# Leveraging biomedical texts in Spanish for drug repurposing, interaction and discovery: normalization of pharmacological substances and chemical compounds to SNOMED CT

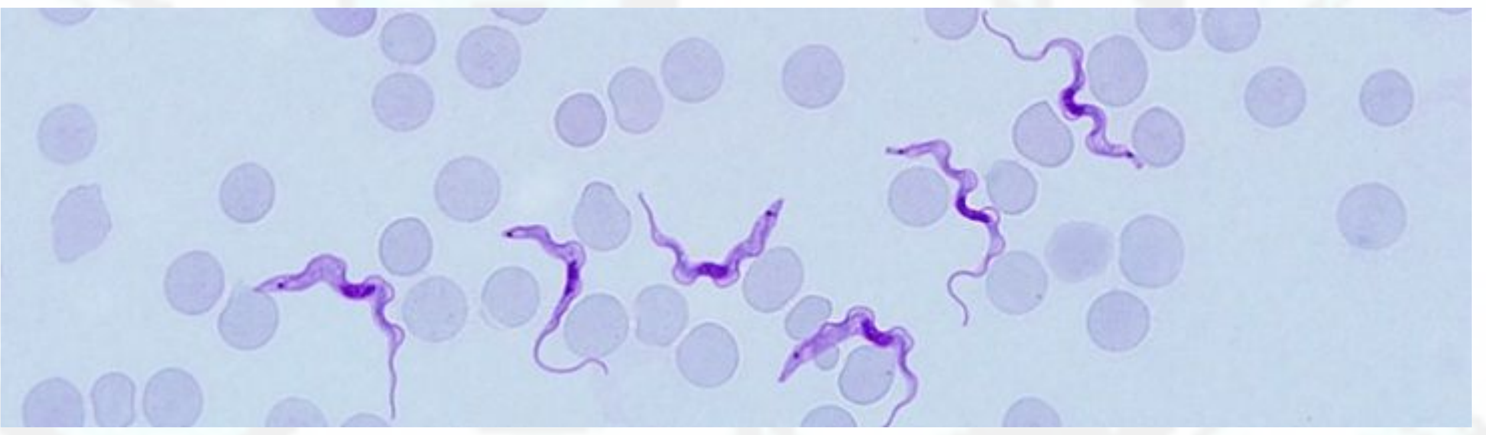
Poster prepared by:
Eulàlia Farré Maduell | Barcelona Supercomputing Center
Salvador Lima López | Barcelona Supercomputing Center
Martin Krallinger | Barcelona Supercomputing Center

Introduction
Methods
Results
Discussion

## Motivation

*Medicines* are essential to treat, prevent and diagnose symptoms and diseases, starting before birth.

However, a large proportion of the elderly and people with comorbidities are currently affected by *polypharmacy*, the simultaneous prescription and use of (too) many medications. Little is known about the interaction of three or more drugs, particularly when metabolization and elimination of drugs are affected (i.e., kidney and liver disease, elderly metabolism, and other).



On the other hand, we lack evidence for *drug efficacy* when faced with new diseases like COVID-19, and little funds address treatment for *orphan diseases* like rare congenital metabolic disorders and sleeping sickness.

*Electronic health records* and *biomedical literature* contain information that can resolve some of these issues. The *recognition* and normalisation of *pharmaceutical drugs/chemical entities* is a critical step toward the subsequent detection of *relations with other biomedically relevant entities* such as genes/proteins, diseases, adverse reactions and unexpected beneficial effects.

Introduction

## PharmaCoNER Gold Standard

- **PharmaCoNER** is the first resource for detecting **chemical**, **drug**, and **gene/protein** entities in Spanish medical documents.
- 1,000 clinical cases from multiple specialties manually annotated and normalized by experts.
- Most mentions normalized to **SNOMED CT** and some to **CHEBI.**

PharmaCoNER

**Web:** https://temu.bsc.es/pharmaconer/
**Dataset:** https://doi.org/10.5281/zenodo.4270157
**Guidelines:** https://doi.org/10.5281/zenodo.3763276

**Citation:** González Aguirre et al, Pharmaconer: Pharmacological substances, compounds and proteins named entity recognition track". In: Proceedings of The 5th Workshop on BioNLP Open Shared Tasks. 2019, pp. 1–10

snomedexpo.org
SNOMED CT EXPO 2022
Sept 29-30, 2022 ✕ Lisbon, Portugal

Poster prepared by:
Eulàlia Farré Maduell | Barcelona Supercomputing Center
Salvador Lima López | Barcelona Supercomputing Center
Martin Krallinger | Barcelona Supercomputing Center

