

# BioCreative VIII – Task 3: Genetic Phenotype Normalization from Dysmorphology Physical Examinations

Davy Weissenbacher<sup>1</sup>, Xinwei Zhao<sup>2</sup>, Jessica R. C. Priestley<sup>2</sup>, Katherine M. Szigety<sup>2</sup>, Sarah F. Schmidt<sup>2</sup>, Karen O'Connor<sup>3</sup>, Ian M. Campbell<sup>2,3,\*</sup>,<sup>†</sup>, Graciela Gonzalez-Hernandez<sup>1,\*</sup>

1. Cedars-Sinai Medical Center West Hollywood, CA, USA; 2. Children's Hospital of Philadelphia, Philadelphia, PA, USA; 3. University of Pennsylvania, Philadelphia, PA, USA; \* Contributed equally as senior authors.

<sup>†</sup> Corresponding author: [campbellim@chop.edu](mailto:campbellim@chop.edu)

## Abstract

The BioCreative VIII Task 3 focuses on normalizing terms mentioned in dysmorphology physical examinations to the Human Phenotype Ontology (HPO) to enable computational analysis geared towards finding correlations between patients with rare genetic diseases, delineate undescribed genetic conditions, or further our understanding of existing ones, among other applications. We made available 3,136 deidentified and manually annotated observations extracted from dysmorphology physical examinations of 1,652 pediatric patients. Task 3 consists of detecting all HPO terms mentioned in an observation and returning the HPO IDs associated with the terms detected. This task is challenging due to discontinuous, overlapping, and descriptive mentions of HPO terms, making strict matching approaches inefficient. The large size and incompleteness of the HPO ontology also prevents the annotation of an exhaustive training set to train conventional multi-class classifiers. A total of 20 teams registered, and 5 teams submitted their predictions. We summarize the corpus, the competing systems, and their results. Using a pre-trained large language model, the top system achieved a .82 F1 score, a score close to human performance, which confirms the recent advance in natural language processing recently commented on the media. The post-evaluation period of the challenge is still open for submission at <https://codalab.lisn.upsaclay.fr/competitions/11351>.

## Motivation

The dysmorphology physical examination is a critical component of the diagnostic evaluation in clinical genetics. This process catalogues often minor morphological differences of the patient's facial structure or body, but it may also identify more general medical signs such as neurologic dysfunction. The findings enable correlation of the patient with known rare genetic diseases. They therefore directly influence clinical diagnosis, the selection of genetic testing, and the interpretation of results---particularly when testing reveal variants of uncertain clinical significance. Beyond the clinic, such information is also useful to researchers attempting to delineate undescribed genetic conditions or to further our understanding of existing ones. Whereas the medical findings are key information, they are nearly always captured within the electronic health record (EHR) as unstructured free text, making it unavailable for downstream computational

analysis. Advanced Natural Language Processing methods are therefore required to retrieve the information from the records.

## Task description and Corpus

Dysmorphology physical examinations are frequently documented in the electronic medical record as a series of organ system observations. To standardize the description and comparison of dysmorphic findings, many clinicians and laboratory professionals utilize the Human Phenotype Ontology (1). This ontology is specially designed for human genetics. The conversion of documented findings into HPO terms is highly labor-intensive and frequently requires advanced training in genetics for the best results.

The BioCreative VIII shared task Track 3 calls for automated systems to extract and normalize the key findings in observations written during dysmorphology physical examinations. Consider the following organ system observation: *EYES: long palpebral fissures with slight downslant. Normal eyebrows.* Two key findings are discussed: *Long palpebral fissure* - HP:0000637 and *Downslanted palpebral fissures* - HP:0000494. A successful system should extract the span of text referring to the key findings, in our example *long palpebral fissures* and *palpebral fissures with slight downslant*, and normalize them to term IDs in the HPO ontology, HP:0000637 and HP:0000494 respectively. The system should ignore the normal finding of *Normal eyebrows*. Both steps, the extraction and the normalization, are particularly difficult on our corpus given the current state of the art in Natural Language Processing.

