Beyond associations: A benchmark Causal Relation Extraction Dataset (CRED) of disease-causing genes, its comparative evaluation, interpretation and application

Nency Bansal,[1,2] Sri Dhinesh R C,[3] Ayush Pathak[4] and Manikandan Narayanan[1,2,5,*]

[1]Department of Computer Science & Engineering, Indian Institute of Technology (IIT) Madras, Chennai, India, [2]Centre for Integrative Biology and Systems Medicine, IIT Madras, Chennai, India, [3]Department of Biotechnology, Alagappa College of Technology, Anna University, Chennai, India, [4]Department of Computer Science & Engineering, Indian Institute of Information Technology (IIIT) Vadodara, Gandhinagar, India and [5]Wadhwani School of Data Science and AI, IIT Madras, Chennai, India
[*]Corresponding author: nmanik@cse.iitm.ac.in

Abstract

Information on causal relationships is essential to many sciences (including biomedical science, where knowing if a gene-disease relation is causal vs. merely associative can lead to better treatments); and can foster research on causal side-information-based machine learning as well. Automatically extracting causal relations from large text corpora remains less explored though, despite much work on Relation Extraction (RE). The few existing CRE (Causal RE) studies are limited to extracting causality within a sentence or for a particular disease, mainly due to the lack of a diverse benchmark dataset. Here, we carefully curate a new CRE Dataset (CRED) of 3553 (causal and non-causal) gene-disease pairs, spanning 284 diseases and 500 genes, within or across sentences of 267 published abstracts. CRED is assembled in two phases to reduce class imbalance and its inter-annotator agreement is 89%. To assess CRED's utility in classifying causal vs. non-causal pairs, we compared multiple classifiers and found SVM to perform the best (F1 score 0.70). Both in terms of classifier performance and model interpretability (i.e., whether the model focuses importance/attention on words with causal connotations in abstracts), CRED outperformed a state-of-the-art RE dataset. To move from benchmarks to real-world settings, our CRED-trained classification model was applied on all PubMed abstracts on Parkinson's disease (PD). Genes predicted to be causal for PD by our model in at least 50 abstracts got validated in textbook sources. Besides these well-studied genes, our model revealed less-studied genes that could be explored further. Our systematically curated and evaluated CRED, and its associated classification model and CRED-wide gene-disease causality scores, thus offer concrete resources for advancing future research in CRE from biomedical literature.

Key words: Causal Relation Extraction (CRE), Annotation, Dataset, Biological Literature Mining, Gene-Disease Dataset, CRED, BioNLP (Biomedical Natural Language Processing)

Introduction

A knowledge of cause-effect relations is crucial in many fields, including biomedicine where the nature of relationship between entities like genes and diseases can influence treatment strategies (i.e., targeting genes that cause rather than respond/correlate to a disease can lead to more effective treatments); economic/social sciences where causal explanations are frequently sought for historical/forecasted events; machine-learning (ML) where side-information on causal vs. correlative features of a ML model can help improve its generalizability and interpretability [1], and to avoid misleading conclusions (such as asthma being suggested as a preventive measure against mortality due to the confounding factor of asthmatics receiving more intensive therapy [2]); and finally natural language processing (NLP) where causal knowledge can improve text summarization and question-answering tasks [3]. However, manually extracting causal relations from vast corpora (e.g., gene-disease relations from the biomedical corpora PubMed with approximately 35 million articles) is impractical; and advanced NLP methods, being increasingly used for automated extraction and summarization of information from literature, need to be extended to perform causal text mining.

Extracting useful information from published articles, specifically RE, is a mature and well-established subfield of NLP. In contrast, the area of CRE from text is in a nascent stage. Some pioneering works on CRE like Sjögren's Syndrome model [4] exist, but it is limited to extracting causal factors of only Sjögren's syndrome and can consider only a single sentence at a time. Challenges remain in extracting across-sentence causality and causal factors of multiple diseases, mainly due to the lack of a diverse benchmark dataset, and also in model interpretability that hinders detection of words that distinguish between causality vs. correlation. To address these challenges, this study (Figure 1) makes three contributions:

1. We have created CRED, a dataset of disease-causing genes extracted from published biomedical abstracts in PubMed [5]. It is a first-of-its-kind dataset as it includes both intra- and inter-sentence gene-disease (G-D) pairs, and spans a diverse set of 500 genes and 284 diseases.
2. To assess the utility of our CRED, we have trained or fine-tuned different state-of-the-art (SOTA) classifiers on CRED and found that the SVM model trained on BioBERT embeddings outperforms with an F1 score of 0.70 to distinguish causal from non-causal G-D pairs.
3. We interpret the predictions of our CRED-trained classification model, and compare it with model interpretation scores of a RE dataset trained model to verify if the CRED model actually focuses on the contexts or words with causal connotations (causal-connotation words) to make correct predictions. Our model indeed pays attention to such causal words and has learned the context well without focusing on the names of specific genes and diseases.
4. We also show applications of our model on real-world data. In the first application, we applied our CRED-trained model to all the abstracts related to Parkinson's disease and found that our model can recapitulate known PD-causing genes from abstracts. We observed that genes predicted to be causal in at least 50 papers were found to be already linked to PD in books [6], [7]. In the second application, we derived a CRED-wide gene-disease causality score.

The results of extensive experiments using our dataset show the promise of CRE from biomedical literature using the models evaluated in this work and pave the way for harnessing our CRED as a benchmark to develop and evaluate new methods for CRE.

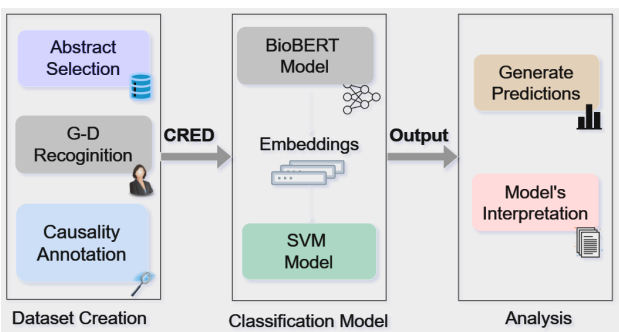

Fig. 1. Overview of our work

Related Work

This section summarizes existing datasets and NLP methods closely related to the CRE task. Since only a few CRE studies exist, many related works discussed here pertain actually to RE, with an emphasis on biomedical RE.

Existing Datasets

There are a few publicly available datasets pertaining to RE from the biomedical literature (see Table 1). A popular dataset for CDR (Chemical Disease Relation) extraction is the BC5CDR (BioCreative V CDR) corpus [12]; containing 1500 abstracts, split into three sets, 500 each for training, validation, and testing. Although BC5CDR is large enough to fine-tune a large language model and contains inter-sentence relations, it does not focus on extracting causal relations (since both mechanistic causal and biomarker correlative type relations between chemicals and diseases are present in this dataset, with no clear labeling of the type of each relation). Other RE datasets include GAD [10] containing gene-disease associations (with no specific mention of causal vs. non-causal associations), and EU-ADR corpus containing pairwise inter-relationships among drugs, disorders and targets, extracted and annotated from medical case reports (including suspected adverse reactions of drugs [11] which precludes statements of causality). BioRED [13] considers relations between various biological entities like chemical, gene, disease, and cell-line; a small fraction (2%) of its relations are causal gene-disease relations as per annotation guidelines (but authors label it as Positive/Negative Correlations, and call the dataset as RE in their paper). Moving to CRE, dataset like ADE (Adverse Drug Effects) [9] was created to extract drug-related adverse effects from medical reports. Although it is a causal relation extraction dataset, it only talks about the effects of the drugs from 1470 medical case reports. SemEval [8] corpus was built to extract causal relations from general natural language and is not specific to biomedical literature. Other biological datasets like Genome Wide Association Studies (GWAS) [14] and DisGeNET [15], talks about G-D association but none of them talk about causality. For CRE, we also considered Sjögren's Syndrome dataset [4]. As it is not publicly available, we requested the dataset from the authors and analysed it. We observed that the dataset considers all the factors that cause Sjögren's Syndrome only. However, it only deals with causal relations within the sentence, not across sentences. It is crucial to consider inter-sentence or within-paragraph relations to incorporate real-world scenarios and to extract more accurate and precise relations.

