Annotation guideline of BioRED

Motivation

The development of the entity recognition and relation extraction methods is highly relying on the manually annotated corpora for the development of the deep learning and machine learning based methods. Most of the published corpora only annotate one or two concepts and/or with only one type of relation between the concepts. However, none of the existing corpora annotated the relations among multiple concepts. Therefore, we aimed to develop the BioRED corpus with multiple popular concepts and the relations among those. In this task, we annotated the relations with either strong or minor association between two entities.

Data sampling

To accelerate the corpus development and improve the quality, we established the corpus on top of the existing high-quality corpora listed in Table 1, instead of building everything from scratch. We randomly selected the articles where two concepts (e.g., chemical and disease) co-occurring in the same sentence. We equally selected the numbers of articles for all concept pairs. We set up the annotation pipeline as a cycle. In each iteration, we randomly sampled a subset with 20 pmids and shared to the curators. Once the curators annotated all entities, we compared their annotations and discussed the differences. We then shared the curators the confirmed set for relation curation. In lack of the annotation guideline of the relations, we briefly summarized the annotation rules by reading the guideline of several relation corpora (Islamaj Doğan et al., 2019; Lee et al., 2016; Li et al., 2016). Curators annotated the relations following the rules in the guideline. We compared the results and discussed the differences again once all curators finished. We summarized our conclusion and updated to the annotation guideline.

Table 1. a summary table of the public corpora

Corpus	# of abstracts	Annotation	# of sampled abstracts	Reference
BC5CDR	1500	Disease, Chemical	204	(Li et al., 2016)
NCBI Disease Corpus	792	Disease	11	(Islamaj Doğan et al., 2014)
tmVar	500	Variant	197	(Wei et al., 2013; Wei et al., 2018)
BC2GN/GNormPlus	694	Gene	48	(Morgan et al., 2008; Wei et al., 2015)
NLM-Gene	550	Gene	140	(Islamaj Doğan et al., 2021)

Annotation tool

The curation is performed using TeamTat (Islamaj Doğan et al., 2020). To accelerate the speed of the manual annotation, we applied the pre-processed annotation by PubTator (Wei et al., 2019) for those concepts that aren't covered by individual corpus. For example, we added genes, species, and variants in PubTator for the articles in BC5CDR corpus. Curators edit/delete/keep the pre-annotation based on their judgment. Curators could access other public resources (e.g., NCBI MeSH, Wikipedia) or even read the full text to clarify the concept spans and identifiers. But if curators can't find sufficient evidence in the abstract to prove the relation between two concepts, the relation wouldn't be annotated even if the evidence is in the full text.

Data format

BioRED used PubTator format (https://www.ncbi.nlm.nih.gov/research/pubtator), which is a plain-text-based format for text and text-bound annotations.

Concept types

To the better understanding, we grouped those highly relevant entities to five concepts types as shown in **Error! Reference source not found.**: (1) Gene: it includes genes, proteins, mRNA and other gene products. (2) Chemical: it contents chemical and drug. (3) Disease: it contents diseases, symptoms, and some disease-related phenotypes. (4) Variant: it contents the genomic/protein variants (including substitutions, deletions, insertions, and others). (5) Species: it includes the species in the hierarchical taxonomy of organisms. (6) CellLine: cell line.

Table 2. a summary table of the concept types and the referred sources in BioRED

Type	Example	Normalized component or identifier	Reference
Gene	ABCA1	19	NCBI Gene
Variant	S276T	p|SUB|S|276|T RS#:2234671	dbSNP
Species	Escherichia coli	562	NCBI Taxonomy
Disease	Congenital hypothyroidism	D003409	MEDIC (a combination of MESH and OMIM)
Chemical	Terbutaline	D013726	MESH (Chemicals and Drugs Category)
CellLine	MCF7/AdrR	CVCL_1452	Cellosaurus

Guideline of the entities

❖ General rules

- ❑ Annotate all the spans of all the six concept types.
- ❑ The full text can be accessed to clarify the concept spans and identifiers.
- ❑ The abbreviation and its long form should be annotated separately if possible. prostaglandin E2 (PGE2) in the text, "prostaglandin E2" and "PGE2" should be both annotated to chemicals with the same identifier (D015232).
- ❑ Annotate both the full name and abbreviation in one entity, if the boundary of the entity covers both of them. For instance, annotate "Deoxyguanosine kinase (dGK) deficiency" entirely to a disease (PMID:19394258).
- ❑ Annotate the composite entity spans (e.g., "MMP-1, -2") entirely. The identifiers of the multiple concepts should be entered separated by ‘‘," with no space (e.g., "4312,4313").
- ❑ Annotate any concept identifier span (e.g., "OMIM 307800" and "rs729302").
- ❑ Annotate the concept spans with misspelling ("HLRRC" in PMID:18366737).

