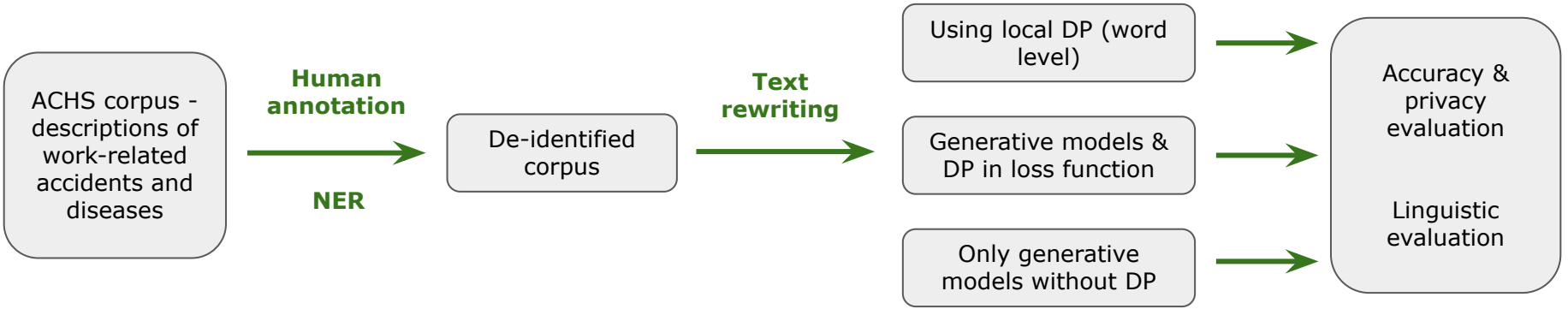
Model	ICD-10	PHI	Clinical Entity	Factuality	Factuality	ADE
	Classification	NER	NER	Classification	NER	Classification
KB-BERT	0.799	0.91	0.803	0.635	0.630	0.183
KB-BERT + Real	**0.833**	**0.941**	0.858	0.732	0.682	**0.199**
KB-BERT + Filtered	**0.833**	0.929	0.854	0.731	0.672	**0.199**
KB-BERT + Pseudo	0.832	**0.941**	**0.861**	**0.736**	**0.684**	0.191

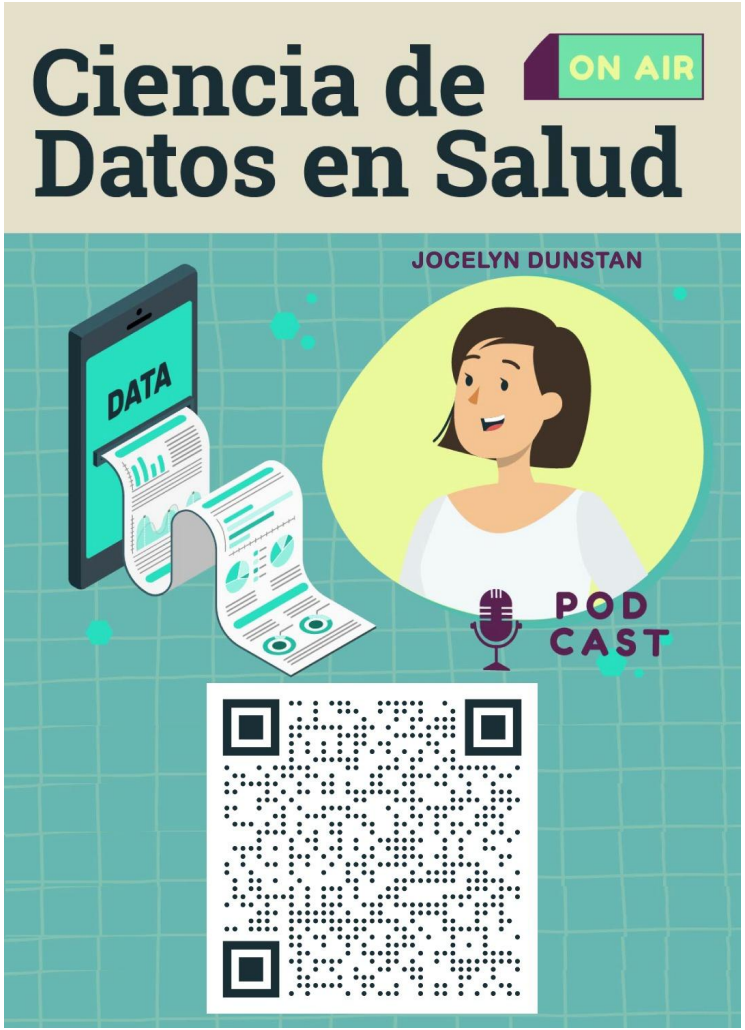
¿Qué estamos tratando de hacer en Chile?

