FAIRClinical: *FAIR-ification of Supplementary Data to Support Clinical Research*

Deliverable D 3.1 The specification of the meta-data enrichment model

Work Package	The specification of the meta-data enrichment model
Lead Beneficiary	Laboratoire Interdisciplinaire des Sciences du Numérique
Contributing Partner(s):	LISN(FR), HES-SO(CH), UNILU (LU), UNNO (UK)
Delivery Date:	31 August 2024
Authors of this Deliverable:	Nona Naderi (FR), Carlos Cuevas Villarmin (FR), Patrick Ruch (CH)
Grant agreement:	CHIST-ERA grant CHIST-ERA-22-ORD-02
Start Date	1 March 2024
Duration	24 months
Acknowledgement	This work is supported by the CHIST-ERA grant CHIST-ERA-22-ORD-02, by the Luxembourg National Research Fund (FNR, INTER/CHIST23/17882238/FAIRClinical), by Swiss National Science Foundation (SNSF, 20CH21_217525), by the Agence Nationale de la Recherche (ANR-23-CHRO-0008-01), and by Engineering and Physical Sciences Research Council (EP/Y036395/1).

1. Introduction

To support patient clinical data discovery and analysis across scientific publications, basic data such as patients demographic data, diseases, and medications need to be identified and harmonized. Scientific publications, however, are available in the form of unstructured text, making much of information not readily available for further reuse and analysis. Therefore, there is an immediate need for automated approaches to extract knowledge from this growing body of literature. To develop and establish reliable machine learning systems, many task specific datasets have been created and made publicly available. Here, we examine the available datasets and their applications for the development of our machine learning approaches for information extraction from clinical case reports.

2. Detailed report on the deliverable

Case reports provide descriptions of patient’s medical problems and clinical management, and they have been considered as important sources for recognizing new diseases, evaluating the benefits and harms of interventions, and medical education (Riley et al., 2022). Well written case reports are expected to specifically provide information regarding the main symptoms of the patients, main clinical findings, diagnoses and interventions and outcomes.

In the following sections, we first describe the entities that are already identified in the SIBiLS (Gobeill et al. 2020) annotation pipeline (Section 2.1), and then, we focus on the entities that we plan to develop machine learning models for their automatic extraction in this project and describe the available resources and datasets (Section 2.2).

2.1 SIBiLS Annotation pipeline

Swiss Institute of Bioinformatics Literature Services (SIBiLS) annotation pipeline identifies some biomedical entities with the help of a set of standardized vocabularies, such as Drugbank (Wishart et al., 2018) for drugs, NCI Thesaurus (Sioutos et al., 2007) for diseases, and neXtProt (Gaudet et al., 2017) for human genes (Gobeill et al. 2020). The evaluation of entity recognition as available in the SIB Literature Services is presented below in Table 1. As shown, these entities can have very contrasted levels of recognition. Some entities are extracted more accurately than the others, for example annotations that are extracted with the help of Anatomical Therapeutic Chemical (ATC) terminology are more accurate (> 90%) compared to those extracted with the help of Drugbank.

Terminology	Nb annotations	True (%)	False (%)
ATC	**94**	**91**	**9**
Chebi	**293**	**79,5**	**20,5**
Covoc (global)	**373**	**100**	**-**
Detection methods	**5**	**100**	**-**
Disprot	**22**	**100**	**-**
Drugbank	266	34,6	65,4
ECO	**8**	**100**	**-**
ENVO	22	45,5	54,5
GO_bp	**74**	**100**	**-**
GO_cc	**21**	**90,5**	**9,5**
GO_mf	**35**	**100**	**-**
ICDO3	**17**	**82,4**	**17,6**
Licence	1	-	100
LOTUS	**36**	**100**	**-**
MDD	18	55,6	44,4
MeSH	**1342**	**89,3**	**10,7**

NCBI Taxonomy (clinic)	**17**	**88,2**	**11,8**
NCBI Taxonomy (full)	110	48,2	51,8
NCI Thesaurus	407	60	40
neXtProt	243	44	56
OTT	216	26,4	73,6
Pubchemmesh	542	26,4	73,6
Uniprot (swissprot)	234	41,5	58,5

Table 1: True and false positive annotations in the SIB Literature Services based on a sample of N=100 PMC articles. Terminologies with good recognition rates are shown in bold.