Introduction
Methods
Results
Discussion

## Corpus Annotation

- **Manually** annotated by domain experts.
- Annotation schema based on previous successful efforts in English: CHEMDNER and BioCreative GPRO track.
- **34 pages annotation guidelines** adapted to Spanish language and domain.
- Corpus consistency analysis using **Inter-annotator Agreement (IAA): 93%**

- Four entity types:
  - ○ **NORMALIZABLES**: chemicals that can be manually normalized to a unique concept identifier (primarily SNOMED CT).
  - ○ **NON_NORMALIZABLES**: chemicals that could not be assigned a unique concept identifier.
  - ○ **PROTEINS**: proteins and genes, including peptides, peptide hormones and antibodies.
  - ○ **UNCLEAR**: general substance class mentions of clinical relevance (e.g. pharmaceutical formulations, vaccines, some predefined substances like alcohol, tobacco or gluten).

## Mention Normalization

- Most mentions were normalized to **SNOMED CT**, with a small subset of them being normalized to **CHEBI**, a dictionary of small chemical compounds.
- SNOMED's "substance" category was the most used, followed by "biological/pharmaceutical product"

| **"albendazol"** → *SCTID: 387558006 | Albendazole (substance) |*

- Some generic/commercial drug names, as well as some proteins, had to be normalized to their family name:

| **"adriamycin"** (commercial name, not in SNOMED as is) → *SCTID: 372817009 | Doxorubicin (substance) |*

| **"botox"** (not in SNOMED as is) → *11894001 |Clostridium botulinum toxin (substance)|*

- CHEBI was used mostly for general chemical groups and drugs with chemical names in their structure which weren't part of SNOMED:

| **"6-metil-Prednisolone"** (not in SNOMED)→ *CHEBI:6888*

- Some very specific mentions could not be found in neither terminology, so they had to be assigned the placeholder NOCODE.

Poster prepared by:
Eulàlia Farré Maduell | Barcelona Supercomputing Center
Salvador Lima López | Barcelona Supercomputing Center
Martin Krallinger | Barcelona Supercomputing Center