Textos clínicos sintéticos

- ACHS ha recopilado un corpus clínico de mil millones de palabras.
- La ley de derechos y deberes de los pacientes podría prohibir el uso de datos para entrenar modelos
- Desde el punto de vista académico es muy interesante utilizar texto clínico real para crear textos sintéticos.
- La correctitud de estos textos se puede medir desde el punto de vista clínico y lingüistico

En colaboración con la Asociación Chilena de Seguridad - ACHS





Gracias por su atención

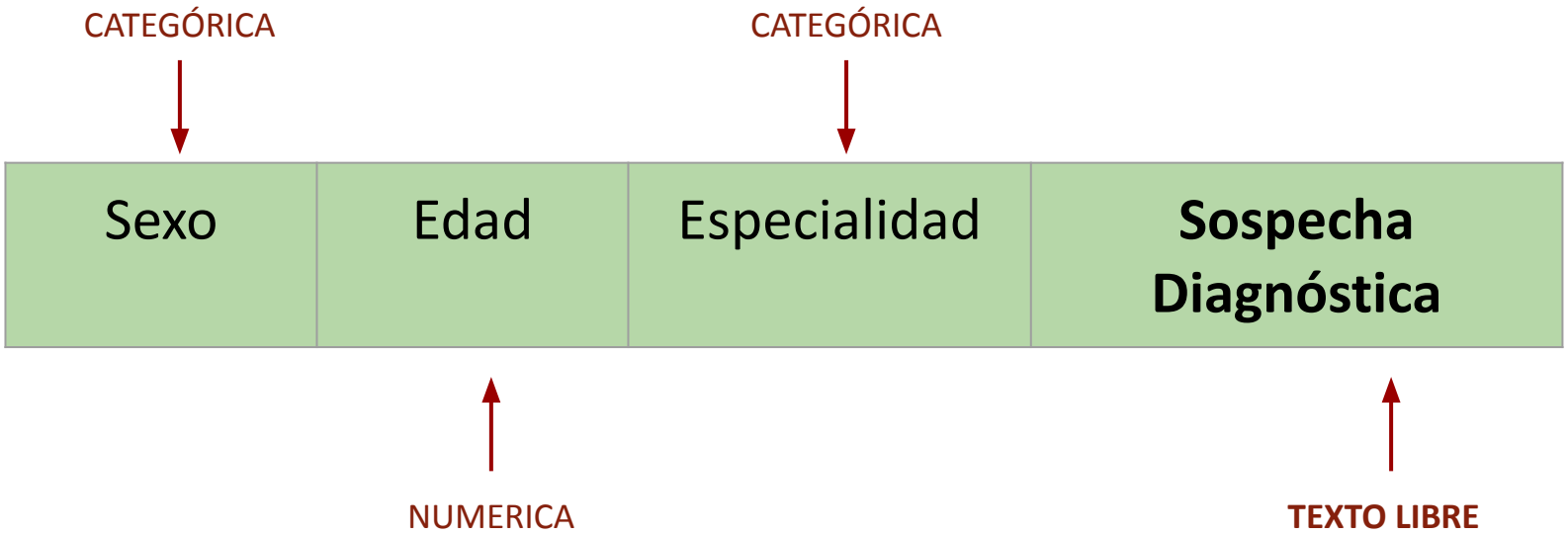
jdunstan@uc.cl

jo_cientifica

Si nos alcanza el tiempo… más de texto clínico en español

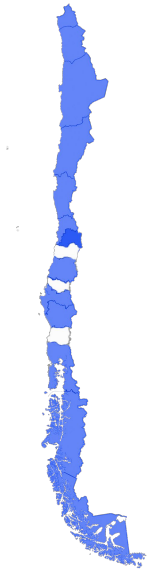
Listas de espera en hospitales públicos chilenos

