Diving into a Large Corpus of Pediatric Notes

Abstract

This paper is about an ongoing project in which we hypothesize that infant colic has causes that can be illuminated by digging into a corpus of pediatric clinical notes. Our ultimate goal is to conduct a large scale study to understand infant colic and potentially other conditions, through Machine Learning on very large, high-dimensional datasets. We present our preliminary exploration of a large corpus of pediatric notes to bring the data in a form amenable to Machine Learning.

1. Motivation

Electronic Health Records (EHRs) are today widely used to record *Clinical Data* reporting on patient healthcare such as visits, diagnoses, labs tests, images, conditions and medications. In a recent study on healthcare informatics (Stead & Lin, 2009), the authors point out that in observing medical personnel in action, a large amount of the time of physicians and nurses was spent in entering data, and much less on reading data. There is thus an enormous, underused and potentially invaluable resource for understanding the prevalence and nature of many health problems. Our main goal is to make this data accessible and useful.

Machine Learning (ML) is a powerful technology for constructing complex models (Mitchell, 1997; Hastie et al., 2009) in very high dimensional spaces that has proven useful in a wide range of arenas. ML coupled with advanced computation capabilities represent a genuine approach to explore EHRs, that wouldn't have been possible many years ago. Successful use of ML requires an understanding of the application domain, and where the data is not already in a form amenable to machine learning, it also depends on assembling comprehensive and trustworthy data.

We consider *infant colic* and *preterm birth*, as two examples of conditions, that can be elucidated through digging into the enormous data being collected in pediatric notes by medical professionals, their prognoses and correlations with many risk factors. For both conditions, longitudinal data about mothers and babies is available in clinical notes and constitute an unexplored "gold mine" that offers an unprecedented opportunity to help understand and diagnose these two conditions. Without loss of generality, we focus in this paper on infant colic, but the framework that we have been building in this pilot study is meant to be extendable to other conditions/diseases.

In order to understand and discover root causes of a poorly-understood condition such as colic, we need to get a clear picture on how this condition is perceived and documented in the clinical notes. More specifically, our immediate goal is to understand what kind of terms and criteria are used by physicians to describe colicky versus non-colicky babies.

The contributions of this paper are as follows:

1. Assembling a large heterogenous corpus of pediatric notes;
2. Designing and implementing a database for this corpus;
3. Tackling *colic*, a poorly-understood infant condition through an initial exploration of the corpus and descriptive statistics;
4. Using topic modeling to dive and explore the notes which showed promise to help label patients;
5. Opening the door to applying ML to other understudied conditions (e.g. prematurity), and more generally, build an initial framework for turning the huge amount of unstructured medical data in EHRs into knowledge.

This paper is organized as follows: Infant colic and the related medical literature are described in Section 2, followed by the related work. In Section 4, we provide challenges of digging into a large heterogenous corpus of pediatric notes along with a description and statistics of our preliminary data. Statistics about colicky babies are provided in Section 5. A step toward building a relational database from the notes is presented in Section 6. Section 7 is about using topic models to explore the notes. We conclude with a summary and future work in Section 8.

Under review by the Learning from Unstructured Clinical Text ICML 2011 Workshop.

2. Infant Colic

Infant colic is defined as persistent inconsolable crying in healthy babies between 2 weeks and 4 months of age, where the baby seems to be in great discomfort and difficult to soothe. It is not a disease but a serious prevalent condition with medical and social consequences, yet its causes remain a mystery for medical research. Prevalence rates of excessive crying vary between definitions. Estimates of the number of affected infants aged 0-6 months who cry 3 or more hours a day, 3 or more days a week, during 3 or more consecutive weeks for no clear cause (Wessel's criteria (Wessel et al., 1967)), range from 2% to 5% (Reijneveld et al., 2001).

Colic is associated with Shaken Baby Syndrome (SBS), infant brain damage that results when a caregiver violently shakes a baby (Barr et al., 2006), (Fujiwara et al., 2009), (Barr et al., 2009). Shaken baby syndrome, highly correlated with colic and crying, affects between 1200 and 1600 babies each year in the US. Median estimates of the number of deaths range from 20-25%, or between 240 and 400 deaths per year in the US. This number is roughly half of all deaths due to child abuse. Nonfatal consequences include visual impairment including blindness, motor and cognitive impairments. Recent studies suggest that excessive crying in infancy can lead to mother postpartum depression (Vik et al., 2009). Finally, colic is costly for healthcare systems, due to various ineffective medications, doctor's office and emergency room visits. The medications doctors prescribe to treat colic or identify its causes often have side effects but don't provide a cure. The medical literature on colic is a mix of hypotheses to explain this mysterious condition based on small datasets. These include lack of bacteria in the intestines, reflux, lactose intolerance, maternal smoking, and parental depression, to cite a few.

