CACRC-Pan: A Concept Annotated Case Report Corpus and Processing Pipeline for Rare Pancreatic Diseases

Christelle MULIMBI [a], Amaury FIERENS [a], Lucile DIERCKX [a,c], Sébastien JODOGNE [a], Siegfried NIJSSEN [a,b,c]
[a] *ICTEAM, UCLouvain, Louvain-la-Neuve, Belgium*
[b] *DTAI, KU Leuven, Leuven, Belgium*
[c] *TRAIL Institute, Belgium*

Abstract. We introduce CACRC-Pan, a corpus of case reports for six rare pancreatic diseases (AIP1-2, CF, HCP, PS and SDS) annotated at two concept levels: token-level symptom labels (HPO-coded spans) and document-level disease labels. This dual annotation enables both standard natural language processing (NLP) tasks and studies on concept bottleneck models, linking interpretable symptom concepts to diagnostic outcomes. We also provide a baseline NLP pipeline combining PubMedBERT for symptom extraction with a lightweight voting classifier for disease inference. Results demonstrate the effectiveness of this simple, interpretable approach, and all resources are released openly to support reproducible research on rare pancreatic diseases.

Keywords. Rare diseases, Pancreas, Named entity recognition, Case report classification, Concept bottleneck models

1. Introduction

A major challenge in modern medicine is the diagnostic odyssey, the prolonged and complex process of reaching an accurate diagnosis. Patients with rare diseases frequently experience years of uncertainty and misdiagnoses, a situation especially critical for those with early symptom onset. Such delays worsen clinical outcomes and impose significant burdens on patients and their families [1].

To address these challenges, researchers have introduced text-based datasets for rare disease diagnosis. The RareDis corpus, for instance, links diseases to clinical manifestations in annotated case reports [2] and benchmarks deep learning–based models for named entity recognition (NER) [3]. More recently, labeled medical records and case reports for automated rare-disease detection have been released using MIMIC-III [4]. Related NLP work includes symptom extraction from electronic health records, patient pre-screening, and retrieval of similar cases [5], aligning with broader surveys of publicly available clinical NLP datasets [6]. Overall, these efforts underscore NLP's promise for rare disease diagnosis and the need for more specialized, domain-specific resources.

A key distinction in this context is between symptom labels and disease labels. For patients and doctors, it can be unsatisfactory to receive a disease prediction from an NLP pipeline without understanding the reasons. A corpus with symptom labels enables two-

step pipelines: first, NLP models recognize symptoms; second, an interpretable model predicts disease labels from the symptom labels—an approach central to recently proposed *concept bottleneck models* [7], which predict outcomes via intermediate, human-interpretable concepts. To promote research in this field, we introduce CACRC-Pan, an open-data corpus covering six rare pancreatic diseases: autoimmune pancreatitis type 1 and 2 (AIP1-2), cystic fibrosis (CF), hereditary chronic pancreatitis (HCP), Pearson syndrome (PS), and Shwachman-Diamond syndrome (SDS). CACRC-Pan is annotated at two concept levels—token-level symptom labels (HPO-coded spans) and document-level disease labels—supporting the training of concept bottleneck models. In our pipeline, we pair a baseline symptom recognition model with a lightweight disease inference module, providing a reproducible framework to study concept-based reasoning and reduce diagnostic uncertainty in rare pancreatic diseases.

2. Methods

In this section, we describe the construction and annotation of CACRC-Pan corpus, followed by the baseline pipeline and evaluation protocol. The corpus consists of publicly available case reports describing six rare pancreatic diseases: AIP1-2, CF, HCP, PS, and SDS. Disease synonyms from Orphanet [8] and associated MeSH terms were used to query PubMed for English human case reports, available freely or via institutional access. Only case-report sections were extracted from each paper. Manual preprocessing removed family-history and explicit-negation sentences and mapped numeric laboratory values to standardized terms (e.g., “amylase 245 IU/L” to “hyperamylasemia”). Reports were annotated via rule-based dictionary matching HPO terms and synonyms [9], augmented with SNOMED CT cross-references. Matches yielded HPO-coded symptom spans, and each report received a document-level ORPHA disease label. All labels were manually checked for terminological accuracy and span consistency, then converted into token-level BIO tags (B-HPO, I-HPO, O) for NER training. Class sizes range from 26 to 129 reports, with most diseases represented by 100 cases. The corpus provides a useful resource for symptom extraction and disease classification in rare pancreatic diseases, along with dual annotation at the symptom and disease levels, enabling research on interpretable, concept-based prediction models such as concept bottleneck architectures. Figure 1 illustrates an entry of the CACRC-Pan corpus.