2.2 Entities of interest for FAIRClinical

We will focus on entities, which are central for the project and which are not recognized with sufficient accuracy or (better) not recognized at all in SIBiLS. We categorize these entities based on three levels of priority: high, medium, and low.

2.2.1 High priority

PICO elements: Given the type of important information that case reports carry as Riley et al., 2022 also highlighted, we will focus on PICO entities – **Population/Problem** (What are the most critical characteristics of the enrolled population? What is the primary disease?), **Intervention** (What is the primary intervention considered?), **Comparator** (To what the intervention is compared?), and **Outcome** (What are the anticipated measures, improvements or effects?). Supervised machine learning approaches rely on annotated datasets for training, so we examine the available datasets to train our models for various entities. Table 2 presents the major datasets that can be used for training our models for the extraction of PICO entities. These datasets are annotated with different annotation criteria depending on their target task. The largest available dataset for PICO is EBM-NLP (Nye et al., 2018), however, only a small part of this dataset is manually annotated, making the quality of the remaining part of the dataset unknown. PICO corpus (Mutinda et al., 2022), on the other hand, was entirely annotated by experienced annotators and provides more detailed annotations for a set of articles on the topic of breast cancer. While these two datasets are annotated at the mention level, PICO sentence corpus (Wallace et al., 2016) provides PICO annotations at the sentence level, making it more appropriate for a sentence classification task rather than a named entity recognition task.

Dataset	Annotation types	Size	domain/Disease
EBM-NLP (Nye et al., 2018)	Participants (age, condition, gender, sample size), Interventions (Behavioral, control, educational, other, pharmacological, physical, psychological, surgical, Outcomes (adverse effects, mental, mortality, other, pain, physical)	5,000 abstracts describing RCTs (of which only 200 are manually annotated)	cardiovascular diseases, cancer, and autism
PICO corpus (Mutinda et al, 2022)	Participants (total-participants, intervention-participants,control-participants age, ethnicity, eligibility, condition, location), Intervention, Control, Outcomes (outcome, outcome-measure, iv-bin-abs, cv-bin, abs, iv-bin-percent, cv-bin-percent, iv-cont-mean, cv-cont-mena, iv-cont-median, cv-bin-median, iv-cont-sd, cv-cont-sd, iv-cont-ql, cv-cont-ql, iv-cont-q3, cv-cont-q3)	1,011 abstracts of randomized controlled trials	breast cancer
EBM-COMET (Abaho et al, 2022)	Clinical outcomes (Physical, Pain, Mental, Mortality and Adverse effects)	300 Randomized Clinical Trial PubMed abstracts	
PICO sentences	Intervention, participant, and	133 articles	

(Wallace et al., 2016)	outcome sentences (2821 sentences)		

Table 2: The publicly available datasets for PICO extraction

Diagnoses and diseases: While PICO datasets provide condition annotations, there are also datasets which focus on diseases and diagnoses. Table 3 presents some of these datasets. These datasets provide identifiers from relevant ontologies and vocabularies, such as MeSH and Disease Ontology and can be used for normalization and grounding of the extracted mentions.

Dataset	Annotation types	Size	Ontology/vocabulary
NCBI-BIO (Doğan et al., 2014)	Disease	793 PubMed abstracts	MeSH[1], OMIM[2]
BC5CDR–BioCreative V Chemical Disease Relation task (CDR) (Li et al., 2016)	Disease , Chemical	1500 Medline abstracts	MeSH
RareDis (Martínez-deMiguel et al., 2022)	Disease Rare disease Symptom	1041 reports	Disease Ontology[3], Orphan Rare Disease Ontology[4], Symptom Ontology[5]
2010-i2b2/VA (Uzuner et al., 2011)	Disease	871 progress reports	-

Table 3: The publicly available datasets for disease and diagnose extraction

Medications, drugs, prescriptions: While intervention annotations in PICO datasets

