Extracting Biomedical Terms from Postpartum Depression Online Health Communities

Sanchari Chowdhuri, MS[1], Sidnei McCrea, BA[2], Dina Demner Fushman, MD, PhD[3], Casey Overby Taylor, PhD[4]

[1] University of Maryland, College Park, MD; [2,4] Johns Hopkins University School of Medicine, Baltimore, MD; [3] Lister Hill National Center for Biomedical Communications, NIH, Bethesda, MD

Abstract

Online health communities play a vital role in supporting patients by connecting them to people with similar health conditions, thereby enabling interactions that may be distinct from those with their healthcare providers. Online health support groups can provide social support for people experiencing clinical conditions such as postpartum depression (PPD). In this paper, we describe our creation of a dataset of annotated PPD discussion forums and a preliminary assessment of topics that are discussed on the forums. Our approach leverages the capabilities of MetaMapLite (MMLite) and the Human Phenotype Ontology (HPO) concept recognition software to identify biomedical terms from BabyCenter.com online health communities. A data extraction pipeline wherein text from discussion forums on the topic of PPD is scraped and structured for annotation by the MMLite and HPO. The final corpus includes 10,584 posts with their all associated comments. Our analysis of the performance of MMLite to annotate biomedical terms relevant to PPD indicated a precision of 86.7%, recall of 81.3%, and AUC of 0.714. We propose a data model illustrating the main topics discussed among PPD forum users. Topics include: exposures, phenotypes, health conditions, behaviors, and timing. This resource has potential to enable investigating previously unexplored experiences with PPD.

Background and Motivation

Postpartum depression (PPD) falls under the purview of perinatal depression, which includes any major or minor depressive episode during pregnancy or 12 months after delivery[1].It is one of the most common ailments associated with childbirth. This treatable medical disorder affects 10% of pregnant women and 13% of women worldwide who have just given birth. This adversely affects functioning of a mother and in turn can negatively impact the baby[2].

Depressive symptoms among pregnant women can include anxiety, insomnia, and changes in mood[3]. Under current clinical guidelines, healthcare providers are encouraged to screen patients at least once during the perinatal period. The Edinburgh Postnatal Depression Scale and the Beck Depression Inventory are some of the diagnostic questionnaires that are used to evaluate perinatal depression[3]. Although pregnant women have numerous encounters with their physicians, they may not report depressive symptoms. A study conducted in Britain found only 12% of mothers reported depressive symptoms to their provider[4]. In fact, 90% of the women knew "something was wrong" but the majority did not seek out professional help[4].

Online health communities can bridge the gap between healthcare providers and their patients. Patients with various diseases use online communities as a source of social support[5]. For example, a study conducted among Japanese women with breast cancer found that women who posted in an online forum reported higher levels of emotional support and decreased feelings of anxiousness[6]. A depression group on Reddit was the basis for a linguistic analysis for researchers Park and Conway. They found that members of the online community were more likely to use positive emotion words than negative emotion words when posting in the forum[5]. Indeed, online health communities can serve as meeting ground for those with diseases to seek out support and advice outside of the healthcare system. Thus, online forums where new parents often seek peer support during the postpartum period have potential to provide a more complete picture of different aspects of the patient which are not captured in clinical notes during patient encounters with their healthcare providers.

For new parents, online health communities can help them to gain peer support regarding intimate postpartum challenges such as parenting, breastfeeding, and postpartum depression[7,8,9,32]. Some benefits to new parents identified by others include the ability to seek and exchange information in a way that allows for "candor (both less harsh and more forthright responses to problems), less negative judgement, reduced obligation to reciprocate support, less relational dependency, more immediate ability to seek support, greater expertise in the network, stigma management, intimacy, access, uninterrupted composition, more expressive communication, and anonymity."[10] Peer support in online health communities provide members with informational, emotional, and instrumental support on health and

disease management. Group members often engage in deep discussions and negotiate differing perspectives of the diseases[11]. In online health communities, patients may also share their strategies for managing the minutiae of their health conditions, describe their illness trajectories, and develop common understanding around disease management[12]. Furthermore, online health communities can provide parenting resources and opportunities for social connections and support[13].

BabyCenter is one of many popular commercial online health communities for parents[14]. This online media company provides information on conception, pregnancy, birth, and early childhood development for parents and expecting parents through eleven country and region-specific properties including websites, apps, emails, print publications, and an online community where parents can connect on a variety of topics[15]. BabyCenter.com allows users to create new forums on a topic, and members of the community can contribute to any forum. Each of these forums hosted on BabyCenter has thousands of users who post questions and interact with other users by adding comments to the posts.

We created a dataset of PPD forums with biomedical term annotations and characterized main topics discussed by forum users. Our approach uses the Human Phenotype Ontology (HPO) concept recognition software[16,31] and MMLite[17] to annotate biomedical terms from BabyCenter online health communities discussing PPD.