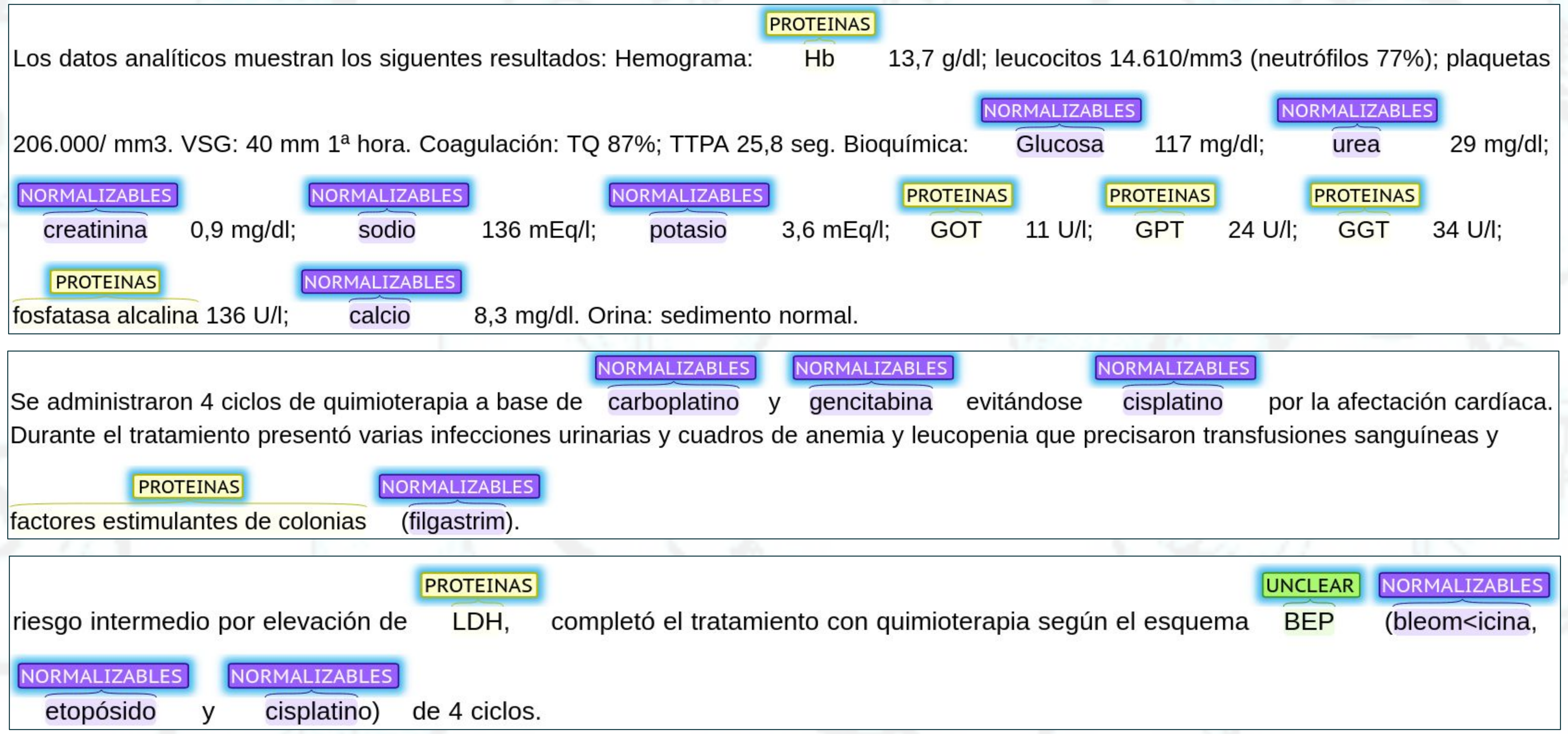
Introduction
Methods
Results
Discussion

# Corpus Statistics

- **1,000 clinical cases** with a total of **7,624 annotations**.
- Most annotations correspond to the classes NORMALISABLES and PROTEINS.