The extraction step is challenging due to the descriptive style of the examinations and their polarity. The observations are short reports where, for conciseness, the mention of a finding can overlap with another finding or be discontinuous. The previous observation is an example of overlapping findings, with the text *palpebral fissures* contributing to both HP:0000637 and HP:0000494 terms. For discontinuous findings, where findings are defined with non-consecutive words, consider *Short nasal bridge* - HP:0003194 in: *NOSE: Short, wide nasal bridge.* Designed extractors had to go beyond the standard sequence labeling approach which, designed to extract contiguous and mutually exclusive terms, fails to capture the discontinuous and overlapping terms. As an additional challenge, the extractor also had to resolve the polarity of the findings, that is, automatically detecting and ignoring normal findings, and only returning the key positive findings.

The normalization step is challenging both due to the large scale of the HPO ontology and unpredictable levels of term detail available. Standard strategies for multi-label classification are designed to assign small sets of classes to input instances. However, to be successful in the task, normalizers had to adapt traditional strategies to assign one term from among the 17,000 terms in the HPO to each key finding detected in an observation. This assignment had to frequently occur without supervision since due to the size of the ontology, any training sets are more likely to not provide examples of use for all terms in the HPO. Furthermore, while specifically designed for human genetics, and constantly improving, the HPO does not have standardized levels of term detail. Consequently, a key finding may need to be matched with a close ancestor in the hierarchy of the ontology, making the default strict matching strategy inefficient since the string of the ancestor will be different from the string of the key finding. For example, there exists both *Naevus flammeus of the eyelid* - HP:0010733 and *Nevus flammeus of the forehead* - HP:0007413, but no

term for the nose, leaving only generic *Nevus flammeus* - HP:0001052 to describe this abnormality of the nose.

Our corpus consists of 3,136 organ system observations extracted from dysmorphology physical examinations of 1,652 pediatric patients evaluated at the Children's Hospital of Philadelphia. We split the dataset into a training set (55%, 1716 observations), a validation set (15%, 454 observations), and a test set (30%, 966 observations). We added 2427 observations in the test set as decoy observations. We automatically de-identified the text of all observations using NLM Scrubber (2) and manually reviewed the text during the annotation process to preserve patient privacy. Four physicians and one medical student annotated all mentions of key positive findings as well as normal findings in the observations. They assigned each finding to its most detailed and unambiguous term in the 2022-06-11 release of the HPO. We speed up the annotation by pre-processing the observations with a baseline system, PhenoTagger (3). Our annotators double-annotated, at least twice, 890 observations in our corpus. In this set, we found an inter-annotator agreement of 0.844 average F1 score. We computed the average with all permutations between the annotators. The annotation discrepancies were automatically resolved by selecting the annotations of the physician with the most years of clinical experience.

## Evaluation

We evaluated the ability of the competing systems to normalize all mentions of key findings in an observation (*Normalization-only*), regardless of whether they could detect the spans of the mentions. We selected the standard Precision, Recall and F1 scores to measure their performance. We also evaluated their ability to both detect the spans of mentions of key findings and normalize them as a supplementary evaluation (*Overlapping Extraction & Normalization*). We selected the overlapping Precision, Recall and F1 scores where a system was rewarded when it extracted the spans or a part of the spans of a labeled key finding mention and correctly assigned the labeled HPO term ID to the mention. Because only the normalization of the key findings in the observation is medically relevant for the physicians, we chose the best system to be the system which achieved the best F1 score on the Normalization-only evaluation.

Multiple baseline systems, freely available and open source, were available to participants to perform our task off the shelf. We evaluated *txt2hpo* (4), *Doc2HPO* (5), *NeuralCR* (6), PhenoBERT (7), and PhenoTagger (3). PhenoTagger was the best-performing system on our test set, with an F1 score of .633 when performing the normalization only, which we reported in Table 1. The authors of PhenoTagger approached the problem with a hybrid method combining dictionary matching and machine learning. The authors compiled a dictionary from the list of all observable terms and their synonyms in the HPO. They used this dictionary to build a training dataset with distant supervision. They trained a BioBERT model with this dataset to classify each n-gram in an observation into an HPO term ID or the special tag None, while retaining term IDs surpassing a predetermined threshold. The outcomes of the dictionary matching and the classifier were subsequently combined to generate the final output.