Existing Methods

Several model architectures have been proposed for information extraction from biomedical texts; notably BioBERT [16], trained on extensive biomedical literature and fine-tuned for Named Entity Recognition (NER), Relation Extraction (RE), and Question Answering (QA) tasks. Other models include SciBERT, [17], BlueBERT [18] and PubMedBERT [19], which were trained on a large corpus from PubMed, and Knowledge-based BERT (K-BERT) [20], limited to NER and QA. They use a graph convolutional neural network for capturing local and non-local dependencies, focusing on simple RE but not CRE. In BERT-GT [21], they have focused on extracting relations between chemicals and diseases from the CDR dataset, In CRE, the Sjögren's Syndrome work leverages causal relations expressed in text, using phrases like "leads-to" and "causes" but is limited to Sjögren's syndrome and CRE extraction within sentence. A survey paper [22] highlights the lack of a benchmark dataset for biomedical causal relation extraction (BCRE). For details, refer to Section A.1 of Appendix.

Table 1. Information on existing RE and CRE datasets. A subset of the total pairs (Tot Pairs) are positive relations (Pos Rel), with positive class referring to causal relationship for CRE and any (including correlative) relationship for RE datasets. Categorizing a dataset as RE vs. CRE is based on the cited publication describing the dataset. Drugs are also viewed as chemicals in EU-ADR and ADE datasets. Targets may also contain genes in EU-ADR dataset. Abbreviations: D - Disease, G - Gene, C - Chemical and V - Variant; # - "Number of" (a number that we couldn't find in the paper or website is marked as NA)

Dataset	**Entity Type**	**CRE/ RE**	**# Tot Pairs**	**# Pos Rel**	**Example text from dataset**
SemEval [8]	General Relations	CRE	10717	1331	*The burst has been caused by water hammer.*
ADE [9]	Drug→Effect	CRE	23516	6821	*The patient developed a rash after taking penicillin.*
Sjögren's Syndrome [4]	Factor→ Sjögren's Syndrome	CRE	383	383	*Hypokalemic paralysis is a rare presentation of Fanconi syndrome (FS) caused by Sjogren's Syndrome.*
GAD [10]	G-D	RE	5330	2756	*The patient had a mutation in the BRCA1 gene, which was associated with an increased risk of breast cancer.*
EU-ADR [11]	Drug-Dis, Target-Drug, Target-Dis	RE	355	261	*The patient was taking methotrexate for rheumatoid arthritis and developed pneumonitis.*
CDR [12]	C-D	RE	12850	3116	*Intravenous administration of lidocaine in a 67-year-old man resulted in profound depression of the sinoatrial nodal pacemakers.*
BioRED [13]	D-G, D-C, D-V, C-C, C-G, G-G, V-C, V-V	RE	NA	6503	*We therefore tested the hypothesis that variability in the HNF-6 gene is associated with subsets of Type II diabetes*

Methodology

This section describes the steps involved in creating our CRE dataset CRED, assessing its utility to perform classification-based causal text mining, and interpreting the classification model's predictions. When creating CRED, we assume a "disease-causative" gene to be defined as a gene that upon perturbation (including over-expression, knockout, or any of its mutations) changes the status or severity of the disease (e.g., leads to manifestation/activation, suppression, or altered severity of the disease).

CRED Creation

A main contribution of our work is in creating a CRED dataset of 5618 causal and non-causal G-D pairs (Table 2), including 1722 intra-sentence and 3896 inter-sentence G-D pairs, which are systematically extracted and annotated based on 267 abstracts in two phases as explained next.

Table 2. CRED Statistics. # represents the number.

Data	**# Abstracts**	**# G-D pairs**	
		Total	Causal
Train before aug	217	2476	303
Train after aug	217	4541	527
Validation	19	462	40
Test	31	615	54
Total before aug	267	3553	397
Total after aug	267	5618	621

Abstract Selection

In phase 1, to select abstracts, we start with a list of gene-disease pairs from DisGeNET repository [15] with a Gene-Disease Association (GDA) score of at least 0.5 for causal relation and some random gene-disease pairs for non-causal relations. Second, we extract abstracts mentioning the gene-disease pair from PubMed and restrict our attention to the top 10 abstracts (in terms of search relevance) for each gene-disease pair. Overall, we select 267 abstracts comprising a total of 3553 (including causal and non-causal) gene-disease pairs, spanning 284 diseases and 500 genes.

G-D Recoginition

After selecting the abstracts, we use PubTator [23; 24] to perform NER of gene and disease entities in the abstract and to group different representations of the same entity to a common ID called a MesH (Medical Subject Heading) ID for diseases and Entrez ID for genes. For example, the different representations of Parkinson's disease, such as PD, Parkinson, or Parkinson's, are all mapped to the same MesH ID D010300.

Causality Annotation

Task Overview: We manually read each selected abstract and annotated each gene-disease pair as causal (1) or non-causal (-1) based on the text. The major problem was that very few abstracts in biology talk precisely about disease-causing genes; associations are more commonly reported in abstracts than causal relations. For example, an abstract mentioning two genes and five diseases may report only one of these ten gene-disease pairs as causal; the other nine pairs are given the non-causal label in our dataset. Figure 2 shows the steps involved in annotating the dataset. Figure 3 shows an example illustrating the CRED creation steps.

Annotation guidelines: Annotation guidelines were formed under the guidance of a senior annotator. Following are the final annotation guidelines followed by the annotators:

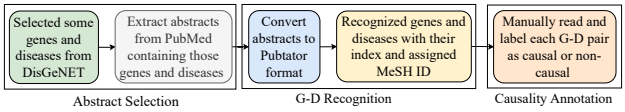

Fig. 2. Steps involved in CRED creation (Phase 1)

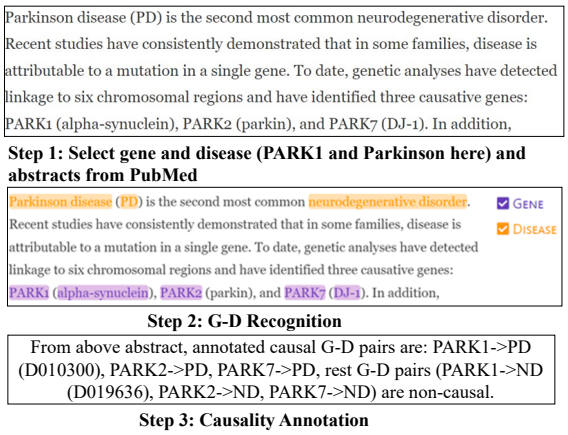

Fig. 3. Easy example (PMID: 15717024) to understand the steps of CRED creation. ND denotes neurodegenerative disorder. For complex example, refer Figure S1 in Appendix

1. The relation should be explicitly mentioned in the abstract and should not depend on pre-acquired knowledge of the disease or gene.
2. The annotations must be performed both within as well as across sentences. The context (the rest of the abstract) can be used for disambiguation.
3. Annotate only causal relationships between the gene and disease (Gene causing a disease). Other types of relationships should not be annotated.
4. Association between genes and disease should not be considered as causal relations. Risk factors and therapeutic targets are not considered causal factors. Causality must be mentioned specifically.
5. Mention of causal relation does not only mean the presence of words like causing, leads to, etc., but also considers the context of the abstract or some biomedical terms like upregulation, perturbation, etc.
6. The gene-disease pair should be labeled as causal (Label=1) if the abstract clearly states that there is a causal relation between genes and diseases and all other relations (association, co-occurrence, disease activating gene, gene activating other genes) are considered as non-causal (Label=-1).
7. Annotators have to check if the recognition of entities by PubTator is correct. If not, then correct the incorrect/unidentified genes and diseases.