- El 75% de la población Chilena está en el sistema público de salud
- Para tener una primera consulta de especialidad hay que entrar en la LE
- Largos tiempos de espera en consultas no-GES

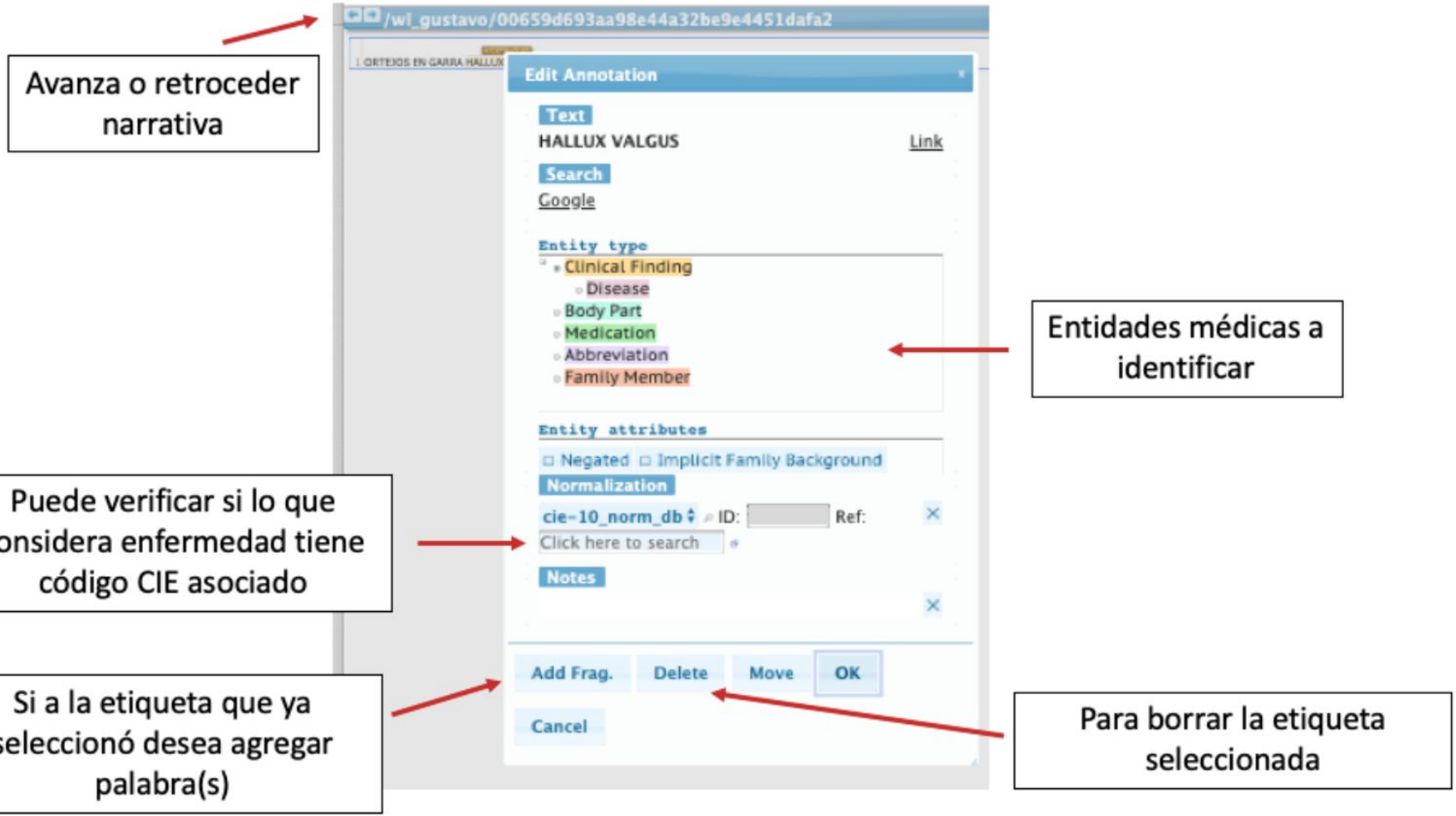


The Chilean waiting list corpus

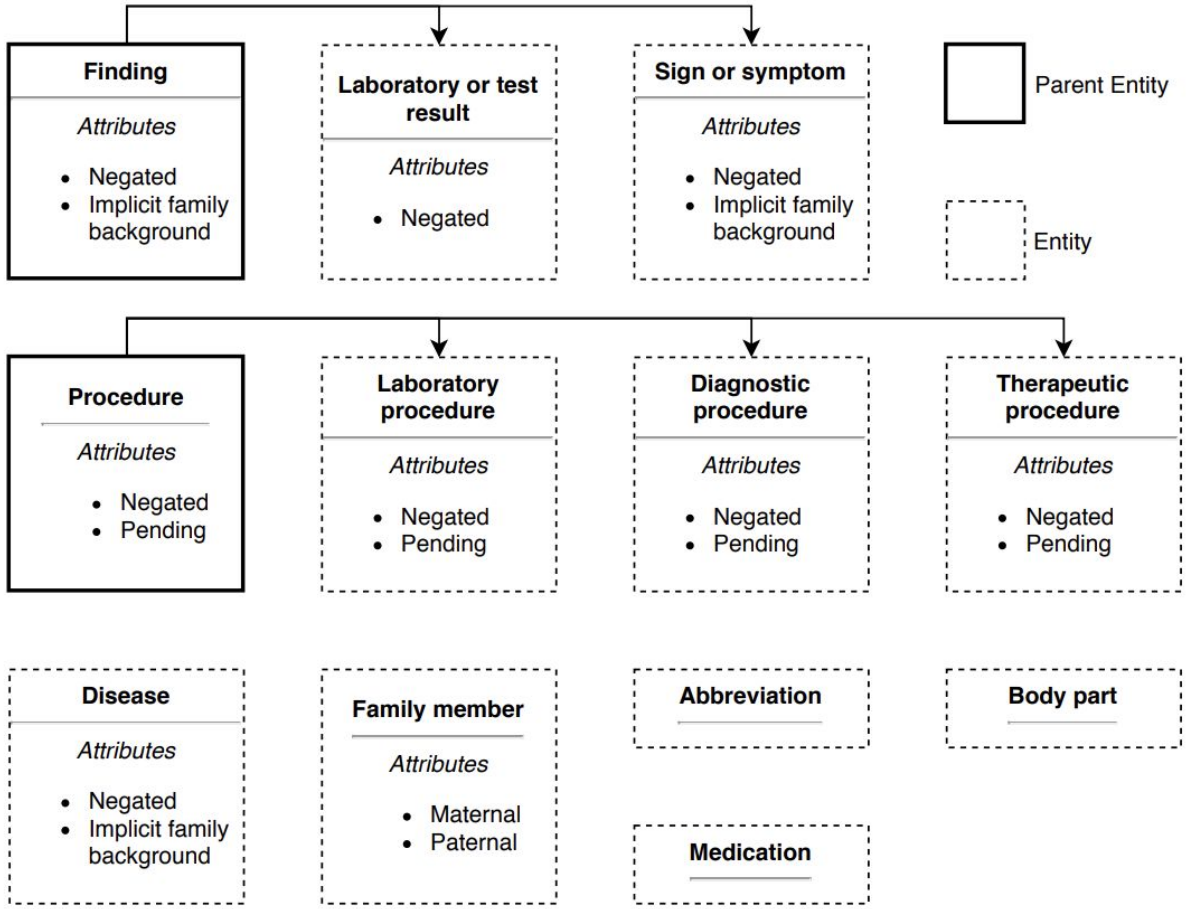
Durante el 2018 conseguimos interconsultas escritas por médicos/as de atención primaria. De las 11 millones de interconsultas se anotaron 10.000



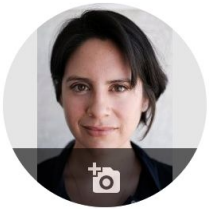
- Annotation guidelines
- Word Embeddings (W2V)


Avanza o retroceder narrativa
/wi_gustavo/00659d693aa98e44a32be9e4451dafa2
Edit Annotation
Text
HALLUX VALGUS
Link
Search
Google
Entity type
Clinical Finding
Disease
Body Part
Medication
Abbreviation
Family Member
Entity attributes
Negated
Implicit Family Background
Normalization
cie-10_norm_db
ID:
Ref:
Click here to search
Notes
Add Frag.
Delete
Move
OK
Cancel
Entidades médicas a identificar
Puede verificar si lo que considera enfermedad tiene código CIE asociado
Si a la etiqueta que ya seleccionó desea agregar palabra(s)
Para borrar la etiqueta seleccionada