❖ Gene

- ❑ The gene spans should be linked to NCBI Gene ID *(*https://www.ncbi.nlm.nih.gov/gene/*)*. Specify the corresponding species and the identifiers to the gene spans.
- ❑ Annotate the gene families including paralogs/orthologs (e.g., bone morphogenetic protein(bmp)), but DO NOT annotate the functional based families (e.g., tumor suppressor gene, protein kinase).

- ❑ Following the previous rule, if at least one of the gene family members mentioned in the text, the identifiers of those members are assigned to the family name (e.g., in "Cytochrome P-450 genes (CYP1A1, CYP2A6, CYP2D6, and CYP2E1)", Cytochrome P-450 genes should be assigned with NCBI Gene: 1543,1548,1565,1571). If no family member is in the text, the first gene reached in the gene family should be assigned (e.g., matrix metalloproteinases to NCBI Gene:4312(MMP1)).

- ❑ In experimental studies, annotate the gene identifier corresponding to the organism/species that is the source of the gene (for example mouse gene transfected into human cells is annotated to record for mouse gene).

- ❑ DO NOT annotate the genes, if they are named for cells and pathways.

❖ Disease

- ❑ The diseases refer to MEDIC which is a combination of MESH (https://meshb.nlm.nih.gov/search) and OMIM (https://omim.org/). We annotate the MESH identifier, if a disease has both the MESH and OMIM identifiers.

- ❑ If no specific disease concept can be reached in MEDIC by the annotated span, we annotate the closest hypernym concept that logically describes the disease span(e.g., X-linked retinitis pigmentosa to D012174 (retinitis pigmentosa) in PMID:17935240). Otherwise, "-" is assigned to the mismatched diseases.

- ❑ Annotate the diseased-related phenotypes but not other (e.g., short ear in PMID:17345627).

- ❑ DO NOT annotate the high-level diseases (e.g., "genetic disorder").
  Excludes:
  - ❑ Cis-acting disease (PMID:12442272)
  - ❑ genetic disorders (PMID:18046082)
  - ❑ familial disorder (PMID:20806042)

- ❑ Annotate "drug addiction" to "disease".

- ❑ Annotate the symptoms to diseases.

- ❑ DO NOT annotate overdose.

- ❑ DO NOT annotate the mental status and phenotypes to diseases unless the term is exactly recorded in MEDIC. e.g., decline in sympathetic tone (PMID:19067809)

- ❑ DO NOT annotate references to biological processes such as "tumorigenesis" or "cancerogenesis".

- ❑ Annotate minimum necessary text spans for a disease. For example, select "hypertension" instead of "sustained hypertension.", unless the span covers the prefix or the suffix can exactly match a specific concept identifier (e.g., "chronic bronchitis" to D029481).

❖ Chemical

- ❑ Chemical spans should be linked to MESH (https://meshb.nlm.nih.gov/search). "-" is assigned to the mismatched chemicals.

- ❑ Annotate the chemical abbreviation with the dose. For example "In addition, GABA content of mice hippocampus treated with GFC75 plus P400 showed an increase of 46.90% when compared with seized mice.", 75 and 400 suffixes are the doses. In this case, we annotate GFC75 and P400 as chemical spans. (PMID:24911645).

- ❑ Annotate the chemical resource for extracts, e.g., "ethanolic extract of Daucus carota seeds (DCE)" in PMID:16755009; "grape seed proanthocyanidin extract" in PMID:11334364.

- ❑ DO NOT annotate antibodies to chemicals. (PMID:17595233) (PMID:20648600)
- ❑ DO NOT annotate saline, placebo, water, and juice.
- ❑ DO NOT annotate residue, free radicals.
- ❑ DO NOT annotate staining chemical (e.g., hematoxylin in PMID:24442316)
- ❑ DO NOT annotate vaccines.

❖ Species

- ❑ The species spans should be linked to NCBI Taxonomy ID (https://www.ncbi.nlm.nih.gov/taxonomy/)
- ❑ DO NOT annotate beyond species rank (e.g., annotate homo sapiens but do not annotate mammalian which is a taxonomy class)

❖ Variant

- ❑ The annotation of the variants follows tmVar annotation guidelines, the variants should be linked to dbSNP (https://www.ncbi.nlm.nih.gov/snp) accession number (RS#) when possible. Otherwise, we annotate the variant components (e.g., wild type, location, and mutant) following the tmVar component format (e.g., V600E to "p|SUB|V|600|E").
- ❑ DO NOT annotate residue of the gene modification.
- ❑ For those variants that cannot be normalized:

  In our observations on the variant recognition/normalization dataset, around 50-60% of the variants cannot be normalized to the specific concept identifiers in dbSNP. For those, we enumerate the components of the variants. Below table listed the normalized components of the curated variant types.