Phase 2 of creation of dataset

In phase 2, we selected 1000 abstracts from PubMed (as explained above) and passed all the G-D pairs (10,283) through our trained model. Our model predicted 735 G-D pairs as causal. Then, both annotators annotated those 735 G-D pairs and found 191 causal G-D pairs, which were added to the training data to reduce the class imbalance. Validation data and test data are kept unchanged in phase 2. As in this phase, we only consider those G-D pairs, which were predicted as causal by our trained model, we discarded all the missing and erroneous entities of PubTator.

Annotators and Inter-Annotator Agreement (IAA):

Two independent in-house annotators (referred to as A1 and A2) provided annotations to curate the causal relationship in each article, following a double-annotation approach. We used the annotations of A1 to prepare the CRED dataset, and the annotations of A1 and A2 to evaluate IAA. The above-mentioned annotation guidelines were followed by both annotators to maintain consistency.

To assess the consistency of annotation among the annotators, we measured the agreement of annotations using Cohen's kappa [25]. The contingency matrix in Table 3 shows the number of agreements and disagreements. We achieved an inter-annotator agreement score of 0.893.

Table 3. Contingency matrix. A1 and A2 refer to two different annotators. (+) and (-) represents causal and non-causal classes respectively

	A1 (+)	**A1 (-)**
A2 (+)	353	43
A2 (-)	32	3709

Classification Model

To assess the utility of our CRED dataset, we want to use it to train a CRE model that can predict if a given G-D pair in an abstract is causal or not.

Data Augmentation:

We augment abstracts with causal relations using techniques like "SynonymAug" from the "nlpaug" library [26] to learn the model well on both positive and negative classes. Results before data augmentation are shown in Table S1 of the Appendix. This involved synonym replacement and spelling variations, ensuring the context remained unchanged, and was applied only to training data. An example of an original and augmented abstract is shown in Figure S2 of Appendix. Statistics of augmented data are mentioned in Table 2.

Data Cleaning and Preprocessing:

To enhance our model's generalizability across various diseases and genes, we use data-cleaning techniques to remove stopwords and substituted disease and gene names with placeholders (@DiseaseTgt$ and @GeneSrc$). This approach focuses the model on context and relationships rather than specific names to help reduce bias and improve performance on unseen diseases and genes.

Classifiers:

We select the SVM classifier based on the performance of different models mentioned in Table 4. Preprocessed and cleaned text is passed as input to BioBERT, and contextualized embeddings generated from BioBERT are passed as input to train and test the SVM classifier (BioBERT+SVM). We took the encoded (contextualized) embeddings of only the gene and disease mentions in the abstract, rather than [CLS] token (special token containing weighted average of all the tokens, generally used for classification tasks) embeddings,

concatenated them, and passed them as input to the SVM classifier.

Hence, for overall training of a classifier, the embeddings are of total dimension 4541×1536, where 4541 is the number of training data points, and 1536 = 2×768 is the concatenation of the 768-long embedding of the gene mention and that of the disease mention in the abstract. We have taken the average of embeddings in an abstract in case of multiple occurrences of the same gene and disease.

Implementation Details:

We train the proposed model using an Amazon Web Services Elastic Computing Cloud (AWS EC2) instance running 64bit Ubuntu operating system on a configuration of 1 NVIDIA GPU with 24 GB memory, 32 vCPUs (Virtual CPUs) with overall 128GB internal memory, and 220 GB externally attached memory. Specification details of all the packages used for this work are mentioned in Table S2 of the Appendix section. It took around 20-25 minutes for models like BioBERT with SVM to train and test on CRED.

Hyper-parameter tuning:

We tune hyper-parameters using Weights and biases [27] to obtain the best performance of the model. All the values of hyper-parameters tried, best values obtained, and results of different values of hyperparameters are mentioned in Figure S3, S4 of the Appendix. We have used the class weights parameter to handle class imbalance. After finalizing the values of all the hyper-parameters, we test the model on unseen test data presented in Section 4.

Model Interpretation

We interpret our model's predictions to analyse the reason behind the model's classification. We calculate the importance score of each input token corresponding to the output. To calculate the importance score of a token, first, we mask that token using "MASK" in the abstract and save the predicted probability score of that abstract, known as the perturbed probability (pert_prob). The importance score (imp_sc) is calculated using Equation 1 inspired by a similar score in [28]. Here, orig_prob refers to the predicted probability of that abstract without masking.

$$imp_sc = max(pert_prob - 0.5, orig_prob - 0.5, 0) \times (orig_prob - pert_prob) \quad (1)$$

The higher the imp_sc of a token, the more that token is responsible for the prediction; it thus reveals the words of the input sentence focused by the model for making the prediction. Negative importance score shows that the word supports making the wrong prediction (since we only calculate these scores for ground-truth causal G-D pairs).

Applications of CRED - Methodological Details

Predicting the causal genes of Parkinson's disease using all relevant papers in PubMed:

To apply our model to real-world data, we selected Parkinson's disease (PD). We searched published articles related to PD from PubMed [5] using the keyword "parkinson's", which were 1,49,945 (as of 11-07-2024). We extracted them in Pubtator format, recognizing all the genes in the abstracts (only 45526 abstracts contained genes). Then, we predict all extracted G-D pairs (1,11,419) (here, the disease is only Parkinson's), using our CRED-trained BioBERT+SVM model, to find all the causal genes of Parkinson's disease in the biomedical literature.

CRED-wide causality score of a G-D pair

To obtain a single causality score of a G-D pair across multiple annotated documents (PubMed abstracts) in CRED, we calculate the ratio of number of papers where the G-D pair is called causal to number of papers where G-D pair is mentioned in CRED.

Evaluation and Results

To assess the utility of our CRE dataset for causal text mining, both in absolute terms and also relative to a SOTA RE dataset CDR, we train and test multiple classifiers on CRED and CDR and present these results in this section. We report performance using precision, recall, and F1-score instead of accuracy to account for class imbalance. We also calculate the area under precision-recall curve (AUPRC) to analyze the performance of SOTA models. We divide our benchmark CRED into training, validation, and testing data, as shown in Table 2. Considering the moderate size of CRED, we apply four-fold cross-validation on training+validation data, in which three folds are used for training, and one fold is used for validation. The validation fold is used for hyper-parameter tuning, provided in Figure S3, S4 in the Appendix.

Comparison with SOTA Methods

Table 4 compares different SOTA methods on CRED and shows that SVM+BioBERT outperformed all other SOTA models. Kindly refer to Section A.2 of Appendix for a detailed explanation of the fine-tuning process and parameter calculation. Parameters are shown in Table S3 of Appendix. Please note that when we tried different BERT-based models for comparison, we fine-tuned existing models on the CRED training dataset and tested them on the same test dataset to enable a fair comparison. All the values of hyper-parameters tried, best values obtained, and results of different values of hyperparameters are mentioned in Figure S5, S6, S7, S8, S9 of the Appendix. In these comparisons, we observe that the BioBERT+SVM model has a good balance of precision and recall and hence gets the highest F1-score for the positive class as well as for the average of both classes. It also outperforms all other models in terms of AUPRC. This emphasizes that the BioBERT+SVM model performs well for both causal and non-causal classes.

Finally, to assess the utility of CRED, we want to check if the BC5CDR dataset, which contains some causal relations, could be used to train a CRE model. Table S4 shows that it is not possible, thereby supporting our efforts to create a CRE dataset. While results shown in Table S4 are evaluated on CRED, we also report out-of-domain results of the CRED-trained model tested on CDR data in Table S5 of the Appendix.

Experiments to assess model robustness

With BioBERT+SVM as the best-performing model, we next assessed the robustness of this model to different training datasets under a four-fold cross-validation framework. We indeed find the model to be robust; see Figure S10.

Another robustness check involved changing the input embeddings of our model, where we tried only CLS token

Table 4. Performance of different SOTA models on CRED. Pos represents the causal class, whereas averages show the average performance of both classes. The "clf" represents the classifier. "FT" represents fine-tuned. "RF" represents Random Forest. 'Pos' denotes the metrics for causal class and 'Avg' contains scores averaged on both classes. Prec, Rec, and F1 denote precision, recall, and F1 score, respectively.