Tools and Terminologies

HPO concept recognition software: This software is a name entity recognition (NER) tool used to identify Human Phenotype Ontology (HPO) terms reported in text documents. We use this software to annotate HPO terms in PPD forum posts (see Section 3.2. Preliminary text analysis). The HPO aims to provide a standardized vocabulary of phenotypic abnormalities encountered in human disease. Each term in the HPO describes a phenotypic abnormality, such as atrial septal defect. The HPO is currently being developed using the medical literature, Orphanet[18], DECIPHER[19], and OMIM[20]. HPO currently contains approximately 11,000 terms (still growing) and over 115,000 annotations to hereditary diseases. The HPO also provides a large set of HPO annotations to approximately 4000 common diseases.

MMLite[17]: MetaMap[21] is a widely-used NER tool that maps biomedical free text to Unified Medical Language System[22] Metathesaurus concepts (includes HPO as well as other biomedical terminologies). MMLite provides near real-time named-entity recognition that is faster than MetaMap and allows users to customize and augment its behavior for specific purposes (e.g., negation detection, restriction by UMLS source and semantic type, etc).

UMLS Tools: The Unified Medical Language System (UMLS)[22] is a set tools that brings together many health and biomedical vocabularies and standards to enable interoperability between computer systems. It has three parts including the Metathesaurus, Semantic Network, and SPECIALIST Lexicon and Lexical Tools. With our use of MMLite, Metathesaurus is leveraged to define terms and codes from many vocabularies including CPT®, MeSH®, RxNorm, HPO, among others. This work also uses the Semantic Network that defines broad categories of biomedical terms (semantic types) in order to tailor MMLite to find categories of specific relevance to this research.

In order to create a dataset of PPD online health community forum posts annotated with biomedical terms we: used web scraping techniques to create a dataset of semi-structured free text posts and comments from BabyCenter.com, used two NER tools to annotate the text, and analyzed the performance of one NER tool to annotate biomedical terms relevant to PPD.

Methods

First, we started with a data extraction process, wherein we created a corpus of semi-processed text posts and performed a preliminary text analysis with the HPO concept recognition software. Next, we configured MMLite to use semantic groups relevant to our goal to extract biomedical terms relevant to PPD. After that, we used MMLite to identify biomedical terms from the semi-processed dataset, a subset of the corpus (100 out of 10,584 posts) that were randomly selected and manually evaluated by two authors (SC and SM). Differences were discussed and resolved by consensus. The manually evaluated posts were used to assess the performance of MMLite to detect biomedical terms relevant to PPD from BabyCenter.com forum posts. Lastly, we proposed a data model to illustrate the major topics covered on PPD forums. All tasks were performed using Python (2.7.12).

Semi-Processed Corpus Creation

We identified three public PPD online health community forums on BabyCenter.com: "Postpartum Depression, Anxiety and Related topics"[23], "Postpartum Depression and Postpartum Anxiety Support Group"[24] and "Postpartum Anxiety Support Group"[25]. The corpus building process started by automating the process of URL building and pagination for every posts and comments for each of the forums. Next, we automated the entire process of scraping posts; their corresponding comments spread across different URLs and combined them into separate raw text files. Each text file (document) contains one post thread (i.e., the main post and comments associated with that post). The extracted text was stored as .txt files.

Preliminary Textual Analysis

Document-level text analysis was conducted, wherein the semi-processed corpus was created from one PPD online health community named "Postpartum Depression, Anxiety and Related Topics[23]." Threads were normalized, tokenized, stemmed, and stop-words were removed. Second, the corpus was screened for antidepressant drugs, bigrams and common words were identified. These antidepressant drugs were extracted from the National Pregnancy Registry for Antidepressants.[26] The frequency of drug and common word mentions are reported. Third, the HPO concept recognition software was used to annotate phenotypes from this corpus. Fourth, we analyzed the frequency and distribution of the HPO terms identified for the corpus and visualized those data as a WordCloud [27]. Lastly, upon completing the preliminary text analyses, we identified major categories of frequently mentioned biomedical terms that may be relevant to PPD.

Expanded Corpus and Approach to NER annotation

After completing the preliminary text analysis of PPD forum posts we expanded our approach to include 2 additional BabyCenter.com forums: "Postpartum Depression and Postpartum Anxiety Support Group"[24] and "Postpartum Anxiety Support Group"[25] from BabyCenter.com. Based upon the major categories of biomedical terms identified in our preliminary text analysis, we also expanded our biomedical term annotation approach. Specifically, we used MMLite that includes HPO terms as well as other biomedical terminologies covered by the Metathesaurus. We also selected semantic groups that included major categories of biomedical terms identified from the preliminary text analysis.

Semi-processed Corpus Annotation using MMLite

In order to identify biomedical terms from PPD forum text, we configured MMLite for selected semantic groups described in Appendix 1. The entire corpus was then annotated using MMLite wherein we created MetaMap Fielded MMI Output for each document present in the corpus along with annotation file for every document in the corpus. We analyzed the average number of MMLite annotations for each document in the corpus. To understand which terms are commonly identified by MMLite's semantic configuration we created a word cloud of biomedical terms identified using MMLite. We also analyzed the frequency of all MMLite identified terms.

