Mining the literature for genes associated with placenta-mediated maternal diseases

Laritza M. Rodriguez, MD, PhD, Stephanie M. Morrison, MPH, Kathleen Greenberg, PhD, Dina Demner Fushman, MD, PhD
Lister Hill National Center for Biomedical Communications, National Library of Medicine, National Institutes of Health, Bethesda, MD

Abstract

Automated literature analysis could significantly speed up understanding of the role of the placenta and the impact of its development and functions on the health of the mother and the child. To facilitate automatic extraction of information about placenta-mediated disorders from the literature, we manually annotated genes and proteins, the associated diseases, and the functions and processes involved in the development and function of placenta in a collection of PubMed/MEDLINE abstracts. We developed three baseline approaches to finding sentences containing this information: one based on supervised machine learning (ML) and two based on distant supervision: 1) using automated detection of named entities and 2) using MeSH. We compare the performance of several well-known supervised ML algorithms and identify two approaches, Support Vector Machines (SVM) and Generalized Linear Models (GLM), which yield up to 98% recall precision and F1 score. We demonstrate that distant supervision approaches could be used at the expense of missing up to 15% of relevant documents.

Introduction

The placenta is the most important organ in human pregnancy. It plays the role of lungs and kidneys for the developing fetus, supplies substrates for its development, and regulates complex immune functions to allow the cohabitation of two different organisms - the mother and the fetus- during the pregnancy. Defects in the placentation process are known to be associated with a wide range of pregnancy related complications such as preeclampsia, uterine growth restriction, and premature rupture of membranes, fetal growth retardation, placenta abruption, spontaneous abortion, and fetal death. The association between defects in the placenta, the placentation process and most of the above-mentioned diseases was established in the 1950's. More recently additional associations between the functionality of the placenta and other maternal/fetal diseases have become evident. For example, in gestational diabetes, adipokins, leptin and TNFα produced in placenta are implicated in gestational insulin resistance and possibly in insulin resistance in the adult life of the fetus through a process known as "fetal programming". The functionality of the placenta influences maternal and fetal health and development and impacts future health for both mother and baby, not only during pregnancy but for the lifetime of both (1) (2) (3) (4). The study of the organ in real time using invasive procedures to sample the placenta and the placenta insertion site can be dangerous for the mother and the fetus, even with the use of guided imaging technology. Safer study approaches during pregnancy have only become available recently with the advancement of genetic and molecular sciences. The full potential of now-available biomedical research tools has not yet been applied to the study of the placenta. Recognizing the importance of multidisciplinary collaboration for the study of the placenta, the National Institutes of Health (NIH) through the National Institute of Child Health and Human Development (NICHD) established the Human Placenta Project (HPP) "to understand the role of the placenta in health and disease and to develop new tools to learn how it develops and functions throughout the pregnancy" (2) .

Discovery of gene pathways and biochemical mechanisms that help explain disease causality can be facilitated by automated extraction of relevant information from the biomedical literature (5). One essential step in understanding complex pathways has been the compilation of disease-specific gene candidates extracted from the literature. Data analysis then makes it possible to create candidate gene assays and translate the results to the wet bench for biochemical research (6).

Our long-term goal is to extract genes, gene pathways, biomarkers, and related events from the human placenta literature to create a specialized human placenta gene repository and to identify pathways that can uncover target genes and gene therapies for pregnancy-related diseases. In this study, we present a machine learning approach to extract and identify sentences describing gene-disease relationships and gene and protein activity from a collection of studies in the placenta literature. We also make the collection of PubMed® (7) abstracts annotated for this study publicly available. In addition, we evaluate if the manually annotated relations among genes, proteins and diseases could have been captured by the rules based on MeSH® indexing – human annotation of MEDLINE® abstracts by NLM indexers.

Background

Perch and Altman point out that all important biomedical knowledge is described in the published research literature, but to operationalize this knowledge we need computational algorithms that can efficiently extract, aggregate, annotate, and store information from the raw text (8).

