Advanced Methods for Big Data Analytics in Women's Health

Mary Regina Boland[†]
Perelman School of Medicine, University of Pennsylvania
Philadelphia, PA, USA
Email: bolandm@pennmedicine.upenn.edu

Karin Verspoor
University of Melbourne
Melbourne, Australia

Maricel G Kann
University of Maryland
Baltimore, MD, USA

Su Golder
University of York
York, UK

Lisa Levine, Karen O'Conner
Perelman School of Medicine, University of Pennsylvania
Philadelphia, PA, USA

Natalia Villanueva-Rosales
University of Texas
El Paso, TX, USA

Graciela Gonzalez-Hernandez
Perelman School of Medicine, University of Pennsylvania
Philadelphia, PA, USA
Email: gragon@pennmedicine.upenn.edu

Women's health is an often-overlooked aspect of medicine. The National Institutes of Health has emphasized the importance of investigating 'sex as a biological variable' in all new research grants. This has placed emphasis once again on the need for more nuanced studies that explore the role of sex as a biological variable on study outcomes. This session sought to elicit participation from researchers with strong backgrounds in women's health and informatics to develop methods that harness big datasets and 'big data techniques' including machine learning and artificial intelligence and apply those tools to women's health questions. Some important questions discussed in this section include Intimate Partner Violence (IPV) and the importance of early identification along with C-section deliveries and the importance of emergency vs. elective procedures.

[†] Work partially supported by the University of Pennsylvania (MRB)

Keywords: Text mining; machine learning; electronic health records; women's health; biomedicine; text mining; natural language processing; precision medicine.

1. Introduction

Recent advances in data science and digital epidemiology have unlocked an unprecedented amount of data for analysis, and uncovered previously unseen sex-specific patterns that point at marked differences in disease symptoms, progression and care that affect women of all ages. In 2016, the NIH published a guidance document[1] and changed its policy for reviewing proposals whereby accounting for "sex as a biological variable" became a required and scorable aspect of the research strategy, highlighting that "an over-reliance on male animals and cells may obscure understanding of key sex influences on health processes and outcomes". Dr. Kathryn Rexrode, chief of the Division of Women's Health at Brigham and Women's Hospital, is quoted[2] as succinctly stating the enormity of the problem: "without the inclusion of women, all the way through from basic research to clinical research, we can't be sure we really have the right answers for 51 percent of the population."

Aside from x-linked inheritable diseases, where women generally are carriers rather than express the disease[3], there are various aspects of women's health that challenge current methods. Recent research shows that variations in physiology may alter the pharmacokinetics or pharmacodynamics that determines drug dosing and effect for women, both in general and particularly during pregnancy[4], as hormonal and other biological differences may influence the impact of drugs, their effectiveness and their side effects. Over two-thirds of women receive prescription drugs while pregnant, with treatment and dosing strategies based on data from healthy male volunteers and non-pregnant women[5]. A paucity of research exists for optimizing prescription usage during pregnancy and more methods are needed that utilize artificial intelligence and machine learning [6-8]. In addition, health processes unique to women, such as pregnancy and pregnancy loss, menstruation and menopause require differential approaches to data representation and analysis. Disorders related to pregnancy and menstruation (such as miscarriage and heavy bleeding, which have a significant impact on women's health) have been recently found to be related to specific genetic mutations and are just being explored[9,10]. Furthermore, it has become clear through numerous recent studies that many diseases (cardiovascular disease, asthma, eating disorders, lung cancer, and autoimmune disorders, among others) impact women differently than men. Advanced data science methods specifically designed for exploring the influence that sex hormones and a women's physiology can have on the pathophysiology of these processes diseases and on their treatment are essential to advance our understanding of key processes in women's health, and, at the same time, the contrast could also shed light on the specific mechanisms that affect men.

This session highlights original research in the form of presentations and papers on the subject of big data and women's health. These include the use of machine learning methods to predict intimate partner violence (IPV) over 1 year before that violence occurs, along with pattern mining to determine patterns of IPV and co-occurrence with other subgroups of IPV, including sexual violence. Another method explores the role of emergency vs. elective C-section deliveries on the study of C-sections as an adverse outcome of delivery. These studies together enable further understanding of processes and diseases that are specific to women or differentially impact women. In harmony with the focus of PSB, the session emphasizes methodological advances and applications in data science, emphasizing reproducibility and validation.

2. Session Summary

The session includes three full-length papers competitively selected for inclusion that are focused on exploring problems associated with the complex problem of intimate partner violence, including patterns and injury prediction (2 distinct papers) and another study focused on deconstructing Cesarean sections into emergency versus elective to better understand this complex health outcome. We selected these important contributions that are applicable to utilize big datasets on studying women's health outcomes.

2.1. *Full-length papers*

In *Co-occurrence Patterns of Intimate Partner Violence,* the authors present a method that learns patterns of survivors of intimate partner violence (IPV) [11]. The main data-source for their study is the National Intimate Partner and Sexual Violence Survey (NISVS). The algorithm then clusters IPV into 5 different subgroups, and the authors compare these algorithm-chosen subgroups to traditional categories of IPV including physical violence, psychological aggression, sexual violence and micro-aggression. An important finding of their pattern analysis and co-occurrence pattern mining is that physical violence often co-occurs with psychological aggression and co-occurs less often with micro-aggression. In addition, the authors found that sexual violence tended to be a mutually exclusive form of IPV. Furthermore, this exclusive nature of sexual violence was so strong that it formed a single connected component in their subsequent network analysis. Overall, the findings from this study underscore the importance of breaking down IPV into type of IPV (e.g., physical violence, psychological aggression, sexual violence and micro-aggression) as these different types of IPV have different co-occurrence patterns and could be important for subsequent studies that link IPV to other health outcomes. The authors results suggest that their method effectively clusters types of IPV patients into subgroups that pertain to the type of IPV experienced by the patient and underscore the importance of co-occurrence patterns in IPV.

