Published online 5 July 2021

https://doi.org/10.1093/nargab/lqab062

RENET2: high-performance full-text gene–disease relation extraction with iterative training data expansion

Junhao Su[†], **Ye Wu**[†], **Hing-Fung Ting**, **Tak-Wah Lam** and **Ruibang Luo**[*]

Department of Computer Science, The University of Hong Kong, Hong Kong, 999077, China

Received March 21, 2021; Revised June 16, 2021; Editorial Decision June 21, 2021; Accepted June 23, 2021

ABSTRACT

Relation extraction (RE) is a fundamental task for extracting gene–disease associations from biomedical text. Many state-of-the-art tools have limited capacity, as they can extract gene–disease associations only from single sentences or abstract texts. A few studies have explored extracting gene–disease associations from full-text articles, but there exists a large room for improvements. In this work, we propose RENET2, a deep learning-based RE method, which implements Section Filtering and ambiguous relations modeling to extract gene–disease associations from full-text articles. We designed a novel iterative training data expansion strategy to build an annotated full-text dataset to resolve the scarcity of labels on full-text articles. In our experiments, RENET2 achieved an F1-score of 72.13% for extracting gene–disease associations from an annotated full-text dataset, which was 27.22, 30.30, 29.24 and 23.87% higher than BeFree, DTMiner, BioBERT and RENET, respectively. We applied RENET2 to (i) ∼1.89M full-text articles from PubMed Central and found ∼3.72M gene–disease associations; and (ii) the LitCovid articles and ranked the top 15 proteins associated with COVID-19, supported by recent articles. RENET2 is an efficient and accurate method for full-text gene–disease association extraction. The source-code, manually curated abstract/full-text training data, and results of RENET2 are available at GitHub.

INTRODUCTION

The association between genes and diseases is essential for developing clinical diagnoses, therapeutic treatments and public health systems for diseases (1). However, the research on gene–disease associations is locked in an enormous volume of biomedical literature. PubMed Central (PMC) (2), a free full-text archive of the biomedical literature, had over 6.6 million articles in 2020. There is a pressing need for accurate and efficient tools to automatically extract gene–disease associations from the literature to improve access to information and support biomedical research (3).

The Relation Extraction (RE) task is critical for extracting gene–disease associations from the literature (4). The task is to determine whether there is an association between a gene–disease pair from a given text. RE is more challenging than the task of finding named entities from texts, namely named entity recognition (NER (5,6)), as it has to incorporate the information from complete sentences (sentence-based) or complete articles (document-based). A wide range of methods, such as BeFree (7), DTMiner (8), LHGDN (9), Thompson *et al.* (10) and Zhou *et al.* (11), employ a sentence-based approach to extracting gene–disease relations. These methods utilize different linguistic and co-occurrence features with machine-learning methods to identify gene–disease relations within each sentence. For example, BeFree (7) applies a shallow linguistic kernel, which uses both a local (orthographic and shallow linguistic features) and global context (trigrams and sparse bigrams) to extract relations from a single sentence. DTMiner (8) improves BeFree by adding a co-occurrence-based ranking module to estimate how closely the pairs are related. Thompson *et al.* (10) designed a sophisticated system to measure sentence complexity. It uses either co-occurring or linguistic patterns to extract gene–disease associations from a single sentence. Zhou *et al.* (11) developed a novel method that integrates the MeSH (Medical Subject Headings) database, term weight and co-occurrence methods to predict gene–disease associations based on the cosine similarity between gene vectors and disease vectors from a sentence. Similar outstanding works including Perera *et al.* (12), Nourani *et al.* (13) and Taha *et al.* (14). However, sentence-based methods can extract relations only within a sentence. In an article, information about gene–disease associations is often spread over multiple sentences. To extract gene–disease associations supported by an article, we need

[*] To whom correspondence should be addressed. Tel: +852 2859 2186; Fax: +852 2559 8447; Email: rbluo@cs.hku.hk

[†] The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

document-based RE methods to understand the context of the whole article.

Existing state-of-the-art document-level RE methods are designed mainly for abstract texts. BioBERT (15) is a comprehensive approach, which applies BERT (16), an attention-based language representation model (17), on biomedical text mining tasks, including NER, RE and question answering (18). BioBERT can extract gene–disease associations from biomedical text by performing classification on both sentences and abstracts. However, because of its attention mechanism, BioBERT constrains its maximum input length to 512 and restricts its application to complete articles. The predecessor to this work, called RENET (19), uses a Convolutional Neural Network (CNN) and a Recurrent Neural Network (RNN) to learn the document representation of gene–disease relations. RENET not only captures the relationship between genes and diseases within a sentence, but also models the interaction of different sentences to understand the context of the whole article. It achieves state-of-the-art performance in gene–disease RE from abstracts. However, RENET uses classification, which is similar to BioBERT, to model the RE problem. This mechanism is not flexible enough to handle more complex relation types, such as ambiguous relations. Moreover, it is designed for abstract texts, not optimized for full-text articles. The existing methods still pose significant challenges to full-text RE (20).

Full-text articles contain much more information than abstract texts. We estimate that a full-text article has an average of 4535 tokens, about 17 times that of an abstract. The average number of gene–disease pairs in a full-text article is estimated to be 197, compared to 5.6 of an abstract. If considering only unique gene–disease pairs, the average of a full-text article is 173, compared to 5.1 of an abstract. This rich content makes gene–disease RE from full text more challenging than from abstracts. The length of full-text articles exceeds the capacity of BioBERT and other BERT-based methods. It reduces the information density of the text, resulting in low prediction accuracy. The long length and large number of gene–disease pairs also make manual curation of a full-text dataset labor-intensive and time-consuming. As a result, there is no publicly available full-text level labeled data for development and evaluation.