Models / Perf Measures	Pos Prec	Pos Rec	Pos F1	Avg Prec	Avg Rec	Avg F1	AUPRC
Random clf as baseline	0.10	0.10	0.10	0.50	0.50	0.50	0.10
FT BioBERT	0.49	0.33	0.40	0.71	0.65	0.67	0.42
FT PubMedBERT	**0.75**	0.22	0.34	**0.83**	0.60	0.65	0.42
FT SciBERT	0.39	**0.50**	0.43	0.67	**0.71**	0.68	0.46
FT BlueBERT	0.41	0.24	0.30	0.67	0.60	0.62	0.36
BioBERT with XGBoost	0.32	0.39	0.35	0.63	0.65	0.64	0.27
BioBERT with RF	0.67	0.22	0.33	0.80	0.61	0.65	0.36
BioBERT with SVM	0.61	0.35	**0.45**	0.78	0.67	**0.70**	**0.47**

embeddings, only G-D embeddings and CLS with G-D embeddings. The performance dipped with CLS and G-D embeddings and was poor with only CLS embeddings, likely because CLS token's embedding contains the weighted average of the complete abstract, which might not be sufficient to distinguish between causal and non-causal G-D pairs in the same abstract.

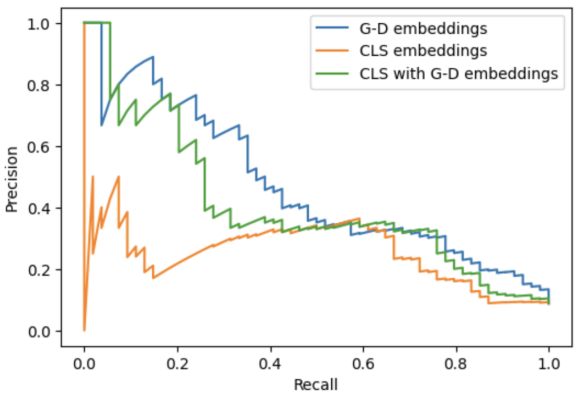

Fig. 4. BioBERT+SVM performance when trained on CLS embeddings, CLS with G-D embeddings and only G-D embeddings.

Interpreting the model's predictions with the help of CRED-trained model

To establish the difference between CRE-trained and RE-trained models and to analyse our model's predictions, we interpret the models' predictions and analyse those words the model focuses on to make the predictions. We observe that the CRED-trained model assigns higher importance scores to causal-connotation words, indicating their greater significance in predicting positive outcomes, compared to the CDR-trained model.

To show the importance scores of causal-connotation words vs associate-related words by CRED-trained model, we selected synonyms of the words "cause" and "associate" (mentioned in section A.3 of the Appendix) and plotted the maximum importance score of the matching words in Figure 5. We can observe that the importance scores of causal-connotation words are higher than associate-related words, which illustrates that our model focuses more on causal-connotation words. We plotted a similar graph for the CDR-trained model in Figure S11 in the Appendix, showing that the CDR-trained model focuses more on associate-related words.

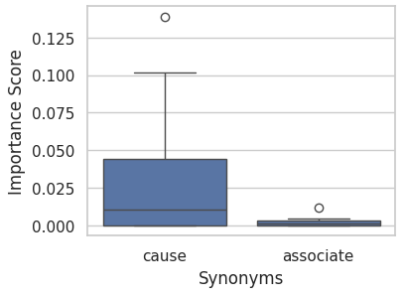

Fig. 5. Comparison of importance scores of "cause" vs. "associate" synonyms for CRED-trained model.

To dive deeper into importance scores, we sort the words based on importance scores and take the top 25 words from CRED-trained and CDR-trained models separately. We extract their importance scores on CRED-trained and CDR-trained models and plot them. Figure 6 shows the comparison of importance scores of CRED-trained and CDR-trained models for the top 25 words of CRED-trained model and Figure S12 in the Appendix for CDR-trained model. We can observe that the top 25 important words of the CRED-trained model contain words like "cause," "caused," etc. Although not all 25 words carry a causal connotation, due to contextualized embeddings that consider the entire context, it still emphasizes certain words with causal implications.

Figure S1 shows an example abstract from our dataset. We constructed the word cloud of this abstract after training the model on the CRED and CDR datasets separately. In Figure 7, LHS and RHS show the word cloud made from the importance score of the CRED-trained and CDR-trained models, respectively.

We can see that, on the left word cloud of Figure 7, the "caused" word is larger in size, representing that the CRED-trained model has given it higher importance. Using prediction interpretation, we can establish the difference between CRE vs RE dataset-trained models. Models trained on CRED emphasize sentence parts that establish causal relations, unlike those trained on datasets like CDR, which may not necessarily prioritize causal-connotation words. Moreover, as we can see, there are no gene names and disease names present in the word cloud; hence, we can verify that our model is not learning specifically about a disease or a gene

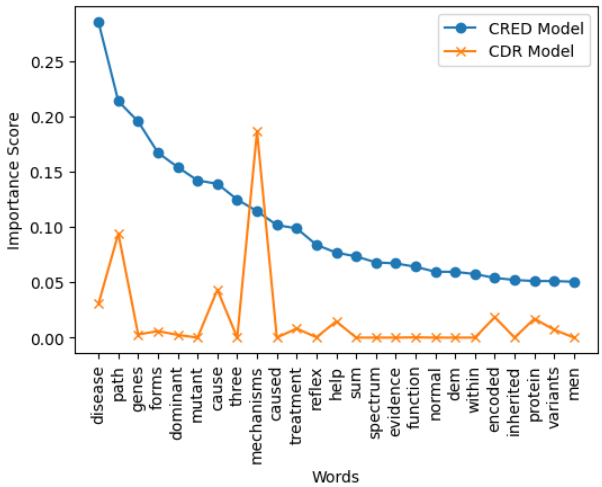

Fig. 6. Comparison of importance scores of top 25 words (from CRED-trained model) in CRED-trained and CDR-trained models.

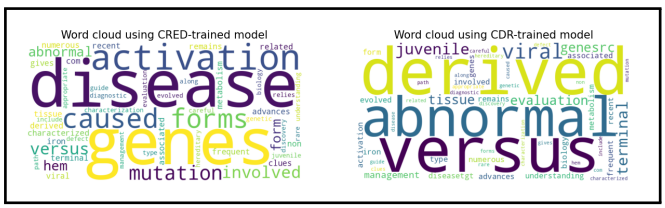

Fig. 7. Example of a word cloud of abstract shown in Figure S1 (after augmentation) based on the importance score of the CRED- and CDR-trained models. Larger word size refers to higher importance-score words focused by the model.

and learning about the context between them. In the right plot of Figure 7, we can observe that it does not capture causal-connotation words like "caused" and instead focuses on words like "chronic", "genes", etc., which does not convey any information on the causality between entities, thereby indicating that the model has not learned about causal relations. These results together demonstrate the utility of CRED over RE datasets in helping models extract causal relations. Note that these model interpretation results are not specific to the BioBERT+SVM model; with other CRED-trained models like BioBERT+XGBoost also, the results are broadly similar (see Figure S13 and S14 in Appendix).

Applications of CRED

CRED-trained model helps predict Parkinson's disease (PD) genes from PubMed abstracts

To test our CRED-trained BioBERT+SVM model in a real-world setting, we applied our model to all abstracts in PubMed related to Parkinson's Disease (PD). We downloaded 1,49,945 such abstracts (in PubTator format as of 11th August, 2024), out of which 45,526 abstracts mentioned at least one gene. Causal predictions for the extracted G-D pairs in these abstracts (1,11,419 G-D pairs, with disease being PD) is shown in Figure 8 (see also Figure S15 in the Appendix, where an enlarged view is shown after removing genes predicted to be causal in more than 25 papers). We see that many genes (10,919) are less studied, with fewer than 100 papers available to establish their causal relation to disease. However, of the remaining 111 genes that are present in at least 100 papers and predicted to be causal in at least 50 papers were found to be already linked to PD in books [6], [7], which were not part of CRED used to train our model. Our CRED-trained classifier can thus recapitulate known PD-causing genes from abstracts not present in CRED and quantify the strength of evidence of the causal relation of these and 11015 other genes to a disease like PD.