In this paper, we use a sample of babies from a large urban hospital to illustrate the type of comprehensive profile we can construct from clinical notes of colicky babies and clinicians' approach to treatments.

3. Related Work

Clinicians convey valuable information about patients, both in the structured and free-text sections of the EHR. Researchers in informatics have investigated ways in which this information can be leveraged for several applications: clinical decision support systems (Demner-Fushman et al., 2010), genome-wide association studies (Kullo et al., 2010; Kho et al., 2011), syndromic surveillance (Hripcsak et al., 2009), pharmacovigilance (Wang et al., 2009), and clinical research (Pakhomov et al., 2007; Himes et al., 2009; Wei et al., 2010).

There are several challenges entailed in processing longitudinal patient information reliably. Our dataset contains a mix of inpatient and outpatient notes, each containing different types of note structures (some with a mix of template- and free-text). While a bag-of-words approach to feature extraction is attractive, one can hope to get valuable information from shallow semantic information derived from the notes. As such, the preprocessing step for feature extraction needs to identify first the document-level structure of the notes (Denny et al., 2009; Li et al., 2010), along with list items and sentence boundaries. Sentences can then be processed to extract medical terms. The identification of medical terms needs to rely on external, established terminologies, such as the UMLS, but there are also many institution-specific terms and abbreviations present in the notes, which are not covered by the UMLS. Traditional clinical NLP tools, like MedLEE (Friedman et al., 2004), leverage an internal lexicon in addition to the UMLS. Because we focus on infant colic, our dataset contains notes about the patients themselves (the infants), and also their mothers. An important processing step for feature extraction is to distinguish which information pertains to the infant and which to the mother. While there has been recent work towards studying co-reference resolution in clinical notes (Savova et al., 2011), little was done to identify whom a clinical event pertains to (the patient or a family member).

4. Assembling the Data

Due to the absence of a national policy that would standardize EHRs, hospitals rely on disparate vendors that use distinct proprietary formats. Through our collaboration with pediatricians at a large urban hospital, we have assembled a dataset of heterogeneous notes and lab reports for over one thousand babies. Different practitioners rely on different subsets of notes, meaning the note types are used in non-standard ways. Hence, we cannot rely on the note types for grouping them into relevant classes. We have access to plain text exports of the notes, with no specification regarding the underlying format. The notes exhibit a wide range of fields consisting of named sections (e.g., `Birth History`), named fields within sections (e.g., `Apgar Score`), and field values of various types, such as measurements and dosages, fixed values from a menu (e.g., `Normal spontaneous vaginal delivery`), or free text (*colicy (sic), but consolable*).