## Systems

Twenty teams registered for the shared task and five submitted three prediction files, which was the maximum number of submission files authorized. We kept the best predictions for each team.

We present the results of each team and summarize the architecture of their best system in Table 1. All systems achieved better performance than the baseline system. Based on a large generative model for the extraction and a combination of generation and dictionary matching for the normalization, the best system (8) achieved an F1 score of 0.82, only 2 points under human performance on this task (i.e. the inter-annotator agreement of 0.844 average F1 score). This confirms the recent technical improvement allowed by Large Language Models (LLMs) largely commented on in the media.

*Table 1: Systems performance (F1: F1-scores; P: Precision; R: Recall) and System summaries (TG: Term Generation; TE: Term Extraction; TN: Term Normalization)*

Team	Normalization-only			Overlapping Extraction & Normalization			System Summary
	F1	P	R	F1	P	R	
Soysal & Robert (8)	.820	.842	.799	.817	.841	.794	TG: fined tuned GPT-4 + TN: exact matching on stems TE: multiple W2NER instances relying on various BERT models + TN: ensemble of Bioformer & dictionary matching TG: fined tuned chatGPT + TN: synonym marginalization (BioSyn) TG: FLAN-T5-XL fined tuned with LoRA + TN: distance similarity (SBERT) and candidates re-ranking with Cross-encoder TN: Ensemble (PhenoTagger & PhenoBERT) + TE: BioLinkBERT Ensemble of a BioBERT multi-class classifier & dictionary matching
Qi et al. (9)	.763	.831	.706	.762	.830	.704	
Kim et al. (10)	.745	.735	.755	.743	.734	.752	
Alhassan et al. (11)	.723	.718	.728	.721	.717	.726	
Lin et al. (12)	.644	.762	.557	.642	.761	.556	
Baseline (3)	.633	.587	.687	.632	.586	.685	

All systems proposed a pipeline approach that divides the process into two subtasks, extraction followed by normalization. The teams adopted two different strategies to handle discontinued and overlapping terms which counted for 16.9% (213/1258) of the terms in our test set. The first strategy unifies the extraction of all terms (continuous and discontinuous) by identifying relations between the tokens. This strategy was successfully implemented by Qi et al. with an ensemble of W2NER (13). The second strategy benefits from the recent improvements of LLMs to generate the list of terms mentioned in an observation. This strategy was the most popular on our task with three of the five competing systems implementing it. All systems except the best system handle explicitly the detection of normal findings by detecting negations in the context of the terms extracted or by training dedicated classifiers. Soysal & Kirk (8), in contrast, taught implicitly their LLMs to ignore normal findings by only presenting key findings during the training of the model. In (14) we found this approach to be very effective since our transformer-based model, trained in a similar way, identified almost perfectly the normal findings in the test set.

Participants normalized the terms extracted with various approaches. Exact matching selects the HPO term which matches exactly a term extracted. It remains the default approach for three out of five systems, but it only normalizes obvious candidates and was used in combination with machine learning which normalized the remaining candidates. We saw a greater variety in the use of machine learning for normalization. The most innovative was to prompt an LLM to generate the HPO preferred term corresponding to the extracted term candidate in (8). Others were normalizing a term candidate either by predicting the closest HPO term with a multi-class classifier (9, 12) or by selecting the closest HPO term to the candidate in an embedding space (10, 11). We will run additional experiments by providing each competing system with the gold standard spans of the HPO terms to evaluate their normalization approaches independently from their extraction

approaches. We will then analyze their predictions to measure the impact of terms unseen in the training set but occurring in the test set as well as HPO terms referred by the clinicians using a description or an explanation, both known to be challenging for current normalizers.