[1] http://www.nlm.nih.gov/mesh/

[2] http://www.ncbi.nlm.nih.gov/omim

[3] http://www.disease-ontology.org

[4] https://www.orpha.net/

[5] https://www.ebi.ac.uk/ols/ontologies/symp

include some medication annotations, there are many mentions of medications and drugs in the texts that require additional resources to be extracted. Table 4 shows the available datasets for the extraction of drugs and chemical entities. Some of these datasets link the annotation mentions to vocabularies and terminologies, such as MeSH and ChEBI, and can be used for the entity normalization task. A few of the datasets, BC5CDR (Li et al., 2016), EU-ADR (Mulligen et al., 2012), N2C2 (Henry et al., 2018), and BIO-RED (Islamaj et al., 2023), were developed for relation extraction tasks. More specifically, BC5CDR not only provides annotations for disease and chemical mentions, but also the relations between these mentions were annotated. EU-ADR provides the relations of drug-disease, target-disease, and target-drug as well as their concept annotations. N2C2 provides annotations for drugs and adverse events, as well as the relations between drug-adverse events. BIO-RED provides annotations for chemical, diseases, genes, variants, species, and cell lines as well as their relations, such as chemical-chemical, chemical-disease, chemical-gene, and gene-disease.

Dataset	**Annotation types (unique mentions)**	**Size**	**Ontology/vocabulary**
BC4CHEMD/CHEMDNER (Krallinger et al., 2015a)	Chemical (19,806)	10,000 PubMed abstracts	–
BC5CDR–BioCreative V Chemical Disease Relation task (Li et al., 2016)	Disease (5818), Chemical (3116)	1500 Medline abstracts	MeSH
CRAFT (Bada et al.,	Chemicals, chemical	67 fulltext biomedical	ChEBI[6]

[6] https://www.ebi.ac.uk/chebi/

2012)	groups	articles	
ChEMFAM (Savery et al., 2020)	Organic (414), inorganic (306) chemicals, amino acids, peptides (299), proteins (537)	200 medline abstracts	MeSH
BC5CHEMD-Patents (Krallinger et al., 2015b)	Chemicals (34,796)	21,000 patent abstracts	–
MedMention (Mohan and Li, 2019)	Organisms, Medical devices, body substance, chemical, finding, clinical attributes, organization, population group, injury or poisoning	4,392 abstracts	UMLS 2017 AA (OMIM, RXNORM, SNOMEDCT_US), NDFRT, NDDF, NCI, NCBI, MeSH, ICD9CM, ICD10CM, ICD10, HPO, HGNC, GO, FMA, CPT)
EU-ADR (Van Mulligen et al., 2012)	Drug, Disease, Target (gene, protein, sequence variants of genes and proteins)	300 Medline abstracts	-
N2C2 (Henry et al., 2018)	Drug, Adverse events	505 discharge summaries	-
BIO-RED (Islamaj et al., 2023)	Chemical, Disease, Gene, Species, Variant, Cell Line	1000 PubMed abstracts	MeSH, NCBI Gene, NCBI Taxonomy, Cellosaurus.

Table 4. The available datasets for chemical and drug annotations.

2.2.2. Medium priority

Genes and proteins: These entities are generally very ambiguous, therefore, the datasets and the models trained for these entities can have various quality levels. Table 5 presents some of the available datasets for protein and gene extraction. GENIA (Kim et al., 2003) was one of the early datasets for name entity recognition, which was also the base for

JNLPBA dataset (Huang et al., 2019), as well as other datasets. This dataset was annotated using GENIA ontology for biological entities.

In order to avoid the ambiguity of DNA, RNA, and proteins, GENETAG corpus (Tananabe et al., 2005) collapses all these annotation types into one. In addition to these datasets, there is also IGN corpus (Dai et al., 2013), which provides gene and gene product annotations for 627 PubMed articles (abstracts and fulltexts) and links the annotations to Gene Entrez (Maglott et al., 2005) IDs. This dataset can also be used for the gene normalization task. More recently, Huang et al., 2020 extended the revised JNLPBA datasets and provided annotations for genes, diseases, and chemicals with links to Entrez, MeSH, and ChEBI and MeSH, respectively. This dataset can also be used for the normalization task.