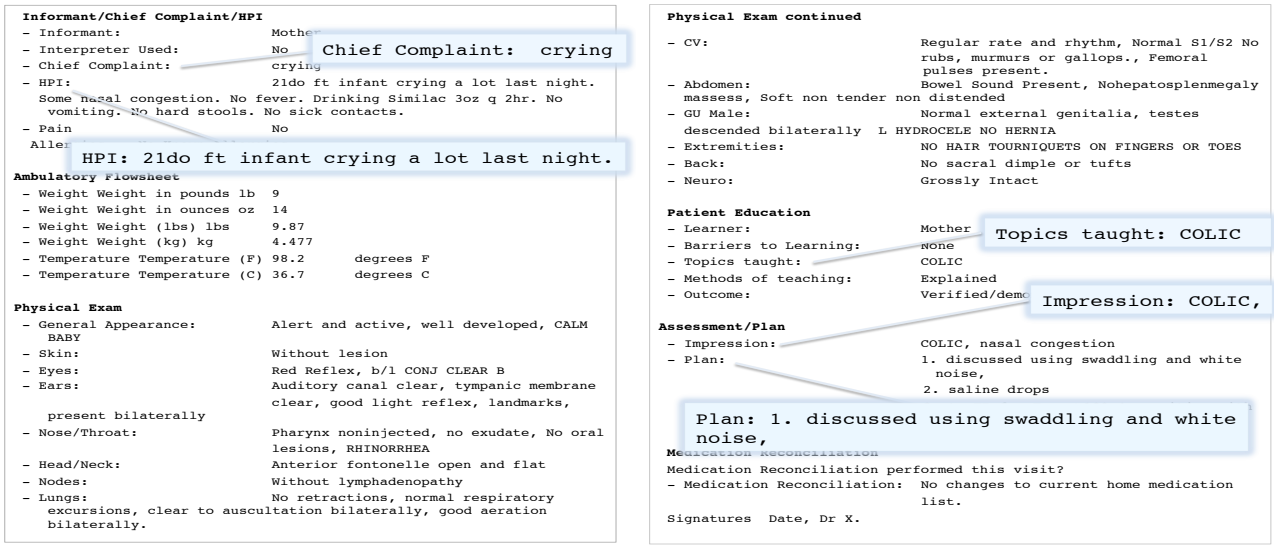

Figure 1. Example of follow up note.

Much of the vocabulary requires domain expertise for interpretation (e.g., `Absolute Nrbc Count`). The free text exhibits idiosyncratic abbreviations and shorthand (e.g., *no PHM now with gm +*), as well as misspellings (e.g., *colickly like his brother*). Because colic is a set of symptoms rather than a disease, practitioners vary in whether they use the term; other terms may be used instead, such as *acid reflux (GERD)*. Further, our goal is to label babies as positive or negative examples of colic (or other conditions). We have conducted initial experiments to automatically identify *topics* that occur across notes. We explore the possibility that by identifying distributions of topics that characterize babies identified as having colic, we can potentially identify additional positive instances among babies not explicitly identified as having colic.

The sample of pediatric notes obtained from [...] hospital spans about two years and a half of data from one single clinic and concerns a population of 1,240 babies. Each baby is described by a set of clinical notes. Some of the data we have acquired is structured (e.g. patient demographics, childbirth conditions); most is recorded in an unstructured and free text format (e.g. notes from physicians). Longitudinal data is available for each baby patient and presented with a set of notes that are grouped in categories. In our sample data we identified 243 types of inpatient and outpatient notes. These represent only a fraction of the large possible note types that healthcare professionals can choose from in the EHR system. Examples of note types in our sample include:

1. Nursing neonatal patient history: this note provides mothers' history (e.g. obstetrical information, past medical history) and a newborn assessment (e.g. physical exam, vital signs);
2. Pediatrics new patient newborn: note from the first visit a patient makes to the clinic;
3. Pediatrics walk-in: a note for an acute care visit by a patient who has either called in or walked in complaining of an acute illness that needs to be seen that day;
4. Pediatrics follows up: a note for a scheduled visit of a patient who has been seen before.

There is a wide variety of notes including inpatient nursery, ancillary, social work, neurology, NICU, surgery, nutrition, OB/GYN delivery, to cite a few. Figure 1 gives an example of note that we deal with. It represents an excerpt of a long pediatric follow up note in which a 21 days old full term baby whose mother complained about *crying* as stated in *Chief complaint* and *HPI (History of Present Illness)* sections. The colic topic was taught to the mother by the pediatrician and the mother received instructions about how to swaddle and use white noise to soothe her baby. The baby was diagnosed with colic in the *Impression* section of the note by the physician.

Note that note structures differ from one type of note to another. We have started a pilot study to analyze this dataset to understand infant colic. We are in the process of acquiring mothers' and babies' data spanning many years of data through 8 OB/GYN and 4 pediatric clinics at [...] hospital. An estimation of the total of pairs (mother, baby) is in the order of ten thousands. Data will be acquired from the institu-

tion's two EHR systems. The first provides the clinical notes while the second includes lab results, diagnoses history and medications.

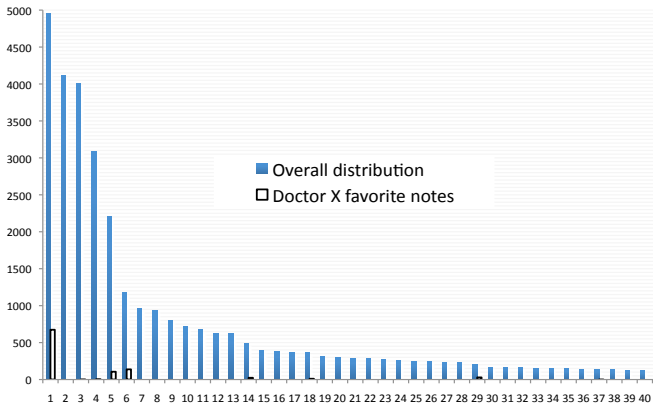

Figure 2. Distribution of the different types of notes present in our corpus.