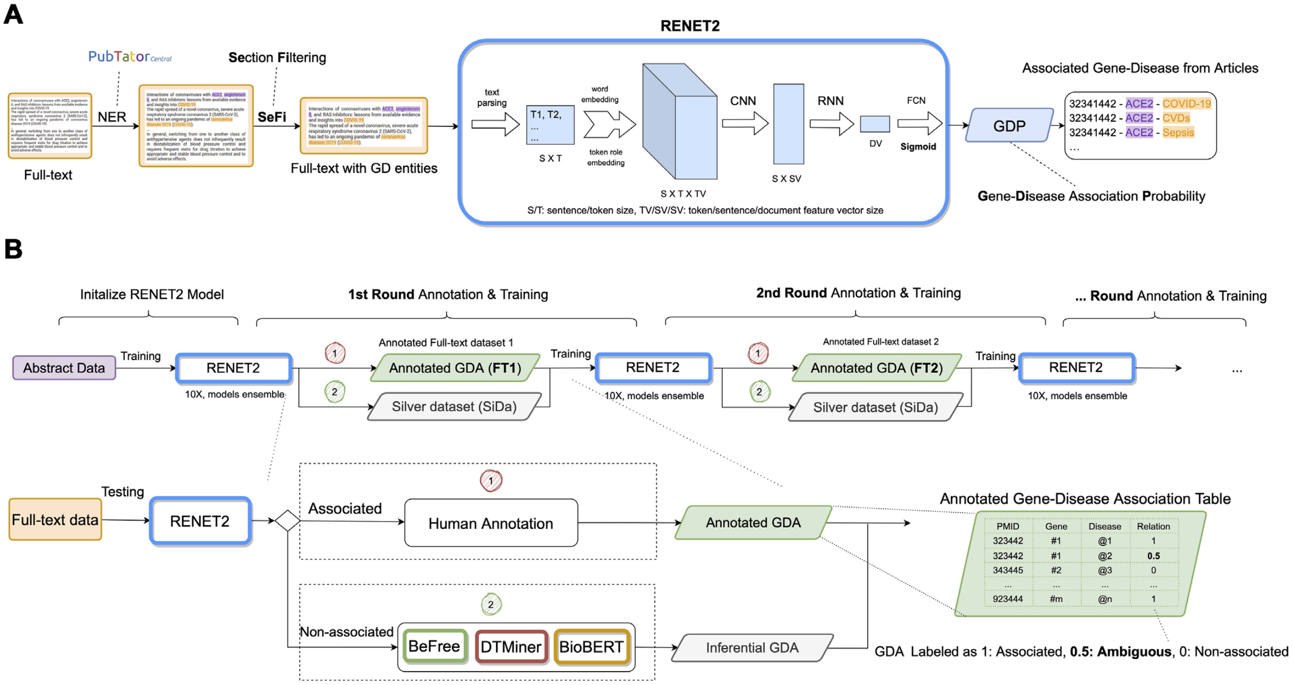
In this paper, we propose RENET2, an accurate and efficient full-text gene–disease RE method with an iterative training data expansion strategy, as shown in Figure 1. In RENET2, we introduced ‘Ambiguous association’, a new relation type to reduce human effort in labeling gene–disease associations, and used a regression-based deep-learning approach to model Ambiguous associations. We applied Section Filtering (SeFi), a novel data-enhancement technique, to reduce the noisy content and improve the information density of the input data for gene–disease RE. We designed a training data expansion strategy, which performs model training, prediction and efficient manual curation iteratively to generate ample high-quality full-text training samples.

We compared the performance of RENET2 with the state-of-the-art methods: BeFree, DTMiner, BioBERT and RENET. We found that RENET2 achieved the best F1 score using a full-text dataset. RENET2 achieved an F1 score of 72.13%, which was 27.22, 30.30, 29.24 and 23.87% higher than BeFree, DTMiner, BioBERT and RENET, respectively. Using RENET2, we analyzed 1 889 558 full-text articles from PMC and extracted 3 717 569 gene–disease-article associations, which is more than five times the number of associations extracted by RENET from abstract data (19). To help medical professionals keep track of advances in COVID-19 research (21,22), we applied RENET2 to the LitCovid articles, a collection of up-to-date publications on COVID-19. We found 1231 proteins associated with the COVID-19 disease and ranked the top 15 proteins according to the number of supporting studies. The source code of RENET2, the extracted gene–disease associations, and the proteins associated with COVID-19 with corresponding articles are available at https://github.com/sujunhao/RENET2.

MATERIALS AND METHODS

Dataset descriptions

We used two dataset levels in this study, one at abstract level and one at full-text level. See Supplementary Table S1 for the data on the annotated gene–disease associations. The abstract level dataset comprises 1000 annotated abstracts and was used (i) as the initial model in iterative training data expansion for training a full-text RENET2 model, and (ii) for benchmarking the effect of adding Ambiguous associations. To construct an annotated abstract level dataset, we started with manually annotating all gene–disease pairs in 500 abstracts from scratch. The 500 abstracts were randomly selected from the 29 192 abstracts included in RENET's training dataset (the 29 192 abstracts contain 57 553 genes, 83 942 diseases and 165 562 gene–disease pairs, in which, 6414, 3807 and 82 255 are unique, respectively). Noteworthy, these 500 abstracts were previously annotated in PubTator, and in our previous study RENET, we used the DisGeNET (23) annotations (DisGeNET is a public collection of gene–disease associations in abstracts collected from multiple sources but without a unified curation criterion). But in this study (RENET2), we decided to reannotate the 500 abstracts from scratch to aim for an even better annotation quality. Our annotation process was conducted by three experts with a Biology, Bioinformatics or Computer Science background, respectively. The workload was distributed evenly. An expert is required to mark an annotation that is not 100% certain as uncertain, thus we designed no overlapping workload between experts as we promote transferring any uncertainties to the whole team. All uncertain annotations were then discussed with and arbitrated by the other two experts. In these 500 abstracts, we found 992 genes, 1391 diseases and 2813 gene–disease pairs, in which, 610, 578 and 2568 are unique, respectively. The number of unique gene–disease pairs against the total is 91.2%. The top three genes are APOE (22 times), TP53 (20) and GSTM1 (19). The top three diseases are Neoplasms (69), Breast Neoplasms (36) and Alzheimer's disease (27). The top three gene–disease pairs are, APOE-Alzheimer's disease (14), TP53-Neoplasms (8) and GSTM1-Rheumatoid Arthritis (8). Using these 500 abstracts, we trained the first RENET2 model and conducted one iteration of training data expansion for


A
PubTator Central
Section Filtering
NER
SeFi
Full-text
Full-text with GD entities
RENET2
text parsing
T1, T2, ...
S X T
word embedding
token role embedding
S X T X TV
CNN
S X SV
RNN
DV
FCN
Sigmoid
S/T: sentence/token size, TV/SV/SV: token/sentence/document feature vector size
GDP
Associated Gene-Disease from Articles
32341442 - ACE2 - COVID-19
32341442 - ACE2 - CVDs
32341442 - ACE2 - Sepsis
...
Gene-Disease Association Probability
B
Initalize RENET2 Model
1st Round Annotation & Training
2nd Round Annotation & Training
... Round Annotation & Training
Abstract Data
Training
RENET2
10X, models ensemble
1
2
Annotated Full-text dataset 1
Annotated GDA (FT1)
Silver dataset (SiDa)
Annotated Full-text dataset 2
Annotated GDA (FT2)
...
Full-text data
Testing
Associated
Human Annotation
Annotated GDA
Non-associated
BeFree
DTMiner
BioBERT
Inferential GDA
Annotated Gene-Disease Association Table
GDA Labeled as 1: Associated, 0.5: Ambiguous, 0: Non-associated

PMID	Gene	Disease	Relation
323442	#1	@1	1
323442	#1	@2	**0.5**
343445	#2	@3	0
...	...	...	...
923444	#m	@n	1

