DATABASE
The Journal of Biological Databases and Curation

LSD600: the first corpus of biomedical abstracts annotated with lifestyle–disease relations

Esmaeil Nourani [1,2], Evangelia-Mantelena Makri[1,3], Xiqing Mao[1], Sampo Pyysalo[4],
Søren Brunak [1], Katerina Nastou [1,5,*], Lars Juhl Jensen[1,6,*]

[1]Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Blegdamsvej 3, Copenhagen 2200, Denmark
[2]Faculty of Information Technology and Computer Engineering, Azarbaijan Shahid Madani University, Tabriz, Iran
[3]Department of Nutrition and Dietetics, Harokopio University, Athens 17676, Attiki, Greece
[4]TurkuNLP group, Department of Computing, Faculty of Technology, University of Turku, Turku 20014, Finland

*Corresponding authors. Katerina Nastou, Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Blegdamsvej 3, Copenhagen 2200, Denmark. E-mail: katerina.nastou@cpr.ku.dk; Lars Juhl Jensen, Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Blegdamsvej 3, Copenhagen 2200, Denmark. E-mail: lars.juhl.jensen@cpr.ku.dk
[5]Present address: Department of Epidemiology Research, Statens Serum Institut, Copenhagen 2300, Denmark
[6]Present address: ZS Discovery, Lottenborgvej 26, 2800 Kongens Lyngby, Denmark.

Abstract

Lifestyle factors (LSFs) are increasingly recognized as instrumental in both the development and control of diseases. Despite their importance, there is a lack of methods to extract relations between LSFs and diseases from the literature, a step necessary to consolidate the currently available knowledge into a structured form. As simple co-occurrence-based relation extraction (RE) approaches are unable to distinguish between the different types of LSF-disease relations, context-aware models such as transformers are required to extract and classify these relations into specific relation types. However, no comprehensive LSF–disease RE system existed, nor a corpus suitable for developing one. We present LSD600 (available at https://zenodo.org/records/13952449), the first corpus specifically designed for LSF–disease RE, comprising 600 abstracts with 1900 relations of eight distinct types between 5027 diseases and 6930 LSF entities. We evaluated LSD600's quality by training a RoBERTa model on the corpus, achieving an F-score of 68.5% for the multilabel RE task on the held-out test set. We further validated LSD600 by using the trained model on the two Nutrition-Disease and FoodDisease datasets, where it achieved F-scores of 70.7% and 80.7%, respectively. Building on these performance results, LSD600 and the RE system trained on it can be valuable resources to fill the existing gap in this area and pave the way for downstream applications.

Database URL: https://zenodo.org/records/13952449

Introduction

Diseases are influenced by both lifestyle and genetic factors [1–4]. Substantial evidence suggests that targeted lifestyle interventions can significantly enhance disease prevention and management in precision medicine [5, 6]. Considering the growing importance of lifestyle factors (LSFs) in disease, attention has been placed towards mining the literature for such relations [7–9]. Despite these efforts to extract and centralize scientific knowledge, the focus is almost exclusively on the relation between diseases and nutrition, and the relations between other LSFs and diseases remain largely unconsolidated and scattered across the scientific literature. To mine LSF–disease relations from the scientific literature, the first step is to detect mentions of these entities in text. Then a straightforward approach to identify pairs of related diseases and LSFs is co-occurrence-based relation extraction (RE). While this will allow the detection of related entities, the nature of their relation will remain unknown. Although this is

a non-issue for gene–disease relations [10], as their nature is mostly causal, the case is completely different when it comes to LSF–disease relations. For example, 'exposure to radiation' is well known to be harmful to health [11], causing diseases like skin cancer [12]. At the same time, in modern medicine, radiotherapy, which also exposes the human body to radiation, is a common method for treating cancer [13]. Therefore, simple RE approaches cannot capture the complexity of LSF–disease relations and it becomes essential to develop more sophisticated approaches that can distinguish between different types of relations—from simple statistical associations (either positive and negative), to more concrete causal, or preventive relations.

The advent of transformer-based language models such as BERT [14] and RoBERTa [15] has revolutionized the field of RE, including domain-specific areas like biomedical RE [16]. However, to optimally use these transformer-based models, dedicated corpora for fine-tuning are required. The focus

of the BioNLP community regarding RE corpora for disease entities has been primarily on disease–gene [17, 18] or disease–chemical (BC5CDR) relations [19]. When it comes to LSF–disease relations, only two relevant corpora exist, namely the ND (Nutrition and Disease) corpus [7] and the FoodDisease corpus [9]. However, there is a notable lack of comprehensive corpora that cover a broader range of lifestyle factors beyond nutrition and different types of LSF–disease relations, highlighting the need for dedicated datasets for training LSF–disease relation extraction models.

In this study, we introduce LSD600, the first corpus specifically focused on LSF–disease relations. LSD600 consists of 600 abstracts annotated with LSF–disease relations, encompassing 1900 relations covering eight different relation types and nine categories of lifestyle factors [20]. We have used LSD600 to train a transformer-based model on the multilabel LSF–disease RE task, which achieved an F-score of 68.5% on our held-out test set. We further validated the model on two independent external corpora of nutrition–disease relations and saw comparable performance. This model represents a first important step in extracting LSF–disease relations from biomedical literature, which can form the basis for knowledge graphs.

Materials and methods

LSF–disease relation corpus

Document selection for annotation

Given that most of the over 37 million scientific articles available in PubMed do not pertain to LSF–disease relations, this project requires a rigorous selection process to identify relevant abstracts where such relations are expected to appear.