- 7,266 mentions (around 95%) normalized to SNOMED CT.
- 29 mentions normalized to CHEBI.
- 329 mentions could not be normalized (NOCODE).

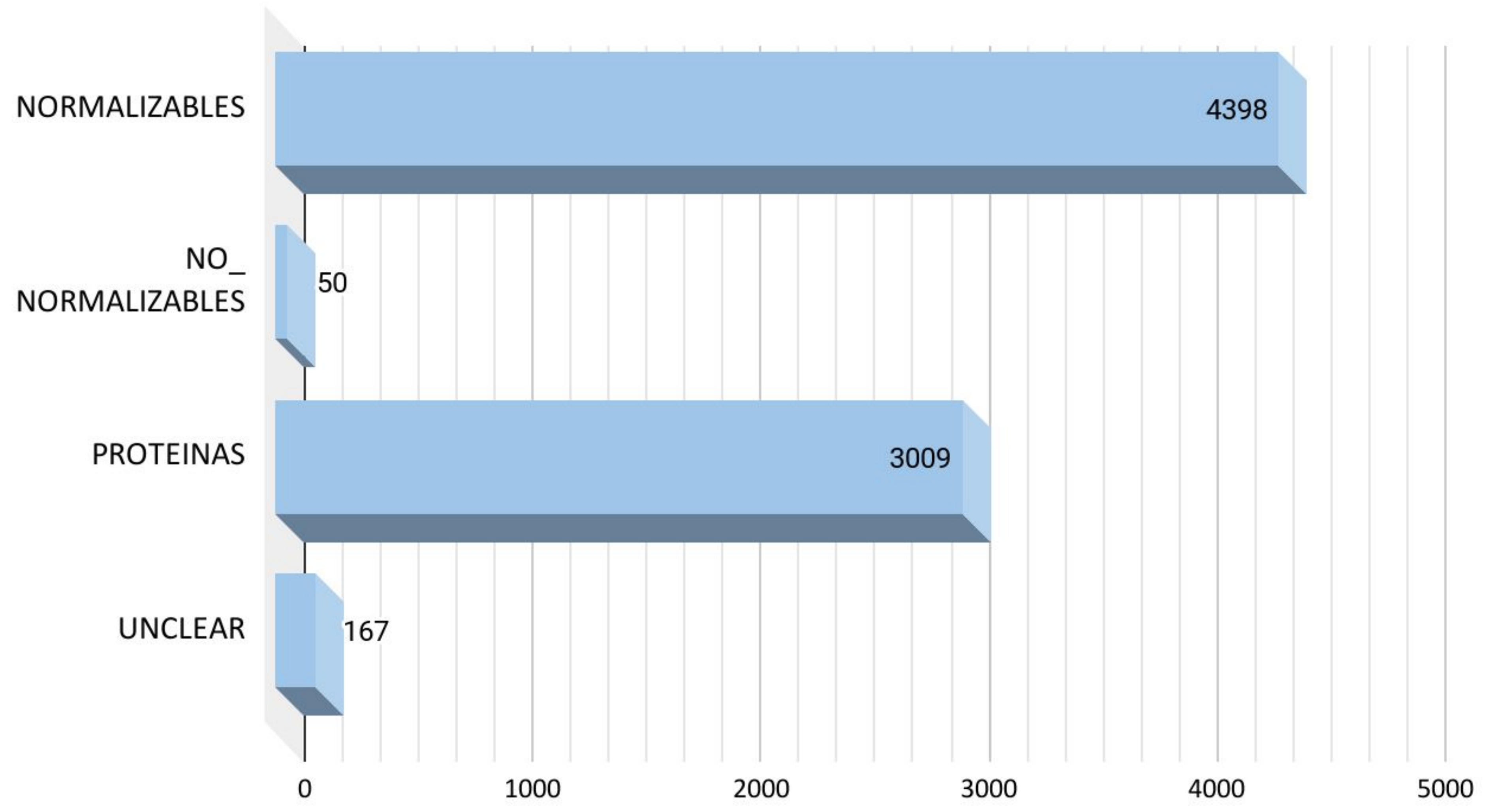
Some annotation examples:


Los datos analíticos muestran los siguentes resultados: Hemograma: Hb 13,7 g/dl; leucocitos 14.610/mm3 (neutrófilos 77%); plaquetas 206.000/ mm3. VSG: 40 mm 1ª hora. Coagulación: TQ 87%; TTPA 25,8 seg. Bioquímica: Glucosa 117 mg/dl; urea 29 mg/dl; creatinina 0,9 mg/dl; sodio 136 mEq/l; potasio 3,6 mEq/l; GOT 11 U/l; GPT 24 U/l; GGT 34 U/l; fosfatasa alcalina 136 U/l; calcio 8,3 mg/dl. Orina: sedimento normal.

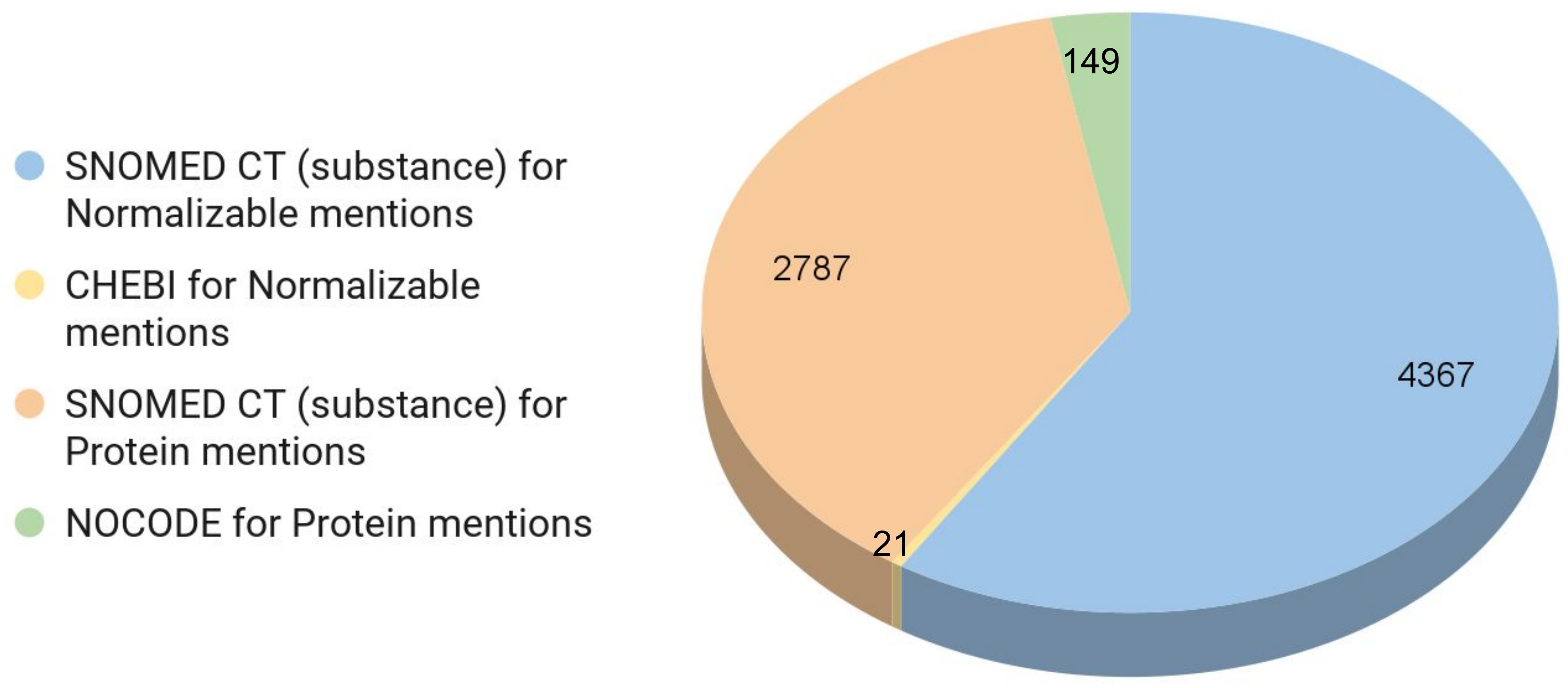
Se administraron 4 ciclos de quimioterapia a base de carboplatino y gencitabina evitándose cisplatino por la afectación cardíaca. Durante el tratamiento presentó varias infecciones urinarias y cuadros de anemia y leucopenia que precisaron transfusiones sanguíneas y factores estimulantes de colonias (filgastrim).

riesgo intermedio por elevación de LDH, completó el tratamiento con quimioterapia según el esquema BEP (bleom<icina, etopósido y cisplatino) de 4 ciclos.