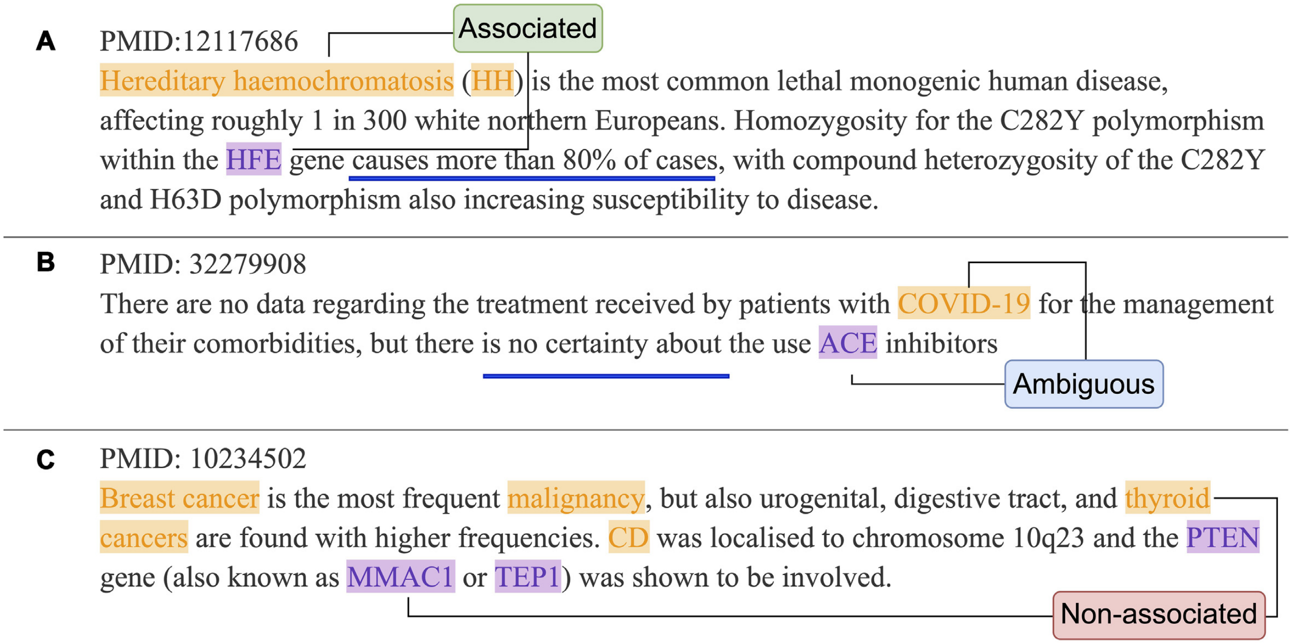
Figure 1. Overview of the RENET2 pipeline. (**A**) Full-text level RE. (**B**) Iterative training data expansion with RENET2 and human annotation.

another 500 abstracts, also selected from the 29 192 abstracts in RENET's training data. Different from the first 500 abstracts that we annotated from scratch, the annotation of another 500 abstracts was assisted by using the 'CURATED', 'INFERRED' and 'ANIMAL MODELS' gene–disease pairs in DisGeNET (those labeled 'LITERATURE' were not used). We regarded the associations found by RENET2 that are also in DisGeNET as true associations, and we manually checked and arbitrated the contradicting associations. The extra 500 abstracts were used only for training, while the fully manually annotated 500 abstracts were used for both training and validation.

The full-text level dataset was constructed starting from using the model trained on the 1000 abstracts to annotate 500 unlabeled full-text articles. The 500 full-text articles are randomly selected from the PMC open-access subset (2), in which only the articles having one or more gene–disease pairs are included. The subset has 1 889 477 articles, and 13 596 260 genes, 15 993 552 diseases and 194 292 671 gene–disease pairs, in which, 136 018, 19 886 and 19 257 344 are unique. The associated predictions were manually curated. The uncertain annotations were dealt with the same way as the abstracts (reviewed and decided by three experts). The non-associated predictions were curated with unanimous support from three other methods: BeFree, DTMiner and BioBERT (see 'Iterative Training Data Expansion' section for more details). Unlike the procedure used in the previous paragraph, in which another 500 abstracts were included in the second iteration, we used the same 500 full-text articles for data expansion of full-text training data expansion because we found a large space for annotation improvement on the same 500 full-text articles. The number of annotations almost doubled in the second iteration on the same 500 full-text articles (Supplementary Table S1). After the second iteration, the 500 full-text articles have 3490 genes, 4342 diseases and 51 642 gene–disease pairs, in which, 2095, 1244 and 46 379 are unique, respectively. The number of unique gene–disease pairs against the total is 89.8%. The top three genes are TNF-α (49 times), IL6 (43) and AKT1 (27). The top three diseases are Breast Neoplasms (180), Infections (122) and Death (116). The top three gene–disease pairs are, TNF-α-Inflammation (34), IL6-Inflammation (33) and TNF-α-Breast Neoplasms (22). Articles use different section names for the same section (e.g. Introduction, Backgrounds, or no section name was used), but were unified by PubTator Central (24) using standardized 'section type identifiers' (25,26). In this study, we relied on the standardized identifiers on selecting target sections.

Definition of the problem

The input of the RE problem is article X, consisting of s tokens $x_1, x_2, \ldots, x_s$. Let $G = \{g_1, g_2, \ldots, g_n\}$ denote the set of gene entities and $D = \{d_1, d_2, \ldots, d_m\}$ denote the set of disease entities in an article. The task of gene–disease RE is, for gene–disease pair $g_i \in G, d_j \in D$, to determine whether the article supports an association relation between g_i and d_j. The task predicts a relation $y(g_i, d_j) \in \{0, 1\}$, where 1 represents an associated relation, and 0 represents a non-associated relation. An associated relation is defined as a pair of gene and disease that has a sort of association, including but not limited to cause, effect, existence and regulation, either positive or negative, as written by the authors (Figure 2A). An association is inferred semantically (i.e. the authors said so), not pragmatically (i.e. the authors can actually mean something else, but we do not care). Non-associated relation is defined as a pair of gene and disease



PMID:12117686
Hereditary haemochromatosis (HH) is the most common lethal monogenic human disease, affecting roughly 1 in 300 white northern Europeans. Homozygosity for the C282Y polymorphism within the HFE gene causes more than 80% of cases, with compound heterozygosity of the C282Y and H63D polymorphism also increasing susceptibility to disease.

PMID: 32279908
There are no data regarding the treatment received by patients with COVID-19 for the management of their comorbidities, but there is no certainty about the use ACE inhibitors

PMID: 10234502
Breast cancer is the most frequent malignancy, but also urogenital, digestive tract, and thyroid cancers are found with higher frequencies. CD was localised to chromosome 10q23 and the PTEN gene (also known as MMAC1 or TEP1) was shown to be involved.

Figure 2. Illustration of the three gene–disease relation types in our dataset. (**A**) Associated relation between *HFE* (gene) and hereditary haemochromatosis (disease) from PMID 12117686 (36). (**B**) Ambiguous association between *ACE* (gene) and COVID-19 (disease) from PMID 32279908 (37). (**C**) Non-association relation between *MMAC1* (gene) and thyroid cancer (disease) from PMID 10234502 (38). Blue underlines are the trigger for the relation type. Gene entities are purple, and disease entities are yellow.

that the authors have semantically said nothing about (Figure 2C).