We started with the 200 abstracts from the LSF200 corpus [20], which were pre-annotated with a wide spectrum of LSF named entities. However, since this corpus was compiled to develop NER systems for LSF detection, many did not contain disease mentions or LSF–disease relations.

We therefore selected 400 additional abstracts from PubMed, which were required to mention at least five LSF and five disease mentions, based on using the JensenLab Tagger (https://doi.org/10.1101/067132) and publicly available LSF [20] and disease [10] dictionaries. Considering the uneven distribution of LSFs in the scientific literature [20], with terms related to nutrition being more common than terms from other categories, we selected at least 30 documents for each LSF category. The within-category selection of documents was done randomly among all documents that satisfied the criteria for number of LSF and disease mentions. The final corpus consists of 600 abstracts and is named LSD600.

The relation types schema

To capture the spectrum of relations between LSFs and diseases, we defined eight different relation types:

1. *Statistical Association*: Statistically significant association between two entity types, requiring the existence of an appropriate control group and the implementation of a statistical test.
2. *Positive Statistical Association*: Subclass of "Statistical Association" where a positive effect is clearly stated.
3. *Causes*: Subclass of "Positive Statistical Association" where causality is clearly implied.

4. *Negative Statistical Association*: Subclass of "Statistical Association" where a negative effect is clearly stated.
5. *Controls*: Subclass of "Negative Statistical Association" where a beneficial impact of the LSF on the disease is stated.
6. *Prevents*: Subclass of "Controls" where the LSF hinders the disease from occurring.
7. *Treats*: Subclass of "Controls" where the LSF has a therapeutic effect on the disease.
8. *No Statistical Association*: Relations where the absence of statistical association is clearly stated in text.

All relation annotations are nondirectional, since in associative relations the direction is undefined, while in relations of impact it is self-evident. For more details and examples on each relation type, please refer to our annotation documentation on Zenodo. Figure 1 shows the eight types of LSF–disease relations in LSD600 as shown in the BRAT RapidAnnotation Tool. Each line and associated color corresponds to a different type of relation: (1) "Statistical Association", (2) "Positive Statistical Association", (3) "Causes", (4) "Negative Statistical Association", (5) "Controls", (6) "Prevents", (7) "Treats," and (8) "No Statistical Association."

Named entity and relation annotation

For LSFs, we followed the definition provided by Nourani *et al.* [20], which defines LSF as encompassing all nongenetic health determinants associated with diseases. This includes a wide range of categories, such as physical and leisure activities, socioeconomic factors, personal care products, cosmetic procedures, sleep habits, mental health practices, and substance use. For disease entities, we used the disease dictionary described in Grissa *et al.*[10], which includes all names and synonyms from Disease Ontology [21] (including syndromes), which has been extended with additional terms from AmyCo [22] and manual additions of missing disease synonyms and acronyms. Symptoms are not considered as diseases.

We pre-annotated LSD600 with disease and LSF entities using the dictionary-based JensenLab tagger (https://www.biorxiv.org/content/10.1101/067132v1). As LSF200 is a corpus with LSF annotations, we only used the tagger to add disease pre-annotations to those abstracts. All pre-annotations were subsequently manually checked and corrected, and relations between them were manually annotated on all 600 abstracts. In the scientific literature, it is common to encounter alternative names referring to identical entities. We systematically annotated these equivalent entities, which is essential for evaluation purposes as it allows relationships stemming from either entity to be recognized as valid [23]. Figure 1 illustrates entity and relation annotation using an example for each relation type.

Manual annotation process and corpus evaluation

High-quality, consistent annotations are a prerequisite for good performance when fine-tuning deep learning classifiers, hence a clear set of guidelines is crucial [24]. Our annotation guidelines were formed and updated during a first round of Inter-Annotator Agreement (IAA) after three researchers (K.N., E.N., and X.M.) individually annotated 30 abstracts and discussed their inconsistencies. A second round of IAA was performed with a fourth annotator (E.M.), who is an
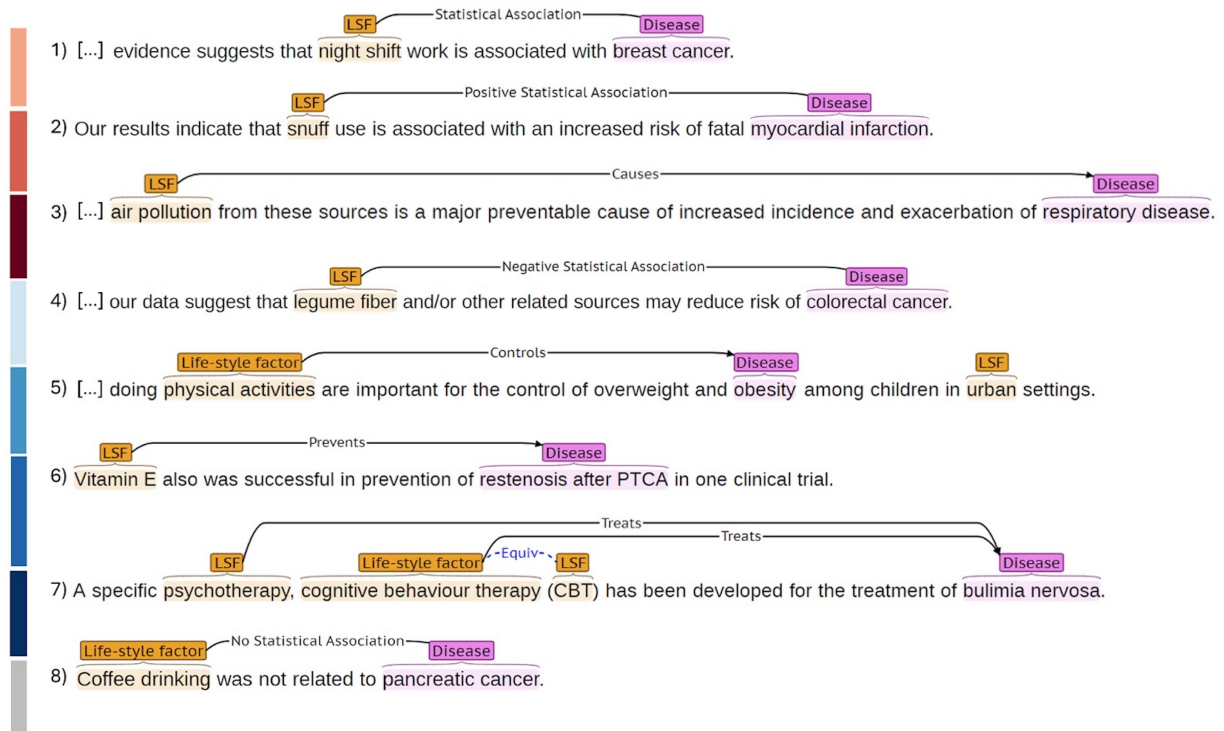
Figure 1. Illustration of the eight LSF–disease relation types in LSD600.