Table 3. a summary table of the subtypes in SequencVariant and its normalized components

Type	Example	Normalized component or identifier
Variant on DNA sequence	c.1922G>A	c|SUB|G|1922|A
	C-to-G transition was identified at nucleotide 857	c|SUB|C|857|G
Variant on protein sequence	R114H	p|SUB|R|114|H
	methionine to threonine substitution at codon 235	p|SUB|M|235|T
RS number	rs763780	rs763780
Allele on DNA sequence	-218G	c|Allele|G|-218
Allele on protein sequence	L638	p|Allele|L|638
	stop codon at position 372	p|Allele|X|372
Nucleotide or acid change	G > C	c|SUB|G||C
	valine for glutamate	p|SUB|V||E

- ❑ The components of the different type of variants are described below:

- ❑ Substitution:
  <Sequence type>|SUB|<wild type>|<mutation position>|<mutant>
  e.g., "c.435C>G" --> "c|SUB|C|435|G"
- ❑ Deletion:
  <Sequence type>|DEL|<mutation position>|<mutant>
  e.g., "c.104delT" --> "c|DEL|104|T"
  e.g., "c.1544-?_2916+?" --> "c|DEL|1544-?_2916+?|"
- ❑ Insertion:
  <Sequence type>|INS|<mutation position>|<mutant>
  e.g., "c.104insT" --> "c|INS|104|T"
- ❑ Insertion + Deletion(indel/delins):
  <Sequence type>|INDEL|<mutation position>|<mutant>
  e.g., "c.2153_2155delinsTCCTGGTTTA" --> "c|INDEL|2153_2155|TCCTGGTTTA"
- ❑ Duplication:
  <Sequence type>|DUP|<mutation position>|<mutant>|<duplication times>
  e.g., "c.1285-1301dup" --> "c|DUP|1285_1301||"
  e.g., "c.1978(TATC)(1-2)" --> "c|DUP|1978|TATC|1-2"
- ❑ Frame shift:
  <Sequence type>|FS|<wild type>|<mutation position>|<mutant>|<frame shift position>
  e.g., "p.Val35AlafsX25" --> "p|FS|V|35|A|25"
  e.g., "p.Ser119fsX" --> "p|FS|S|119||"
- ❑ <Sequence type>:
  c: DNA sequence
  r: RNA sequence
  g: Genome sequence
  p: Protein sequence
  m: Mitochondrial sequence
- ❑ <wild type> / <mutant>:
  A,T,C,G: DNA nucleotide
  C,I,S,Q,M,N,P,K,D,T,F,A,G,H,L,R,W,V,E,Y,X: Amino acid

❖ CellLine

- ❑ Annotate cell line identifiers to the "species of origin" of the concept page in Cellosaurus (https://web.expasy.org/cellosaurus/). "-" is assigned to the mismatched cell lines.

❖ Conflict cases among different concept types

- ❑ [Gene/Disease] If the disease and gene share the same name, e.g., adrenoleukodystrophy (ALD) gene, annotate the concept following the below cases:
  - Adrenoleukodystrophy gene > GeneOrGeneProduct
  - Adrenoleukodystrophy patient > DiseaseOrPhenotypicFeature
  - ALD gene > GeneOrGeneProduct
  - ALD patient > DiseaseOrPhenotypicFeature
  - adrenoleukodystrophy (ALD) gene > GeneOrGeneProduct
  - adrenoleukodystrophy (ALD) patient > DiseaseOrPhenotypicFeature

- ❑ Annotate the concept with longer spans. For instance, Deoxyguanosine kinase (dGK) deficiency (MESH:C580039) instead Deoxyguanosine kinase (in PMID:19394258).

- ❑ Annotate the prefix of the corresponding species to the gene span separately (e.g., "human leukocyte antigen (HLA)-DRB1" to "human" to species and "leukocyte antigen (HLA)-DRB1" to gene).

- ❑ Annotate the virus spans (e.g., HIV-1 and HBV) to species. But annotate the infection of the virus to disease (e.g., "HIV-1 infected" maps to MESH: D015658).

- ❑ Some of the proteins can be chemicals for particular usage or purpose. We annotate the organic compounds to proteins. While, we annotate the proteins or molecules as the biologic medical products to treat diseases .

- ❑ Annotate the alleles to variants rather than chemicals, DO NOT annotate the amino acids to chemicals.