Performance Assessment of MetaMap Lite to Identify Biomedical Terms from Pregnancy & PDD Online Health Forum communities

For our analysis of performance, we randomly selected a sample of 100 text files out of our entire corpus. Manual annotation was conducted by two annotators separately wherein each annotator independently went through pre-annotated files to check for missing and correctly assigned terms and identified biomedical terms (treatments, phenotypes and health conditions) relevant to PPD that were missed by MMLite as False Negatives, identified terms that were correctly assigned a biomedical concept by MMLite as True Positives, and assigned terms which were wrongly assigned to a different concept classification as False Positives. These assignments were used to estimate the precision and recall of MMLite to identify biomedical terms relevant to PPD from PPD online health community forum text.

Precision is a measure of result relevancy. Precision refers to proportion of positive identifications that were correct. Recall is a measure of how many truly relevant results are returned. Recall refers to proportion of actual positives that were correctly identified. The precision-recall curve shows the tradeoff between precision and recall for different threshold. A high area under the curve represents both high recall and high precision, where high precision relates to a low false positive rate, and high recall relates to a low false negative rate. High scores for both show that the MMLite is returning accurate results (high precision), as well as returning a majority of all positive results (high recall).[28]

The performance of MMLite was measured using area under the ROC curve (AUC). We plotted a ROC curve of the true positive rate versus the false positive rate for every possible classification threshold. The true positive rate is the rate of occurrence that MMLite assigns the correct concept. The false positive rate is the rate of occurrence that MMLite assigns the incorrect concept. Cohen's kappa was also calculated to estimate inter-rater agreement between the two reviewers.

Results

Results from semi-processed corpus creation & preliminary text analysis

The initial semi-processed corpus from the "Postpartum Depression, Anxiety and Related Topics" forum consisted of 10028 text files post threads. The average number of words present per post was 208. After removing stop words, the average number of words per post was 114. Figure 1 shows the most frequent words, antidepressant medications, and bigrams. Figure 2 shows the WordCloud of HPO terms identified the corpus.