Ambiguous association

To better model the relationship between genes and diseases, and to represent relationships that are difficult for an expert to decide (because they may be 'Semantically Ambiguous' or 'Incomprehensible'), we introduced an additional relation type, called Ambiguous association, which is defined as an uncertain association between a gene and a disease. 'Semantically Ambiguous' means that authors themselves said an association is uncertain. Figure 2B has shown a case of 'Semantically Ambiguous'. 'Incomprehensible' means our experts can neither confirm nor deny an association given the texts. 'Incomprehensible' is not uncommon due to a large variety in publication qualities. Instead of representing an Ambiguous association as a new class, we use a probability score, GDP (Gene–disease Association Probability). $p(g_i, d_j) \in [0,\ 1]$ to model the new relations. 'Non-associated' and 'Associated' are still represented by 0 and 1. We use 0.5 to represent an Ambiguous association, as it is between Non-associated and Associated. The rationale of using Ambiguous association is that an uncertain annotation should neither be considered Associated nor Non-associated in model training. We used Ambiguous associations in the model training stage to improve the model's generalization capability. In the prediction stage, we extract Associated gene–disease pairs by computing $\hat{y}(g_i,\ d_j)$ as:

$$\hat{y}\left(g_i,\ d_j\right) = \begin{cases} 1, & if\ p(g_i,\ d_j) > 0.5 \\ 0, & \text{otherwise} \end{cases},\ g_i \in G, d_j \in D \quad (1)$$

Overview of the RENET2 pipeline

An overview of the RENET2 pipeline is shown in Figure 1A. First, we conducted NER to identify the gene (G) and disease (D) entities in an article. For this task, we utilized PubTator Central (24), a state-of-the-art automated concept annotation tool, designed for biomedical full-text articles. It applied cutting-edge machine learning and deep learning techniques for concept disambiguation to improve NER accuracy.

Based on the NER results, we applied SeFi to reduce the input article's noise. After the preprocessing steps, the data was fed to the RENET2 model for training and prediction. RENET2 produced a probability score, the GDP score, for each gene–disease pair in the article. The GDP scores allowed us to extract the gene–disease associations from the articles.

Section Filtering (SeFi)

SeFi is a technique designed to filter noisy content from full text for RE. We observed that paragraphs containing no gene–disease pairs (i.e. containing only one type or no gene/disease entities), did not provide information on gene–disease relations mentioned at the full-text level. For example, method sections that discussed experiment settings but not specific genes or diseases were not helpful for identifying any gene–disease associations. Based on this observation, we designed a simple filtering technique, called SeFi, to improve data quality for full-text level RE. The idea of SeFi is to delete paragraphs in each section without any gene and disease entity pair information. It is regarded as a preprocessing module for RENET2. To conduct SeFi, (i) we found all gene and disease entities and paragraph information using PubTator Central, and (ii) deleted paragraphs that did not have any gene and disease entity pairs in each section. We used PMC's standardized 'section type identifiers' available from PubTator Central (PTC) as the section names for each article. We found that this technique improved RE performance in full text. See the 'Results' section.

RENET2 model

RENET2 was built and expanded using the RENET framework. Each word in the RENET2 model is represented by a word vector combined with a one-hot feature vector. The word vector captures the semantic features of a word, while the feature vector denotes whether a word is a target gene, target disease, non-target gene, or non-target disease. For predicting any pair of a gene and disease, the words that contain a gene or disease that match the gene or disease, are marked at target gene or target disease in the vector. Other words that contain a gene or disease are marked at non-target gene or non-target disease otherwise. Then a document-level representation of the target gene and disease is computed in two steps: (i) from word representation to sentence representation, using a CNN, and (ii) from sentence representation to document representation, using an RNN. Finally, a Feed Forward Neural Network is applied to calculate the Gene–disease Association Score. For the detailed network architecture and hyperparameters, see Supplementary Note. The computation of RENET2 neural networks is represented as:

$$\hat{p}\left(g_i,\ d_j\right) = \ \text{sigmoid}\left(\phi\left(g_i,\ d_j\right)\right) \tag{2}$$

where $\phi(g_i,\ d_j)$ is the learned document representation for $(g_i,\ d_j)$. Note that we use a sigmoid activation function to compute the probability of association.

To incorporate Ambiguous associations for training, RENET2 models treat extraction as a regression problem. RENET2 uses the mean square errors (MSE) function as the loss function:

$$\text{MSE} = \frac{1}{nm}\sum_{i\,=\,1}^{n}\sum_{j\,=\,1}^{m}\left(p_{g_i,\ d_j} - \ \hat{p}_{g_i,\ d_j}\right)^2 \tag{3}$$

where $m,\ n$ is the number of gene and disease entities in the article, respectively.

We implemented RENTE2 using Pytorch (27). We estimated that the average token number in a sentence from a biomedical full-text article to be about 27, which is consistent with Lippincott *et al.* (28). We configured the maximum token number in a sentence to be 54 to cover most cases. We configured the maximum number of input sentences as 1000 in a full text. The number was empirically determined, making RENET2 capable of handling a maximum of 54 000 (54 × 1000) tokens, which is more than the number of tokens in most full-text articles. More settings and hyperparameters are available in the Supplementary Note.

Model ensemble

RENET2 uses the ensemble technique (29) to boost its performance in full-text RE. The ensemble of RENET2 models is done by training $\theta \in \mathbb{N}^+$ RENET2 models and integrating their prediction results. A gene–disease pair is predicted as Associated if ≥50% of the RENET2 models predict it as Associated. For ensembling θ RENET2 models, the ensembled relation type $\hat{y}_{\text{ensemble}}$ of a gene and disease pair is computed as:

$$\hat{y}_{\text{ensemble}}\left(g_i,\ d_j\right) = \begin{cases} 1, & \Sigma_{k=0}^{\theta}\ \hat{y}_k\left(g_i, d_j\right) \geq \frac{\theta}{2} \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

where $\theta \in \mathbb{N}^+$, and $\hat{y}_k(g_i, d_j)$ is the relation type prediction of the k^{th} model. We set $\theta\ =\ 10$ as default in RENET2.

Iterative training data expansion

RENET2, like many other deep-learning methods with supervised learning, requires ample high-quality training samples for good performance. But because of the large number of gene–disease pairs and long text length, manual labeling of full-text articles is labor-intensive and often far too expensive. In RENET2, we developed 'iterative training data expansion' to make full-text labeling faster. Instead of manually labeling every full-text article from scratch, the basic idea of the method is to have an expert curate a small number of articles initially labeled by a machine model, and then use the curated labels to train a larger machine model to be used in the next iteration. With the improved performance of the machine model, less and less curation effort is expected to be required in subsequent rounds. The slowdown or even reverse of the machine model performance would denote a stop in curating more machine-labeled results for training. In the field of image classification, a similar idea was used to construct LSUN (30), a large-scale image-classification dataset.

The workflow of iterative training data expansion is depicted in Figure 1B. At the beginning of each iteration, we use the best existing RENET2 model to predict the gene–disease associations from a number of full-text articles. The initial RENET2 model can be from training on a curated public dataset such as DisGeNET, which contains only abstracts, or from the last iteration. Based on the prediction results, all gene–disease pairs are processed using the following two steps:

(i) The gene–disease pairs predicted as Associated undergo a manual annotation process. By manually checking each predicted Association, we determine whether it is an Association, Ambiguous association, or Non-association. This step ensures the precision of the positive predictions to be fed to the next iteration. We found this curation process much more efficient than manual-labeling from scratch because (a) most false positive labels are significant to humans, and (b) for labels that trigger hesitation, we simply label them as Ambiguous associations. The annotated gene–disease pairs are regarded as a gold dataset and are used for both training and evaluation.