To develop approaches for the automated extraction of genes and gene pathways studied in the human placenta, it is necessary to identify specific genes, their mechanisms of action, and related diseases in the human placenta literature. Having information on already-identified cause/effect relationships between genes and diseases should expedite the extraction process. Extraction of this information from the literature will allow the discovery of novel relationships as new genes are identified in the human placenta. Some hereditary aspects of such associations have already been explored in the clinical setting, where women who have developed pregnancy-induced hypertension or gestational diabetes appear to have a higher incidence and family history of chronic hypertension and Type II diabetes, respectively (1). Similar approaches to automated extraction of information from the literature have been useful in Alzheimer disease research: literature-based gene enrichment and prioritization tools were used for the discovery of novel genes related to Alzheimer's disease (AD) (9) and protein-protein interaction networks constructed using data from literature sources, among others, were advantageous in providing objective prioritization of disease-gene candidate criteria (10).

The understanding of biological systems requires not only the ability to extract entities such as cells, proteins, genes, diseases, etc., but also to establish associations between different identified entities and cause/effect relationships (11). It is necessary to understand the role of each entity in a given relationship to establish a cause/effect association. Reports in the literature and shared natural language processing tasks have focused on DNA methylation (12), protein modifications (13), and gene expression mentions in anatomical locations (14), among others. Automatic sentence classification has been used to support evidence based medicine (15) , for automatic extraction of clinically useful sentences in clinical evidence resources (16), and for automatic retrieval of abstracts on randomized controlled trials (17).

The goal of the current study is to develop approaches to facilitate automatic text identification of gene associations, their mechanisms of action, and their effects on diseases by employing automated extraction of entities, their roles, and the relationships reported in the human placenta literature. In this work, we present the annotated collection of documents needed for supervised machine learning approaches for extraction of the above information and a comparison of supervised and distantly supervised approaches for extraction of contexts that contain this information.

Methods

In our previous short communication **(**18**)**, we described the dataset obtained for the study. In this section, we summarize information about the data and focus on the information extraction methods.

Data

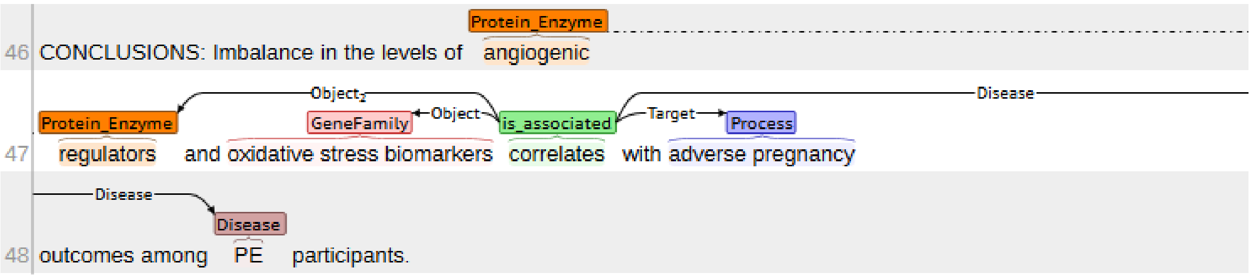
The corpus was downloaded from PubMed on December 8, 2015, searching for: placenta AND (human OR woman OR women) AND (gene OR genes OR biomarker* OR polymorphism* OR enzyme*) AND (gestational diabetes OR hypertension OR preeclampsia OR pre-eclampsia OR eclampsia OR SGA OR growth restriction OR preterm OR HELLP OR acute fatty liver OR DVT OR anemia OR placenta abruption OR placenta previa OR stillbirth OR miscarriage) in the articles that have abstracts. The terms preeclampsia, hypertension, and other pregnancy and fetus-related diseases were included in the search because they are the most common causes of maternal-fetal morbidity in pregnancy. The search retrieved 428 MEDLINE citations. The search was specifically aimed at retrieving human placenta-related studies; however, the manual review revealed that only 300 documents referred to human placenta, while the remaining 128 were human placenta gene studies in animal models. The 128 documents with mentions of gene studies in animal placentas were excluded from the study.