expert on the field of lifestyle factors, by annotating a new set of 30 abstracts to further update and solidify the final guidelines. To evaluate the quality of the guidelines and the corpus, an F1-score was calculated. The annotation guidelines support multilabel annotations, and the RE task was framed as a multilabel relation classification problem.

Tasks were assigned to annotators at the abstract level, with X.M annotating 400 abstracts and E.M annotating the remaining 200. To ensure consistency in the annotations, E.M also reviewed and updated the annotations for the 400 abstracts initially annotated by X.M. We annotated relations across sentences, ensuring that the connections between sentences were informed by the surrounding text. For a statistical association to be annotated, it must be statistically significant and the existence of an appropriate control group must be stated. To determine if an association is positive or negative, we considered both textual descriptions and any stated odds ratios, risk ratios, hazard ratios, or coefficient values. Hypothetical statements are not annotated as either. We annotated both direct and indirect relations that were stated, not inferred, as mentioned in detail in the guidelines.

The BRAT Rapid Annotation Tool [25] was used to perform all the annotations.

Transformer-based relation extraction
Relation extraction pipeline
The transformer-based system employed for relation extraction was adapted from a previously developed binary relation extraction tool, which has proven effective for RE in past applications [26, 27]. As described in these earlier papers, we frame RE as a multilabel classification task, aiming to determine which relation types (if any) exist for a pair of candidate named entities in the text. The system uses an architecture comprising a pretrained transformer encoder and a decision layer with a sigmoid activation function. It can leverage pretrained language models and accepts training, validation, and prediction data in both BRAT standoff and a custom JSON format, supporting extensive hyper-parameter optimization. Evaluation metrics are computed after each training epoch for hyper-parameter tuning. The system is trained for a predefined number of epochs, choosing the model weights that gave the highest F1-score on the development set.

As no separate sentence boundary detection is used, the system can train on and predict cross-sentence relations at the document level. To inform the model about which pair of named entities in the text to predict relation labels for, the entities are replaced by 'unused' tokens from the model. This masking approach prevents the model from learning based on the actual entities rather than the surrounding context. For additional details on data preprocessing and the pipeline, please refer to Mehryary *et al.* [26].

Experimental setup
In our experiments, we partitioned the LSD600 corpus into distinct training, development, and test sets. The system was trained on the training set, and hyperparameters were optimized by grid search to obtain the best F1-score on the development set. The test set was used only once for prediction and evaluation of the best model, after determining the optimal hyper-parameters.

Validation on external corpora
In addition to evaluating our model on the held-out test set from LSD600, we validated it on two external corpora, namely the ND (Nutrition and Disease) corpus [7] and the FoodDisease corpus [9].

The relation types in these corpora are not completely aligned with LSD600. To calculate performance of our model on these corpora, we mapped the relation types of the model to their relation types. The ND corpus uses only "Related" versus "Unrelated" labels. We collapsed all subclasses of "Statistical Association" in our hierarchy into one category, which can be mapped to "Related." "No Statistical Association" along with no extracted relation is considered as "Unrelated." For the FoodDisease corpus, the existing relation types are called "Cause" and "Treat"; however, these do not correspond to our classes with the same name. Instead, we collapsed all subclasses of "Positive Statistical Association" and mapped it to *Cause*. Similarly, we collapsed all subclasses of "Negative Statistical Association" and mapped it to "Treat." Since we do not retrain the model, we evaluate it on the entirety of both corpora.

Results and discussion
Corpus statistics

As its name suggests, LSD600 comprises 600 abstracts annotated with LSF–disease relations. Unlike most existing RE corpora, we do not limit the annotations to intra-sentence relations, with 16% of relations in LSD600 being cross-sentence. This should be compared to the <5% observed in several other biomedical RE corpora [26–30]. Despite the inherent difficulty of cross-sentence relation annotation, the final IAA F1-score was 82.1%, demonstrating that the annotation quality of LSD600 is high.