BRAT


Finding
Attributes
Negated
Implicit family background
Laboratory or test result
Attributes
Negated
Sign or symptom
Attributes
Negated
Implicit family background
Parent Entity
Entity
Procedure
Attributes
Negated
Pending
Laboratory procedure
Attributes
Negated
Pending
Diagnostic procedure
Attributes
Negated
Pending
Therapeutic procedure
Attributes
Negated
Pending
Disease
Attributes
Negated
Implicit family background
Family member
Attributes
Maternal
Paternal
Abbreviation
Medication
Body part

Artículos publicados



Jocelyn Dunstan

Universidad Católica de Chile
No verified email - Homepage
Clinical NLP

Scholar
Página del grupo
Twitter
Linkedin

Journals

BMC Public Health (2019)

Revista Med. Chile (2021)

Revista Med. Clinica Las Condes (2022)

Clinical Dermatology (2021)

ACM Healthcare (2022)

BMC Med. Inf. Dec. Mak (2021)

Conference Proceedings

EMNLP Clinical Workshop (2020)

ACL Clinical Workshop (2022)

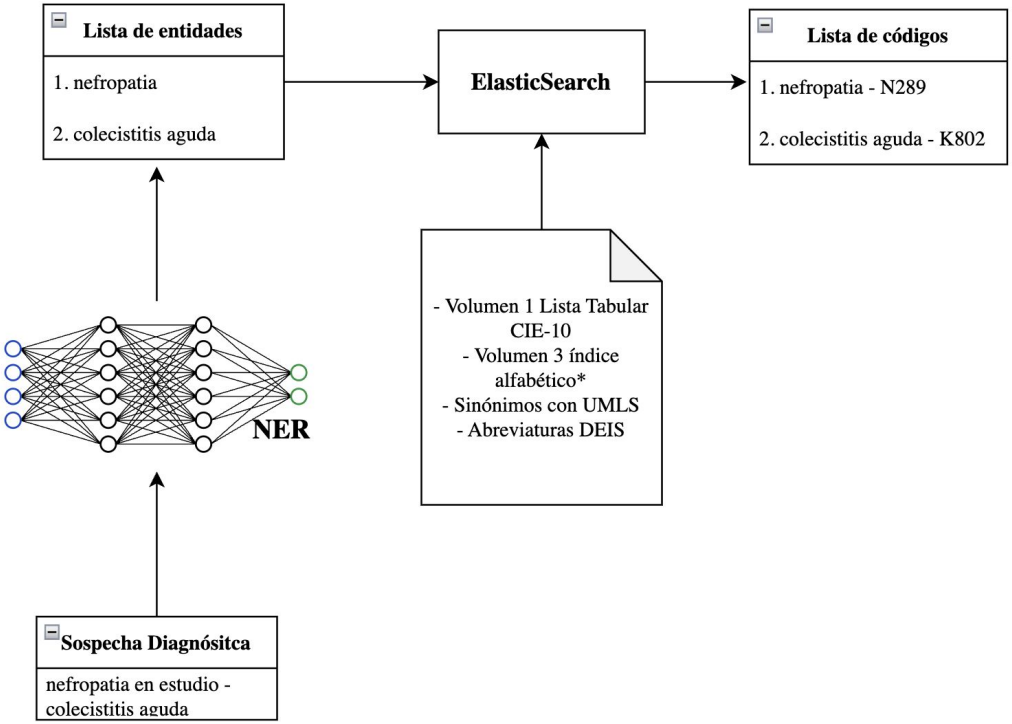
Coling (2022)

EMNLP Clinical Workshop (2022)

Usando PLN ¿Podemos mejorar el manejo de listas de espera?