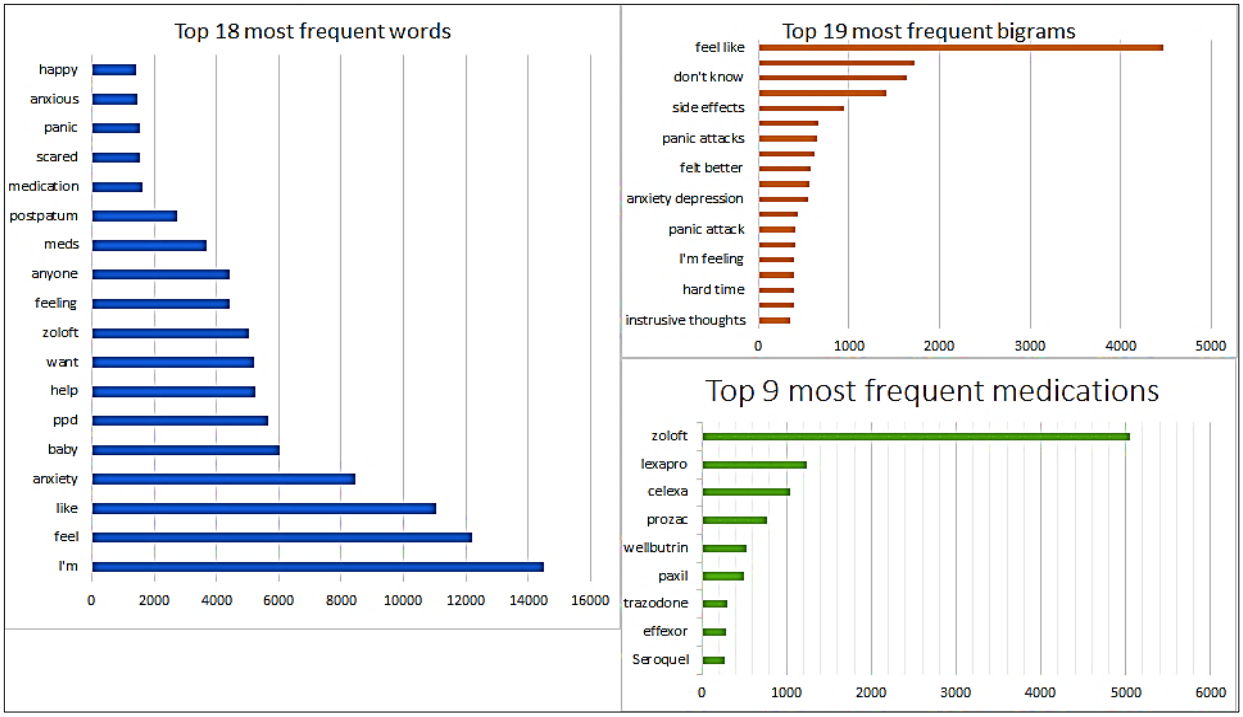

Figure 1: Preliminary text analysis summary

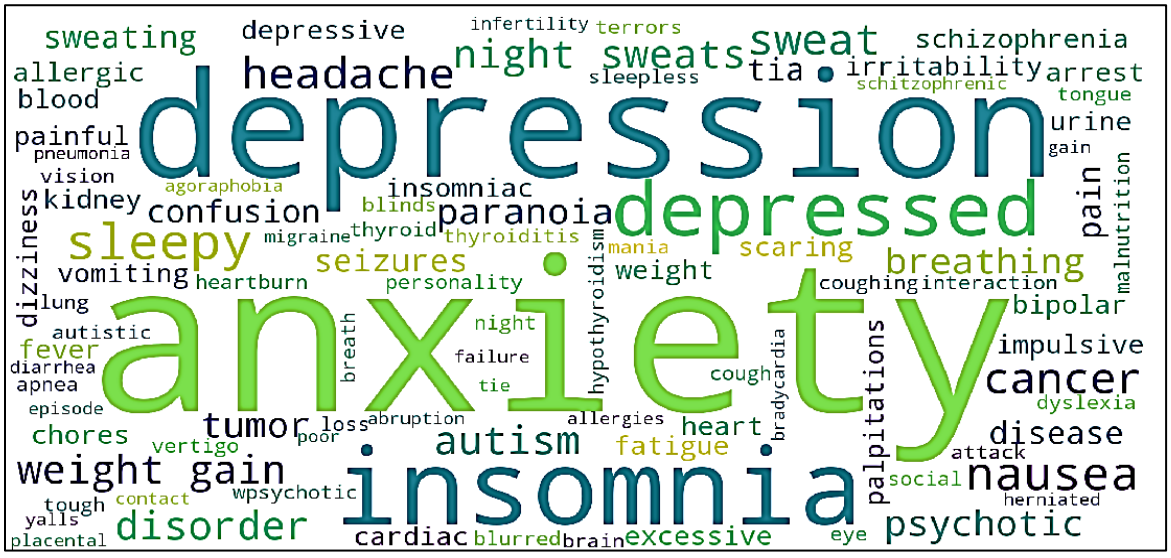

Figure 2: WordCloud of terms identified using the Human Phenotype Ontology (HPO) concept recognition software.

The preliminary text analysis provided us with a guideline to look for specific categories of biomedical concepts. For example, words like "medication", "counseling" and medication names e.g., Zoloft fall under an **EXPOSURE** category that includes both medical and non-medical treatments. Words like "nausea", "headache" and "irritation" may be grouped under a **PHENOTYPE** category that includes symptoms, clinical features and behaviors. Lastly, words like "depression", and "allergy" can be grouped under a **HEALTH CONDITION** category that includes present and previous health conditions. These categories represented major topics of discussion by users of PPD online health communities. In summary, based on preliminary text analysis, three major topics were identified: **EXPOSURE**, **PHENOTYPE**, and **HEALTH CONDITION** (See Figure 3).

Results from configuring MMLite to use UMLS Semantic Groups

When configuring MMLite to annotate biomedical terms relevant to PPD, UMLS semantic groups were selected to include major categories identified in Figure 3. Specific UMLS semantic groups are summarized in Appendix 1 and were included in the MMLite Configuration.

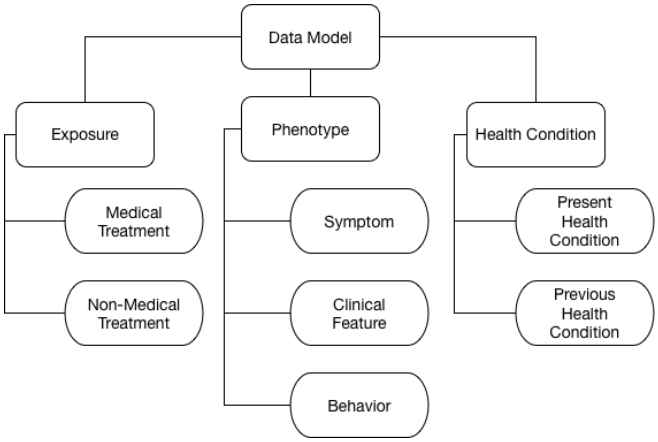

Figure 3: Major categories of biomedical concepts.

Results from expanding the semi-processed corpus to include MMLite biomedical term annotations

Our expanded corpus drawing from 3 online forums included 10,584 threads. MMLite generated annotation files consists of those terms which belong to the current semantic group selection along with their start and end position, name of the semantic type and UMLS Concept Unique Identifier (CUI).-The average number of annotations (words identified by MMLite) per thread for the current corpus was 26 words.

The most frequent biomedical terms annotated by MMLite across all threads are shown in Table 1. Figure 4 shows the corresponding WordCloud. The dataset is available on a public github repository (https://github.com/translational-informatics/ppd-named-entity-dataset).