- ❑ The rule for antagonist/agonist/blocker/inhibitor: If the sentence mentions a drug/chemical and then tells the target gene, we annotate both the chemical and the gene.
    - e.g., "However, successful anticoagulation was achieved by administration of direct thrombin inhibitor, argatroban." in PMID:16000134. We annotate Negative_Correlation for thrombin and argatroban.
- ❑ Following the previous rule, we annotate the mention to a gene if the inhibitor/blockage/agonist is part of the gene name.
    - e.g., "plasminogen activator inhibitor-1" in PM ID:16167916
- ❑ Following the previous rule, if "<gene> inhibitors/agonists" is mentioned in the context of a family of compounds, without mentioning a specific drug, We annotate the entire mention ("<gene> inhibitors/agonists") as a chemical.
    - "Methadone dose, presence of cytochrome P-450 3A4 inhibitors" in PM ID:16801510
- ❑ We don't annotate functions of genes that are described as inhibitors or activators of processes.
    - We don't annotate "coagulation inhibitors" in PMID:17003923

❖ Others

- ❑ Annotate spans with morphological variations such as adjectives. For instance, annotate "hypertensive" to "hypertension", "psychiatric" to "Mental Disorders" and "VLCAD-deficient" to "VLCAD-deficiency".

- ❑ The suffixes of the gene/protein (e.g., "kinase" of "dGK kinase", "gene" of "ESR1 gene") are ignored unless the meaning is changed without the suffix (e.g., "gene" in "Becker muscular dystrophy gene").

- ❑ DO NOT annotate the general terms such as: disease, syndrome, deficiency, complications, gene, drug, protein, nucleotide, etc. However, the disease terms such as pain, cancer, tumor, and death should be retained.

Guideline of the relation pairs

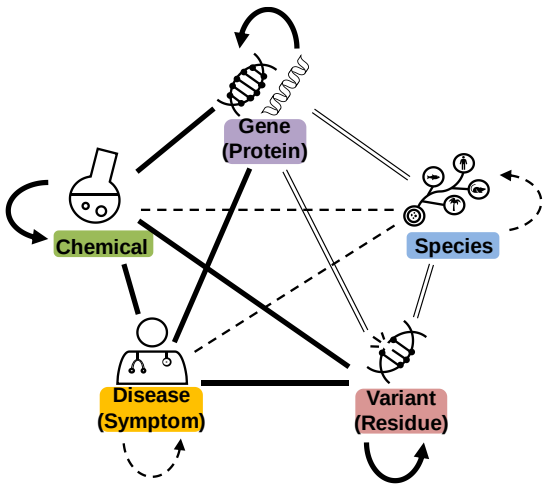

Figure 1. Categorized relations between concepts. The patterns of the lines between the concepts present the categories: (▬) Popular associations: The concept pairs are frequently discussed in the biomedical literature. (═) Corresponding gene/species: The corresponding gene (or species) can simply be recognized during the variant (or gene) identifier mapping. (---) Rarely discussed associations: Some other relation types are rarely discussed in the biomedical text.

❖ General rules

- ❑ We only annotate the (1) Popular associations and the (2) Corresponding gene/species in Figure 1.

- ❑ Co-occurrence in the same sentence is not required.

- ❑ DO NOT annotate a confirmed irrelevance. (e.g., "Nonetheless, NBS1 gene heterozygosity is not a major risk factor for lymphoid malignancies in childhood and adolescence" in PMID:16152606)

- ❑ Annotate the associations with explicit statements in the text.

- ❑ The full text is NOT allowed to access. DO NOT annotate if the statement in the abstract is not clear. (e.g., PMID:14722929)

- ❑ For Co-treatment, we annotate "Negative_Correlation" between the two chemicals and the disease, and a "Cotreatment" relation between the two chemicals. For an example in PMID:16720068, "Possible neuroleptic malignant syndrome related to concomitant treatment with paroxetine and alprazolam.".

- ❑ For Drug_Interaction, we annotate "Positive_Correlation" between the two chemicals and the disease, and a "Drug_Interaction" relation between the two chemicals.

❖ Disease-chemical

- ❑ Annotate the association of chemicals and disease due to the "toxicity" and "drug-induced", but not annotate the chemicals with the disease concept-"toxicity" directly.
    - PMID:19728177
      "RESULTS: Our patient was admitted to the MICU after being found unresponsive with presumed toxicity from acetaminophen which was ingested over a 2-day period. ... In patients with FHF and cerebral edema from acetaminophen overdose, prolonged therapeutic hypothermia could potentially be used as a life saving therapy and a bridge to hepatic and neurological recovery."
      We annotate the associations between acetaminophen and FHF/cerebral edema due to the toxicity, but we do not annotate acetaminophen and toxicity.