¿Podemos hacer un uso secundario de los datos? (por ej, para estimar incidencia de enfermedades)

Codificación automática de enfermedades



Automatic Coding at Scale: Design and Deployment of a Nationwide System for Normalizing Referrals in the Chilean Public Healthcare System

Fabián Villena
Center for Mathematical Modeling
& Department of Computer Sciences
University of Chile
fabian.villena@uchile.cl

Matías Rojas
Center for Mathematical Modeling
University of Chile
matias.rojas.g@ug.uchile.cl

Felipe Arias
Center for Mathematical Modeling
University of Chile
felipe.arias.t@ug.uchile.cl

Jorge Pacheco
Dept. of Statistics and Health Information
Chilean Ministry of Health
jorge.pacheco@minsal.cl

Paulina Vera
Dept. of Statistics and Health Information
Chilean Ministry of Health
paulina.vera@minsal.cl

Jocelyn Dunstan
Dept. Computer Science & IMC
Pontifical Catholic University of Chile
jdunstan@uc.cl

https://aclanthology.org/2023.clinicalnlp-1.37/

Caso de éxito

8260/3 | Morfologia C73.9 | Topografia

III.- Tiroidectomía total: Cáncer del tiroides de 6 mm. de eje mayor con los caracteres de un

8260/3 | Morfologia

8010/3 | Morfologia C73.9 | Topografia

carcinoma diferenciado papilar con esclerosis del estroma con infiltración de cápsula tiroidea .