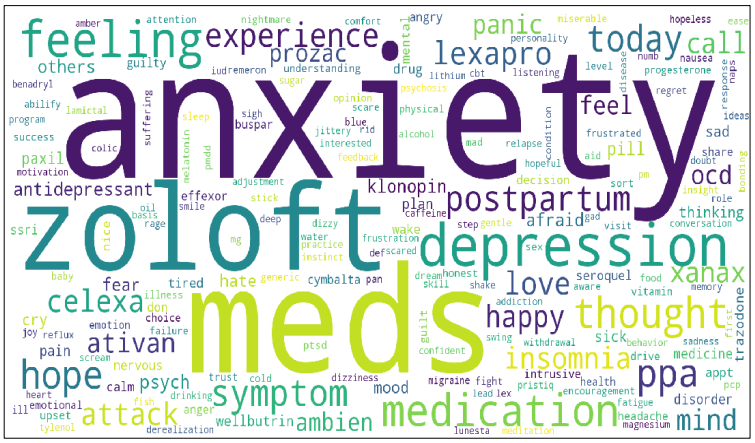

Figure 4: WordCloud of the MMLite terms in the corpus.

Table 1: Frequently occurring MMLite Terms

Words	Frequency
anxiety	24119
Meds	15266
Zoloft	14602
Depression	12487
Feeling	6453
Medication	5688
PPA	5343
Hope	5193
Thoughts	4870
Today	4493
Postpartum	4384
Symptoms	4050
Lexapro	3579
Love	3395
Experience	3324
Celexa	3298
Happy	3190
insomnia	3125

Results from assessing the performance of MMLite to identify biomedical terms

Findings from reviewing MMLite annotations of 100 randomly selected documents from the entire corpus indicated a precision of 87.6% and a recall 81.3%. Figure 5 illustrates the ROC curve with an AUC of 0.71. The kappa score between the two reviewers was 0.96.

Upon manual review of false positive and false negative results, we identified three main themes: wrong annotation of words in "short form" (e.g., "pls" for "please"), misspelled terms, and biomedical terms that did not fit selected semantic groups.

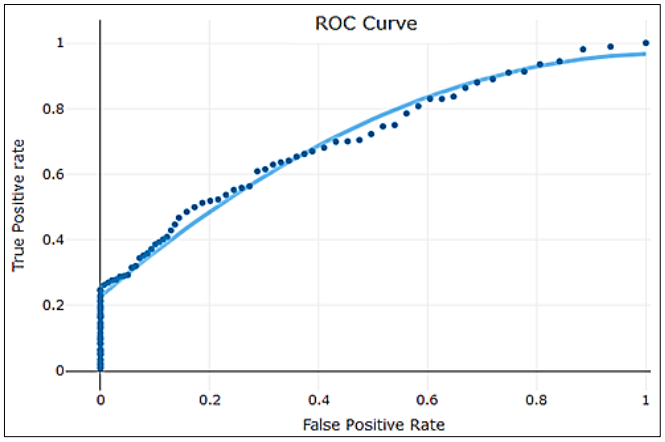

Figure 5: ROC Curve for MMLite Performance

Wrong annotation of word in "short form":

Since the corpus has conversational text, many short forms were misinterpreted by MMLite, therefore contributing several false positives in the corpus. Examples include:

- The term "meds," contextually used to refer to medications, has been wrongly interpreted by MMLite as Microcephaly, Epilepsy, And Diabetes Syndrome (Meds).
- "AD," the acronym for anti-depressant has been missed by MetaMap's annotation.
- The term "pls," which refers to short form of "please," has been wrongly interpreted as Papillon-Lefevre Disease.

Misspelled Terms:

MMLite could not identify misspelled terms. For example, in some posts misspelled drug names such as Ativan and Trazadone existed and thus were missed by MMLite.

Semantic Group Coverage:

The initial semantic groups as shown in Table 1 were efficient in annotating terms regarding treatments, phenotypes, and health conditions; however, there were subsets of biomedical terms relevant to PPD that were not recognized. We proposed additional semantic groups that my help to better identify those terms (see Appendix 2).

Other Considerations:

The term "Park" which was used in the context of open space or garden has been mis-annotated as Parkinson's Diseases. PPA is annotated as Primary Progressive Aphasia (disorder) / Phosphonoacetic Acid or Phenylpropanolamine instead of Post-Partum Anxiety. The term "rash" has been wrongly annotated as skin irritation whereas the term was contextually used for hasty decisions.

Discussion

Semi-processed corpus creation and preliminary text analysis

From the preliminary text analysis of the corpus, we found that the most commonly occurring words can be divided into three main categories: treatment (medication, Zoloft), phenotype (anxiety, happy, panic), and health conditions (ppd), as shown in the Figure 3 data model. Our initial WordCloud of biomedical terms identified that using the HPO concept recognition software, however, illustrated that biomedical terms related to treatment and health conditions were largely missing (See Figure 2). These findings supported our decision to expand our approach to use MMLite to annotate a broader range of biomedical terms of potential relevance to PPD.