Dataset	Annotation types	Size
GENETAG (Tananabe et al., 2005)	Gene, protein sentences	20,000 Medline sentences
JNLPBA (Huang et al., 2019)	DNA, RNA, Protein, Cell line, Cell type	2000 Medline abstracts
GENIA (Kim et al., 2003)	DNA, RNA, Protein, Cells, Tissue, Chemical, Organism	2404 Medline abstracts
GPRO (Pérez-Pérez et al., 2017)	Gene and proteins	21,000 patent abstracts
EBED (Huang et al., 2020)	Gene (Entrez), Disease (MeSH), Chemical (ChEBI and MeSH)	3200 abstracts, 400 paragraphs, 300 figure legends, 300 patents

Table 5: The available datasets for protein and gene annotations.

licensing models: Carbon et al., 2019 compiles 56 resources for licenses used in the biomedical domain that can be used for the purpose of our project.

2.2.3. Low priority

Low priority entities are entities that are already extracted with a precision of 80% or over with SIBiLS pipeline and we will not focus on them in this project.

References:

- Abaho, M., Bollegala, D., Williamson, P. R., & Dodd, S. (2022). Assessment of contextualised representations in detecting outcome phrases in clinical trials. *arXiv preprint arXiv:2203.03547*.

Bada, M., Eckert, M., Evans, D., Garcia, K., Shipley, K., Sitnikov, D., Baumgartner, W.A., Cohen, K.B., Verspoor, K., Blake, J.A. & Hunter, L. E. (2012). Concept annotation in the CRAFT corpus. *BMC bioinformatics*, *13*, 1-20.

-Carbon, S., Champieux, R., McMurry, J. A., Winfree, L., Wyatt, L. R., & Haendel, M. A. (2019). An analysis and metric of reusable data licensing practices for biomedical resources. *PloS one*, *14*(3), e0213090

-Dai, H. J., Wu, J. C. Y., & Tsai, R. T. H. (2013). Collective instance-level gene normalization on the IGN corpus. *PLoS One*, *8*(11), e79517.

- Doğan, R. I., Leaman, R., & Lu, Z. (2014). NCBI disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, *47*, 1-10.

-Gaudet, P., Michel, P. A., Zahn-Zabal, M., Britan, A., Cusin, I., Domagalski, M., Duek, P.D., Gateau, A., Gleizes, A., Hinard, V. and Rech de Laval, ... & Bairoch, A. (2017). The neXtProt knowledgebase on human proteins: 2017 update. *Nucleic acids research*, *45*(D1), D177-D182.

- Gobeill J., Caucheteur D., Michel P.A., Mottin L., Pasche E., Ruch P. (2020). SIB Literature Services: RESTful customizable search engines in biomedical literature, enriched with automatically mapped biomedical concepts. *Nucleic Acids Res*.

48(W1):W12-W16. doi: 10.1093/nar/gkaa328. PMID: 32379317; PMCID: PMC7319474.

-Henry, S., Buchan, K., Filannino, M., Stubbs, A., & Uzuner, O. (2020). 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *Journal of the American Medical Informatics Association*, *27*(1), 3-12.

-Huang, M. S., Lai, P. T., Tsai, R. T. H., & Hsu, W. L. (2019). Revised JNLPBA corpus: a revised version of biomedical ner corpus for relation extraction task. arXiv. *arXiv preprint arXiv:1901.10219*.

-Huang, M. S., Lai, P. T., Lin, P. Y., You, Y. T., Tsai, R. T. H., & Hsu, W. L. (2020). Biomedical named entity recognition and linking datasets: survey and our recent development. *Briefings in Bioinformatics*, *21*(6), 2219-2238.

-Islamaj, R., Wei, C. H., Lai, P. T., Luo, L., Coss, C., Gokal Kochar, P., ... & Lu, Z. (2024). The biomedical relationship corpus of the BioRED track at the BioCreative VIII challenge and workshop. *Database*, *2024*, baae071.

-Kim, J. D., Ohta, T., Tateisi, Y., & Tsujii, J. I. (2003). GENIA corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics*, *19*(suppl_1), i180-i182.

-Krallinger, M., Rabal, O., Leitner, F., Vazquez, M., Salgado, D., Lu, Z., ... & Valencia, A. (2015a). The CHEMDNER corpus of chemicals and drugs and its annotation principles. *Journal of cheminformatics*, *7*, 1-17.