(ii) The gene–disease pairs that are predicted as Non-associated are cross-validated by other methods, in our case, BeFree, DTMiner and BioBERT. We regard a gene–disease pair as inferential Non-associated if all methods predict it as Non-associated. This step ensures the precision of the negative predictions to be fed to the next iteration. As these pairs are cross validated by other methods, but not manually inspected, they constitute a Silver Dataset (SiDa).

We then enter the next iteration and train a new RENET2 model using both the gold and SiDa. Depending on the available resources and goal, iteration can be stopped if there is a time constraint or satisfactory performance is reached.

Evaluation metrics

We use precision, recall and F1 score metrics to evaluate the model performance at each training iteration. Recall can be simply calculated as (annotated associations being correctly predicted)/(total annotated associations). However, calculating precision is more complicated, because we cannot assume all gene–disease pairs in our full-text dataset are annotated, so some true-positive may be misclassified as false-positive. To avoid these misclassifications, we regard an associated gene–disease pair prediction as true-positive if it (i) also exists in our validation dataset and with matched prediction, or (ii) is predicted as associated by all four methods: RENET2, BeFree, DTMiner and BioBERT. Note that we are not saying that a gene–disease association predicted positive by all four state-of-the-art methods is absolutely correct, but we consider that (i) there is a relatively small chance of an association predicted positive by four fundamentally different methods being incorrect, and (ii) an association categorized as false-positive in all four methods will have the same effect on changing the precision. The same evaluation methods and metrics were applied to benchmarking all tools. Each experiment was repeated five times with randomly picked 80% training and 20% validation data (i.e. 5-fold cross validation).

RESULTS

Performance using Ambiguous associations

We expected that if the use of Ambiguous associations in training could improve the performance of RE, it must hold true with abstracts because they usually have the highest density of gene and disease entities. We benchmarked the use of Ambiguous associations with the abstract dataset, and we found it effective in improving both precision and recall. The results are shown in Supplementary Table S2. RENET2 achieved a 2.40% higher F1 score (71.55% against 69.15%) when Ambiguous associations were used. Compared with RENET2's predecessor RENET, which works only with abstracts, RENET2 achieved a 2.77% higher F1 score (71.55% against 68.78%).

Performance of iterative training data expansion

We compared the performance of RENET2 at different iterations to evaluate the effect of iterative training data expansion. We trained models using the annotations and curations obtained at the end of each iteration and tested them against our final full-text dataset, i.e. the 500 full-texts with two rounds of curation. The results, shown in Table 1, verified the effectiveness of the iterative training data expansion strategy. With less and less human effort put into each iteration, both precision and recall continued to improve. A leap was observed on recall (from 20.02 to 72.04%) when we switched from using abstracts to full-texts for training, suggesting a substantial difference between abstracts and full texts for extracting gene–disease associations, and the essentiality of having a method designed for full-text RE. From our observation, full texts have, on average, 17 times more tokens, and 34 times more unique gene–disease pairs than abstracts.

Table 1. Comparison of RENET2's performance at different training data expansion iterations

Training dataset	Testing dataset	Precision	Recall	F1 score
500 abstracts	500 full texts second curation (5-fold cross validation, 80% training; 20% validation)	0.6024	0.2539	0.3573
1000 abstracts		0.6505	0.2002	0.3062
500 full texts first curation		0.6765	0.7204	0.6977
500 full texts second curation		**0.7062**	**0.7371**	**0.7213**

The best result in each column is in bold.

Comparison of RENET2 models with different training settings

We introduced three new techniques to RENET2: Model Ensemble (ENS), SeFi and SiDa. But how these techniques work alone and together remained to be studied, so we compared RENET2's performance with different technique combinations using the full-text dataset. The results are shown in Table 2. We observed that all techniques improved model performance when used alone. SeFi resulted in the most significant improvement in RENET2, since all F1 scores when SeFi was not used were lower than those using SeFi. Using SeFi alone resulted in a 7.62% increase in the F1 score, indicating that filtering out noisy text is critical for full-text RE. In addition, RENET2 excludes the method section for training by default. The influence of each section is analyzed in detail in a later section. ENS alone improved F1 score by 4.05%. SiDa improved F1 score by 1.41%, but it had a different impact on precision and recall, increasing precision by 4.68%, but decreasing recall by 3.25%. This matched our expectation that the additional Non-associated labels from the SiDa would reduce false positive predictions, but increase false negative predictions to a certain extent. Thus, if SiDa is used for better precision, other techniques are needed to compensate for decreasing recall. When all three techniques were used, the F1 score improved by 11.49% (16.45% higher precision and 4.84% higher recall). Therefore, we will use all three techniques in RENET2 by default in future analyses.

Table 2. Comparison of RENET2's performance on the full-text dataset when trained with different settings

Ensemble	SeFi	SiDa	Precision	Recall	F1 score	F1 score increment
N	N	N	0.5417	0.6887	0.6064	-
Y	N	N	0.5791	0.7327	0.6469	4.05%
N	Y	N	0.6465	0.7233	0.6826	7.62%
N	N	Y	0.5885	0.6562	0.6205	1.41%
Y	N	Y	0.6079	0.6946	0.6484	4.19%
N	Y	Y	0.6888	0.6906	0.6897	8.32%
Y	Y	N	0.6746	**0.7606**	0.7150	10.86%
Y	Y	Y	**0.7062**	0.7371	**0.7213**	11.49%

The best result in each column is in bold. ENS: Ensemble, SeFi: Section Filtering, SiDa: Silver Dataset, F1 increment: increase in F1 score compared to the basic setting.

Table 3. Comparison of different methods for full-text gene–disease RE

Method	Precision	Recall	F1 score
BeFree	0.3152	0.7808	0.4491
DTMiner	0.2761	0.8624	0.4183
BioBERT	0.2803	0.9128	0.4289
RENET	0.3681	0.7002	0.4826
RENET2	**0.7062**	0.7371	**0.7213**
RENET2 high-sensitivity mode	0.5518	**0.9217**	0.6903

The best result in each column is in bold.

Comparison between RENET2 and other methods

To the best of our knowledge, RENET2 is the first open-source method optimized for full-text gene–disease RE. We compared RENET2 to three state-of-the-art gene–disease RE methods: BeFree, DT-Miner, BioBERT and RENET. For BeFree and DTMiner, the best pretrained models were downloaded and used for benchmarking. For BioBERT, we used a BioBERT-based model fine-tuned on the GAD (7) dataset. For RENET, we used pretrained model described in the RENET paper. To ensure a fair comparison, all methods used PTC for the NER steps.

The results are shown in Table 3. RENET2 outperformed the other four methods by a significant margin, achieving 70.62% precision, which was 39.09, 43.01, 42.59 and 33.81% higher than BeFree, DTMiner, BioBERT and RENET, respectively. For overall performance, RENET2 achieved a 72.13% F1 score, which was 27.22, 30.30, 29.24 and 23.87% higher than BeFree, DTMiner, BioBERT and RENET, respectively. The four methods underperformed on precision, partially because they are sentence-based and could not leverage multi-sentence context to sift out non-conclusive associations.