Expanding corpus to include MMLite biomedical term annotations

Our use of MMLite enabled annotating a range of biomedical terms related to PPD. Apart from clinical symptoms, individual behaviors and social behaviors appear to also be important topics related to PPD. This finding is consistent with findings of other indicating that mental process and social determinants are important indicators of PPD[29, 30].

Assessing the performance of MetaMap Lite to identify phenotype annotations

Expanding our approach to use MMLite to identify a broader range of biomedical terms has potential to introduce additional noise. Our assessment of MMLite helps shed light on the degree to which some biomedical terms relevant to PPD may be missed or wrongly assigned in the final corpus.

A common pattern seen among false positive results (both missed and wrongly assigned) included colloquial short forms. This occurred due to MMLite assigning the wrong biomedical concept (e.g., "meds" wrongly annotated to Microcephaly, Epilepsy, and Diabetes Syndrome (MEDS)). In addition, for terms relevant to PPD that were missed such as timing of symptoms (e.g., today, 2 months, postpartum), we proposed a revised data model that includes behavior and timing as separate categories. (Figure 6).

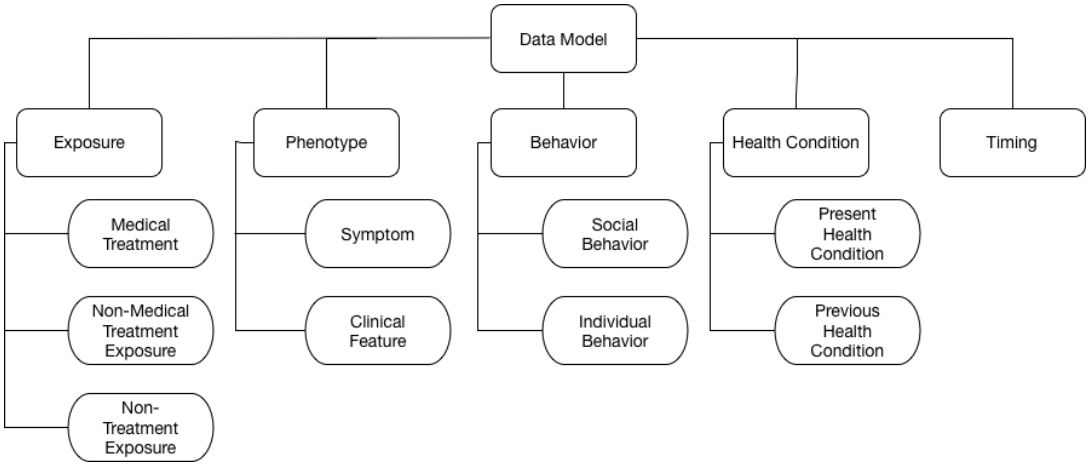

Figure 6: Revised Data Model

Approaches to avoid some of these false negative results we encountered are to customize the vocabulary for PPD and to use spell check. We can also avoid some missed annotations that appear to be important topics to PPD by expanding the semantic groups to include those proposed in Appendix 2. These revisions have potential to improve both precision and recall.

Conclusion

Patient narratives and conversational text captured on online health forums are different from clinical text captured in a healthcare setting. This work provides a corpus of 10,584 semi-processed texts extracted from postpartum depression online discussion forums and with biomedical terms annotations of reasonable accuracy (estimated precision of 86.7% and recall of 81.3 %). Major topics discussed included: exposures, phenotypes, health conditions, behaviors, and timing. This dataset has potential to enable investigating previously unexplored experiences with postpartum depression.

Acknowledgements

We would like to thank Stella Liang for compiling medications relevant to postpartum depression.

References

1. Gavin NI, Gaynes BN, Lohr KN, Meltzer-Brody S, Gartlehner G, Swinson T. Perinatal depression: a systematic review of prevalence and incidence. Obstetrics & Gynecology. 2005 Nov 1;106(5):1071-83.
2. Maternal and child mental health [Internet]. World Health Organization. World Health Organization; 2016 [cited 2018May2]. Available from: http://www.who.int/mental_health/maternal-child/en/
3. Committee on Obstetric Practice. The American College of Obstetricians and Gynecologists Committee Opinion no. 630. Screening for perinatal depression. Obstetrics and gynecology. 2015 May;125(5):1268.
4. Whitton A, Warner R, Appleby L. The pathway to care in post-natal depression: women's attitudes to post-natal depression and its treatment. Br J Gen Pract. 1996 Jul 1;46(408):427-8.
5. Park A, Conway M. Longitudinal changes in psychological states in online health community members: understanding the long-term effects of participating in an online depression community. Journal of medical Internet research. 2017 Mar;19(3).