ID	Text	HPOcode	Term	Start	End	ORPHAcode
[AIPF0001] [PMID25524485]	A 22-year-old man presented with pruritus of 3 weeks duration. He complained of mild abdominal pain, nausea, vomiting, and dark urine. Physical examination disclosed icteric sclerae. Initial laboratory tests showed an elevated transaminases with elevated serum AST, 70 IU/L and elevated serum ALT, 161 IU/L, a cholestasis with hyperphosphatasemia, 937 IU/L and total bilirubin 3.1 mg/dl, direct bilirubin 2.0 mg/dl, amylase 73 U/L, and a hyperlipasemia, 258 U/L. Serum IgG4 level was within normal range. Abdominal computed tomography showed diffuse enlarged pancreatic head and a 1.3 cm–sized enhancing lymph node adjacent to the distal common bile duct. Microscopic examinations of the resected pancreas showed diffuse lymphoplasmacytic infiltration with acinar destruction, interstitial storiform fibrosis and obliterative phlebitis. Immunohistochemical staining showed abundant IgG4-positive cells (>50 cells/high-power field [HPF]). In addition to these findings, granulocytic epithelial lesions were seen in the interlobular pancreatic ducts.	HP:0000989	pruritus	33	41	ORPHA:280302

Figure 1. Annotated case report from the CACARC-Pan corpus

The NER component of our baseline pipeline detects tokens corresponding to HPO-coded symptoms. Long reports are split into 512-tokens chunks, and token labels are aligned with span annotations to build input-label pairs. We evaluated BlueBERT, BioBERT, ClinicalBERT, PubMedBERT, and SciBERT [10], tuning hyperparameters

(batch size, learning rate) by grid search with a class-weighted loss. Evaluation is conducted at the entity level (span-level) using precision, recall, and F_2 scores.

The outputs of the NER component are then used to infer the most likely disease label among the six ORPHA codes. Given a case p, let T_p be the NER-detected symptom terms and HPO_p their corresponding codes. For each disease d, Orphanet/ORDO provides HPO_d, the set of HPO symptom codes associated with d, and their preferred HPO labels T_d. Our lightweight classifier computes four similarity metrics, each quantifying how closely a set T_p of case report symptoms matches each set T_d of disease symptoms. A voting scheme then combines two (bi-source) or three (tri-source) metrics into a final prediction.

Specifically, the four computed similarity metrics are the following. The *term-dense* metric (**W2V**) computes the cosine similarity between Word2Vec-like embeddings [11]. Here, Word2Vec was trained on lists of patient terms by representing each case report p and each disease d as the mean vector of their respective symptom terms T_p and T_d. The *Jaccard-based* metric (**J**) corresponds to the standard Jaccard similarity between the two sets HPO_p and HPO_d. The *frequency-weighted* metric (**FW**) leverages Orphanet/ORDO frequency annotations (via Orphadata), weighting each disease-symptom link by its reported occurrence: *obligate* = 1.00, *very frequent* = 0.95, *frequent* = 0.65, *occasional* = 0.20, *very rare* = 0.05. Finally, the *term-sparse* metric (**TF**) computes cosine similarity between TF-IDF vectors: A TF-IDF vectorizer is fit on disease profiles (each T_d concatenated into a short document) and applied to the patient profile (T_p).

A weighted voting scheme combines the corresponding metric scores into the final prediction, with weights α, β, and γ manually tuned under the constraints $\alpha+\beta = 1$ for bi-source and $\alpha+\beta+\gamma = 1$ for tri-source. We used an 85/15 stratified train/test split by ORPHA code. Together, the NER model and weighted voting classifier form our baseline NLP pipeline for rare pancreatic disease classification.