- ❑ Annotate the relation between the target disease and the measure level of the molecules/chemicals in blood test with correlation, "Compared with control patients, CRF and ESRD patients had higher preoperative serum creatinine levels, a greater percentage of patients with hepatorenal syndrome.", the pairs of "CRF, creatinine" and "ESRD, creatinine" is annotated. (PMID:11773892)

- ❑ DO NOT annotate the association of disease to the staining chemicals. For instance, "The extent of neuronal injury was determined by 2,3,5-triphenyltetrazolium staining." (PMID:1711760) and "Renal lesions were analyzed in hematoxylin and eosin, periodic acid-Schiff, and Masson's trichrome stains. SRL-treated rats presented proteinuria and NGAL (serum and urinary) as the best" (PMID:24971338)

- ❑ DO NOT annotate the association of any disease with the chemical concept "analgesia" in "patient-controlled analgesia (PCA) pump". For instance, "patient-controlled analgesia (PCA) pump" (PMID:9672936)

- ❑ DO NOT annotate toxicity with its chemical. For instance, do not annotate "unresponsive with presumed toxicity from acetaminophen" (PMID:19728177)

- ❑ Annotate the chemicals which prevents the chemical-induced disease (PMID:24971338+18503483)

❖ Disease-gene

- ❑ DO NOT annotate the prognostic factor of genes or its variants to the target disease.

- ❑ Annotate the negative correlation (e.g., knockdown gene to the target disease). For instance, annotates Gankyrin and HCC in "Gankyrin expression in the tumor microenvironment is negatively correlated with progression-free survival in patients undergoing sorafenib treatment for HCC." (PMID:28777492)

- ❑ DO NOT annotate the symptoms of the target disease to the correlated variants or genes, unless the clear statement between the symptom and variant is provided.

❖ Disease-variant & Disease-gene

- ❑ Annotate the corresponding diseases with the variant when it is observed in patients.
- ❑ Annotate the diseases with the associated genes.
- ❑ Annotate the corresponding gene of the variant with the variant-associated disease.
- ❑ Annotate the association between the variants (and corresponding gene) and the diseases belonging to the target disease. For instance, annotate the association between "autosomal recessive disease" and "272gly----stop" (PMID:1671881).
- ❑ Annotate the minor association (e.g., the observation of the association on few patients).
- ❑ Annotate the associations in the particular races.
- ❑ Annotate the association between disease and variant if the inheritance of the genetic disease is confirmed by the variant.
- ❑ DO NOT annotate the symptoms of the target disease to the correlated variants or genes, unless the clear statement between the symptom and variant is provided. (ex in PMID: 1952108)

❖ Gene-gene

- ❑ Annotate the pairs of the members in the protein complex, e.g., EPO/EPOR (PMID:22808010).
    - In "We cloned the cDNA of three remaining human NADH:ubiquinone oxidoreductase subunits of this IP fraction: the NDUFS2 (49 kDa), NDUFS3 (30 kDa), and NDUFS6 (13 kDa) subunits.", all the pairs between two of the NDUFS2, NDUFS3, and NDUFS6 are annotated. (PMID:9647766)
- ❑ DO NOT annotate the "is_a" association between two genes. For instance, DO NOT annotate "breast and ovarian cancer gene" and "BRCA1" (PMID:8944024).
- ❑ Annotate protein-protein interaction for singling proteins and the receptors. For instance, "Epo-R signaling proteins (Akt, STAT5, p70s6k, LYN, and p38MAPK)" (PMID:27640183)

❖ Gene-chemical

- ❑ Annotate the chemical with the gene receptor. For instance, "antipsychotic drugs that have a high affinity with the D2 receptor" (PMID:16867246)

❖ Chemical-chemical

- ❑ DO NOT annotate the "is_a" association between two chemicals. For instance, "Nefiracetam is a novel pyrrolidone derivative" (PMID:8829135)
- ❑ Annotate the chemical-chemical association for the chemical contributes to the other for treatment, side effects and drug-drug interaction. In "the midline serotonin B3 cells in the medulla contribute to the hypotensive action of methyldopa", the pair of serotonin and methyldopa is annotated. (PMID:2422478)

❖ Chemical-variant

- Annotate the chemical with the variant, if the chemical binding ability of the gene impaired by the variant. For instance, “This variant (apoE Guangzhou) may cause a marked molecular conformational change of the apoE and thus impair its binding ability to lipids.” (PMID:18046082)

- Annotate the chemical with the variant, if different mRNA stability between different alleles affected by the chemical. For instance, “After transfection and inhibition of transcription with actinomycin D, analysis of mRNA turnover failed to reveal differences in mRNA stability between A118 and G118 alleles, indicating a defect in transcription or mRNA maturation." (PMID:16046395)

❖ Others

- DO NOT annotate the pairs of concepts with negative conclusions (e.g., “not significant”, “no correlation was observed”). For instance, “For most of the variants identified in the Kenyan and Sudanese study population, a causative association with NSARD appears to be unlikely” is not annotated.