The 600 abstracts that make up the corpus were randomly partitioned on the document-level into a training set (60%), a development set (20%), and a held-out test set (20%). The corpus contains a total of 1900 manually annotated relations, which are distributed over eight relation types that are organized into a hierarchy. This makes the LSD600 corpus both larger and more comprehensive than existing corpora (Table 1). The breakdown of LSF–disease relations across corpus partitions and relation types is shown in Fig. 2. The left side of the figure shows our relation schema, emphasizing superclasses and subclasses. The bar chart on the right side shows the number of examples for each relation type in the corpus. We distributed the corpus into train:development:test sets in a 60:20:20 ratio, with each bar showing the breakdown for each relation in the different sets. By far, the most frequent relation type was "Positive Statistical Association" (32%), which reflects that the literature primarily focuses on lifestyle risk factors and that it is harder to demonstrate a causal relation.

Figure 3 shows the distribution of the eight relation types across high-level Disease Ontology [21] categories with at least 50 relations. Each bar represents one disease category, with colors indicating different relation types. For the category disease of cellular proliferation (primarily cancers), the corpus mainly contains "Positive Association" and "Causes" relations. This aligns with the existing knowledge, as many LSFs are risk factors for cancer. In contrast, disease of mental health has the highest fraction of "Treats" relations, which makes sense given that certain LSFs, such as exercise or social engagement, can play a positive role in mental health. Lastly,

Table 1. Comparison of LSD600 to existing corpora used for RE system validation

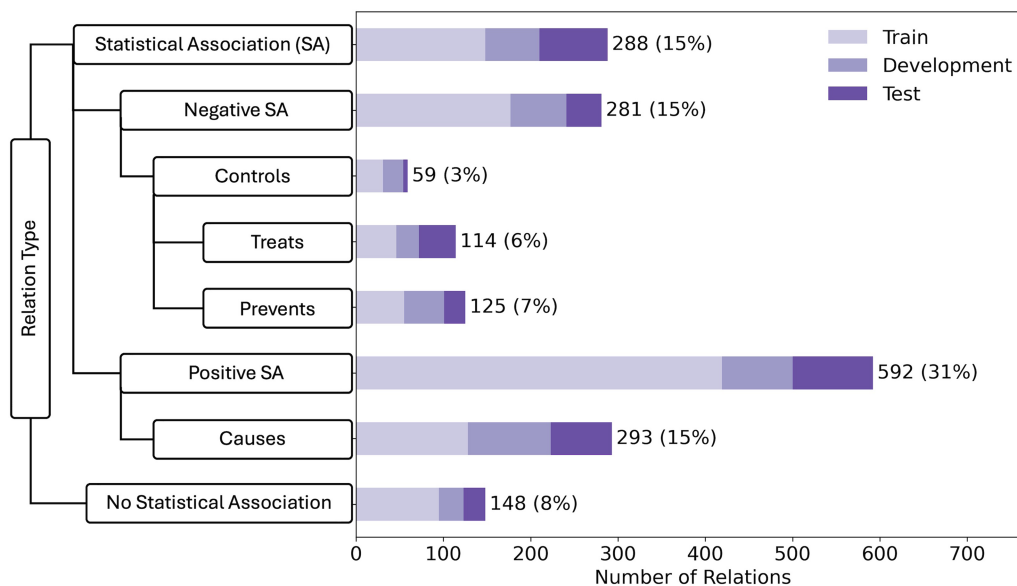| Corpus	LSF/Nutrient		Disease		Relations	Cross-sentence relations	Relation types
	Mentions	Unique	Mentions	Unique			
LSD600	6930	2281	5027	956	1900	311	8
ND	234	NA	278	NA	513	0	2
FoodDisease	608	343	608	314	464	0	2