3. Results

This section presents the performance of our baseline NLP pipeline on the CACRC-Pan corpus. Table 1 compares the considered biomedical Transformer models for the NER symptom extraction. PubMedBERT achieved the best balance, as recall and F_2 are prioritized over precision. In the rare disease context, this ensures that most true symptoms are captured, reducing the risk of missing critical information while maintaining false positives at a manageable level.

Table 1. Token-level symptom NER. Model selection is focused on recall and F_2.

Model	Batch	Learning rate	Recall (%)	F_2 (%)	Precision (%)
BlueBERT (PubMed)	1	3e-5	86.7	83.9	74.5
BioBERT	1	2e-5	90.8	89.9	**86.6**
ClinicalBERT	1	3e-5	90.6	89.1	83.8
PubMedBERT	1	2e-5	**93.6**	**90.8**	81.3
SciBERT (cased)	4	2e-5	92.8	**90.8**	83.6

For disease inference, because CACRC-Pan is class-imbalanced (26-129 reports per disease), we report macro- and weighted-averaged recall and F_1. The macro average computes the metric independently for each class and then takes the unweighted mean, treating all diseases equally. The weighted average uses class support as weights, reflecting overall performance under imbalance.

Table 2 reports the results for rare disease inference, assuming perfect symptom identification in case reports. Among individual metrics, the frequency-weighted metric performs best, followed by the term-dense and Jaccard-based ones. However, using a weighted combination improves performance: Bi-source score combining the frequency-weighted and term-dense metric surpasses any single metric, and tri-source score adding the term-sparse metric achieves the highest F_1 and recall. When used alone, the term-sparse metric performs poorly and is therefore omitted from Table 2, as are other suboptimal fusion variants.

The final baseline for the full pipeline combines PubMedBERT with the tri-source fusion score, combining the frequency-based, term-dense, and term-sparse metrics. End-to-end evaluation yields an F_1 of 87.5% (macro) and 90.4% (weighted), with recall of 88.4% and 90.1%, respectively. The overall accuracy of 90.1% indicates the robustness of the symbolic layer to small extraction errors.

Table 2. Performance of rare disease inference on the test set. This table reports the four individual metrics ("W2V" – term-dense, "FW" – frequency-weighted, "J" – Jaccard-based, and "TF" – term-sparse), as well as the fusion score.

		Tri-source (FW+W2V+TF)	**Tri-source (J+W2V+TF)**	**Bi-source (FW+W2V)**	**Bi-source (J+W2V)**	**FW**	**W2V**	**J**
Macro	F_1	**83.6**	80.6	81.9	76.2	78.3	68.7	65.3
	Recall	**85.7**	84.0	84.1	80.1	80.9	70.8	70.7
Weighted	F_1	**86.2**	84.9	85.5	80.5	80.6	74.0	70.3
	Recall	**85.8**	84.3	85.0	80.0	80.2	73.2	68.9
Weights		$\alpha = 0.1$, $\beta = 0.2$, $\gamma = 0.7$	$\alpha = 0.3$, $\beta = 0.4$, $\gamma = 0.3$	$\alpha = 0.1$, $\beta = 0.9$	$\alpha = 0.6$, $\beta = 0.4$	–	–	–

4. Discussion

This work introduces a concept-annotated corpus of case reports and a reproducible benchmark for classifying textual case reports of rare pancreatic diseases. The dual annotation enables the development of interpretable models relying on meaningful clinical concepts. Our baseline NLP pipeline combines a BERT-like encoder tuned for high sensitivity, essential in the rare disease context, with a tri-source fusion score for disease identification. The best-performing score fusion prioritized the term-sparse metric ($\gamma = 0.7$), based on cosine similarity over TF-IDF vectors, suggesting that sparse, disease-specific terms are useful for discrimination. While macro scores are lower than weighted ones due to class imbalance, their high values indicate that even the rarest classes are well represented.

Although these findings demonstrate the value of CACRC-Pan and its baseline NLP pipeline, they also highlight directions for further improvement. The corpus is limited to six conditions, English-only case report, and positive cases only. Preprocessing removes negated and family-history sentences, which may bias the distribution and underestimate real-world complexity. NER evaluation at the entity (span) level may not fully reflect patient-level utility, and fusion weights could be sensitive to distribution shifts. In addition, Word2Vec is trained on task-specific term lists, which may limit coverage. Ultimately, clinical, multi-site validation will be necessary to assess the real-world impact of NLP for rare disease diagnosis.