- In the case that a disease is induced by a chemical, and the other concepts (e.g., chemical, or protein) treats/affects the induced disease, we annotate the three pairs by following examples. For instance:
    - “In addition, there is convincing clinical evidence that monotherapy with continuous subcutaneous apomorphine infusions is associated with marked reductions of preexisting levodopa-induced dyskinesias.” (PMID:11009181)
        - Positive_Correlation between levodopa and dyskinesias
        - Negative_Correlation between apomorphine and dyskinesias
        - Negative_Correlation between apomorphine and levodopa
    - Absence of PKC-alpha attenuates lithium-induced nephrogenic diabetes insipidus. (PMID:25006961)
        - Positive_Correlation between lithiumand nephrogenic diabetes
        - Positive_Correlation between PKC-alpha and dyskinesias
        - Positive_Correlation between PKC-alpha and lithium
    - Characterization of a novel BCHE "silent" allele: point mutation (p.Val204Asp) causes loss of activity and prolonged apnea with suxamethonium. (PMID:25054547)
        - Positive_Correlation between p.Val204Asp and apnea
        - Negative_Correlation between apnea and suxamethonium
        - Association between p.Val204Asp and suxamethonium
        - Association between BCHE and apnea
        - Association between BCHE and suxamethonium
    - Curcumin prevents maleate-induced nephrotoxicity (PMID:25119790)
        - Negative_Correlation between Curcumin and nephrotoxicity
        - Positive_Correlation between maleate and nephrotoxicity
        - Negative_Correlation between Curcumin and maleate

- Annotate the previous reported association, even the conclusion demonstrates the association is not significant. (PMID:15824163)

Guideline of the relation types

We collected a list of relation types between two concepts (e.g., upregulation between two genes). Further, we merged some of the relation types to narrow down the curation complexity and increase the number of instances in each relation type.

Table 4. A mapping of the directional and nondirectional relation types

Concept1	Concept2	Relation type (Directional)	Relation type (Nondirectional)
Gene	Gene	Upregulation	Positive_Correlation
Gene	Gene	Downregulation	Negative_Correlation
Gene	Gene	Regulation	Association

Gene	Gene	Positive_Correlation	Positive_Correlation
Gene	Gene	Negative_Correlation	Negative_Correlation
Gene	Gene	Bind	Bind
Gene	Gene	Modification	Association
Gene	Gene	Association	Association
Chemical	Gene	(C->G) Exhibition (G->C) Response	Positive_Correlation
Chemical	Gene	(C->G) Suppression (G->C) Resistance	Negative_Correlation
Chemical	Gene	Association	Association
Chemical	Gene	Receptor	Bind
Chemical	Gene	Chem_Motification	Association
Chemical	Variant	Association	Association
Chemical	Variant	Resistance	Negative_Correlation
Chemical	Variant	Reponse (and sensitivity)	Positive_Correlation
Gene	Disease	Positive_Correlation	Positive_Correlation
Gene	Disease	Negative_Correlation	Negative_Correlation
Gene	Disease	Association	Association
Variant	Disease	Cause	Cause
Variant	Disease	Association	Association
Chemical	Disease	Treatment	Negative_Correlation
Chemical	Disease	Induce	Positive_Correlation
Chemical	Disease	Association	Association
Chemical	Chemical	Cotreatment	Cotreatment
Chemical	Chemical	Inhibition	Negative_Correlation
Chemical	Chemical	Increase	Positive_Correlation
Chemical	Chemical	Drug_Interaction	Drug_Interaction
Chemical	Chemical	Association	Association
Chemical	Chemical	Comparison	Comparison

❖ Disease-chemical

★ Positive Correlation

- ❑ Chemical-induced disease.
- ❑ Chemical (or Higher dose of the chemical) causes a higher risk of the disease.
- ❑ Disease causes the increase of the chemical measured level.
- ❑ The level of the chemical and the risk of the disease present a positive correlation.
- ❑ Chemical exposures during development alter disease susceptibility later in life.

★ Negative Correlation

- ❑ The disease-treated chemical/drug.
- ❑ Disease causes the decrease of the chemical measured level.

- ❑ The chemical/drug drops down the susceptibility of the disease.

★ Association

- ❑ A safety drug of the potential disease (PMID:20722491 - capecitabine(C110904) - hepatic and renal dysfunctions(D008107|D007674))
- ❑ The associations of the pairs which cannot be categorized to positive/negative correlation.
- ❑ The associations without clear description.

❖ Disease-gene

- ❑ If an association between a variant and a disease is confirmed, the corresponding gene should associate with the disease by "Association" type.

★ Positive Correlation

- ❑ The overdose of protein causes the disease.
- ❑ The knockout gene prevents the disease.
- ❑ The side effect of protein (drug) causes the disease

★ Negative Correlation

- ❑ The protein (drug) is used to treat/prevent the disease.
- ❑ Lack of the protein causes the disease.
- ❑ The knockout gene causes the disease.