Manual Annotation

We set out to annotate gene and protein activity events and all biomarkers, genes, and disease entities involved in the events, along with the roles the genes play in the events. We based our annotation schema on those developed for event annotation in the GENIA corpus (19) and for gene-drug relationships extraction (20). Table 1 presents the entities and events annotated in our corpus, along with their counts. Events were marked as negated or speculation when clear negation or hedging was stated in the text. To differentiate between genes and proteins with the same name, we annotated entity mentions involved in *increased* or *decreased levels* or *amount of* events as *protein*. We annotated mentions involved in *increased* or *decreased expression* or *activity of* events as *gene*. We used BRAT online annotation tool (21) installed in a secured server and configured with the annotation schema designed for this study.

The four authors all with biomedical informatics experience and training--two physicians (DF and LR), a biologist (SM), and a cellular and molecular scientist (KG)--annotated the first 20 documents individually, reconciled the differences together, and finalized the guidelines for annotation. The remaining documents were annotated individually by two annotators each, and the differences were reconciled in pairs. We computed inter-annotator agreement using F1-score, as proposed by Hripcsak and Rothschild (22). The inter-annotator agreement between the pairs of annotators was fair. The agreement between the clinician and the geneticists was higher (58.5% on average), than between the geneticists whose F1-score was 40.3%. The better agreement between the clinician and the geneticists could be explained by learning, as for annotating the second half of the documents we paired the geneticists with the clinician, so the agreement between the geneticists was measured earlier in the process. The final modest scores reflect the difficulty of the task. The gold standard should not be affected by the modest initial agreement, as all differences were discussed and carefully reconciled in group meetings.

The resulting collection of annotated documents was then used for machine learning experiments to extract and identify sentences containing information about biomarkers and the activity of genes and proteins in the human placenta that are associated with diseases in different stages of pregnancy. Figure 1 shows an example of a fully annotated sentence.



CONCLUSIONS: Imbalance in the levels of angiogenic regulators and oxidative stress biomarkers correlates with adverse pregnancy outcomes among PE participants.

Figure 1: Annotation of protein activity mention associated with disease in human placenta study. In this example, *correlates* is an event of type *is_associated*, and *angioneic regulators* and *oxidative stress biomarkers* are entities of the type *protein_enzyme* and *gene_family* respectively. These entities have the roles of objects of the *correlates* event.

We explored two distant supervision approaches on the same collection as an alternative to manual annotation. We explored using the NLM PubTator web-based tool (23) to automatically annotate entity types: *Disease*, *Species*, *Mutation*, and *Gene*. We then compared the use of these automatically derived annotations as features in machine learning to manual annotation-based features.

We also explored exploiting MeSH indexing. Jimeno Yepes et al. demonstrated the use of MeSH headings and subheadings in extraction of gene-disease relationships (24). In the MeSH indexing process, NLM indexers assign headings (terms that describe the topic of the document) and subheadings (qualifiers) from the MeSH controlled vocabulary to MEDLINE citations. The subheading creates coordination among the headings, e.g., in the document PMID: 25305692 the indexing terms Pre-Eclampsia/genetics* and 14-3-3 Proteins/genetics* potentially indicate involvement of 14-3-3 Proteins in Pre-Eclampsia (disease); the asterisk indicates that both headings are the major topics in the article. For each one of the 300 documents we: 1) extracted indexing terms with the same subheadings; 2) derived indexing-based rules that establish associations among entities involved in the events presented in Table 1 and roughly correspond to these events; and 3) evaluated the relations against our gold standard as described below. Note that this approach is complementary to the initial PubMed retrieval strategy, which is recall-oriented and designed to find as many potentially relevant articles as possible. The MeSH indexing rules are geared towards finding specific relations in the set of the retrieved articles.