Cálculo de Incidencia de Psoriasis

CED CLINICAL AND EXPERIMENTAL DERMATOLOGY
THE EDUCATIONAL JOURNAL OF THE BRITISH ASSOCIATION OF DERMATOLOGISTS

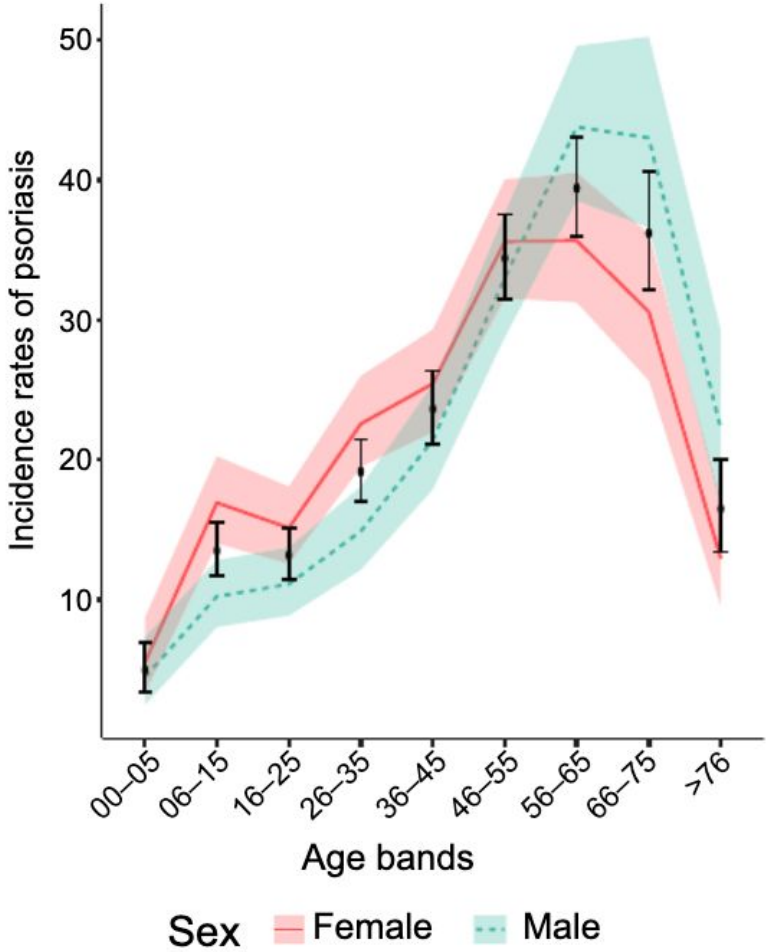
ORIGINAL ARTICLE

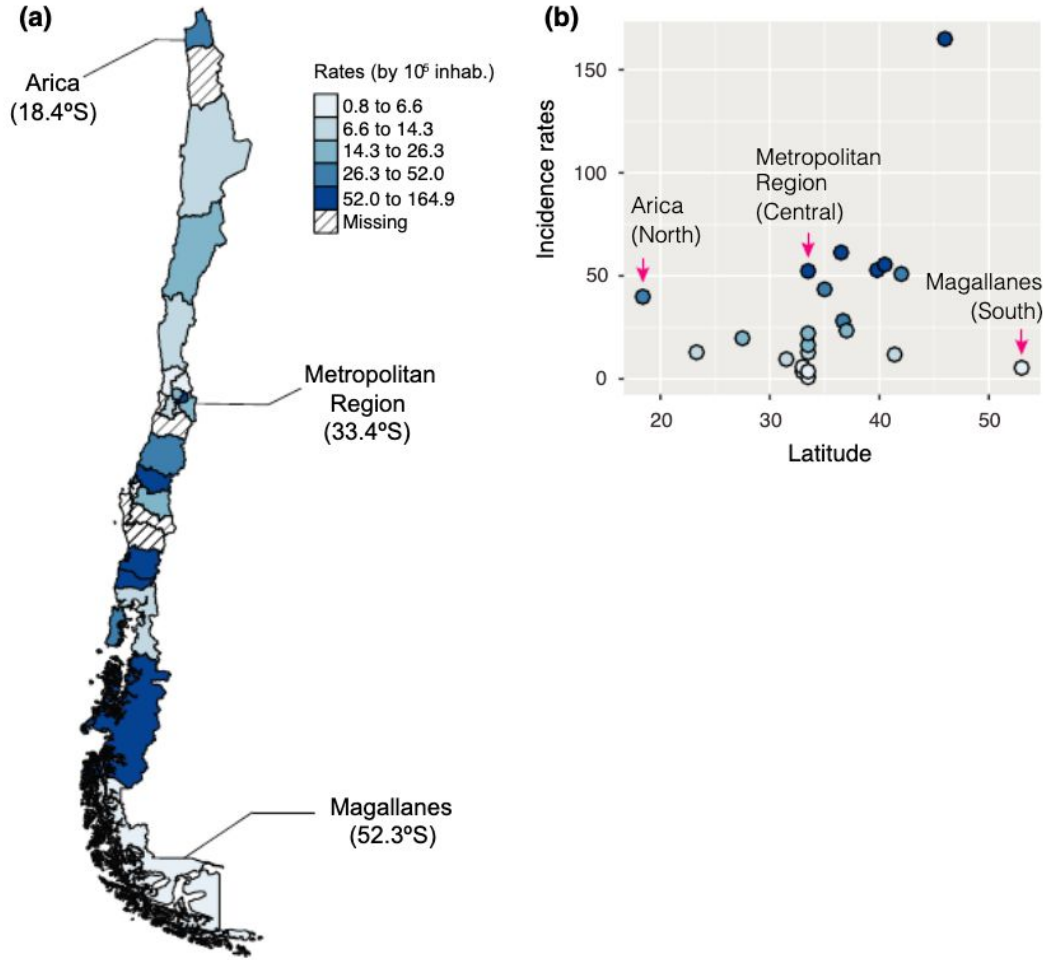
The incidence of psoriasis in Chile: an analysis of the national Waiting List Repository

C. Lecaros, J. Dunstan, F. Villena, D.M. Ashcroft, R. Parisi, C.E.M. Griffiths, S. Härtel, J.T. Maul, C. De la Cruz

First published: 29 April 2021 | https://doi.org/10.1111/ced.14713

https://onlinelibrary.wiley.com/doi/abs/10.1111/ced.14713


Incidence rates of psoriasis
50
40
30
20
10
00–05
06–15
16–25
26–35
36–45
46–55
56–65
66–75
>76
Age bands
Sex
Female
Male


(a)
Arica
(18.4ºS)
Rates (by 10⁵ inhab.)
0.8 to 6.6
6.6 to 14.3
14.3 to 26.3
26.3 to 52.0
52.0 to 164.9
Missing
Metropolitan
Region
(33.4ºS)
Magallanes
(52.3ºS)
(b)
Incidence rates
150
100
50
0
Arica
(North)
Metropolitan
Region
(Central)
Magallanes
(South)
20
30
40
50
Latitude

Conclusiones

- IA tiene variadas aplicaciones en medicina
- PLN apoya la extracción de información clave desde textos y dictado por voz
- PLN en medicina tiene características propias y es necesario crear recursos lingüísticos y computacionales para apoyar su uso en países que no hablan inglés
- El avance del área requiere el acceso a datos anonimizados y el apoyo a iniciativas interdisciplinarias