Figure 2. LSD600 statistics for the eight distinct LSF–disease relation types.
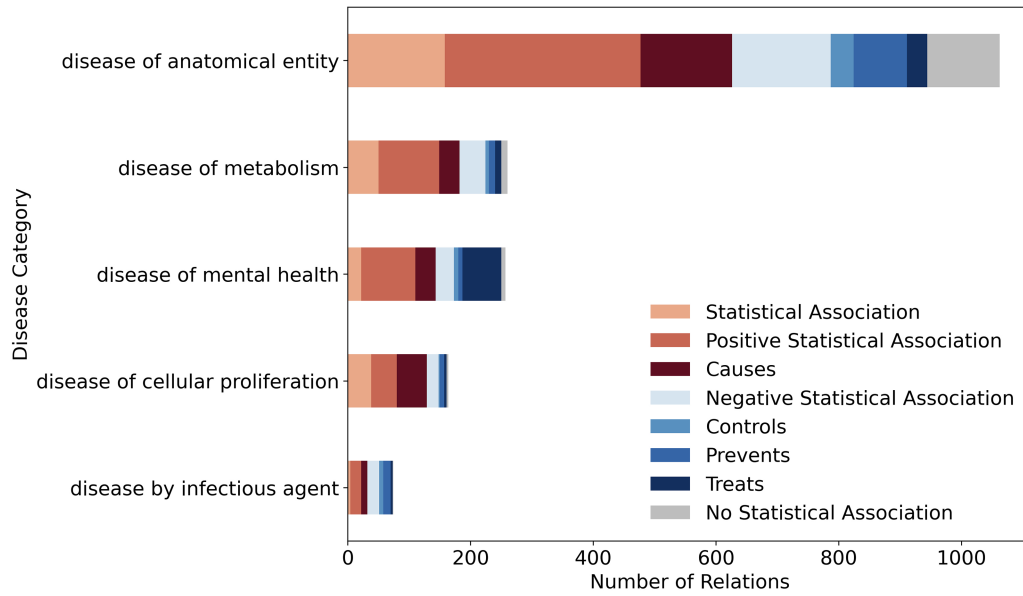
Figure 3. LSD600 statistics for eight distinct LSF–disease relation types per disease category.
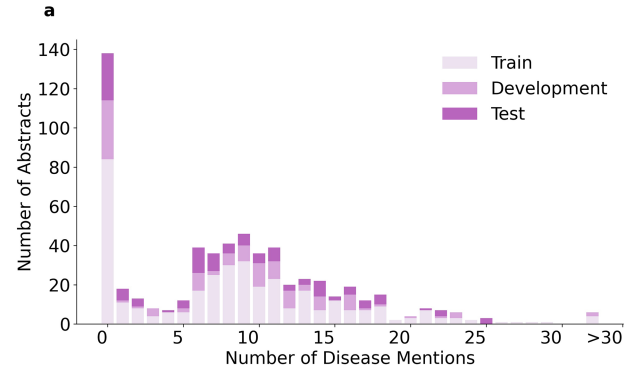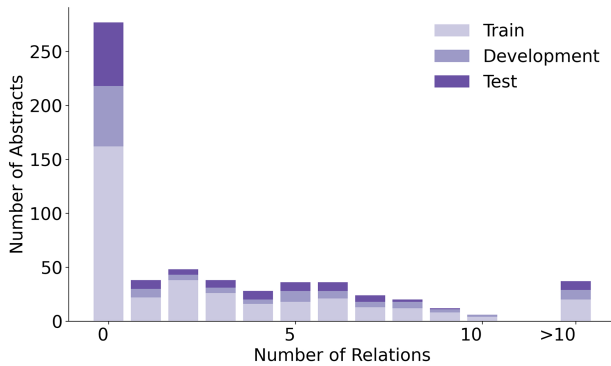

Figure 4. An overview of the distribution of relations in LSD600.

Figure 5. Distribution of disease and LSF mentions in LSD600 abstracts.

for disease by infectious agent, the emphasis is on prevention and control rather than direct causation, highlighting lifestyle's role in managing infections rather than acting as a primary cause.

The 1900 relations in LSD600 involve 5027 diseases and 6930 LSF entities. Figure 4 provides an overview of the distribution of relations per abstract, while Fig. 5 shows the distributions of disease and LSF mentions per abstract, with bar chart (a) showing how many abstracts contain how many disease mentions and bar chart (b) showing how many abstracts contain how many LSF mentions. Additional details are available in Supplementary Tables 2, 3, and 4.

In 276 abstracts, no relations were annotated. This outcome can be primarily attributed to two factors: First, our guidelines specify that we only annotate relations that provide factual information. Second, there are 138 abstracts with no annotated disease mentions. A total of 137 out of these 138 abstracts originate from the LSF200 NER corpus, which focused on LSFs and did not require any diseases to be mentioned. Only 29 abstracts have no LSF mentions. This too is primarily because of LSF200, in which documents were selected randomly from specific journals. However, a portion of abstracts without entity mentions come from the additional
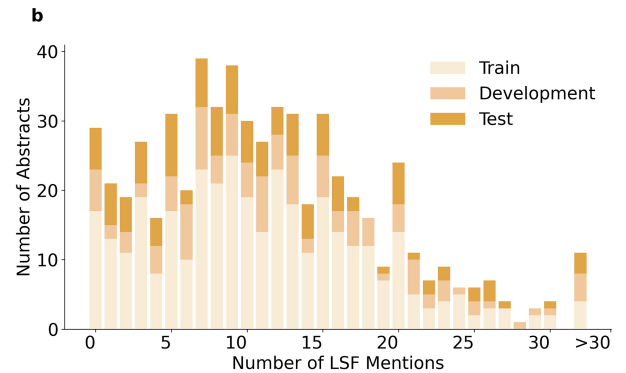
400 abstracts, not just LSF200. Despite requiring at least five LSF and five disease mentions pre-annotated by the Jensen-Lab Tagger, some abstracts have zero mentions after manual correction of false positives.