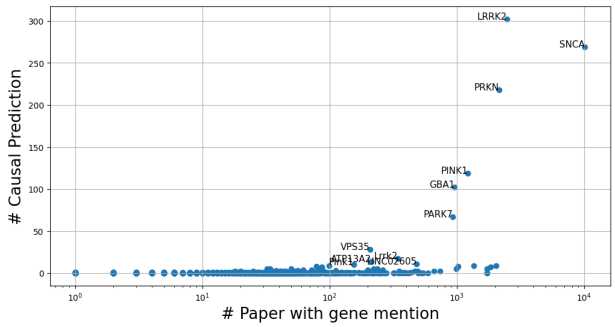

Fig. 8. Application of our model to predict PD-causing genes. The plot shows the number of PubMed abstracts in which the gene and PD are co-mentioned vs. the number of abstracts in which the gene is predicted as causal for PD by our model.

CRED-wide causality score for each G-D pair

As a G-D pair can be mentioned in multiple PubMed abstracts, and each abstract may support the pair as either causal or non-causal, we have a key problem of deriving a dataset-wide causality score of a G-D pair. We take the ratio of number of papers where a G-D pair is called causal to number of all papers where G-D pair is mentioned from all documents (PubMed abstracts) annotated in CRED to derive a single causality score for each G-D pair.

Figure 9 shows the gene-disease network generated from the top 20 G-D pairs based on the causality score from CRED. This figure reveals genes like *FMR1*, which causes multiple diseases like Tremors, Ataxia, etc., with high causality score. We expand this graph in Figure 10 to include top 250 G-D pairs. In this graph, we can observe more causal genes that are common across diseases. For instance, *SCN1A* is causal for Epilepsy, Autism Spectrum Disorder and Migraine with Aura. CRED can thus be helpful in determining the causal genes of not only a particular disease but also common genes that cause different diseases. Such genes can aid target drug discovery efforts pertaining to multiple diseases.

Conclusion

This work proposes CRED, an intra/inter-sentence causal relation extraction (CRE) dataset between genes and diseases from biomedical literature. This work is motivated by the scarcity of available annotated CRE datasets, which have hindered the development of a biomedical CRE model. Our benchmark CRED covers 267 abstracts and 5618 gene-disease pairs (after augmentation), out of which 621 are causal. To demonstrate the utility of our dataset, we train an SVM model with BioBERT embeddings of CRED. We test other SOTA methods on our dataset to establish the unique value our CRED

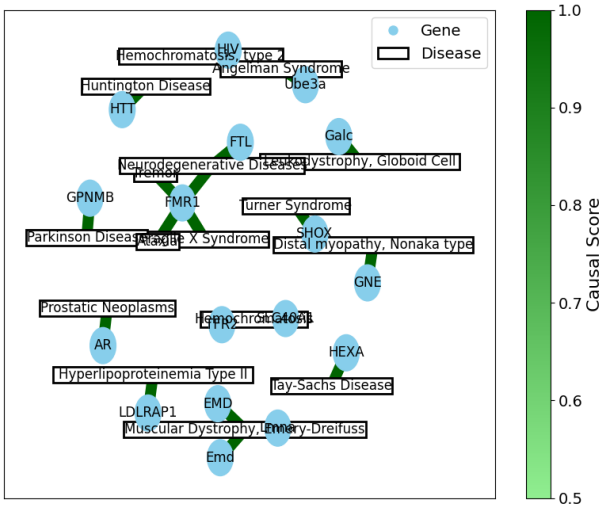

Fig. 9. Graph of top 20 causal G-D pairs; edges thickness and color shown are proportional to the CRED-wide causality score.

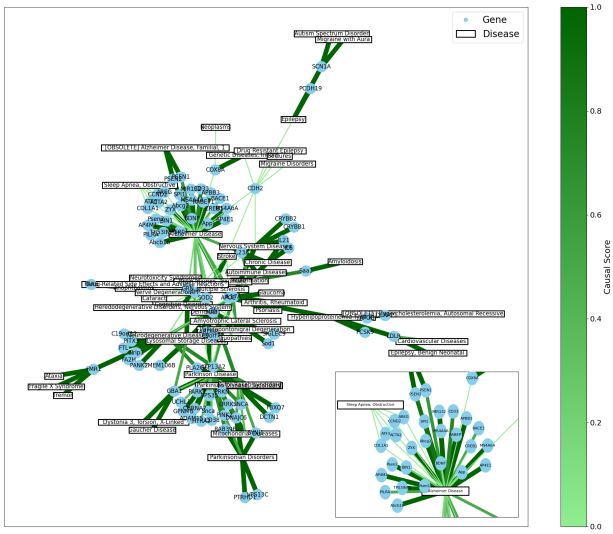

Fig. 10. Graph of top 250 causal G-D pairs; edges thickness and color shown are proportional to the CRED-wide causality score.

dataset brings to the research community. We also show the interpretations of the predictions of the model to show that the model is focusing on words with causal connotations for making the predictions. This also shows that our dataset is not biased toward a few keywords, disease names, or gene names. We also apply our model to real-world data. We applied it to all the abstracts on Parkinson's disease and observed that genes predicted to be causal in at least 50 papers were found to be already linked to PD in books. In another application, we derived a CRED-wide gene-disease causality score. It can be used to establish the causality of a G-D pair with multiple contradicting documents. Our CRE Dataset is a novel causal benchmark dataset that can pave the way for further research on language models for causal relation extraction.

Competing interests

No competing interest is declared.

Author contributions statement

NB and MN conceived and formulated this study, including its methodology, analyses and applications. NB performed all key analyses in the manuscript and interpretation of the results with inputs from MN. NB and SDRC annotated the PubMed abstracts in the CRED dataset. AP conducted analyses related to applications of CRED. NB and MN wrote and reviewed the manuscript. MN guided and supervised the study.

Acknowledgments

The authors thank members of the Bioinformatics and Integrative Data Science (BIRDS) group at IIT Madras for their valuable feedback during presentations of this work. We would also like to thank Adwait Parsodkar and Ananya Sai for their feedback on the manuscript. This work is supported by the Wellcome Trust/DBT grant IA/I/17/2/503323 awarded to MN.

References

1. B. Schölkopf, *Probabilistic and Causal Inference: The Works of Judea Pearl*, ch. Causality for Machine Learning, p. 765–804. New York, USA: Association for Computing Machinery, 2022.
2. B. Glocker, M. Musolesi, J. Richens, and C. Uhler, "Causality in digital medicine," *Nature Communications*, vol. 12, no. 1, 2021.
3. B. Drury, H. G. Oliveira, and A. de Andrade Lopes, "A survey of the extraction and applications of causal relations," *Natural Language Engineering*, vol. 28, no. 3, pp. 361–400, 2022.
4. P. D. Gujarathi, S. K. R. G. Reddy, V. M. B. Karri, A. R. Bhimireddy, A. S. Rajapuri, M. Reddy, M. Sabbani, B. Cheriyan, J. VanSchaik, T. Thyvalikakath, and S. Chakraborty, "Note: Using causality to mine Sjögren's Syndrome related factors from medical literature," in *COMPASS '22: ACM 2022* (E. Bainomugisha, W. Brunette, N. Dell, and A. Sharma, eds.), pp. 674–681, ACM, 2022.
5. PubMed, "PubMed website." `https://pubmed.ncbi.nlm.nih.gov/`, 2023. Accessed: 2023-02-10.
6. S. A. Factor and W. J. Weiner, *Parkinson's disease: diagnosis and clinical management*. Demos Medical Publishing, 2 ed., 2007. ISBN: 978-1-933864-00-6.
7. S. Lesage and J. Trinh, eds., *Parkinson's Disease: Genetics and Pathogenesis*. MDPI Books, 2023. ISBN: 978-3-0365-8881-0.
8. I. Hendrickx, S. N. Kim, Z. Kozareva, P. Nakov, D. Ó. Séaghdha, S. Padó, M. Pennacchiotti, L. Romano, and S. Szpakowicz, "Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals," in *Proceedings of the 5th International Workshop on Semantic Evaluation, SemEval@ACL 2010* (K. Erk and C. Strapparava, eds.), pp. 33–38, The Association for Computer Linguistics, 2010.
9. H. Gurulingappa, A. M. Rajput, A. Roberts, J. Fluck, M. Hofmann-Apitius, and L. Toldo, "Development of a

benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports," *Journal of Biomedical Informatics*, vol. 45, no. 5, pp. 885–892, 2012.