5. Conclusion

This paper introduces CACRC-Pan, an open-data, symptom focused corpus of case reports for six rare pancreatic diseases, with an open-source, reproducible NLP pipeline[1]. The corpus is doubly annotated at the symptom and disease levels, enabling concept based and interpretable learning approaches, such as concept bottleneck models. The pipeline frames symptom detection as a NER task using a PubMedBERT encoder, chosen to prioritize high recall and F_2 scores in order to reduce the risk of missing clinically informative symptoms. Detected symptoms are then integrated through a lightweight multi-metric scoring scheme to rank the most likely underlying disease, and end-to-end experiments demonstrate that this simple design provides a solid baseline for future benchmarking.

Future work will assess the pipeline on data augmented with disease-free patients and common diseases exhibiting comparable symptom coverage, to better approximate real-world conditions. We also plan to strengthen case-report classification with more recent dense representation methods, such as BERT-based models, rather than Word2Vec embeddings. Finally, we will extend both the corpus and the pipeline to account for negation, temporality, and family history, and to incorporate normalized laboratory values as structured features, enabling more realistic rare-disease inference.

References

[1] F. Faye, C. Crocione *et al.*, "Time to diagnosis and determinants of diagnostic delays of people living with a rare disease: results of a Rare Barometer retrospective patient survey," *European Journal of Human Genetics*, vol. 32, no. 9, pp. 1116-1126, Sep. 2024.

[2] I. Segura-Bedmar, D. Camino-Perdones, and S. Guerrero-Aspizua, "Exploring deep learning methods for recognizing rare diseases and their clinical manifestations from texts," *BMC Bioinformatics*, vol. 23, no. 1, p. 263, Jul. 2022.

[3] C. Martínez-deMiguel, I. Segura-Bedmar, E. Chacón-Solano, and S. Guerrero-Aspizua, "The RareDis corpus : A corpus annotated with rare diseases, their signs and symptoms," *Journal of Biomedical Informatics, vol. 125, p. 103961, Jan. 2022.*

[4] M. Rolando, V. Raggio, H. Naya, L. Spangenberg, and L. Cagnina, "A labeled medical records corpus for the timely detection of rare diseases using machine learning approaches," *Scientific Reports*, vol. 15, no. 1, p. 6932, Feb. 2025, publisher: Nature Publishing Group.

[5] N. Garcelon, A. Burgun, R. Salomon, and A. Neuraz, "Electronic health records for the diagnosis of rare diseases," *Kidney International*, vol. 97, no. 4, pp. 676-686, Apr. 2020.

[6] Y. Gao, D. Dligach, L. Christensen, et al., "A Scoping Review of Publicly Available Language Tasks in Clinical Natural Language Processing," *Journal of the American Medical Informatics Association*, vol. 29, no. 10, pp. 1797–1806, 2022, doi: 10.1093/jamia/ocac127.

[7] P. W. Koh, T. Nguyen, Y. S. Tang, S. Mussmann, E. Pierson, B. Kim, and P. Liang, "Concept bottleneck models," in *International conference on machine learning*. PMLR, 2020, pp. 5338-5348.

[8] A. Rath, A. Olry, F. Dhombres, M. M. Brandt, B. Urbero, and S. Ayme, "Representation of rare diseases in health information systems: the Orphanet approach to serve a wide range of end users," *Human Mutation*, vol. 33, no. 5, pp. 803-808, May 2012.

[9] S. Köhler *et al.*, "The Human Phenotype Ontology in 2021," *Nucleic Acids Research*, vol. 49, no. D1, pp. D1207-D1217, Jan. 2021.

[10] B. Wang, Q. Xie *et al.*, "Pre-trained Language Models in Biomedical Domain: A Systematic Survey," *ACM Comput. Surv.*, vol. 56, no. 3, pp. 55:1-55:52, Oct. 2023.

[11] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and their Compositionality," in *Advances in Neural Information Processing Systems,* vol. 26. Curran Associates, Inc., 2013.

[1]Corpus and code available at: https://forge.uclouvain.be/AmauryFierens/cacrc-pan