- Krallinger, M., Rabal, O., Lourenço, A., Perez, M. P., Rodriguez, G. P., Vazquez, M., ... & Valencia, A. (2015b). Overview of the CHEMDNER patents task. In *Proceedings of the fifth BioCreative challenge evaluation workshop* (pp. 63-75).

-Li, J., Sun, Y., Johnson, R. J., Sciaky, D., Wei, C. H., Leaman, R., Davis, R., Mattingly, A.P., Wiegers, C.J., & Lu, Z. (2016). BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database*, *2016*.

-Maglott, D., Ostell, J., Pruitt, K. D., & Tatusova, T. (2005). Entrez Gene: gene-centered information at NCBI. *Nucleic acids research*, *33*(suppl_1), D54-D58.

- Martínez-deMiguel, C., Segura-Bedmar, I., Chacón-Solano, E., & Guerrero-Aspizua, S. (2022). The RareDis corpus: a corpus annotated with rare diseases, their signs and symptoms. *Journal of Biomedical Informatics*, *125*, 103961.

-Mohan, S., & Li, D. (2019). Medmentions: A large biomedical corpus annotated with UMLS concepts. *arXiv preprint arXiv:1902.09476*.

- Mutinda, F., Liew, K., Yada, S., Wakamiya, S., & Aramaki, E. (2022, November). PICO corpus: a publicly available corpus to support automatic data extraction from biomedical literature. In *Proceedings of the first Workshop on Information Extraction from Scientific Publications* (pp. 26-31).

- Nye, B., Li, J. J., Patel, R., Yang, Y., Marshall, I. J., Nenkova, A., & Wallace, B. C. (2018, July). A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. In *Proceedings of the conference. Association for Computational Linguistics. Meeting* (Vol. 2018, p. 197). NIH Public Access.

-Pérez-Pérez, M., Rabal, O., Pérez-Rodríguez, G., Vazquez, M., Fdez-Riverola, F., Oyarzabal, J., ... & Krallinger, M. (2017). Evaluation of chemical and gene/protein entity recognition systems at BioCreative V. 5: the CEMP and GPRO patents tracks.

- Riley, D. S., Barber, M. S., Kienle, G. S., Aronson, J. K., von Schoen-Angerer, T., Tugwell, P., Kieny, H., Helfand, M., Altman, D. G., Sox, H., Werthmann, P. G., Moher, D., Rison, R. A., Shamseer, L., Koch, C. A., Sun, G. H., Hanaway, P., Sudak, N. L., Kaszkin-Bettag, M., … Gagnier, J. J. (2022). CARE guidelines for case reports: explanation and elaboration document. Translation into Russian. *Digital Diagnostics*, *3*(1). https://doi.org/10.17816/DD105291

-Savery, M. E., Rogers, W. J., Pillai, M., Mork, J. G., & Demner-Fushman, D. (2020). Chemical entity recognition for MEDLINE indexing. *AMIA Summits on Translational Science Proceedings*, *2020*, 561.

-Sioutos, N., de Coronado, S., Haber, M. W., Hartel, F. W., Shaiu, W. L., & Wright, L. W. (2007). NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information. *Journal of biomedical informatics*, *40*(1), 30-43.

-Tanabe, L., Xie, N., Thom, L. H., Matten, W., & Wilbur, W. J. (2005). GENETAG: a tagged corpus for gene/protein named entity recognition. *BMC bioinformatics*, *6*, 1-7.

-Uzuner, Ö., South, B. R., Shen, S., & DuVall, S. L. (2011). 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, *18*(5), 552-556.

-Van Mulligen, E. M., Fourrier-Reglat, A., Gurwitz, D., Molokhia, M., Nieto, A., Trifiro, G., ... & Furlong, L. I. (2012). The EU-ADR corpus: annotated drugs, diseases, targets, and their relationships. *Journal of biomedical informatics*, *45*(5), 879-884.

-Wallace, B. C., Kuiper, J., Sharma, A., Zhu, M. B., & Marshall, I. J. (2016). Extracting PICO sentences from clinical trial reports using supervised distant supervision. *Journal of Machine Learning Research*, *17*(132), 1-2.

-Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., ... & Wilson, M. (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic acids research*, *46*(D1), D1074-D1082.