10. À. Bravo, J. P. González, N. Queralt-Rosinach, M. Rautschka, and L. I. Furlong, "Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research," *BMC Bioinformatics*, vol. 16, pp. 55:1–55:17, 2015.
11. E. M. van Mulligen, A. Fourrier-Réglat, D. Gurwitz, M. Molokhia, A. Nieto, G. Trifirò, J. A. Kors, and L. I. Furlong, "The EU-ADR corpus: Annotated drugs, diseases, targets, and their relationships," *Journal of Biomedical Informatics*, vol. 45, no. 5, pp. 879–884, 2012.
12. J. Li, Y. Sun, R. J. Johnson, D. Sciaky, C. Wei, R. Leaman, A. P. Davis, C. J. Mattingly, T. C. Wiegers, and Z. Lu, "Biocreative V CDR task corpus: a resource for chemical disease relation extraction," *Journal of Biological Databases and Curation*, vol. 2016, 2016.
13. L. Luo, P.-T. Lai, C.-H. Wei, C. N. Arighi, and Z. Lu, "BioRED: a rich biomedical relation extraction dataset," *Briefings in Bioinformatics*, vol. 23, no. 5, p. bbac282, 2022.
14. GWAS, "GWAS website." `https://www.genome.gov/genetics-glossary/Genome-Wide-Association-Studies/`, 2023. Accessed: 2023-01-10.
15. DisGeNET, "DisGeNET website." `https://www.disgenet.org/search/`, 2023. Accessed: 2023-01-25.
16. J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.
17. I. Beltagy, K. Lo, and A. Cohan, "Scibert: A pretrained language model for scientific text," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP Hong Kong, China, 2019* (K. Inui, J. Jiang, V. Ng, and X. Wan, eds.), pp. 3613–3618, Association for Computational Linguistics, 2019.
18. Y. Peng, S. Yan, and Z. Lu, "Transfer learning in biomedical natural language processing: an evaluation of bert and elmo on ten benchmarking datasets," *arXiv preprint arXiv:1906.05474*, 2019.
19. Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, "Domain-specific language model pretraining for biomedical natural language processing," *ACM Tansactions on Computing for Healthcare*, vol. 3, no. 1, pp. 2:1–2:23, 2022.
20. W. Liu, P. Zhou, Z. Zhao, Z. Wang, Q. Ju, H. Deng, and P. Wang, "K-BERT: Enabling language representation with knowledge graph," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 03, pp. 2901–2908, 2020.
21. P. Lai and Z. Lu, "BERT-GT: cross-sentence n-ary relation extraction with BERT and graph transformer," *Bioinformatics*, vol. 36, no. 24, pp. 5678–5685, 2021.
22. A. Akkasi and M. Moens, "Causal relationship extraction from biomedical text using deep neural models: A comprehensive survey," *Journal of Biomedical Informatics*, vol. 119, p. 103820, 2021.
23. PubTator, "PubTator website." `https://www.ncbi.nlm.nih.gov/research/pubtator/`. Accessed: 2023-02-15.
24. C.-H. Wei, H.-Y. Kao, and Z. Lu, "PubTator: a web-based text mining tool for assisting biocuration," *Nucleic Acids Research*, vol. 41, no. W1, pp. W518–W522, 2013.
25. P. K. Bhowmick, A. Basu, and P. Mitra, "An agreement measure for determining inter-annotator reliability of human judgements on affective text," in *Coling 2008: Proceedings of the workshop on Human Judgements in Computational Linguistics* (R. Artstein, G. Boleda, F. Keller, and S. Schulte im Walde, eds.), (Manchester, UK), pp. 58–65, Coling 2008 Organizing Committee, 2008.
26. E. Ma, "NLP Augmentation." `https://github.com/makcedward/nlpaug`. Accessed 24-01-2024.
27. WandB, "Weights & Biases." `https://wandb.ai/site`. Accessed: 15-09-2023.
28. B. Alipanahi, A. Delong, M. T. Weirauch, and B. J. Frey, "Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning," *Nature Biotechnology*, vol. 33, no. 8, pp. 831–838, 2015.
29. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (J. Burstein, C. Doran, and T. Solorio, eds.), (Minneapolis, USA), pp. 4171–4186, Association for Computational Linguistics, 2019.
30. S. K. Sahu, F. Christopoulou, M. Miwa, and S. Ananiadou, "Inter-sentence relation extraction with document-level graph convolutional neural network," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (A. Korhonen, D. Traum, and L. Màrquez, eds.), (Florence, Italy), pp. 4309–4316, Association for Computational Linguistics, 2019.

Appendix A.1: Existing methods

Many model architectures have been proposed for information extraction from text data. One of the State-of-the-art models, BioBERT [16], has been trained using the BERT model [29] architecture, with a large amount of biomedical literature data and fine-tuned for three downstream tasks, namely, Named Entity Recognition (NER), Relation Extraction (RE), and Question Answering (QA). Several other models like SciBERT [17] and PubMedBERT [19] have been proposed for information extraction tasks. PubMedBERT, also based on BERT, was trained on a large corpus of biomedical research papers, including abstracts from PubMed and full-text articles from PubMed Central. This corpus contains over 100 million words, making it one of the largest pre-training corpora for biomedical text. Knowledge-based BERT (K-BERT) [20] has also been proposed, which takes knowledge bases for prediction. They have limited it to NER and QA. Another model, like the inter-sentence relation extraction model [30], also focuses on the inter-sentence relation extraction task. They have built a labeled edge graph convolutional neural network model on a document-level graph. The graph is constructed using various inter and intra-sentence dependencies to capture local and non-local dependency information. Although they achieved good performance for simple relation extraction, they did not discuss causal relation extraction. In CRE, we have Sjögren's Syndrome [4] work. In this work, the authors present a model that leverages causal relations expressed in the text to understand factors related to Sjögren's syndrome. The authors use specific word phrases like "leads-to" and "causes" to extract causal relationships from text related to Sjögren's syndrome. Although they have worked on CRE, their work is limited to only one disease, Sjögren's syndrome, and they extracted CR only within the sentence. Moreover, we have read the Survey paper [22]. The paper provides a comprehensive literature review of the causal relation extraction problem tailored to the biomedical domain, describing datasets used for CRE tasks. The authors also mention that no benchmark dataset is available for biomedical causal relation extraction (BCRE). Some works talk about causality but have concerning limitations due to a lack of data.

Appendix A.2: Fine-tuning process and parameter calculation

Fine-tuning BioBERT: For fine-tuning BioBERT classifier on the whole abstract embeddings (of length 512 × 768), we fine-tune only the parameters of a single hidden-layer neural network classifier that comprises 512 hidden neurons, each of which is connected to 768 neurons of BioBERT's output embedding layer. Embedding of each token (768 dimensions) is reduced to 1 dimension using average pooling, and 512 elements are passed to the output layer. The output from the hidden layer (512 elements) is transformed into the one final output (for binary classification). All parameters of the base BioBERT model are frozen. The total number of parameters (weights and biases) in this fine-tuned neural network classifier stacked at the end of the base BioBERT model can then be calculated as follows:
From the input to the hidden layer: 768 (input features) × 512 (neurons) weights + 512 biases = 393,728 parameters.
From the hidden to the output layer (a sigmoid function): 512 (neurons in hidden layer) weights + 1 bias = 513 parameters.
Total parameters = 393,728 (hidden layer) + 513 (output layer) = 394,241 parameters.
So, the total parameters of Fine-tuned BioBERT classifier is 110 M + 394,241, out of which we trained 394,241 parameters during the fine-tuning process.