Relation extraction
System evaluation

We fine-tuned the RoBERTa-large-PM-M3-Voc-hf model on the development set using a grid search of the parameter
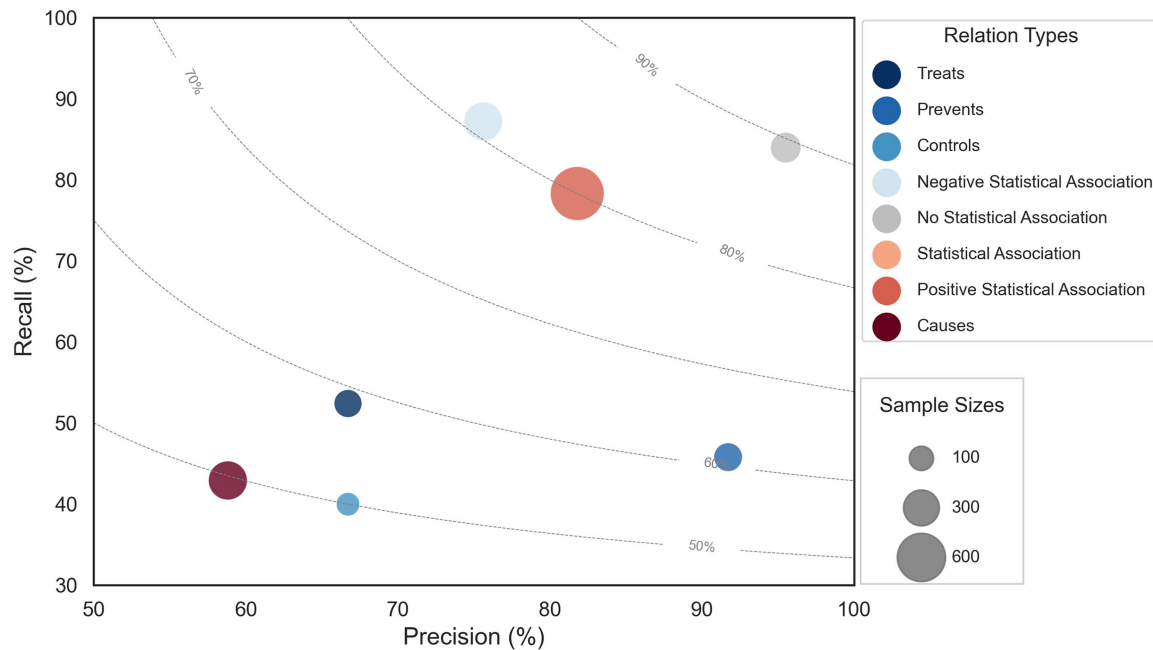
Figure 6. Performance plot for relation extraction across different relation types.

Table 2. Manual error analysis on the held-out test set

Error category	FN count	FP count	Total count	Total (%)
Cross-sentence relationship	61	0	61	26.9
Convoluted text excerpt	30	25	55	24.2
Model error	21	24	45	19.8
Co-reference resolution	16	8	24	10.6
Rare keyword	13	7	20	8.8
Annotation error	10	9	19	8.4
Ambiguous keyword	1	2	3	1.3
Total	152	75	227	100

combinations in Supplementary Table 5. After each epoch, the evaluation is checked, and only the best performing epoch is reported. The best hyperparameters were Maximum Sequence Length (MSL) = 180, Learning Rate (LR) = 5e-5, and Maximum number of training epochs = 75. This resulted in 75.5% precision, 65.2% recall, and 70.0% F-score on the development set. The best model was used for the final prediction on the held-out test set, yielding 77.3% precision, 61.6% recall, and 68.5% F-score. The performance on individual LSF–disease relation types is shown in Fig. 6. The marker size corresponds to the number of examples in the corpus and the dotted lines represent different F-score contours. The size of each circle represents the number of relations in the corpus, and the color corresponds to each of the eight different relation types.

Manual error analysis

We manually classified the errors made by the best model on the test set in seven categories (Table 2). The most prevalent category of errors is "Cross-sentence relations," which refers to any error that crosses sentence boundaries. The model fails to extract 61 of the 77 such relations in the test set, which accounts for more than a quarter of

all errors and 40% of the FNs. This partially explains the poor recall for "Causes" relations, as 40% of all such annotations in the test set span sentence boundaries. This problem is not primarily due to the maximum sequence length of the model—which only eight examples exceed—but is rather due to cross-sentence RE being inherently hard for current models [31, 32]. However, cross-sentence relations do not explain the difference in recall between "Statistical Association" and "Positive Statistical Association" (Fig. 6), which have 20% and 16% cross-sentence relations, respectively. This difference is instead explained by differences in the inherent complexity with which relations are described.

"Convoluted text excerpt" refers to sentences that are written in a way that makes them challenging to understand also for a human. "Co-reference resolution" errors are closely related to "convoluted text excerpt" errors, but refer specifically to cases where determiners (e.g. *this, its*) cause difficulty in identifying which entities are being discussed. "Ambiguous keywords" refers to sentences in which the model likely got confused by words/phrases that can have multiple meanings. For example, in the sentence "Dietary management is important for those with celiac disease," the word "important" does not give an unambiguous clue to the relation type. "Rare keyword" errors are caused by words/phrases that have a clear meaning but for which there were insufficient training examples for the model to learn them. These four error categories cause 29% of all annotated "Statistical Association" relations to be missed. This number is only 9% for "Positive Statistical Association," thus explaining the difference in recall between the two.

"Model errors" comprise wrong predictions that do not belong to any of the categories above. For example, the model could be asserting hypotheses as a fact or simply misinterpreting perfectly comprehensible phrases, while it also has the tendency to overestimate hedging expressions or inferred relations. This type of error is to a large extent responsible for the

poor precision observed, e.g. "Causes." However, this issue may not be as severe as it appears as FPs account for only about 30% of all errors.

Lastly, we found only 19 "annotation errors" where manual inspection revealed that the model's predictions are correct and the corpus annotations were wrong. Correcting these, and recalculating the performance metrics, resulted in a 2.1% improvement, increasing the F-score from 66.4% to 68.5%.

Label confusion analysis

Another way to understand the errors made by the model is to look whether different relation types get confused with each other, and if so, which ones get confused. The vast majority (82%) of FNs are not caused by label confusion, but are simply completely missed by the model. Almost half of these are cross-sentence relations. Similarly, 63% of the FPs are relations predicted by the model where there should be no relation.