Using the default mode of RENET2, we observed lower recall than the other three methods (73.71% against 78.08, 86.24 and 91.28% of BeFree, DTMiner and BioBERT). To accommodate some usage scenarios that favor recall over precision, RENET2 also provides a high-sensitivity mode. Using this mode, we achieved the best recall (92.17%) of all methods, while having a 3.10% lower F1 score than the default RENET2. In the high-sensitivity mode, the use of the SiDa for training was disabled, and the ENS's voting strategy was modified to 'predict positive as long as one out of ten models support it'.

Studying the importance of difference full-text sections for relation extraction

We studied the importance of different sections for full-text RE. We summarized the count of gene–disease associations in each section using the full-text dataset. The results are shown in Figure 3A. We found that introduction, discussion and abstract were the three most informative sections for gene–disease RE. The findings highlight that using abstract alone is insufficient for gene–disease RE because a large portion of gene–disease associations are from the other sections. We also found that few associations were found in the method section.

To better understand the relationship between the different sections, we measured the overlapping rate of associations found in them. A heat map is shown in Figure 3B. The overlapping rate of two sections is computed as overlap $(A,\ B) = |A \cap B|\ / \min(|A|,\ |B|)$, where A and B are the gene–disease associations found in the two sections. We found that the highest overlapping rate was between abstract, result and discussion. This indicates that many gene–disease associations from the abstract can also be cross-validated in the result or discussion section. Although we have not made use of this discovery in RENET2, it could lead to even better precision in our future investigations.

To understand how iterative training data expansion improved the recall rate in different sections, we applied three RENET2 models trained with data from different iterations to the full-text dataset. The recall breakdown by section is shown in Figure 3C. We found that the two full-text models performed much better than the abstract model on the abstracts themselves, so we believe that a higher overall number of tokens and a more diverse corpus can comprehensively improve the performance of gene–disease RE.

We also performed an experiment with one-section left out to train RENET2 to see the effect of each section on the prediction power of RENET2. The results are shown in Figure 3D. In spite of removing a section from the training dataset, all models were trained using RENET2's default setting (i.e. ENS, SeFi and SiDa enabled). When all sections were used for training, the F1 score was 71.55%. When the method section was left out, the F1 score increased slightly by 0.58%. For this reason, we left out the method section in model training by default in RENET2. We found that leaving out any sections other than method decreased the F1 score from 1.45% to 11.68%. The findings show that leaving out the introduction section severely deteriorated the performance (F1 score from 71.55 to 59.87%), indicating the importance of having the introduction section when training a model for full-text gene–disease RE.

Application 1: Large-scale full-text gene–disease relation extraction

We applied RENET2 on a large-scale biomedical literature database to build a collection of gene–disease associations from existing studies. We used RENET2 to extract gene–disease associations from all full-text articles available in the PMC open-access subset (downloaded on August 2020), which has more than 2.75 million full-text articles. It is the largest collection of full-text articles available for download and text mining (2). After filtering out articles without any

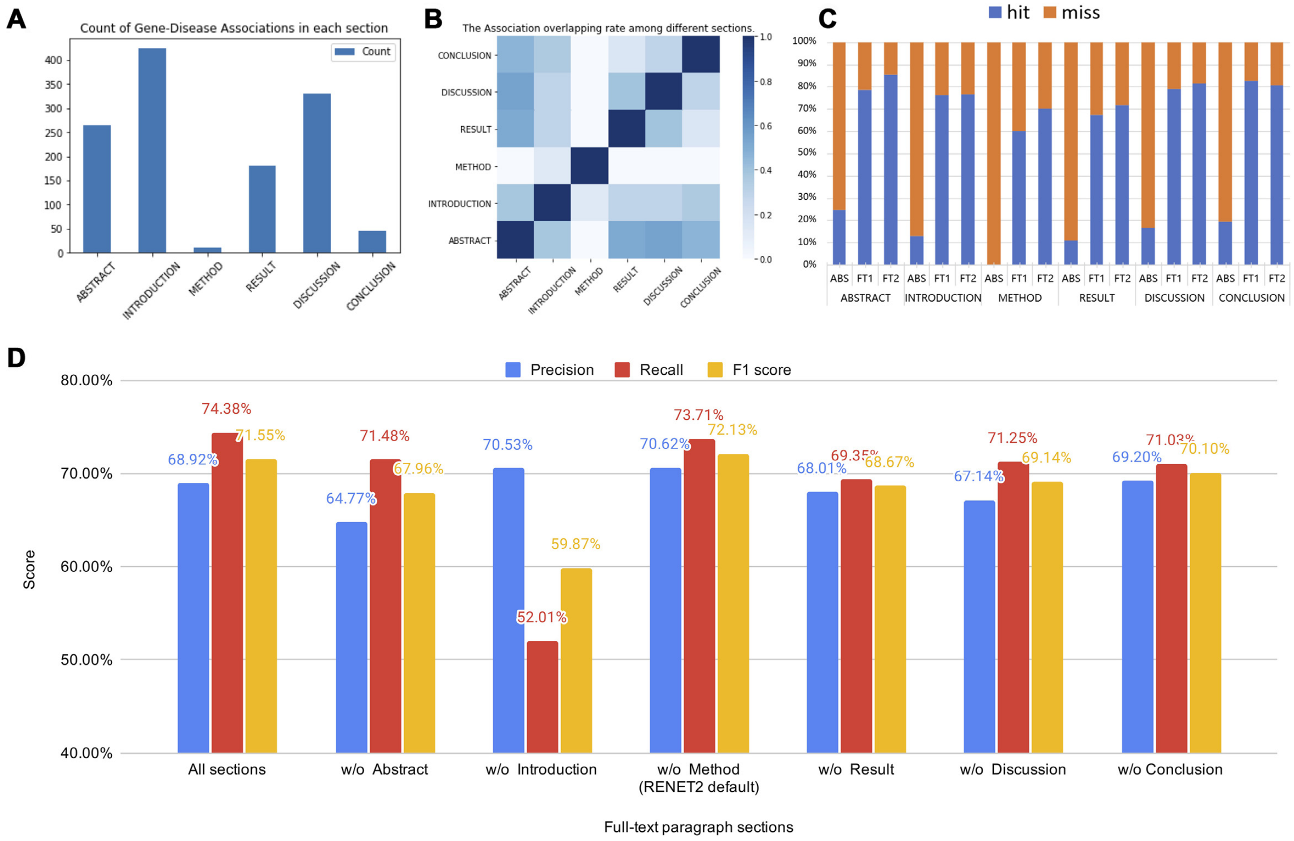


Figure 3. Gene–disease RE in different full-text sections and their effect on RENET2 models. (**A**) The count of 'Associated' predictions in different sections. (**B**) The association overlapping rate among different sections. The darker color indicates a higher overlap rate. (**C**) The recall rate of multiple RENET2 models on different sections. ABS: model trained with 1000 abstracts, FT1: 500 full texts 1st curation, FT2: 500 full texts second curation. (**D**) RENET2's performance after removing different sections from training.

gene–disease pair, we applied RENET2 on the 1 889 558 remaining articles. We found 3 717 569 gene–disease associations from the articles in 14.65 wall-clock hours using a computing cluster with 49 NVIDIA GeForce GTX 1080 Ti GPU cards (detailed statistics in Supplementary Table S3). This was more than five times the number of associations extracted from abstracts by RENET (19). The scripts for running RENET2 and the entire set of extracted associations are available in RENET2's GitHub repo.

