Diving into a Large Corpus of Pediatric Notes

Ansaf Salleb-Aouissi[1], Ilia Vovsha[1], Anita Raja[3], Axinia Radeva[1], Hatim Diab[1], Rebecca Passonneau[1], Faiza Khan Khattak[1], Ronald Wapner[2], Mary McCord[2]

[1] Center for Computational Learning Systems
Columbia University 475 Riverside Drive MC 7717
New York, NY 10115 USA

[2] Columbia University College of Physicians and Surgeons
630 West 168th Street, New York,
NY 10032 212-305-CUMC

[3] The Cooper Union
30 Cooper Square
New York, NY 10003

Motivation of Infant Colic

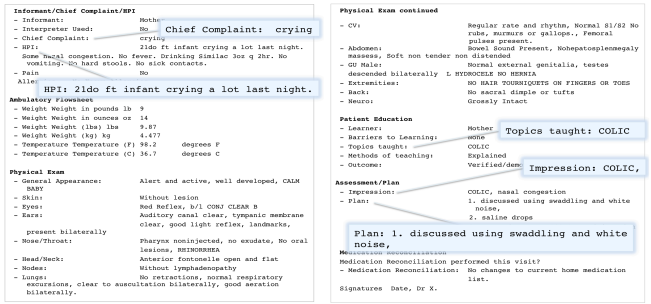
- Infant colic: a medical condition characterized by baby crying for 3+ hours per day, for 3+ days per week, for 3+ weeks.
- Colic affects between 2% and 5% of infants.
- Colic has a strong correlation with mother postpartum depression and Shaken Baby Syndrome. This accounts for between 240 and 400 deaths per year in the United States.

http://letopusa.files.wordpress.com/2011/03/colic.jpg

Sample of Pediatric Notes

EHR Pediatric notes:
Heterogeneous corpus of pediatric notes collected from the New York Presbyterian Hospital.
2.5 years of data:

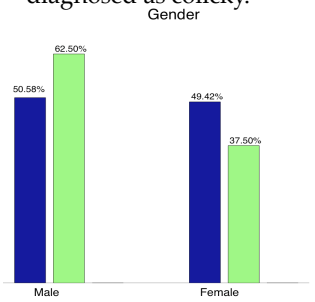
- #babies: 1,240
- #colicky babies: 40
- #note types: 243
- #notes: 34,069
- #notes per baby: 1 – 258

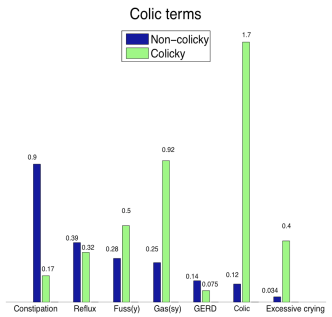


Statistics and Topic Models

Results:

- 63% of colicky babies were male (51% of all babies were male).
- Constipation was noted 4 times as often in non-colicky than in colicky babies.
- Excessive crying was noted 10 times as often in colicky than in non-colicky babies.
- Topic Modeling (machine learning approach) discovers the topics discussed in the pediatric notes. Topics can help label babies with colic even if they were never diagnosed as colicky.





Motivation of Preterm Birth (PTB)

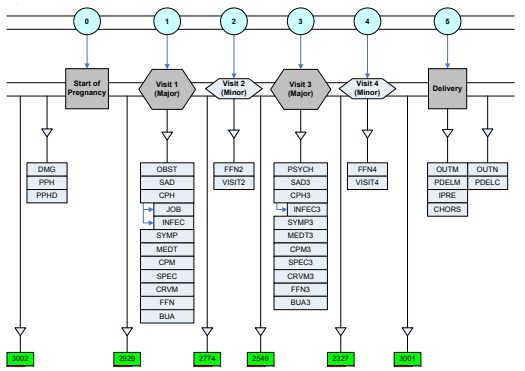
- Birth of a baby before 37 completed weeks of gestation
- Over 26 billion dollars are spent annually PTB
- Rate: About 12-13% of infants born preterm in the US.
- Previous research: Focused on individual risk factors
- Goal: Develop a prediction system that combines well-known risk factors using machine.

Picture of a 23 weeks preemie in an incubator (source: March of Dimes)

Preterm Prediction Study Data

Data: Observational prospective study Performed by NICHD. 2,929 of participating women were followed at 24, 26, 28 and 30 weeks gestation:

- #spontaneous PTB < 32 weeks: 50
- #spontaneous PTB < 35 weeks: 129
- #spontaneous PTB < 37 weeks: 309
- #Indicated PTB < 37 weeks: 124



Results:
We used Support Vector Machines and obtained an average of sensitivity and specificity in predicting PTB of 57% and 68% respectively, well above the 21% for sensitivity and 30% for specificity reported in the literature on this data.

Conclusion & Future Work

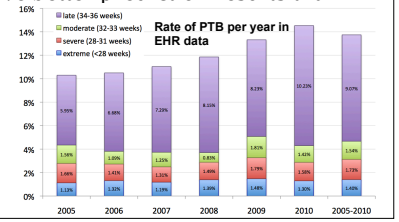
- Infant colic and of preterm birth are both exciting data science problems.
- EHR is a rich source of information, but the ability to harness it is forthcoming.
- We can understand better baby colic and sort out different cases of baby crying.
- Prediction of Preterm Birth is not elusive, we achieve better prediction results than any previous study.

Future work:

- Explore a larger EHR data: We collected a 5-year snapshot of EHR data from the NYPH.

Period: 01/2005 to 10/2011.
Population: 43,000 women and 35,000 babies.

- Use social media data and parental blogs.



Acknowledgments: This project is funded by the Executive Vice President for Research's Research Initiatives in Science & Engineering (RISE) program and NSF #1454855. IRB-AAAF2852 and IRB-AAAJ2054

Contributors: David Waltz, Tara Randis, Harriet McGurk, Noemie Elhadad, Ashish Tomar, Ashwath Rajan.