Relation types being confused with each other accounts for 55 out of 227 (24%) of all errors (FP + FN). The relations more commonly confused are "Prevents" being predicted as "Negative Statistical Association," "Statistical Association" being predicted as "Causes," and "Causes" being predicted as "Positive Statistical Association."

Although accurate relation extraction at the most specific level of granularity is ideal, a RE system is already useful if it can identify whether a relation exists or not and whether it is positive or negative. At this level of granularity, only 8 of the 55 remain and the F-score on the test set improves by 4.9% from 68.5% to 73.4%. Supplementary Table 6 provides full statistics about label confusion by the model for all predicted relations in the test set.

Validation results on external corpora

To further validate the RE system, we extended our evaluation beyond the held-out test set from the LSD600 corpus to include the only publicly available corpora for lifestyle factors: the Nutrition and Disease (ND) corpus and the Food-Disease corpus (Table 1). Despite being limited to nutrition and not being perfectly aligned with our corpus in terms of relation types, these corpora provided a valuable benchmark for assessing the generalizability of our system.

To use our RE system to make predictions on these corpora, we mapped the relation types from LSD600 to the relation types in each of the two other corpora (see "Materials and methods" section for details). For the ND corpus, the RE system achieved 92.0% precision, 27.3% recall, and 42.2% F-score. To understand the very low recall, we performed an error analysis. Most of the FNs turned out to be relations that should not be annotated according to our guidelines. For example, the ND dataset included hypothetical statements and relations between diseases and metabolic biomarkers. When excluding the FNs that fall outside the scope of our annotation guidelines, the recall is 57.5% and the F-score is 70.7%.

For the FoodDisease corpus, the RE system achieved a precision of 94.3%, and a recall of 54.8%, resulting in an F-score of 69.3%. Similar to the ND corpus, an error analysis was performed to exclude false negatives due to

out-of-scope relations. This led to a 70.7% recall and 80.7% F-score.

The excellent precision achieved on both corpora indicates that the relations identified by our system as LSF–disease relations are perfectly accepted by the standards of these other corpora. These performance results are especially promising considering that we did not train our model on these external datasets but used the model trained on LSD600.

Conclusions

In this study, we present LSD600, the first of its kind, manually curated, LSF–disease corpus for relation extraction. We believe LSD600 will be an invaluable resource for the development and benchmarking of RE methods. In contrast to existing corpora, LSD600 encompasses all aspects of LSFs and annotates eight different relation types between these LSFs and diseases. Our annotations are also not limited to sentence boundaries in the text and include 16% cross-sentence relations. Additionally, we present an LSF–disease RE system trained using the LSD600 corpus, which achieved an F-score of 68.5% on the held-out test set. We further validated the model on external corpora, achieving promising performance. This RE system can be a valuable tool for extracting LSF–disease relations and has many downstream applications, given the crucial role of LSFs in disease onset, development, and management.

Supplementary data

Supplementary data is available at *Database* online.

Funding

Data availability

The LSD600 corpus and the fine-tuned transformer-based RE system are available under open licenses. The LSD600 corpus, including annotation guidelines, can be accessed on Zenodo at https://zenodo.org/records/13952449. Additionally, the Zenodo page includes two tables: one containing the 600 abstracts in the corpus with several metadata fields, and another consolidating 1900 manually annotated relations, along with the LSF and disease entity mentions for each relation. The implementation source code for the RE system is available on GitHub at https://github.com/EsmaeilNourani/LSF_Disease_RE.

References

1. Byrne S, Boyle T, Ahmed M *et al*. Lifestyle, genetic risk and incidence of cancer: a prospective cohort study of 13 cancer types. *Int J Epidemiol* 2023;**52**:817–26. https://doi.org/10.1093/ije/dyac238

2. Olsson T, Barcellos LF, Alfredsson L. Interactions between genetic, lifestyle and environmental risk factors for multiple sclerosis. *Nat Rev Neurol* 2017;**13**:25–36. https://doi.org/10.1038/nrneurol.2016.187

3. Sun Y, Yuan S, Chen X *et al*. The contribution of genetic risk and lifestyle factors in the development of adult-onset inflammatory bowel disease: a prospective cohort study. *Official J Am Coll Gastroenterol | ACG* 2023;**118**:511. https://doi.org/10.14309/ajg.0000000000002180

4. Said MA, van de Vegte YJ, Zafar MM *et al*. Contributions of interactions between lifestyle and genetics on coronary artery disease risk. *Curr Cardiol Rep* 2019;**21**:89. https://doi.org/10.1007/s11886-019-1177-x

5. Mutie PM, Giordano GN, Franks PW. Lifestyle precision medicine: the next generation in type 2 diabetes prevention?. *BMC Med* 2017;**15**:171. https://doi.org/10.1186/s12916-017-0938-x

6. Mansour S, Alkhaaldi SMI, Sammanasunathan AF *et al*. Precision nutrition unveiled: gene–nutrient interactions, microbiota dynamics, and lifestyle factors in obesity management. *Nutrients* 2024;**16**:581. https://doi.org/10.3390/nu16050581

7. Fu C, Pan X, Wu J *et al*. KG4NH: a comprehensive knowledge graph for question answering in dietary nutrition and human health. *IEEE J. Biomed. Health Inform* 2023;1–12. 10.1109/JBHI.2023.3338356

8. Dang LD, Phan UTP, Nguyen NTH. GENA: A knowledge graph for nutrition and mental health. *J Biomed Inform* 2023;**145**:104460. https://doi.org/10.1016/j.jbi.2023.104460

9. Cenikj G, Strojnik L, Angelski R *et al*. From language models to large-scale food and biomedical knowledge graphs. *Sci Rep* 2023;**13**:7815. https://doi.org/10.1038/s41598-023-34981-4

10. Grissa D, Junge A, Oprea TI *et al*. Diseases 2.0: a weekly updated database of disease–gene associations from text mining and data integration 2022;**2022**:baac019.