Statistic	Value
Total number of babies	1,240
Number of colicky babies	40
Number of premature babies	86
Number of note types	243
Total number of notes	34,069
Minimum number of notes per baby	1
Maximum number of notes per baby	258
Average number of notes per baby	27.5
Minimum time span of notes in days	1
Maximum time span of notes in days	862
Average number of days	317.1

Table 1. Some Statistics on the notes

Statistics about the notes are provided in Table 1 that shows the number of babies, notes overall, the minimum, maximum and average number of notes per baby along with the time span in days.

Because of the complexity of the EHR systems and the different possibilities to enter clinical data, each healthcare practitioner has his/her own favorites notes selected among the different types of notes in the EHR system. Hence, the need to process *all* of the notes given that a same clinical information can be documented by 2 different users in different types of notes. The distribution of the different types of notes overall as opposed to those used by a given practitioner is provided in Figure 2. Note that the most frequent note used, `Pediatric_follow_up_note`, which is as described earlier a note about a scheduled visit, is also the most frequently used by this doctor. The second most frequent note `Miscellaneous_nursing_note` is not used as it is mostly written by nurses. The third and fourth top-used notes, `Ancillary_note` and `AMB_Care_Triage_Telephone_Triage_Form` are not frequently used either by this doctor as they

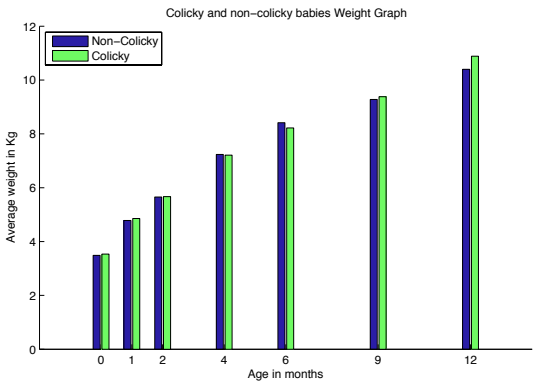

Figure 3. Avg. weights for colicky vs. non-colicky babies.

are not about clinical data but ancillary hospital services and triage of patient calls. The next most frequent notes, `Pediatric_Walk-in_Note` and `Pediatric_Newpatient_Newborn_Note` are obviously highly used by this doctor and contain valuable clinical information about patients.

5. Infant Colic Statistics

From the data of 1,240 babies only 40 had "Colic" in their impression section which means only 40 were diagnosed to be Colicky by the doctor. The statistics for these 40 colicky as compared to the 1,200 non-colicky is shown in Figure 3, Figure 4 and Figure 5. Among the colicky babies 62.50% were male while in the non-colicky babies the number of male and female was almost equal. For both cases the percentage of Normal Vaginal delivery was higher than C-Section. Among the colicky babies 10.52% were "Exclusively" breast-fed while for the non-colicky percentage was 12.37%. A large number of the babies was both formula and breast-fed for both colicky and non-colicky case. Pediatrician do use different terms for colic. Considering the terms *Constipation, Reflux, Fuss(y), Gas(sy), GERD, Colic and Excessive crying* as possible proxies for colic, we computed the average number of these terms per colicky and non-colicky baby. Numbers show that Colic, Gas(sy), Fuss(y), and Excessive crying terms are present at higher rates in the colicky babies notes. The average number of times baby care-givers visited the hospital without an appointment was approximately the same for colicky and non-colicky babies, with a slight increase of the number of calls in the colicky babies population. Finally, in order to compare the growth of colicky and non-colicky babies, we plotted the average weight for the two populations at birth, then at 1, 2, 4, 6, 9, and 12

