A Perspective on Understanding Infantile Colic

Ansaf Salleb-Aouissi, Axinia Radeva, Rebecca Passonneau
Ashish Tomar and David Waltz
Columbia University
Center for Computational Learning Systems
{{ansaf, axinia, atomar, waltz}@ccls, becky@cs}.columbia.edu

Mary McCord and Harriet McGurk
Columbia University, College of
Physicians and Surgeons
mm26@mail.cumc.columbia.edu
hem2@columbia.edu

Noémie Elhadad
Columbia University
Biomedical Informatics
noemie@dbmi.columbia.edu

Introduction We propose to study infant colic, a widespread condition that has potentially serious consequences, to develop a rich database documenting possible causes, and a statistical Machine Learning (ML) approach to modeling them. Infant colic remains a mystery for medical research despite its prevalence. Colic occurs in 5-40% of births, depending on the criteria, although other sources are more conservative [11]. The medical community deems colic as inconsolable crying in healthy babies between 2 and 16 weeks old [10]. It often occurs in the late afternoon or evening. According to the Wessel's criteria [12], a baby is said to be colicky if the crying persists for at least 3 hours, 3 days per week for 3 consecutive weeks. Within these criteria, there remain a wide range of behaviors, depending on severity and persistence of the symptoms. Treatment varies substantially from doctor to doctor but babies are often given different medications to treat colic or identify its causes, often with side effects but no cure.

Research indicates that crying, especially in the first 4 months of age, is an important stimulus for Shaken Baby Syndrome (SBS) [3, 8], serious brain damage babies suffer as a result of a caregiver's brutal shaking. Recent studies also suggest that excessive crying in infancy has a small but significant adverse effect on cognitive development [13], and is associated with higher risks of mental health problems later in life [5]. Finally, colic is costly for healthcare systems [9], including various ineffective medications, doctor's office and emergency room visits, and even hospitalization. Prevention programs such as the PURPLE project [2] aim to educate parents with colicky babies; it has proven effective in reducing SBS, but not with the hoped-for magnitude.

Despite an abundant literature, colic remains unexplained with no reliable treatment. We hypothesize there are discoverable root causes for colic. Most of the studies to date have been conducted on relatively small samples of data, are low dimensional and often draw conclusions from minimal statistics, e.g. correlating mothers' smoking status with colic. We would like to conduct the *first truly large scale study* to tackle this problem through ML on a very large, high-dimensional database of infants. We assume the underlying root cause for colic is complex, possibly a combination of variables, or distinct syndromes. Machine Learning is a powerful technology for constructing complex models in very high dimensional spaces and has proven useful in a wide range of applications, including learning from medical databases. Cutting-edge ML methods are likely to demystify this medical issue through the discovery of patterns not detectable by other means.

Discovering the root causes behind colic will have an outstanding and broad impact worldwide: (1) It will have a positive impact on parents, who will enjoy their babies minus the stress from excessive crying; (2) It will affect pediatricians who, thanks to our research, will be able to respond to parent questions, diagnose colic, prescribe the right treatments and overall provide enhanced care; (3) It will lower the number of kids with lifelong problems and death caused by Shaken Baby Syndrome; (4) It will improve the Electronic Medical Records (EMR) systems: learning from EMRs can be

of interest to many medical applications, and will especially benefit research on other understudied medical conditions; (5) Finally, it will help reduce the economic costs inherent to colic.

Our chance of success is high: (1) we have significant expertise in ML and in assembling and mining large-scale and high-dimensional databases across multiple domains, including in the areas of epilepsy and electrical grid; (2) as parents and physicians who deal and have dealt with colicky babies, we are highly motivated to conduct this study; (3) we provide a cross-disciplinary team including clinicians and experts in ML and medical informatics.

Data Our research is under IRB approval number IRB-AAAF2852. The raw data will be provided from multiple sources and multiple EMR systems, e.g. New York-Presbyterian Hospital/Weill Cornell Medical Center and private practices affiliated with Columbia University. Our study will cover patients from birth to 6 years and will cover as well the mother's past health history and pregnancy information. We have got a preliminary data set of 1,500 electronic medical records of baby patients between 0 to 4 months of age from New York-Presbyterian Hospital (NYPH). Some of the data is structured (e.g. patient demographics, childbirth conditions); however most of patient information is recorded in an unstructured data format (e.g. notes from physicians and nurses), free text, which will require some heavy preprocessing. Longitudinal data is available for each baby patient which is observed on multiple occasions over time. Each patient is presented with a set of notes which are grouped in different categories. In our sample, we identified twenty eight different categories, e.g.

1. Nursing neonatal patient history: this note provides mothers history (e.g. obstetrical information, past medical history) and 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.
5. Ancillary: a note used to document most phone calls.