11. Kaszuba-Zwoińska J, Gremba J, Gałdzińska-Calik B *et al*. Electromagnetic field induced biological effects in humans. *Przegl Lek* 2015;**72**:636–41

12. Watson M, Holman DM, Maguire-Eisen M. Ultraviolet radiation exposure and its impact on skin cancer risk. *Semin Oncol Nurs* 2016;**32**:241–54. https://doi.org/10.1016/j.soncn.2016.05.005

13. Miłowska K, Grabowska K, Gabryelak T. Applications of electromagnetic radiation in medicine. *Postepy Hig Med Dosw* 2014;**68**:473–82. https://doi.org/10.5604/17322693.1101572

14. Devlin J, Chang M-W, Lee K *et al*. BERT: pre-training of deep bidirectional transformers for language understanding. 2018.

15. Liu Y, Ott M, Goyal N *et al*. (2019) RoBERTa: a robustly optimized BERT pretraining approach. arXiv preprint, arXiv ID: arXiv:1907.11692.

16. Yang X, Yu Z, Guo Y *et al*. Clinical relation extraction using transformer-based models 2021. arXiv preprint, arXiv ID: arXiv:2107.08957.

17. Su J, Wu Y, Ting H-F *et al*. RENET2: high-performance full-text gene–disease relation extraction with iterative training data expansion. *NAR Genomics Bioinform* 2021;**3**:lqab062. https://doi.org/10.1093/nargab/lqab062

18. Doughty E, Kertesz-Farkas A, Bodenreider O *et al*. Toward an automatic method for extracting cancer- and other disease-related point mutations from the biomedical literature. *Bioinformatics* 2011;**27**:408–15. https://doi.org/10.1093/bioinformatics/btq667

19. Li J, Sun Y, Johnson RJ *et al*. BioCreative V CDR Task Corpus: a resource for chemical disease relation extraction. *Database* 2016, baw068

20. Nourani E, Koutrouli M, Xie Y *et al*. Lifestyle factors in the biomedical literature: an ontology and comprehensive resources for named entity recognition. *Bioinformatics* 2024;**40**:btae613. https://doi.org/10.1093/bioinformatics/btae613

21. Schriml LM, Mitraka E, Munro J *et al*. Human disease Ontology 2018 update: classification, content and workflow expansion. *Nucleic Acids Res* 2019;**47**:D955–D962. https://doi.org/10.1093/nar/gky1032

22. Nastou KC, Nasi GI, Tsiolaki PL *et al*. AmyCo: the amyloidoses collection. *Amyloid* 2019;**26**:112–17. https://doi.org/10.1080/13506129.2019.1603143

23. Kim J-D, Ohta T, Pyysalo S *et al*. Overview of BioNLP'09 shared task on event extraction. In Tsujii J (ed.), *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*, 1–9. Boulder, Colorado: Association for Computational Linguistics, 2009.

24. Neves M, Ševa J. An extensive review of tools for manual annotation of documents. *Briefings Bioinform* 2021;**22**:146–63. https://doi.org/10.1093/bib/bbz130

25. Stenetorp P, Pyysalo S, Topić G *et al*. Brat: a Web-based Tool for NLP-Assisted Text Annotation. Segond F (ed.), In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 102–07. Avignon, France: Association for Computational Linguistics, 2012.

26. Mehryary F, Nastou K, Ohta T *et al*. STRING-ing together protein complexes: corpus and methods for extracting physical protein interactions from the biomedical literature. STRING-ing together protein complexes: corpus and methods for extracting physical protein interactions from the biomedical literature. *Bioinformatics* 2024;**40**:btae552

27. Nastou K, Mehryary F, Ohta T *et al*. RegulaTome: a corpus of typed, directed, and signed relations between biomedical entities in the scientific literature. RegulaTome: a corpus of typed, directed, and signed relations between biomedical entities in the scientific literature. *Database* 2024;**2024**:baae095. https://doi.org/10.1093/database/baae095

28. Björne J, Heimonen J, Ginter F *et al*. (2009) Extracting complex biological events with rich graph-based feature sets. *Proceedings of the Workshop on BioNLP Shared Task - BioNLP'09*, Association for Computational Linguistics, Boulder, Colorado, p. 10.

29. Miranda-Escalada A, Mehryary F, Luoma J *et al*. Overview of drugprot task at biocreative VII: data and methods for large-scale text mining and knowledge graph generation of heterogenous chemical–protein relations. *Database* 2023;**2023**:baad080. https://doi.org/10.1093/database/baad080

30. Mehryary F, Björne J, Salakoski T *et al*. *Potent pairing: ensemble of long short-term memory networks and support vector machine for chemical-protein relation extraction* **2018**;2018: bay120

31. Yao Y, Ye D, Li P *et al*. DocRED: a large-scale document-level relation extraction dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 764–77. Florence, Italy: Association for Computational Linguistics. 2019.

32. Nachtegael C, De Stefani J, Cnudde A *et al*. DUVEL: an active-learning annotated biomedical corpus for the recognition of oligogenic combinations. *Database* 2024;**2024**:baae039. https://doi.org/10.1093/database/baae039