The following rules presented in the form *subheading* (*Heading type*, *Heading type*) capture associations corresponding to the events. For example, the rule *genetics (disease, chemical)* corresponds to association and finer grained events involving genes and proteins. In document PMID 11766889 this rule captures *genetics (Pre-Eclampsia, HLA Antigens)*, which approximately captures the *decreased_level_of* event manually annotated in "A deficit in levels of the HLA-G3 transcript was observed in mild pre-eclampsia compared to normal placentas."

- enzymology(disease, anatomy)
- enzymology(anatomy, cellular structure)
- biosynthesis(chemical, chemical)
- genetics(disease, chemical)
- genetics(disease, cellular structure)
- genetics(Physiological Phenomena, disease)
- immunology(anatomy, disease)
- immunology (Physiological Phenomena, disease)
- metabolism(disease, chemical)
- metabolism(anatomy, disease)
- metabolism(Physiological Phenomena, disease, chemical)
- pathology| physiopathology (disease, anatomy)
- pathology| physiopathology (anatomy, chemical)
- physiology (anatomy, chemical)
- physiology (Physiological Phenomena, chemical)
- physiology (Physiological Phenomena, anatomy)

We evaluated the correspondence between the associations captured by the above rules and the manually annotated gold standard (GS) as follows: 1) the MeSH-based association corresponds to at least one event in the GS (True Positives), 2) the events in the GS are not captured in MeSH indexing (False Negatives), 3) No events related to a disease are annotated in GS and the rules produced no associations (True Negatives), 4) A MeSH-based association has no corresponding events in GS (False Positive).

Machine Learning

We used R statistical software core package for data preprocessing and machine learning experiments, and the packages RTextTools, E1071, and MxNet for the machine learning experiments (25) (26) (27).

In the machine learning (ML) experiments, we used sentences with entities and events (963), sentences with only entity mentions (246), and a random sample of 200 sentences with no entity mentions or events. We used WordtoVec filter to process the sentences eliminating English stop words, numbers, punctuation, and extra white spaces.

For classification purposes we used supervised machine learning algorithms known to perform well on text with settings briefly described below. We used a stepwise gradual approach adding features to measure performance with each of the ML algorithms. The features used for the stepwise approach were: bag of words, manually annotated events, entities, roles and attributes (negation and speculation). To test the distant supervision approach based on automatic annotation of entities, we used entities identified using PubTator, PubTator types and PubTator codes.

We used the R implementation of *Support Vector Machine* (SVM) (28) based on LibSVM with default settings for the radial kernel.

Glmnet: an R package that fits a generalized linear model via penalized maximum likelihood (29). We used the default settings with alpha set to 0.2, which makes it closer to ridge penalty.

LogitBoost: is an algorithm in R adapted to add a convex optimization to Boosting (30) that uses decision stumps, one node decision trees as weak learners (31).

Max Entropy: R implementation of Low-memory Multinomial Logistic Regression with Support for Text Classification. The regularizers were turned off in our experiments.

Neural Networks: (NNet) we used an R implementation of the single layer neural network with the default settings (32).

Evaluation

We evaluated Recall, Precision and F1 scores for extraction of sentences containing entities and events using 10-fold cross validation on the set of 963+246 + 200 sentences labeled respectively as containing events, entities only, and not relevant. We evaluated Sensitivity and Specificity for gene-protein/disease relations in MeSH indexing.

Results

A summary of the manual annotation of the entities and events in 300 documents is shown in Table 1.