## Conclusion

In this paper, we presented the results of the BioCreative VIII Task 3 which challenges participants to extract and normalize key findings in 3,136 observations from dysmorphic examinations. Given an observation, the task consists of detecting the spans of all key findings mentioned and returning the list of their corresponding IDs in the Human Phenotype Ontology. This task is challenging due to 1) discontinue, overlapping, and often descriptive mentions of HPO terms making strict matching approaches inefficient and, 2) the large size and incompleteness of the ontology preventing the annotation of an exhaustive training set to train conventional multi-label classifiers. All 5 systems competing resolved the task with a pipeline approach, with most systems relying on LLMs to generate the terms mentioned in the observations. The top system achieved .82 F1 score when normalizing the terms, a score which reaches human performance and confirms the recent advance in natural language processing largely commented on the media.

## Fundings

IMC was supported by grant K08-HD111688 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development. GGH and DW were partially supported by grant R01LM011176 from the National Library of Medicine, and by grant R01AI164481 from the National Institute of Allergies and Infectious Diseases.

## References

1. Köhler, S., Gargano, M., Matentzoglou, N. et al. (2021) The Human Phenotype Ontology in 2021. *Nucleic Acids Research*, **49**, D1, pp. D1207-D1217.
2. <https://lhncbc.nlm.nih.gov/scrubber/>. Last access November 6, 2023.
3. Luo, L., Yan, S., Lai, PT. et al. (2021) PhenoTagger: a hybrid method for phenotype concept recognition using human phenotype ontology. *Bioinformatics*, **37**, 13, pp. 1884-1890.
4. <https://github.com/GeneDx/txt2hpo/>. Last access November 6, 2023.
5. Liu, C., Sampaio Peres Kury, F., Li, Z., et al. (2019) Doc2Hpo: a web application for efficient and accurate HPO concept curation. *Nucleic Acids Research*, **47**, W1, pp. W566-W570.
6. Arbabi, A., Adams, D. R., Fidler, S., Brudno, M., (2019) Identifying Clinical Terms in Medical Text Using Ontology-Guided Machine Learning. *JMIR Med Inform*, **7**, 2, pp. e12596.

7. Feng, Y., Qi, L., Tian, W., (2023) PhenoBERT: A Combined Deep Learning Method for Automated Recognition of Human Phenotype Ontology. IEEE/ACM Transactions on Computational Biology and Bioinformatics, **20**, 2, pp. 1269-1277.
8. Soysal, E., and Roberts, K., UTH-Olympia@BC8 Track 3: Adapting GPT-4 for Entity Extraction and Normalizing Responses to Detect Key Findings in Dysmorphology Physical Examination Observations. Proceedings of the BioCreative VIII Challenge Evaluation Workshop.
9. Qi, J., Luo, L., Yang, Z., and Lin, H. (2023) DUTIR-BioNLP@BC8 Track 3: Genetic Phenotype Extraction and Normalization with Biomedical Pre-trained Language Models. Proceedings of the BioCreative VIII Challenge Evaluation Workshop.
10. Kim, H., Kim, C., Sohn, J., Beck, T., Rei, M., Kim, S., Simpson, I., Posma, J.M., Lain, A., Sung, M., and Kang, J. (2023) Advancing Phenotype Named Entity Recognition and Normalization for Dysmorphology Physical Examination Reports. Proceedings of the BioCreative VIII Challenge Evaluation Workshop.
11. Alhassan, A., Schlegel, V., Aloud, M., Batista-Navarro, R., and Nenadic, G. (2023) DiscHPO@BC8 Track 3: Recognising and Normalising Continuous and Discontinuous Genetic Phenotypes Using T5 Variants and Sentence-Transformers Models. Proceedings of the BioCreative VIII Challenge Evaluation Workshop.
12. Lin, Y-J., Feng, Z., and Kao, H-Y. (2023) IKMLab@BC8 Track 3: Sequence Tagging for Position-Aware Human Phenotype Extraction with Pre-trained Language Models. Proceedings of the BioCreative VIII Challenge Evaluation Workshop.
13. Li, J., Fei, H., Liu, J., et al. (2022) Unified Named Entity Recognition as Word-Word Relation Classification. Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI-22), pp. 10965-10973.
14. Weissenbacher, D., Rawal, S., Zhao, X., et al. (2023) PhenoID, a language model normalizer of physical examinations from genetics clinical notes. medRxiv [Preprint].