Application 2: Finding proteins associated with COVID-19

With the fast expansion of COVID-19 research, it is getting harder and harder for researchers to keep up with the latest progress. RENET2 provides an efficient way for users to track associated proteins and pinpoint the subset of the literature of interest. We applied RENET2 to extract proteins associated with COVID-19 from the LitCovid (21) articles. LitCovid is a curated literature hub for tracking up-to-date scientific information about COVID-19. It had 73 654 articles in the dataset dated June 2020. After filtering out articles without any protein names, we applied RENET2 on the 19 368 remaining articles, and we found 1231 proteins that are reported to be associated with COVID-19 in at least one article. The results are shown in Figure 4. The top 15 proteins are *ACE2*, *IL-6*, *CRP*, *Spike*, *ACE*, *TNF-alpha*, *TMPRSS2*, *IL-1beta*, *ORF1a/b*, *Fibrinogen*, *CD8*, *Mpro*, *AST*, *IL-10* and *IFN-gamma*. The findings are consistent with Yeganova *et al.* (31). The scripts for running RENET2 and the results are available in RENET2's GitHub repo.

DISCUSSION

In this paper, we introduced RENET2, a deep-learning-based RE method to extract full-text gene–disease associations from full-text articles. RENET2 can use Ambiguous associations for training and multiple techniques can be applied, including ENS, SeFi and a SiDa, to boost its performance on full text. The new iterative training data expansion method proved to be effective in improving RE performance, while reducing human effort. In our experiments, RENET2 significantly outperformed state-of-the-art methods on full-text gene–disease RE. We demonstrated RENET2's utility using two applications. We applied RENET2 to the PMC open-access subset, which includes over a million full-text articles, and extracted over three million gene–disease associations. We applied RENET2 to the fast-expanding pool of COVID-19 research articles and ranked the top 15 proteins verified in another more systematic study. The source code and the results of this study are publicly available in GitHub.

Some practical challenges remain, leaving room for the further development of RENET2. First, the upper-bound accuracy of RE is capped by the accuracy of NER. Our

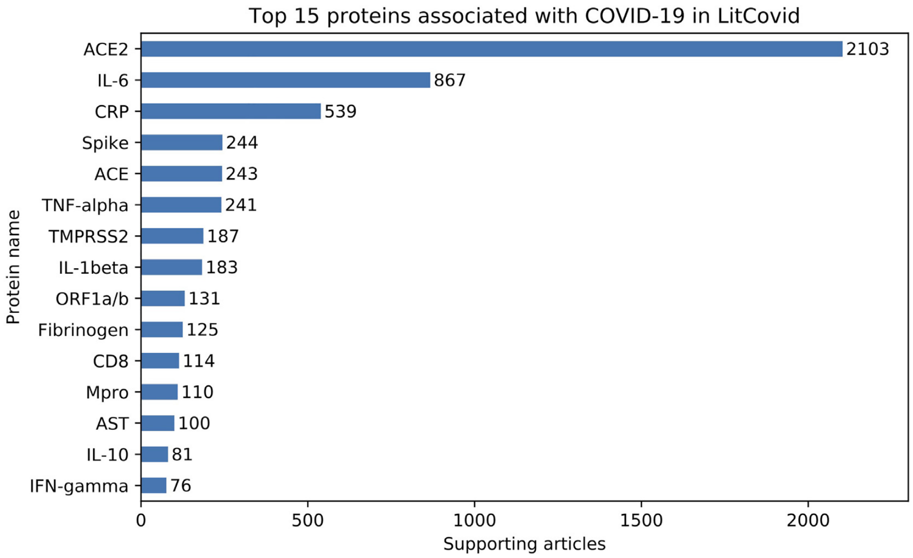


Figure 4. The top 15 proteins RENET2 found associated with COVID-19 in the LitCovid articles. The labels show the number of articles that support the protein's association with COVID-19.

study used PTC, which is the best system to date for NER. However, we still found a small number of NER results to be inconsistent and erroneous, such as incomplete gene and disease entities, and failure to disambiguate gene and disease acronyms. We estimate that the accuracy of the RENET2 model can be improved by at least 8% if the NER annotation is error-free. We expect to see in the near future that deep-learning methods could improve the accuracy of NER. Second, the current model is limited to computing one gene–disease pair association at a time, which results in a waste of computation resources (32). Different gene–disease pairs from the same article share most of their contexts. A significant amount of computing time and resources can be saved if multiple gene–disease pairs are processed in a single computation step. One of the possible solutions is to use a graph-based network structure to represent genes and diseases as vertices, and relationships as edges (33) for model training and inference. We also hope to incorporate deep language representation models into full-text RE in our future research. Recent studies show that deep language representation models, such as ELMo (34) and BERT (16), can develop strong language understanding capability by pre-training on large-scale unlabeled corpora. We aim to solve the input length limitation as well as a few other limitations of applying deep language representation models using DocBERT (35) as an example to improve the performance of full-text gene–disease RE.

DATA AVAILABILITY

The source-code and the manually curated abstract/full-text training data and results of RENET2 are available at https://github.com/sujunhao/RENET2.

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

ACKNOWLEDGEMENTS

We thank Zhenxian Zheng for his critical comments on model structure and data annotation.

FUNDING

HKSAR Government [ECS 27204518, TRS T21-705/20-N to R.L.]; HKU [URC 17208019 to H.T.].
Conflict of interest statement. None declared.

REFERENCES

1. Little,J., Bradley,L., Bray,M.S., Clyne,M., Dorman,J., Ellsworth,D.L., Hanson,J., Khoury,M., Lau,J., O'Brien,T.R. *et al.* (2002) Reporting, appraising, and integrating data on genotype prevalence and gene–disease associations. *Am. J. Epidemiol.*, **156**, 300–310.
2. Roberts,R.J. (2001) PubMed Central: the GenBank of the published literature. *Proc. Natl Acad. Sci. U.S.A.*, **98**, 381–382.
3. Kilicoglu,H. (2018) Biomedical text mining for research rigor and integrity: tasks, challenges, directions. *Brief. Bioinform.*, **19**, 1400–1414.
4. Bach,N. and Badaskar,S. (2007) A review of relation extraction. *Literat. Rev. Lang. Stat. II*, **2**, 1–15.
5. Nadeau,D. and Sekine,S. (2007) A survey of named entity recognition and classification. *Lingvist. Investig.*, **30**, 3–26.
6. Habibi,M., Weber,L., Neves,M., Wiegandt,D.L. and Leser,U. (2017) Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics*, **33**, i37–i48.
7. Bravo,À., Piñero,J., Queralt-Rosinach,N., Rautschka,M. and Furlong,L.I. (2015) Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research. *BMC Bioinformatics*, **16**, 55.
8. Xu,D., Zhang,M., Xie,Y., Wang,F., Chen,M., Zhu,K.Q. and Wei,J. (2016) DTMiner: identification of potential disease targets through biomedical literature mining. *Bioinformatics*, **32**, 3619–3626.
9. Bundschus,M., Dejori,M., Stetter,M., Tresp,V. and Kriegel,H.-P. (2008) Extraction of semantic biomedical relations from text using conditional random fields. *BMC Bioinformatics*, **9**, 207.
10. Thompson,P. and Ananiadou,S. (2017) Extracting gene-disease relations from text to support biomarker discovery. In: *Proceedings of the 2017 International Conference on Digital Health*. pp. 180–189.