BioBERT + SVM: Total number of parameters of BioBERT+SVM is 110 M (base BioBERT) + 4541 (total 4541 training data points leading to 4541 parameters in the Dual SVM problem, with 711 of them being non-zero and corresponding to support vectors). We have reported the difference in the number of parameters of both models in Table S3 in the Appendix.

Appendix A.3: Synonyms

Synonyms of cause used are:'causation', 'do', 'inducer', 'induction', 'make', 'inducement', 'stimulate', 'stimulus', 'inducive', 'get', 'causative', 'have', 'cause', 'induce', 'stimulation', 'inducing', 'causation', 'impact', 'developement', 'get', 'make', 'case', 'movement', 'induce', 'stimulus', 'causal', 'crusade', 'do', 'inducer', 'inducement', 'causal-agent', 'campaign', 'causative', 'grounds', 'reason', 'stimulation', 'drive', 'stimulate', 'cause', 'induction', 'causa', 'inducive', 'effort', 'suit', 'inducing', 'caused', 'causes'.

Synonyms of associate used are: associate','associations', 'connective', 'associatory', 'fellow', 'tie-in', 'connexion', 'associate-degree', 'associable', 'familiar', 'affiliate', 'companion', 'fellowship', 'link', 'linked', 'support', 'related', 'consort', 'colligation', 'relation', 'association', 'comrade', 'connect', 'associative', 'linkage', 'connection', 'link-up', 'correlate', 'correlative', 'correlativity', 'correlated', 'correlation'

Appendix B: Supplementary Figures

The management and understanding of hereditary hemochromatosis have evolved with recent advances in iron biology and the associated discovery of numerous genes involved in iron metabolism. HFE-related (type 1) hemochromatosis remains the most frequent form, characterized by C282Y mutation homozygosity. Rare forms of hereditary hemochromatosis include type 2 (A and B, juvenile hemochromatosis caused by HJV and HAMP mutation), type 3 (related to TFR2 mutation), and type 4 (A and B, ferroportin disease). The diagnostic evaluation relies on comprehension of the involved pathophysiologic defect, and careful characterization of the phenotype, which gives clues to guide appropriate genetic testing.

Fig. S1. Example of an abstract (PMID: 25064704) from our dataset. In this abstract, annotated causal G-D pairs are: HJV→Juvenile hemochromatosis, HAMP→Juvenile hemochromatosis. All other G-D pairs are annotated as non-causal.

Original abstract:
Genetic studies have led to major discoveries in the pathogenesis of various neurodegenerative diseases. Ubiquitin-positive familial frontotemporal dementia was recently found to be caused by mutations in the Progranulin gene (PGRN), and the major constituent of the inclusions, TDP-43, was subsequently identified.

Augmented abstract:
genetic studies have shown the important discoveries in the pathogenesis of different neurodegenerative diseases. ubiquitin - positive familial frontotemporal dementia was discovered in be caused by mutations in the progranulin gene (pgrn), where main constituent of the inclusions, tdp - 43, was subsequently recognized.

Fig. S2. Example showing original and its augmented abstract (PMID:18322368) from CRED

Hyper Para	Values	Best
Kernel	RBF, Linear, Poly	Poly
C	0.1, 1, 10	10
Degree	6, 10, 20, 30	10
Gamma	Scale, Auto	Scale

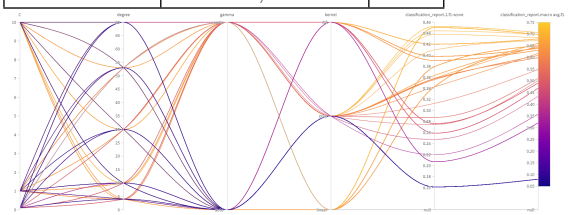

Fig. S3. Hyper parameters tuning for SVM

Hyper Parameters	Values	Best
Learning Rate	0.01, 0.1, 0.2	0.1
Max Depth	5, 10, 20, 40	40
No. of estimators	100, 200, 400	400
Gamma	0.1, 0.2, 0.3	0.1
Scale pos weight	100, 200, 500	500

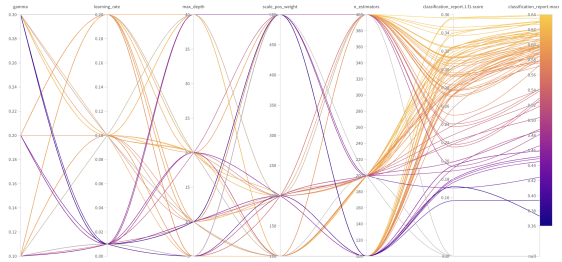

Fig. S4. Hyper parameters tuning for XG Boost

Hyper Parameters	Values	Best
No. of jobs	2,5,10	5
Max Depth	10, 50 , 100	10
No. of estimators	100, 200, 400	100
Max leaf nodes	100, 200, 500	100
Max features	Auto, Sqrt	Sqrt

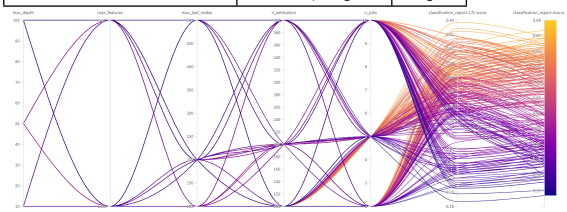

Fig. S5. Hyper parameters tuning for Random Forest

Hyper Para	Values	Best
Learning Rate	2e-5, 3e-5, 5e-5	3e-5
No. of train epochs	1, 2, 3, 4	4
Weight decay	0.01, 0.05, 0.1	0.05
Warmup Steps	0, 500, 1000	1000

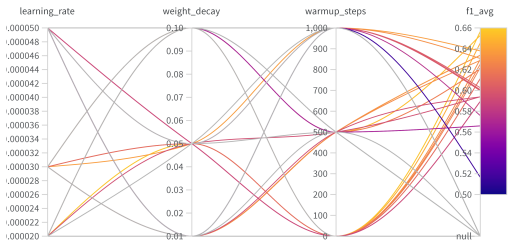

Fig. S6. Hyper parameters tuning for BioBERT

Hyper Para	Values	Best
Learning Rate	'2e-5, 3e-5, 5e-5	2e-5
No. of train epochs	1, 2, 3, 4	3
Weight decay	0.01, 0.05, 0.1	0.05
Warmup steps	0, 500, 1000	0

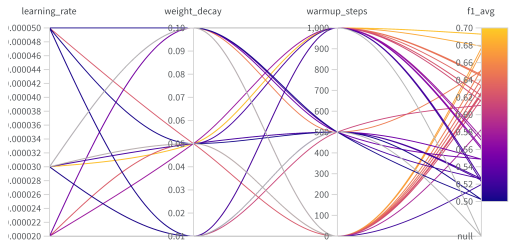

Fig. S7. Hyper parameters tuning for PubMedBERT

Hyper Para	Values	Best
Learning Rate	'2e-5, 3e-5, 5e-5	3e-5
No. of train epochs	1, 2, 3, 4	2
Weight decay	0.01, 0.05, 0.1	0.05
Warmup steps	0, 500, 1000	0

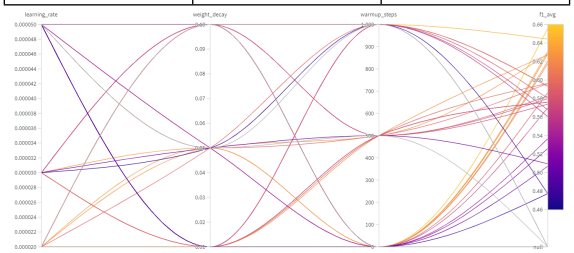

Fig. S8. Hyper parameters tuning for SciBERT

Hyper Para	Values	Best
Learning Rate	'2e-5, 3e-5, 5e-5	3e-5
No. of train epochs	1, 2, 3, 4	3
Weight decay	0.01, 0.05, 0.1	0.05
Warmup steps	0, 500, 1000	0