PharmaCoNER Mention Type Distribution



Entity Normalization Statistics



Introduction

Methods

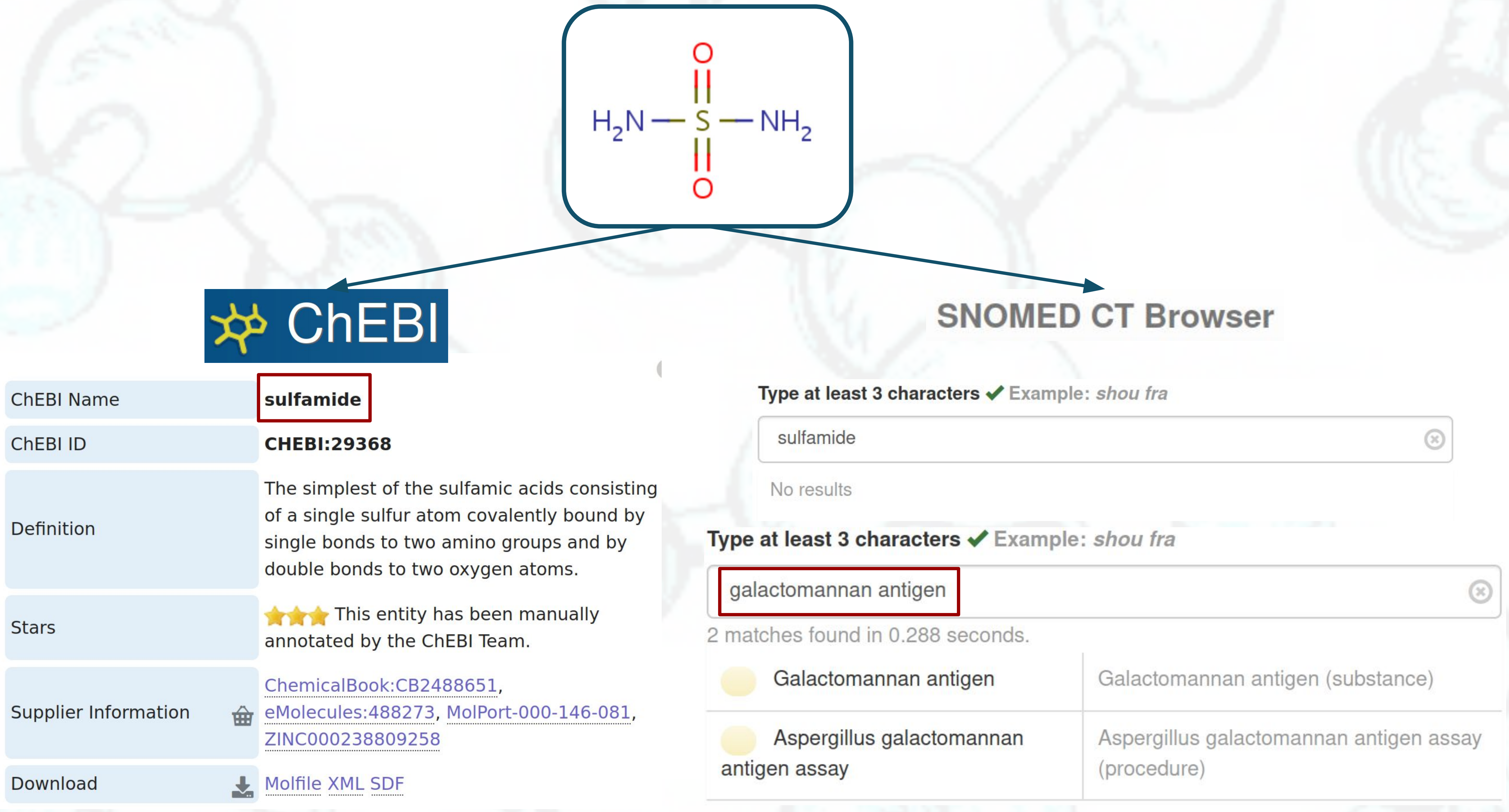
Results

Discussion

# Leveraging biomedical texts in Spanish for drug repurposing, interaction and discovery: normalization of pharmacological substances and chemical compounds to SNOMED CT

Poster prepared by:
Eulàlia Farré Maduell | Barcelona Supercomputing Center
Salvador Lima López | Barcelona Supercomputing Center
Martin Krallinger | Barcelona Supercomputing Center

## Discussion



ChEBI

| | |
|---|---|
| ChEBI Name | **sulfamide** |
| ChEBI ID | **CHEBI:29368** |
| Definition | The simplest of the sulfamic acids consisting of a single sulfur atom covalently bound by single bonds to two amino groups and by double bonds to two oxygen atoms. |
| Stars | ★★★ This entity has been manually annotated by the ChEBI Team. |
| Supplier Information | ChemicalBook:CB2488651, eMolecules:488273, MolPort-000-146-081, ZINC000238809258 |
| Download | Molfile XML SDF |

SNOMED CT Browser

**Type at least 3 characters** ✓ Example: *shou fra*

sulfamide

No results

**Type at least 3 characters** ✓ Example: *shou fra*

galactomannan antigen

2 matches found in 0.288 seconds.

| | |
|---|---|
| Galactomannan antigen | Galactomannan antigen (substance) |
| Aspergillus galactomannan antigen assay | Aspergillus galactomannan antigen assay (procedure) |

- SNOMED CT is overall a valid ontology for pharmaceutical compounds and substances, since it was able to provide a code for a majority of our mentions.
- Even then, there are some gaps for very specific content that can be closed using specialized resources like CHEBI.
- SNOMED CT presents the advantage of containing many other concepts related to medicines, for instance, disorders, findings and events.

## Conclusions

- Specialized resources in Spanish are needed to increase the usefulness of text mining tools in the biomedical domain.
- PharmaCoNER can be used to train and evaluate automatic systems that detect chemicals, proteins and genes in Spanish and normalize them to SNOMED CT.

## Future **Directions**

- **Relation Extraction:** chemical compounds and pharmacological substances should be related to symptoms, diseases and their polarities (if that symptom or disease means improvement or worsening).
- **Information discovery:** results should also aim at drug repurposing and discovery, in particular toward orphan diseases.

## Acknowledgements

- Special thanks to the PharmaCoNER original authors: Aitor Gonzalez Agirre, Ander Intxaurrondo and Obdulia Rabal
- PharmaCoNER was promoted through the collaboration between the Spanish Plan for the Advancement of Language Technology (Plan TL) and the BSC/CNIO.