The most commonly mentioned genes were *LEP*, *PHLDA2*, *FLT1*, *PGF*, *STOX1*, and *VEGF*. The *LEP* gene, which provides instructions for making a protein called leptin, is involved in the regulation of body weight, and in our collection it was associated with pre-eclampsia (33) (34). The *PHLDA2* gene, generally associated with tumor growth suppression and placenta growth, was associated with placenta insufficiency and fetal growth retardation in our collection (35) (36). The *FLT1*, *VEGF*, and *PGF* (placenta growth factor) genes, known to be associated with development of embryonic vasculature, the regulation of angiogenesis, cell survival, and cell migration, were associated with pre-eclampsia and intrauterine growth restriction in our collection (37) (38) (39) . The most frequent events were associations between gene mentions and the diseases preeclampsia and intrauterine growth retardation followed by HELLP syndrome and preterm delivery.

Extraction of sentences containing entities and events:

The results of classifying sentences as containing gene-placental mediated events, entities only, and not relevant are shown in Table 2. Adding manually annotated entities improved the results for all classifiers by at least 60%. For the automatically extracted entities, the boost was much smaller, and the difference in the improvements provided by the automated and manual entity annotation was preserved even when other manual features were added to PubTator annotations. The second visible improvement in the results is associated with adding the event annotations as features, whereas the roles and the modality features either do not contribute anything or slightly worsen the results.

Table 1. Event Types and Entities in manual annotation

Event Types	**#**	**Entities**	**#**
is_associated	260	Gene	1324
increases_activity	214	Disease	1189
decrease_activity	151	Protein_Enzyme	633
increases_levels_of	142	Organ	374
is_different	106	Tissue	184
decrease_levels_of	93	PhysiologicProcess	139
is_expressed	72	Cell	110
affects_expression	49	GeneFamily	104
affects_modifies	33	Process	74
regulates	22	PhysiologicState	29
cause	19	Cell_Component	26
inhibits	16	Substance	10
is_active	10	Chromosome	4
activates	7	**Total instances**	**4200**
stimulates	4		
synthesizes	3		
suppresses	2		
induces_expression	2		
interacts	1		
Total instances	**1206**		

Table 2. Contribution to sentence-level classification results of progressively adding entities, events, roles and attributes features to the bag of words. R – Recall, P – Precision. PubTator entities+ includes manually annotated event, roles and attributes.

	Bag of words			+ entities			+events			+roles			+attributes			PubTator entities			PubTator entities+		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
SVM	0.70	0.55	0.56	0.87	0.83	0.85	**0.98**	**0.98**	**0.98**	0.95	0.95	0.95	0.95	0.95	0.95	0.70	0.64	0.66	0.85	0.85	0.85
LogitBoost	0.53	0.53	0.52	0.87	0.82	0.84	0.94	0.99	0.96	0.96	0.95	0.95	0.96	0.95	0.95	0.65	0.63	0.64	0.80	0.79	0.79
GLMNet	0.60	0.51	0.53	0.83	0.75	0.78	**0.97**	**0.99**	**0.98**	0.89	0.88	0.88	0.89	0.88	0.88	0.65	0.55	0.58	0.80	0.79	0.79
NNetWork	0.47	0.48	0.47	0.74	0.69	0.70	0.93	0.99	0.96	0.76	0.74	0.74	0.77	0.74	0.75	0.56	0.55	0.55	0.80	0.76	0.78
MaxEnt	0.59	0.56	0.57	0.84	0.73	0.76	0.92	0.98	0.95	0.87	0.82	0.83	0.87	0.82	0.83	0.65	0.61	0.62	0.78	0.76	0.77

Evaluation of MeSH Indexing-based event extraction:

Not surprisingly, MeSH headings contained the following diseases included in our search: *preeclampsia*, *HELLP syndrome*, *intrauterine growth retardation*, *gestational diabetes* and *premature rupture of membranes*. The chemicals included: *Nitric Oxide Synthase*, *Plasminogen Activators*, *Nucleic Acids*, *Nucleotides*, and *Nucleosides* among others. The most common subheadings were: *genetics*, *metabolism*, *physiology*, *pathology*, and *enzymology*. Three documents did not have MeSH indexing. One was published prior to the date indexing started for the publishing journal; the two others are in journals not indexed for MEDLINE. The non-indexed documents were discarded from the calculations. The sensitivity for the MeSH based association rules was 0.87 and specificity was 0.97. Table 3 presents the evaluation results.