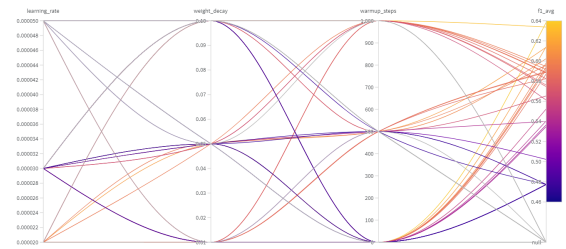

Fig. S9. Hyper parameters tuning for BlueBERT

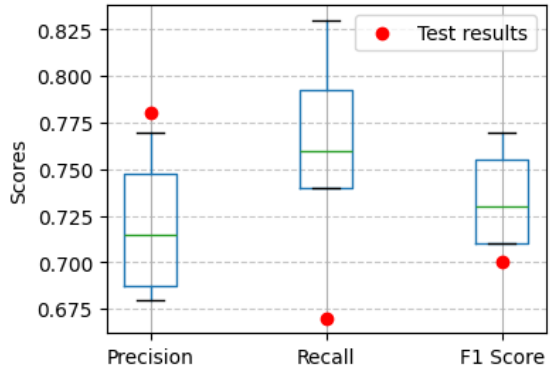

Fig. S10. Performance of BioBERT+SVM model trained on different three-fold combinations and tested on held-out validation fold and test data.

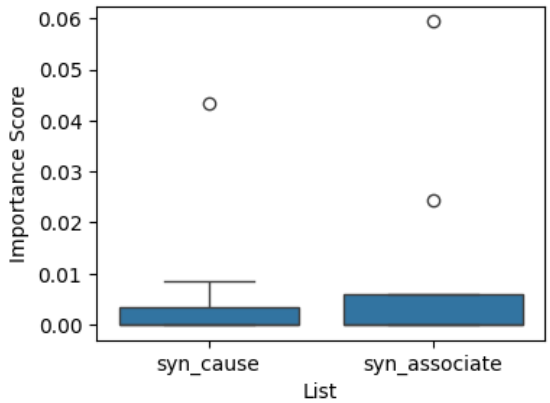

Fig. S11. Comparison of importance scores of causal-connoted words and associate words for CDR trained model. Here syn_cause represents the synonym of the word "cause," and syn_associate represents the synonym of the word "associate".

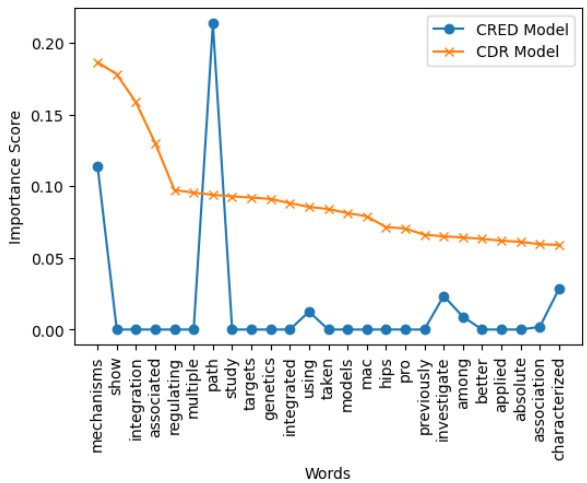

Fig. S12. Comparison of importance scores of top 25 words (from CDR-trained model) in CRED-trained and CDR-trained SVM models.

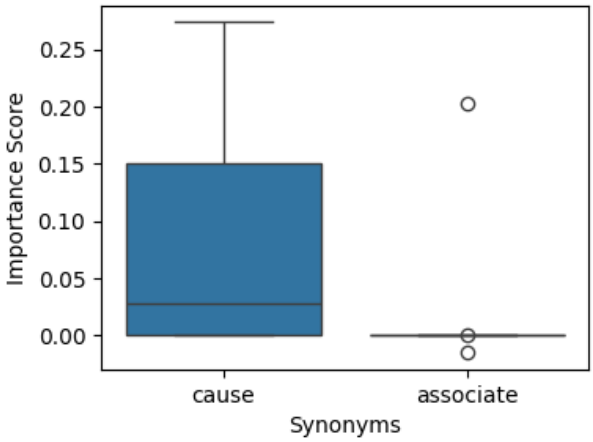

Fig. S13. Comparison of importance scores of "cause" and "associate" synonym words for CRED trained XGBoost model.

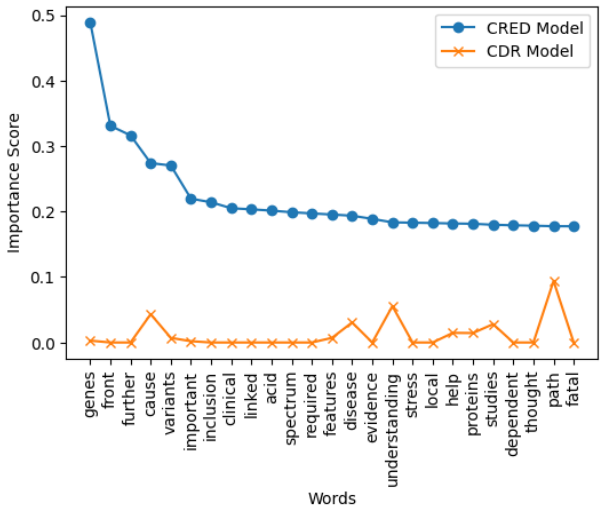

Fig. S14. Comparison of importance scores of top 25 words (from CRED-trained model) in CRED-trained and CDR-trained XGBoost models.

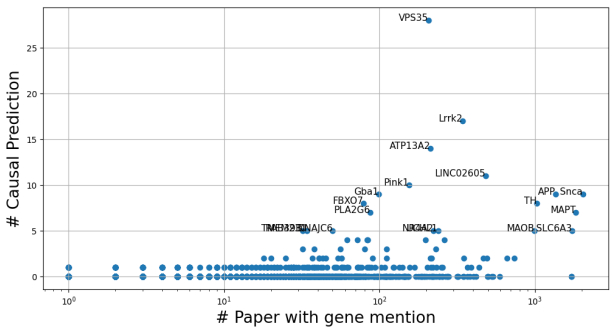

Fig. S15. Parkinson's related genes for 0-2000 occurences.

Appendix C: Supplementary Tables

Table S1. Performance of BioBERT+SVM model on data before augmentation.

	Positive class			Average		
	Prec	**Rec**	**F1**	**Prec**	**Rec**	**F1**
SVM	0.35	0.35	0.35	0.64	0.64	0.64

Table S2. Software/Packages specifications

Package/Software	Version
matplotlib	3.8.0
nltk	3.8.1
numpy	1.26.0
python	3.10.0
scikit-learn	1.3.1
scipy	1.11.3
seaborn	0.13.0
torch	2.0.1
tqdm	4.66.1
wandb	0.15.11
wordcloud	1.9.2
xgboost	2.0.0

Table S3. Parameters of models

	Total Params	Trained Params
BioBERT	110 M + 394754	394754
BioBERT + SVM	110 M + 711	711

Table S4. Performance of CDR-trained and CRED-trained BioBERT+SVM model, evaluated on CRED test data. For convenience CRED values are copied from Table 4

	Positive class			Average		
	Prec	**Rec**	**F1**	**Prec**	**Rec**	**F1**
CDR	0.41	0.39	0.40	0.68	0.67	0.67
CRED	0.61	0.35	0.45	0.78	0.67	0.70

Table S5. Performance of BioBERT+SVM model. Tr-Te represents training data - test data. So, CDR-CRED represents that the model is trained on CDR dataset and tested on CRED. CRED-CDR represents the vice-versa. Prec, Rec, and F1 denote precision, recall, and F1 score respectively

	Positive class			Average		
Tr-Te	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**
CDR-CRED	0.14	0.44	0.21	0.54	0.59	0.52
CRED-CDR	0.27	0.65	0.39	0.58	0.62	0.55