6. Setoyama Y, Yamazaki Y, Namayama K. Benefits of peer support in online Japanese breast cancer communities: differences between lurkers and posters. Journal of medical Internet research. 2011 Oct;13(4).
7. Cowie GA, Hill S, Robinson P. Using an online service for breastfeeding support: what mothers want to discuss. Health Promotion Journal of Australia. 2011;22(2):113-8.
8. Evans M, Donelle L, Hume-Loveland L. Social support and online postpartum depression discussion groups: A content analysis. Patient education and counseling. 2012 Jun 1;87(3):405-10.
9. Salonen AH, Kaunonen M, Åstedt-Kurki P, Järvenpää AL, Isoaho H, Tarkka MT. Effectiveness of an internet-based intervention enhancing Finnish parents' parenting satisfaction and parenting self-efficacy during the postpartum period. Midwifery. 2011 Dec 1;27(6):832-41.
10. Walther JB, Boyd S. Attraction to computer-mediated social support. Communication technology and society: Audience adoption and uses. 2002 Jan 1;153188.
11. Mamykina L, Nakikj D, Elhadad N. Collective sensemaking in online health forums. InProceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems 2015 Apr 18 (pp. 3217-3226). ACM.
12. Huh J, Ackerman MS. Collaborative help in chronic disease management: supporting individualized problems. InProceedings of the ACM 2012 conference on Computer Supported Cooperative Work 2012 Feb 11 (pp. 853-862). ACM.
13. Stevens NR, Hamilton NA, Wallston KA. Validation of the multidimensional health locus of control scales for labor and delivery. Research in nursing & health. 2011 Aug;34(4):282-96.
14. Sinton P. E-Tailing's Rising Stars / BabyCenter is one of many riding the boom in Internet sales [Internet]. SFGate. San Francisco Chronicle; 1999 [cited 2018May2]. Available from: https://www.sfgate.com/business/article/E-Tailing-s-Rising-Stars-BabyCenter-is-one-of-2950217.php
15. Company Information [Internet]. BabyCenter. BabyCenter; 2018 [cited 2018May9]. Available from: https://www.babycenter.com/about
16. Köhler S, Doelken SC, Mungall CJ, Bauer S, Firth HV, Bailleul-Forestier I, Black GC, Brown DL, Brudno M, Campbell J, Fitzpatrick DR. The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. Nucleic acids research. 2013 Nov 11;42(D1):D966-74.
17. Demner-Fushman D, Rogers WJ, Aronson AR. MetaMap Lite: an evaluation of a new Java implementation of MetaMap. Journal of the American Medical Informatics Association. 2017 Jan 27;24(4):841-4.
18. Rath A, Olry A, Dhombres F, Brandt MM, Urbero B, Ayme S. Representation of rare diseases in health information systems: the Orphanet approach to serve a wide range of end users. Human mutation. 2012 May;33(5):803-8.
19. Firth HV, Richards SM, Bevan AP, Clayton S, Corpas M, Rajan D, Van Vooren S, Moreau Y, Pettett RM, Carter NP. DECIPHER: database of chromosomal imbalance and phenotype in humans using ensembl resources. The American Journal of Human Genetics. 2009 Apr 10;84(4):524-33.
20. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic acids research. 2005 Jan 1;33(suppl_1):D514-7.
21. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. InProceedings of the AMIA Symposium 2001 (p. 17). American Medical Informatics Association.
22. UMLS Terminology Services -- Home [Internet]. U.S. National Library of Medicine. National Institutes of Health; [cited 2018May2]. Available from: https://uts.nlm.nih.gov/home.html
23. Postpartum Depression, Anxiety and Related Topics [Internet]. BabyCenter Community. BabyCenter; [cited 2018May9]. Available from: https://community.babycenter.com/groups/a15325
24. Postpartum Depression and Postpartum Anxiety Support Group [Internet]. BabyCenter Community. BabyCenter; [cited 2018May9]. Available from: https://community.babycenter.com/groups/a6742129
25. POSTPARTUM ANXIETY SUPPORT GROUP [Internet]. BabyCenter Community. BabyCenter; [cited 2018May9]. Available from: https://community.babycenter.com/groups/a6718921
26. National Pregnancy Registry for Antidepressants [Internet]. MGH Center for Women's Mental Health. [cited 2018May18]. Available from: https://womensmentalhealth.org/clinical-and-research-programs/pregnancyregistry/antidepressants/
27. Mueller A. word clouds in Python--wordcloud 1.3 documentation [Internet]. GitHub. Github ; 2018 [cited 2018May9]. Available from: https://github.com/amueller/word_cloud
28. Precision-Recall [Internet]. 1.4. Support Vector Machines - scikit-learn 0.19.1 documentation. Scikit-learn Developers ; [cited 2018May7]. Available from: http://scikit-learn.org/stable/auto_examples/model_selection/plot_precision_recall.html
29. Barbadoro P, Cotichelli G, Chiatti C, Simonetti ML, Marigliano A, Di Stanislao F, Prospero E. Socio-economic determinants and self-reported depressive symptoms during postpartum period. Women & health. 2012 May 1;52(4):352-68.
30. Stuart-Parrigon K, Stuart S. Perinatal depression: an update and overview. Current psychiatry reports. 2014 Sep 1;16(9):468.