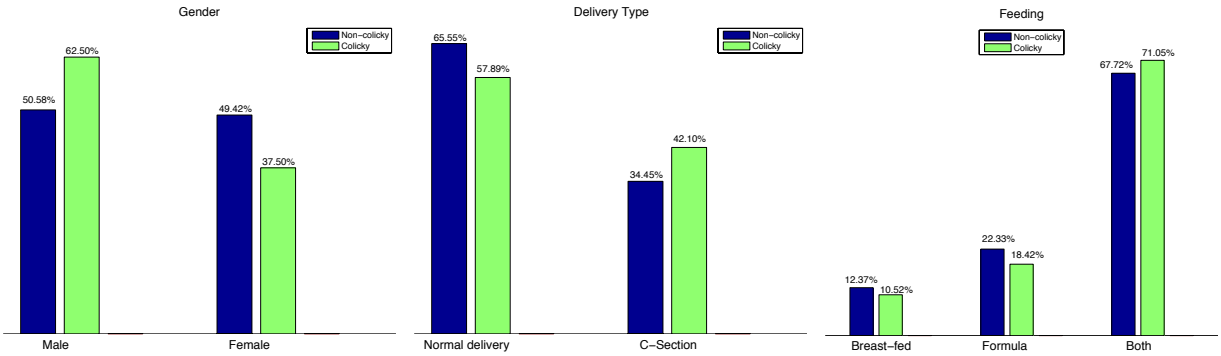

Figure 4. Distribution of gender, delivery type and feeding at 1 month among Colicky and Non-colicky babies.

months, which represent the routine visit schedule at the clinic where the weight check is done. We note the growth of the colicky and non-colicky babies is almost the same which supports the known fact that colicky babies do grow normally as their peers. Note that the information extraction regarding the weights was particularly challenging, given the presence of outliers and missing values.

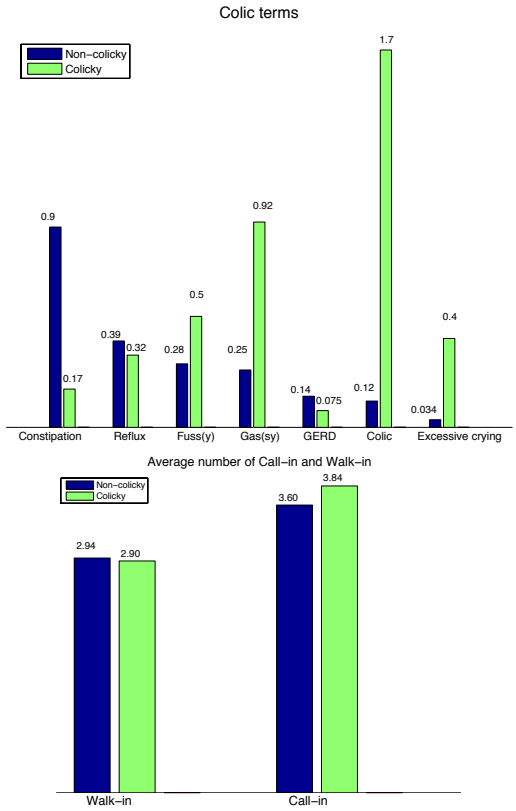

Figure 5. Distribution of colic terms and avg. number of call-in and walk-in notes in Colicky & Non-colicky babies.

6. A Database for the Notes

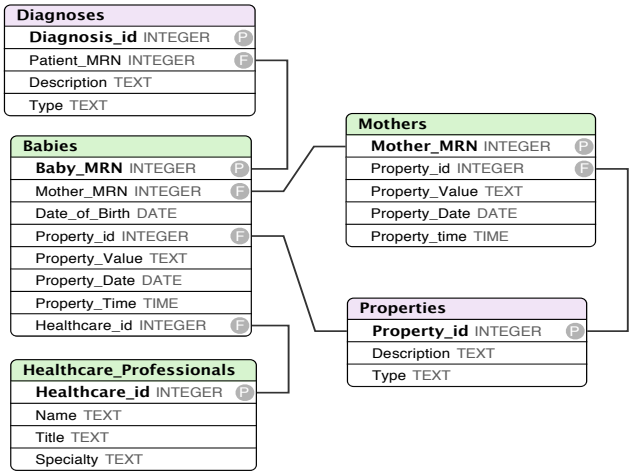

Figure 6. Conceptual database modeling our pediatric longitudinal data.

We started building a database for the clinical pediatric corpus. Having a relational database representing the notes will make it easier to query and organize the notes. On the other hand, it will also facilitate data preparation for Machine Learning and Data Mining techniques. We designed the conceptual model shown in Figure 6. The *Babies* table captures information about the baby and the time series of the recorded properties such as the weight and height. Table *Healthcare professional* includes physicians, nurses, social workers. The *Mothers* table captures information about the mother health and pregnancy information. As for the babies, all the properties of the mother are included in that table. The *Diagnoses* contains all diagnoses used by the physicians. We are currently in the process of populating some of these tables. Specifically, properties such as weights, sex, gender, race, gestational age, date of birth, delivery type, Apgar

scores and anesthesia type are included on the *Babies* and *Properties* tables. Healthcare_Professionals is also populated with the physicians, nurses, social workers, speech and hearing therapists etc.