11. Zhou,J. and Fu,B.-q. (2018) The research on gene–disease association based on text-mining of PubMed. *BMC Bioinformatics*, **19**, 37.
12. Perera,N., Dehmer,M. and Emmert-Streib,F. (2020) Named entity recognition and relation detection for biomedical information extraction. *Front. Cell Dev. Biol.*, **8**, 673.
13. Nourani,E. and Reshadat,V. (2020) Association extraction from biomedical literature based on representation and transfer learning. *J. Theor. Biol.*, **488**, 110112.
14. Taha,K., Davuluri,R., Yoo,P. and Spencer,J. (2021) Personizing the prediction of future susceptibility to a specific disease. *PLoS One*, **16**, e0243127.
15. Lee,J., Yoon,W., Kim,S., Kim,D., Kim,S., So,C.H. and Kang,J. (2020) BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, **36**, 1234–1240.
16. Devlin,J., Chang,M.-W., Lee,K. and Toutanova,K. (2018) BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv doi: https://arxiv.org/abs/1810.04805?source=post_page, 24 May 2019, preprint: not peer reviewed.
17. Vaswani,A., Shazeer,N., Parmar,N., Uszkoreit,J., Jones,L., Gomez,A.N., Kaiser,L. and Polosukhin,I. (2017) Attention is all you need. In: *Advances in neural information processing systems*. pp. 5998–6008.
18. Simpson,M.S. and Demner-Fushman,D. (2012) Biomedical text mining: a survey of recent progress. In: *Mining Text Data*. Springer, Boston, MA, pp. 465–517.
19. Wu,Y., Luo,R., Leung,H.C.M., Ting,H.-F. and Lam,T.-W. (2019) Renet: A deep learning approach for extracting gene-disease associations from literature. In: *International Conference on Research in Computational Molecular Biology*. Springer, Cham, pp. 272–284.
20. Dai,H.-J., Chang,Y.-C., Tsai,R.T.-H. and Hsu,W.-L. (2010) New challenges for biological text-mining in the next decade. *J. Comput. Sci. Tech.*, **25**, 169–179.
21. Chen,Q., Allot,A. and Lu,Z. (2020) Keep up with the latest coronavirus research. *Nature*, **579**, 193–193.
22. Chen,Q., Allot,A. and Lu,Z. (2021) LitCovid: an open database of COVID-19 literature. *Nucleic Acids Res.*, **49**, D1534–D1540.
23. Piñero,J., Ramírez-Anguita,J.M., Saüch-Pitarch,J., Ronzano,F., Centeno,E., Sanz,F. and Furlong,L.I. (2020) The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res.*, **48**, D845–D855.
24. Wei,C.-H., Allot,A., Leaman,R. and Lu,Z. (2019) PubTator Central: automated concept annotation for biomedical full text articles. *Nucleic Acids Res.*, **47**, W587–W593.
25. Comeau,D.C., Wei,C.-H., Islamaj Doğan,R. and Lu,Z. (2019) PMC text mining subset in BioC: about three million full-text articles and growing. *Bioinformatics*, **35**, 3533–3535.
26. Kafkas,Ş., Pi,X., Marinos,N., Morrison,A. and McEntyre,J.R. (2015) Section level search functionality in Europe PMC. *J. Biomed. Semant.*, **6**, 7.
27. Paszke,A., Gross,S., Massa,F., Lerer,A., Bradbury,J., Chanan,G., Killeen,T., Lin,Z., Gimelshein,N., Antiga,L. *et al.* (2019) Pytorch: An imperative style, high-performance deep learning library. *Advances inneural information processing systems*, **32**, 8026–8037.
28. Lippincott,T., Séaghdha,D. and Korhonen,A. (2011) Exploring subdomain variation in biomedical language. *BMC Bioinformatics*, **12**, 212.
29. Rokach,L. (2010) Ensemble-based classifiers. *Artif. Intel. Rev.*, **33**, 1–39.
30. Yu,F., Seff,A., Zhang,Y., Song,S., Funkhouser,T. and Xiao,J. (2015) Construction of a large-scale image dataset using deep learning with humans in the loop. arXiv doi: https://arxiv.org/abs/1506.03365v1, 04 June 2016, preprint: not peer reviewed.
31. Yeganova,L., Islamaj,R., Chen,Q., Leaman,R., Allot,A., Wei,C.-H., Comeau,D.C., Kim,W., Peng,Y., Wilbur,W.J. *et al.* (2020) Navigating the landscape of COVID-19 research through literature analysis: a bird's eye view. arXiv doi: https://arxiv.org/abs/2008.03397, 11 September 2020, preprint: not peer reviewed.
32. Zhong,Z. and Chen,D. (2020) A frustratingly easy approach for joint entity and relation extraction. arXiv doi: https://arxiv.org/abs/2010.12812, 23 March 2021, preprint: not peer reviewed.
33. Peng,N., Poon,H., Quirk,C., Toutanova,K. and Yih,W.-t. (2017) Cross-sentence n-ary relation extraction with graph lstms. *Trans. Assoc. Comput. Linguist.*, **5**, 101–115.
34. Peters,M.E., Neumann,M., Iyyer,M., Gardner,M., Clark,C., Lee,K. and Zettlemoyer,L. (2018) Deep contextualized word representations. In: *Proceedings of NAACL-HLT*. pp. 2227–2237.
35. Adhikari,A., Ram,A., Tang,R. and Lin,J. (2019) Docbert: bert for document classification. arXiv doi: https://arxiv.org/abs/1904.08398v3, 22 August 2019, preprint: not peer reviewed.
36. Timms,A., Sathananthan,R., Bradbury,L., Athanasou,N. and Brown,M. (2002) Genetic testing for haemochromatosis in patients with chondrocalcinosis. *Ann. Rheum. Dis.*, **61**, 745–747.
37. Gracia-Ramos,A.E. (2020) Is the ACE2 overexpression a risk factor for COVID-19 infection? *Arch. Med. Res.*, **51**, 345–346.
38. Nelen,M.R., Kremer,H., Konings,I.B., Schoute,F., van Essen,A.J., Koch,R., Woods,C.G., Fryns,J.-P., Hamel,B. and Hoefsloot,L.H. (1999) Novel PTEN mutations in patients with Cowden disease: absence of clear genotype–phenotype correlations. *Eur. J. Hum. Genet.*, **7**, 267–273.