31. Groza T, Köhler S, Moldenhauer D, Vasilevsky N, Baynam G, Zemojtel T, Schriml LM, Kibbe WA, Schofield PN, Beck T, Vasant D. The human phenotype ontology: semantic unification of common and rare disease. The American Journal of Human Genetics. 2015 Jul 2;97(1):111-24.
32. Romano AM. A changing landscape: implications of pregnant women's internet use for childbirth educators. The Journal of perinatal education. 2007;16(4):18.

Appendix 1. Selected UMLS Semantic Groups

Data Model mapping	**UMLS Semantic Groups [17]**	**UMLS Semantic Type Classification [17]**	**Definition of UMLS Semantic Group [17]**	**Example (*following words are mapped to the semantic group*)**
clinical condition /phenotype	mobd	Mental or Behavioral Dysfunction	A clinically significant dysfunction whose major manifestation is behavioral or psychological. These dysfunctions may have identified or presumed biological etiologies or manifestations.	PMDD, ADHD, anxiety, Postpartum depression, autism, baby blues, depression OCD (Obsessive compulsive Disorder), mood swing, PTSD, Panic attacks *Words depicting an ongoing mental or behavioral problem*
phenotype	sosy	Sign or Symptom	An observable manifestation of a disease or condition based on clinical judgment, or a manifestation of a disease which is experienced by the patient.	Pain, depression symptoms, vomit, nervous, dizzy, insomnia, sore *Words depicting physical symptoms*
phenotype	cgab	Congenital Abnormality	An abnormal structure or one that is abnormal in size or location, present at birth or evolving over time because of a defect in embryogenesis.	Monster, laryngomalacia, IBS (Irritable Bowel Syndrome), Ichthyosis Bullosa of Siemens
phenotype	acab	Acquired Abnormality	An abnormal structure or one that is abnormal in size or location, found in or deriving from a previously normal structure.	breakdown
clinical condition	dsyn	Diseases or Syndrome	Any of the psycho-social activities of humans or animals that can be observed directly by others.	Heart attack, Premenstrual Dysphoric Disorder Migraine disorder Infantile Neuroaxonal etc *Words depicting diseases and syndromes*
phenotype (behavior - revised data model)	bhvr	Behavior	Any of the psycho-social activities of humans that can be observed directly by others.	Sex, visit
treatment	clnd	Clinical Drug	A pharmaceutical preparation as produced by the manufacturer. The name usually includes the substance, its strength, and the form.	
treatment	orch	Organic Chemical	The general class of carbon-containing compounds, usually based on carbon chains or rings, and containing hydrogen with or without nitrogen, oxygen, or other elements	Xanax, Zoloft, Celexa, Lexapro, PPA (Phenylpropanolamine) etc. *Words depicting medicine names*
treatment	phsu	Pharmacologic Substance	A substance used in the treatment or prevention of pathologic disorders.	Xanax, Zoloft, Celexa, PPA *Words depicting medicine names*
phenotype (behavior - revised data model)	inbe	Individual Behavior	Behavior exhibited by a human that is not a direct result of interaction with other members of the species, but may affect on others.	Honesty, cry, yearn, choice, failure, self-talk, cheating *Words depicting individual behavior*

phenotype	menp	Mental process	A physiologic function involving the mind or cognitive processing.	Happy, angry, mad, jittery, fear love, mourn *Words depicting emotion*
treatment	bodm	Biomedical or Dental Material	A substance used in biomedicine or dentistry predominantly for its physical, as opposed to chemical properties.	Pills, tablets
phenotype/ behavior	socb	Social Behavior	Behavior that is direct result of the interaction of humans with their fellows. This may include anti-social behavior.	Smile, adoption, funeral *Words depicting social behavior*

Appendix 2. Additional UMLS Semantic Groups Following Assessment of MMLite Performance

Data Model mapping	**UMLS Semantic Group [17]**	**UMLS Semantic Type Classification [17]**	**Definition of UMLS Semantic Group [17]**	**Example**
exposures that are not treatments	aapp	Amino acids and chains of amino acids connected by peptide linkages.	Amino acids and chains of amino acids connected by peptide linkages.	soy, pitocin
phenotype	clna	Clinical Attribute	An observable property or state of an organism of clinical interest.	partum
phenotype	fndg	Finding	That which is discovered by direct observation of an organism attributes, including the clinical history of the patient.	stress, muscle twitches, feeling unhappy
exposures that are not treatments	hops	Hazardous or Poisonous Substance	A substance of concern because of its potentially hazardous or toxic effects.	heroine
exposures that are not treatments	inch	Inorganic Chemical	Chemical elements and their compounds, excluding the hydrocarbons and their derivatives	magnesium, water
clinical condition	neop	Neoplastic Process	A new and abnormal growth of tissue in which the growth is uncontrolled and progressive	brain tumor, brain cancer
timing	tmco	Temporal Concept	Concept pertaining to time or duration.	Today
treatment	topp	Therapeutic or Preventive Procedure	A procedure, method, or technique designed to prevent or treat disease,	medication change
treatment	vita	Vitamin	A substance, usually an organic chemical complex, presents in natural products or made synthetically.	Vitamin B Vitamin D