Future Work Figure 1 gives a general overview of our project. The research objective of this project is to conduct extensive ML research on Colic data collected from New York Presbyterian Hospital and further from data collected via social networks and questionnaires. Our data challenges the traditional ML framework in many ways: (1) Sample Bias: e.g. most of the patients coming to NYPH are Latino American; (2) Complex data: data is by nature longitudinal and relational; (3) Uncertain labels: we will need to define the colic label and/or the severity of colic from the EMRs.

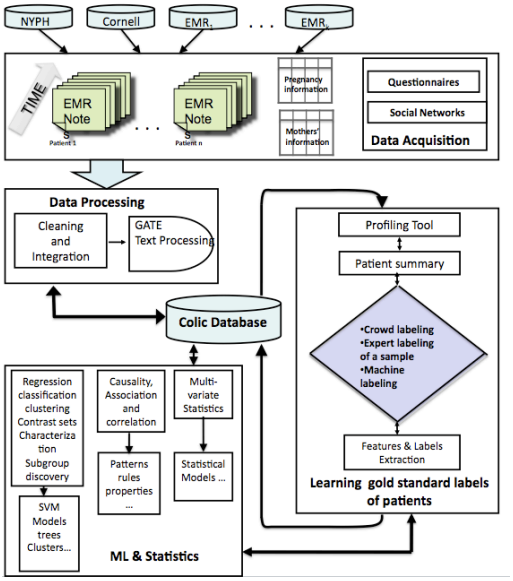

Figure 1: Overview of the project

There is a wealth of ML methods that would help visualize and explore colic data that may lead to more understanding of this condition. Besides research papers, we will dig into excellent books on Machine Learning and modern multivariate statistical techniques (e.g. [6, 7]). We will tackle the problem of understanding the colic condition from both descriptive and-predictive angles. In addition to applying existing ML methods, our specific research aims include: (1) to develop principled methods to learn what features and labels can be produced from structured information in combination with unstructured free text documents (e.g. medical records), and to accurately model domain entities (e.g. infants and parents) using this information, (2) to advance the research of extracting meaningful patterns from data, using correlation and association analysis [1, 4] and complex hypothesis generation together with evaluation measures, in this

case to identify causal patterns, (3) design an active ML framework to mix expert and non-expert labels in a principled way in a domain where collecting labels is uncertain and expensive.

References

[1] Rakesh Agrawal, Tomasz Imielinski, and Arun N. Swami. Mining association rules between sets of items in large databases. In Peter Buneman and Sushil Jajodia, editors, *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, Washington, D.C., May 26-28, 1993*, pages 207–216. ACM Press, 1993.

[2] RG Barr, FP Rivara, and et al. Barr, M. Effectiveness of educational materials designed to change knowledge and behaviors regarding crying and shaken-baby syndrome in mothers of newborns: a randomized, controlled trial. *Pediatrics*, 123:3:972 980, 2009.

[3] Ronald G. Barr, Roger B. Trent, and Julie Cross. 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.

[4] Sergey Brin, Rajeev Motwani, and Craig Silverstein. Beyond market baskets: Generalizing association rules to correlations. In Joan Peckham, editor, *SIGMOD 1997, Proceedings ACM SIGMOD International Conference on Management of Data, May 13-15, 1997, Tucson, Arizona, USA*, pages 265–276. ACM Press, 1997.

[5] Mary Brown, Ralf G Heine, and Brigid Jordan. Health and well-being in school-age children following persistent crying in infancy. *Journal of Paediatrics and Child Health*, 45(5):254–262, 2009.

[6] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference and prediction*. Springer, 2 edition, 2009.

[7] Alan Julian Izenman. *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*. Springer Publishing Company, Incorporated, 2008.

[8] S. Levitzky and R. Cooper. Infant colic syndrome maternal fantasies of aggression and infanticide. *Clinical Pediatrics*, (39):395400, 2000.

[9] S Morris, I St James-Roberts, J Sleep, and P Gillham. Economic evaluation of strategies for managing crying and sleeping problems. *Arch Dis Child*, 84:15–19, 2001.

[10] P Rautava, H Helenius, and L. Lehtonen. Psychosocial predisposing factors for infantile colic. *British Medical Journal*, 307:600604, 1993.

[11] Sijmen A. Reijneveld, Emily Brugman, and Remy A. Hirasing. Excessive infant crying: The impact of varying definitions. *Pediatrics*, 108(4):893–897, 2001.

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

[13] Dieter Wolke, Gabriele Schmid, Andrea Schreier, and Renate Meyer. Crying and feeding problems in infancy and cognitive outcome in preschool children born at risk: A prospective population study. *Journal of Developmental and Behavioral Pediatrics*, 30(2):226–238, 2009.