★ Association

- ❑ The associations of the pairs which cannot be categorized to previous relation types.
- ❑ The associations without clear description.
- ❑ The functional gene prevents the occurrence of the disease.
- ❑ Protein deficiency.

❖ Disease-variant

★ Positive Correlation

- ❑ The variant increases the risk of the disease.
- ❑ Significant frequency (p-value) of the disease with the specific allele.

- ❑ The variant causes the gene to be either over-express or non-functional and further causes disease (or raises the disease susceptibility).
- ❑ Disease is caused by protein deficiency, and the variant is responsible for the deficiency.
- ❑ The variant plays a role in genetic predisposition to the disease.
- ❑ The variant has a significant contribution to the disease.
- ❑ Annotate "Positive_Correlation" to the pair of "founder mutation" and the target disease. (PMID:10788334,19394258)

★ Negative_Correlation

- ❑ The variant decreases the risk of the disease.

★ Association

- ❑ The variant observed from a number of patients.
- ❑ The variant is associated with a lower prevalence of the disease or is responsible for the lower disease susceptibility
- ❑ The associations of the pairs which cannot be categorized to "Cause" relation type.
- ❑ The associations without clear description.

❖ Gene-gene

★ Positive_Correlation

- ❑ Two genes present the positive correlation in gene expression results.
- ❑ Gene A is a transcription factor of the gene B, and gene A upper regulates the gene B.
- ❑ Two genes present a positive correlation in any way.

★ Negative_Correlation

- ❑ Two genes present the negative correlation in gene expression results.
- ❑ Gene A is a transcription factor of the gene B, and gene A down regulates the gene B.
- ❑ Two genes present a negative correlation in any way.

★ Bind

- ❑ physical interaction between two proteins.
- ❑ Protein A binds the promoter of gene B.
- ❑ Two or more proteins in a complex.

- ❑ Annotate “bind” to the protein and its protein receptor. ("androgen receptor" and "androgen" in PMID:15599941)
- ❑ If the binding causes a positive or negative correlation, annotate the association to Positive_Correlation/Negative_Correlation.
- ❑ DO NOT annotate the protein bind on a gene promoter region.

★ Association

- ❑ Annotate the modification (e.g, phosphorylation, dephosphorylation, acetylation ,deacetylation and other modifications) to association.
- ❑ The associations of the pairs which cannot be categorized to any other association types.
- ❑ The associations without clear description.

❖ Gene-chemical

- ❑ If an association between a variant and a chemical is confirmed, the same association type should be assigned to the corresponding gene of the variant and the chemical.

★ Positive_Correlation

- ❑ The chemical causes a higher expression of the gene.
- ❑ Higher gene expression causes higher sensitivity of the chemical.
- ❑ The variant triggers the chemical adverse effects or causes the side effect worse.
- ❑ The gene causes the chemical over-response.
- ❑ Two genes present a positive correlation in any way.

★ Negative_Correlation

- ❑ The chemical causes a lower expression of the gene.
- ❑ Higher gene expression causes lower sensitivity (resistance) of the chemical, and vise versa.
- ❑ The variant has a protective role for the development of the chemical adverse effects.
- ❑ The gene causes the chemical resistance.
- ❑ Two genes present a negative correlation in any way.

★ Association

- ❑ The associations of the pairs which cannot be categorized to any other association types.
- ❑ The associations without clear description.

★ Bind

- ❑ A chemical binds the promoter of a gene.
- ❑ A protein is the chemical receptor.
- ❑ If the binding causes a positive or negative correlation, annotate the association to Positive_Correlation/Negative_Correlation.

❖ Chemical-chemical

- ★ Positive_Correlation (between A and B)
- ❑ Chemical A increases the sensitivity of the chemical B.
- ❑ Chemical A increases the treatment/inducing effectiveness of the chemical B to a disease.
- ❑ Chemical A increases the effectiveness of the gene activation caused by chemical B.
- ★ Negative_Correlation (between A and B)
- ❑ Chemical A decreases the sensitivity of the chemical B.
- ❑ Chemical A decreases the treatment/inducing effectiveness of the chemical B to a disease.(PMID:25080425 betaine attenuates isoproterenol-induced acute myocardial injury in rats.)
- ❑ Chemical A decreases the side effects of the chemical B. (PMID:24587916 LF(D007781) - Dexamethasone(D003907))
- ❑ Chemical A decreases the gene activation caused by chemical B. (PMID:17035713 Caspase activation by cisplatin was inhibited by CAA)
- ★ Association
- ❑ Annotate the chemical conversion to association type. (PMID:17391797 that catalyzes the conversion of phosphatidylethanolamine to phosphatidylcholine.)
- ❑ The associations of the pairs which cannot be categorized to any other association types.
- ❑ The associations without clear description.
- ★ Drug_Interaction
- ❑ A pharmacodynamic interaction occurs when two chemicals/drugs are given together.
- ★ Cotreatment
- ❑ Combination therapy.
- ★ Conversion
- ❑ A chemical converse to the other chemical.