Table 3. Comparison of manual annotations (GS) and associations based on MeSH indexing rules for document-level classification

MeSH-GS Category	# documents
Rule-based associations capture at least one event annotated in GS (True Positive)	256
GS event is not captured by any rule (False Negative)	36
No events related to a disease are annotated in GS and the rules produced no associations (True Negative)	4
A rule-based association has no corresponding events in GS (False Positive)	1
Documents with no MeSH indexing	3

Discussion

Our results demonstrate that an SVM or a GLM model can be used to reliably identify sentences containing information about genes associated with placenta-mediated disorders. Our collection provides a means to further extract several types of events, such as *increased/decreased activity* and *increased/decreased levels of*. Additionally, our exploration of MeSH indexing shows that MeSH headings used in conjunction with subheadings can be used to identify documents that establish gene or protein activity in certain diseases. MeSH controlled vocabulary does not include specific gene or protein terms, but it can be used in conjunction with entity annotation tools such as PubTator.

The field of genetic studies in the human placenta in real time has emerged in recent years due to availability of tissue testing on smaller sample sizes, which decreases the need for invasive intrauterine procedures that can be deleterious to the fetus and the mother. The Human Genome Project has made it possible to make important advancements in the study of genetic related diseases; knowledge grows every day on gene activity and physiology (40). With an estimated 19,000-20,000 human protein-coding genes, the task of finding genes associated with disease in the human placenta would be daunting without automated support. We have demonstrated that it is possible to use machine learning algorithms to extract knowledge from the literature in this field. As the literature grows and reports increase in size and volume, it is necessary to make use of automatic tools to guide the wet-bench research. Likewise, automatic extraction of genes, gene disease relationships, and activity allows mapping to annotated genetic databases for knowledge extraction and further guidance.

We show that even sophisticated literature searches alone are not sufficient to extract only human placenta studies: at least 30% of the studies retrieved by our search were researching human tissues in animal models. Our results also show that although distant supervision using the state-of-the-art tools such as PubTator for named entity recognition is possible, about 15% of the documents with gene mentions will be missed using this approach. Similarly, relying on MeSH indexing alone will also result in missing 13% of the relevant documents. However, MeSH indexing is an alternative source of high quality information that we plan to combine with the machine learning approaches in the future.

Limitations and Future Work

The largest limitation of the study is the size of the corpus and the paucity of some events, such as *activates*, *stimulates*, and *suppresses*. The collection, however, reflects the state of this area of research: a search on PubMed in January, 2017 retrieved 488 documents; only 60 more than the study corpus. The genetic human placenta literature is growing at a relatively slow pace, but we anticipate that as the knowledge increases, the rate will also increase. The specific events extracted from the identified sentences can be used to inform future research projects investigating gene pathways related to the identified gene-disease associations and potential precision medicine approaches for the mitigation of placenta-mediated disorders.

Conclusion

Our study demonstrates the possibility of automatic extraction of sentences containing information about genes associated with placenta-mediated disorders in the human placenta literature. We compare the performance of several well-known supervised ML algorithms and identify two approaches, SVM and GLM, which yield up to 98% recall precision and F1 score. We demonstrate that distant supervision approaches based on MeSH indexing and automatic extraction of entities could be used at the expense of missing up to 15% of relevant documents. Our annotation schema, guidelines, and the annotated documents will be available at https://bionlp.nlm.nih.gov/PlacentaCollection.