7. Topic Models

Topic modeling is an unsupervised machine learning approach to discover the topics discussed in a corpus. Unsupervised methods do not require labeled data, which is advantageous when labels are costly to acquire, as in a domain requiring significant expertise. The intuition behind topic modeling is that when doctors or nurses prepare to write medical records, they first have in mind a set of topics to address. They fill in the EHRs using words associated with the different topics. Topic modeling identifies which words have the greatest probability of occurring together, and posits an abstract topic that conditions these probabilities.

We create a single document for each patient, which concatenates the content of all notes of that patient. Therefore we end with 1,240 documents. After generating the topic models for these documents, each document can be represented as a subset of the total topics, each in a proportion dependent on the content words. To preprocess the documents, we strip all the non-content words, and only keep the free text. Words and characters that are removed include section and field names, person names, punctuation, digits and stop-words. After pre-processing, we end up with 4,518,148 tokens representing 33,421 distinct words.

A topic model consists of a probability distribution over topics, and then for each topic, the probability of each word in the vocabulary. The parameters behind the probability distributions are treated as *latent* variables. By analyzing on a set of observations (words in the documents), it is possible to recover the latent structures of the generative model. The particular model we use is based on Latent Dirichlet Allocation (LDA) with Gibbs Sampling. For the experiment, we use the Topic Modeling module of MALLET (McCallum, 2002), a machine learning toolkit for natural language processing tasks.

We perform exploratory data analysis to discover the topics for the word *colic* (and variants such as *colicky*). The resulting model depends on the investigator's choice of k, the number of topics to discover.

How fast to converge? The convergence rate of the probabilistic topic model depends on both the size of the dataset and the number of topics (k). Figure 7 shows the rate of convergence for different values of k. The model of a larger k (dash line) often has a slower convergence rate. For this dataset, models usually converge after 6,000 iterations.

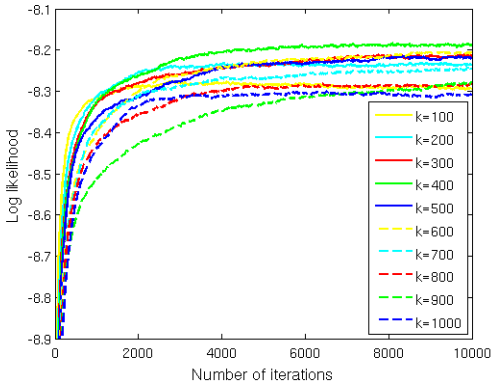

Figure 7. Convergence of the model and log-likelihood. The model of a larger k (dash line) often has a slower convergence rate than a smaller k (solid line). For this dataset, models usually converge after 6,000 iterations.

Topic index	# of *colic* tokens	ranking
344	302	1
127	6	62
326	4	86
366	2	51
53	1	216
236	1	270
322	1	210

Table 2. Topics of word *colic*. Summary of the topics related to word *colic* when K=400 after 10,000 iterations.

How many topics? When k is small, topics are very general. The word *colic* becomes more likely when k is large, and the topics are more fine-grained. Here we present two methods to choose the value of k. One is to use log-likelihood of the model, which is keyword independent. The other relies on a weighted ranking of a keyword of interest, e.g., *colic*, which we now illustrate.

Suppose we have a topic model at hand with $K = 400$. Table 2 shows all the topics where *colic* has a non-zero probability. There are a total of 317 tokens of the word *colic*. The majority (302) are assigned to topic 344, where *colic* is the word with the highest probability. Six tokens are assigned to topic 127, and *colic* ranks 62^{nd} among the words in this topic. In sum, Table 2 illustrates that any word w will have different ranks across the set of topics generated by a given topic model. To investigate the prominence of a specific word w within a given topic model, we calculate a normalized rank for w with respect to the model, which we refer to as its ranking index. For example, we can calculate the rank