❖ Chemical-variant

★ Positive Correlation

❑ The chemical causes a higher expression of the gene because of the specific variant.

❑ The variant causes higher sensitivity of the chemical.

❑ The variant causes the chemical over-response.

❑ The gene and the chemical present a positive correlation in any way.

★ Negative Correlation

❑ The chemical causes a lower expression of the gene because of the specific variant.

❑ The variant causes lower sensitivity of the chemical.

❑ The variant causes the chemical resistance.

❑ The gene and the chemical present a negative correlation in any way.

❑ The variant presents a protective role to the chemical adverse effect (or chemical caused disease).

★ Association

❑ The associations of the pairs which cannot be categorized to positive/negative correlation.

❑ The associations without clear description.

❑ variant located on a chemical specific binding site, e.g., the sequence variant c.465G>T encodes a conservative amino acid substitution, p.Glu155Asp, located in EF-hand 4, the calcium binding site of GCAP2 protein. (PMID:21405999)

Guideline of the novel triage

Each relation should be assigned one of two content bins:

- "Novel" is used for relations that are related to the main point or novelty of the abstract. Any information that would be part of the results or conclusions of the paper is considered novel.
- "No" is for relations that are background information, typically providing context for the abstract, such as results of previous studies or relevant details that are needed to understand why the paper is important.

Reference

Islamaj Doğan, R., Kim, S., Chatr-Aryamontri, A., Wei, C.-H., Comeau, D. C., Antunes, R., et al. (2019). Overview of the BioCreative VI Precision Medicine Track: mining protein interactions and mutations for precision medicine. *Database, 2019*. https://doi.org/10.1093/database/bay147

Islamaj Doğan, R., Kwon, D., Kim, S., & Lu, Z. (2020). TeamTat: a collaborative text annotation tool. *Nucleic acids research, 48*(W1), W5-W11. https://doi.org/10.1093/nar/gkaa333

Islamaj Doğan, R., Leaman, R., & Lu, Z. (2014). NCBI disease corpus: a resource for disease name recognition and concept normalization. *Journal of Biomedical Informatics, 47*, 1-10. https://doi.org/10.1016/j.jbi.2013.12.006

Islamaj Doğan, R., Wei, C.-H., Cissel, D., Miliaras, N., Printseva, O., Rodionov, O., et al. (2021). NLM-Gene, a richly annotated gold standard dataset for gene entities that addresses ambiguity and multi-species gene recognition. *Journal of Biomedical Informatics, 118*, 103779. https://doi.org/10.1016/j.jbi.2021.103779

Lee, K., Lee, S., Park, S., Kim, S., Kim, S., Choi, K., et al. (2016). BRONCO: Biomedical entity Relation ONcology COrpus for extracting gene-variant-disease-drug relations. *Database, 2016*. https://doi.org/10.1093/database/baw043

Li, J., Sun, Y., Johnson, R. J., Sciaky, D., Wei, C.-H., Leaman, R., et al. (2016). BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database, 2016*, baw068. https://doi.org/10.1093/database/baw068

Morgan, A. A., Lu, Z., Wang, X., Cohen, A. M., Fluck, J., Ruch, P., et al. (2008). Overview of BioCreative II gene normalization. *Genome biology, 9*(2), 1-19. https://doi.org/10.1186/gb-2008-9-s2-s3

Wei, C.-H., Allot, A., Leaman, R., & Lu, Z. (2019). PubTator central: automated concept annotation for biomedical full text articles. *Nucleic acids research, 47*(W1), W587-W593. https://doi.org/10.1093/nar/gkz389

Wei, C.-H., Harris, B. R., Kao, H.-Y., & Lu, Z. (2013). tmVar: a text mining approach for extracting sequence variants in biomedical literature. *Bioinformatics, 29*(11), 1433-1439. https://doi.org/10.1093/bioinformatics/btt156

Wei, C.-H., Kao, H.-Y., & Lu, Z. (2015). GNormPlus: an integrative approach for tagging genes, gene families, and protein domains. *BioMed research international, 2015*, 918710. https://doi.org/10.1155/2015/918710

Wei, C.-H., Phan, L., Feltz, J., Maiti, R., Hefferon, T., & Lu, Z. (2018). tmVar 2.0: integrating genomic variant information from literature with dbSNP and ClinVar for precision medicine. *Bioinformatics, 34*(1), 80-87. https://doi.org/10.1093/bioinformatics/btx541