References

1. Brosens I, Pijnenborg R, Vercruysse L, Romero R. The “Great Obstetrical Syndromes” are associated with disorders of deep placentation. Am J Obstet Gynecol. 2011 Mar;204(3):193–201.
2. Guttmacher AE, Spong CY. The human placenta project: it’s time for real time. Am J Obstet Gynecol. 2015 Oct;213(4 Suppl):S3-5.
3. Guttmacher AE, Maddox YT, Spong CY. The Human Placenta Project: placental structure, development, and function in real time. Placenta. 2014 May;35(5):303–4.
4. Lacroix M, Kina E, Hivert M-F. Maternal/Fetal Determinants of Insulin Resistance in Women During Pregnancy and in Offspring Over Life. Curr Diab Rep. 2013 Apr;13(2):238–44.
5. Cohen PR. DARPA’s Big Mechanism program. Phys Biol. 2015 Jul;12(4):045008.
6. Guttmacher AE, Jenkins J, Uhlmann WR. Genomic medicine: who will practice it? A call to open arms. Am J Med Genet. 2001;106(3):216–22.
7. Pubmed Bethesda (MD): National Library of Medicine (US). [1946] - [Internet]. [cited 2015 Nov 13]. Available from: http://pubmed.gov/
8. Percha B, Altman RB. Learning the Structure of Biomedical Relationships from Unstructured Text. PLoS Comput Biol. 2015 Jul;11(7):e1004216.
9. Furniss SK, Yao R, Gonzalez G. Automatic Gene Prioritization in Support of the Inflammatory Contribution to Alzheimer’s Disease. AMIA Summits Transl Sci Proc. 2014;2014:42–7.
10. Browne F, Wang H, Zheng H. A computational framework for the prioritization of disease-gene candidates. BMC Genomics. 2015;16(Suppl 9):S2.
11. Pyysalo S, Ohta T, Miwa M, Cho H-C, Tsujii J, Ananiadou S. Event extraction across multiple levels of biological organization. Bioinforma Oxf Engl. 2012 Sep 15;28(18):i575–81.
12. Ohta T, Pyysalo S, Ananiadou S, Tsujii J. Pathway Curation Support as an Information Extraction Task. In 2011. Available from: http://www.nactem.ac.uk/papers/Ohta_LBM_2011.pdf
13. Pyysalo S, Ohta T, Miwa M, Tsujii J. Towards exhaustive protein modification event extraction. In: Proceedings of BioNLP 2011 Workshop. Portland, Oregon: Association for Computational Linguistics; 2011. p. 114–23.