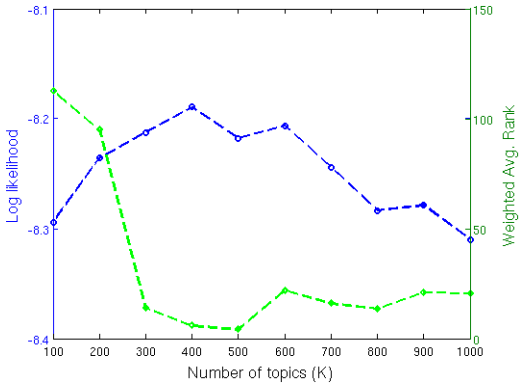

Figure 8. Number of topics. The blue line is the log-likelihood (higher is better). The green line is the weighted average of the ranking of word *colic* (lower is a higher rank, also a larger probability). The choice of K balances high likelihood of the model against probability of the word of interest. In this case, K=400 is chosen for further analysis.

# of Topics (K)	Avg. rank for *colic*	Avg. rank for *non-colic*
400	38.82	54.61

Table 3. Average rankings of *colic* topics between babies with word *colic* in notes and babies without word *colic*.

of *colic* in the topic model shown in Table 2. First we weight the rank of colic for each topic by its frequency in that topic. Then we sum the weighted ranks, and divide by the total frequency. We calculate the $rank_z^w = \frac{1\times302+62\times6+86\times4+51\times2+216\times1+270\times1+210\times1}{302+6+4+2+1+1+1} = 5.73$. The result is the *ranking index* of *colic* in this topic model, meaning that of all words in these seven topics, *colic* has a rank of around 6. Formally, we define the ranking index of a word w:

$$RankingIndex_w = \frac{\sum_z rank_z^w \times N_z^w}{\sum_z N_z^w},$$

where $rank_z^w$ is the ranking of word w in topic z. N_z^w is the count of word w are assigned to topic z.

We choose k based on the tradeoff between log-likelihood (keyword independent) and the estimated ranking of the keyword (keyword dependent). Figure 8 shows the change in log-likelihood and ranking index as k increases. $K = 400$ has the highest log-likelihood and a high rank for *colic*.

What do the *colic* topics look like? We use a word cloud to visualize the words and their probabilities for each topic. Figure 9 displays the prominent *colic* topic (index 344; see Table 2). Babies where variants of the term *colic* appear in the notes will necessarily have topic profiles where the *colic* topic has a higher rank, compared with non-colicky babies. Table 3 summarizes the average of the ranking of all *colic* topics for both colicky and non-colicky babies.

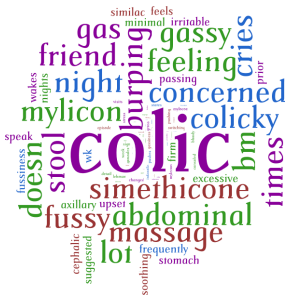

Figure 9. Word cloud for topic in which the most probable word is *colic* generated with K=400 and 10,000 iterations. Higher probabilities are indicated with larger font size. T

A baby that is not referred to in the notes as having colic can nevertheless have a highly ranked *colic* topic. We speculate that high rank of the *colic* topic, rather than explicit presence of colic terms in the notes, can be used to assign positive labels to babies. For example, we observe one baby where the word *colic* does not appear in the notes, but where the *colic* topic (#344) is the 11th highest ranking topic.

8. Summary and Future Work

Our study started by assembling and analyzing a sample of infants' patient records. The goal of this pilot study is to develop an initial hypothesis through exploratory data analysis that can drive the assembly of an initial database, and provide the basis for a supervised machine learning approach. Supervised machine learning depends on labeled data, meaning that the entities of interest (infant patients) are assigned labels indicating whether they are positive or negative examples of the phenomenon under study (colic). The labels are used for training a machine learning model on a portion of the data, and for evaluating the performance of the model on a test set of data. We plan on using topic models to help us label babies with a colic label. We also look forward to acquire more data about babies as well as mothers' health history and pregnancy information that might provide insights on colic causes.

Acknowledgments. This project is partially funded by [...] and is being conducted under IRB-[...]. Many thanks to Prof K.R. for advice on building a relational database and to professor D.R. for valuable feedback.

References

Barr, RG, Rivara, FP, and Barr, M, et al. Effectiveness of educational materials designed to change knowl-

edge and behaviors regarding crying and shaken-baby syndrome in mothers of newborns: a randomized, controlled trial. *Pediatrics*, 123:3:972 980, 2009.

Barr, Ronald G., Trent, Roger B., and Cross, Julie. Age-related incidence curve of hospitalized shaken baby syndrome cases: Convergent evidence for crying as a trigger to shaking. *Child Abuse and Neglect*, 30:1:7–16, 2006.

Demner-Fushman, D., Chapman, W., and McDonald, C. What can natural language processing do for clinical decision support? *J Biomed Inform*, 42(5): 760–772, 2010.

Denny, J., Spickard, A., Johnson, K., Peterson, N., Peterson, J., and Miller, R. Evaluation of a method to identify and categorize section headers in clinical documents. *J Am Med Inform Assoc*, 16(6):806–815, 2009.

Friedman, C., Shagina, L., Lussier, Y., and Hripcsak, G. Automated encoding of clinical documents based on natural language processing. *J Am Med Inform Assoc*, 11(5):392–402, 2004.

Fujiwara, T, Barber, C, Schaechter, J, and Hemenway, D. Characteristics of infant homicides: findings from a u.s. multisite reporting system. *Pediatrics*, 124(2): 210217, 2009.

Hastie, Trevor, Tibshirani, Robert, and Friedman, Jerome. *The elements of statistical learning: data mining, inference and prediction*. Springer, 2009.

Himes, B., Dai, Y., Kohane, I., Weiss, S., and Ramoni, M. Prediction of chronic obstructive pulmonary disease (copd) in asthma patients using electronic medical records. *J Am Med Inform Assoc*, 16(3):371–379, 2009.

Hripcsak, G., Soulakis, N., Li, L., Morrison, F., Lai, A., Friedman, C., Calman, N., and Mostashari, F. Syndromic surveillance using ambulatory electronic health records. *J Am Med Inform Assoc*, 16(3):354–361, 2009.

Kho, A., Pacheco, J., Peissig, P., Rasmussen, L., Newton, K., Weston, N., Crane, P., Pathak, J., Chute, C., andI. Kullo, S. Bielinski, Li, R., Manolio, T., Chisholm, R., and Denny, J. Electronic medical records for genetic research: Results of the emerge consortium. *Sci Transl Med*, 3(79), 2011.

Kullo, I., Fan, J., J, J. Pathak, Savova, G., Ali, Z., and Chute, C. Leveraging informatics for genetic studies: use of the electronic medical record to enable a genome-wide association study of peripheral arterial disease. *J Biomed Inform*, 17(5):568–574, 2010.

Li, Y., Gorman, S. Lipsky, and Elhadad, N. Section classification in clinical notes using a supervised hidden markov model. In *Proc. ACM Int. Health Informatics Symposium (IHI)*, pp. 744–750, 2010.

McCallum, Andrew Kachites. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu, 2002.

Mitchell, Tom M. *Machine Learning*. McGraw-Hill, New York, 1997.

Pakhomov, S., Weston, S., Jacobsen, S., Chute, C., Meverden, R., and Roger, V. Electronic medical records for clinical research: application to the identification of heart failure. *Am J Manag Care*, 13(6): 281–288, 2007.

Reijneveld, Sijmen A., Brugman, Emily, and Hirasing, R. A. Excessive Infant Crying: The Impact of Varying Definitions. *Pediatrics*, (108):893–897, 2001.

Savova, G., Chapman, W., Zheng, J., and Crowley, R. Anaphoric relations in the clinical narrative: corpus creation. *J Am Med Inform Assoc*, Epub, 2011.

Stead, William W. and Lin, Herbert S. *Computational Technology for Effective Health Care: Immediate Steps and Strategic Directions*. The National Academies Press Washington, D.C., 2009.

Vik, T, Grote, V, Escribano, J, Socha J, Verduci, E, Fritsch, M, Carlier, C, von, Kries R, and Koletzko, B, European Childhood Obesity Trial Study Group. Infantile colic, prolonged crying and maternal postnatal depression. *Acta Paediatr*, 98(8):1344–1348, 2009.

Wang, X., Hripcsak, G., Markatou, M., and Friedman, C. Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study. *J Am Med Inform Assoc*, 16(3):328–337, 2009.

Wei, Wei-Qi, Tao, Cui, Jiang, Guoqian, and Chute, Christopher. A high throughput semantic concept frequency based approach for patient identification: A case study using type 2 diabetes mellitus clinical notes. In *AMIA*, pp. 857–861, 2010.

Wessel, M. A., Cobb, J. C., Jackson, E. B., S., Harris G., Jr., and Detwiler, A. C. Paroxysmal fussing in infancy, sometimes called "colic,". *Pediatrics*, 14 (421), 1967.