14. Gerner M, Nenadic G, Bergman CM. An exploration of mining gene expression mentions and their anatomical locations from biomedical text. In: Proceedings of the 2010 Workshop on Biomedical Natural Language Processing. Uppsala, Sweden: Association for Computational Linguistics; 2010. p. 72–80.
15. Kim SN, Martinez D, Cavedon L, Yencken L. Automatic classification of sentences to support Evidence Based Medicine. BMC Bioinformatics. 2011;12(Suppl 2):S5–S5.
16. Morid MA, Fiszman M, Raja K, Jonnalagadda SR, Del Fiol G. Classification of Clinically Useful Sentences in Clinical Evidence Resources. J Biomed Inform. 2016 Apr;60(C):14–22.
17. Chung GY. Sentence retrieval for abstracts of randomized controlled trials. BMC Med Inform Decis Mak. 2009;9(1):10.
18. Rodriguez L, Morrison S, Greenberg K, Demner Fushman D. Towards automatic discovery of Genes related to Human Placenta. AMIA: NIH, NLM; 2016.
19. Kim J-D, Ohta T, Tsujii J. Corpus annotation for mining biomedical events from literature. BMC Bioinformatics. 2008;9(1):10.
20. Percha B, Garten Y, Altman RB. Discovery and explanation of drug-drug interactions via text mining. Pac Symp Biocomput Pac Symp Biocomput. 2012;410–21.
21. Stenetorp P, Pyysalo S, Topić G, Ananiadou S, Aizawa A. Normalisation with the brat rapid annotation tool. In: Proceedings of the 5th International Symposium on Semantic Mining in Biomedicine. Zürich, Switzerland; 2012.
22. Hripcsak G. Agreement, the F-Measure, and Reliability in Information Retrieval. J Am Med Inform Assoc. 2005 Jan 31;12(3):296–8.
23. Wei C-H, Kao H-Y, Lu Z. PubTator: A PubMed-like interactive curation system for document triage and literature curation. In: BioCreative 2012 workshop. 2012. p. 20–4.
24. Jimeno Yepes A, Verspoor K. Mutation extraction tools can be combined for robust recognition of genetic variants in the literature. F1000Research. 2014;3:18.
25. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2016. Available from: https://www.R-project.org/
26. Jurka TP, Collingwood L, Boydstun AE, Grossman E, Atteveldt W van. RTextTools: Automatic Text Classification via Supervised Learning [Internet]. 2014. Available from: https://CRAN.R-project.org/package=RTextTools
27. Chen T, Kou Q, He T. mxnet: MXNet [Internet]. 2015. Available from: https://github.com/dmlc/mxnet/R-package
28. Chang C-C, Lin C-J. LIBSVM: A Library for Support Vector Machines. ACM Trans Intell Syst Technol. 2011 May;2(3):27:1–27:27.
29. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw. 2010;33(1):1–22.
30. Schapire RE. A Brief Introduction to Boosting. In: Proceedings of the 16th International Joint Conference on Artificial Intelligence - Volume 2 [Internet]. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.; 1999. p. 1401–1406. (IJCAI'99). Available from: http://dl.acm.org/citation.cfm?id=1624312.1624417
31. Dettling M, Bühlmann P. Boosting for tumor classification with gene expression data. Bioinformatics. 2003;19(9):1061.
32. Venables WN, Ripley BD. Modern Applied Statistics with S [Internet]. Fourth. New York: Springer; 2002. Available from: http://www.stats.ox.ac.uk/pub/MASS4
33. Hogg K, Blair JD, von Dadelszen P, Robinson WP. Hypomethylation of the LEP gene in placenta and elevated maternal leptin concentration in early onset pre-eclampsia. Mol Cell Endocrinol. 2013 Mar 10;367(1–2):64–73.
34. Kaartokallio T, Cervera A, Kyllönen A, Laivuori K, Kere J, Laivuori H, et al. Gene expression profiling of pre-eclamptic placentae by RNA sequencing. Sci Rep. 2015 21;5:14107.
35. McMinn J, Wei M, Schupf N, Cusmai J, Johnson EB, Smith AC, et al. Unbalanced placental expression of imprinted genes in human intrauterine growth restriction. Placenta. 2006 Jul;27(6–7):540–9.
36. Ishida M, Monk D, Duncan AJ, Abu-Amero S, Chong J, Ring SM, et al. Maternal inheritance of a promoter variant in the imprinted PHLDA2 gene significantly increases birth weight. Am J Hum Genet. 2012 Apr 6;90(4):715–9.
37. Nishizawa H, Pryor-Koishi K, Kato T, Kowa H, Kurahashi H, Udagawa Y. Microarray analysis of differentially expressed fetal genes in placental tissue derived from early and late onset severe pre-eclampsia. Placenta. 2007 Jun;28(5–6):487–97.

38. Zhou Y, McMaster M, Woo K, Janatpour M, Perry J, Karpanen T, et al. Vascular endothelial growth factor ligands and receptors that regulate human cytotrophoblast survival are dysregulated in severe preeclampsia and hemolysis, elevated liver enzymes, and low platelets syndrome. Am J Pathol. 2002 Apr;160(4):1405–23.
39. Unal ER, Robinson CJ, Johnson DD, Chang EY. Second-trimester angiogenic factors as biomarkers for future-onset preeclampsia. Am J Obstet Gynecol. 2007 Aug;197(2):211.e1-4.
40. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001 Feb 15;409(6822):860–921.