In *Intimate Partner Violence and Injury Prediction from Radiology Reports,* **Chen et al.** present an algorithm to predict which patients will experience injury as a result of IPV[12]. Because there are different types of IPV and not all IPV results in an injury to the partner, this method would be useful in determining *a priori* what patients will be likely to experience injuries as a result of IPV. This study differs from the previous study in that **Chen et al.'s** algorithm utilizes data from a large academic hospital's violence prevention support program from Jan 2013 - Jun 2018. For information on the subsequent injuries, the authors also had access to the patients' radiology reports. The authors develop a machine learning model assess IPV patients for risk of injury. Their method was successfully able to predict IPV 1.34 years before entrance into a violence prevention program with 95% sensitivity and 71% specificity. There are future plans to deploy their model as a clinical risk model for early detection of IPV.

In *Not All C-sections are the same: Investigating Emergency vs. Elective C-section Deliveries,* **Canelón et al.** present a method that utilizes Electronic Health Records (EHR) data to breakdown Cesarean sections (C-sections) into emergency vs. elective C-sections [13]. This breakdown is important because C-sections are often deemed an 'adverse outcome' across the board. However, there can be situations where it is the best outcome for a particular patient. Therefore, detailing out the important difference between a patient with an elective or planned C-section (e.g., in the case of a patient with complex comorbidities) versus an emergency C-section (e.g., as the result of an amniotic fluid embolism) is important when determining if the C-section is an adverse delivery outcome or not. In this study, the authors confirm that they adequately capture the differences between emergency and elective C-section by comparing the rates on weekday versus weekend, observing the expected drop in elective C-sections on the weekends. In addition, they modeled emergency deliveries in general as an adverse outcome and found that the following patient characteristics increased the risk of an emergency delivery: preterm birth, being younger than 25, identifying as Black/African American, Asian, or Other/Mixed, after adjusting for pregnancy number and C-section number for each patient. Interestingly, later pregnancies and repeat cesareans decreased the risk of an emergency delivery, and identifying as White, Hispanic, and Native Hawaiian/Pacific Islander patients appeared to lower the risk of an emergency delivery. The same risk factors and trends were found also for Cesarean deliveries (when looking at emergencies as the outcome) except that Asian patients did not have an increased risk of an emergency delivery in the C-section population, and Native Hawaiian/Pacific Islander patients did not have a reduced risk in this group. Overall, modeling the relationship between emergency vs. elective deliveries is important to understanding the relationship between other comorbidities and risk factors for C-sections. In addition, it is important for breaking down C-sections into those that are likely adverse events (e.g., emergencies) versus those that are due to comorbidities or other patient health issues (e.g., elective or planned).

3. Discussion

Informatics and 'Big Data Analytics' algorithms as applied and developed specifically for women's health questions such as those presented in this session enable novel approaches of existing data from diverse sources including EHR and survey data sources. These methods can be used for early prediction of IPV (over 1-year before violence occurs) and these methods have potential to be implemented in clinics for early identification of at-risk patients. Before these methods can be implemented, care must be taken that these machine learning algorithms have not 'learned' any features or other signals that may be indicative of patterns of care that may be biased against women or other minority or otherwise disadvantaged groups. However, the work presented in this session does represent important first steps towards early risk prediction for a complex issue such as IPV.

Overall, the research presented in this session focuses on different clinical questions that pertain to women and women's health, including IPV and C-section as an adverse outcome following delivery or birth. The studies presented explore the complexity and the need to take these larger groups (either IPV or C-sections) and further break them down into meaningful subclusters, in the case of IPV that would be breaking it down into physical violence vs. sexual violence and so forth. In the case of C-sections, it requires breaking it down into emergency vs. elective C-sections. This highlights the complexity of these outcomes and the importance of developing novel informatics algorithms to study these important women's health outcomes. The overarching goal will be to use these findings and algorithms to improve clinical care in the form of enhanced understanding of risk factors or to predict patients at risk for IPV for early identification at the point of care.

References

1. NIH. Consideration of Sex as a Biological Variable in NIH-funded Research [Internet]. . *Available from:* https://orwh.od.nih.gov/sites/orwh/files/docs/NOT-OD-15-102_Guidance.pdf. 2015.
2. Esposito L. 7 Major Gaps in Women's Health Research [Internet] Available from: https://health.usnews.com/health-care/patient-advice/slideshows/7-major-gaps-in-womens-health-research. 2017.
3. X-Linked Recessive Disorders [Internet] Available from: https://www.sciencedirect.com/topics/neuroscience/x-linked-recessive-disorders. 2020.
4. Louis GMB, Yeung E, Kannan K, et al. Patterns and Variability of Endocrine-disrupting Chemicals During Pregnancy: Implications for Understanding the Exposome of Normal Pregnancy. *Epidemiology.* 2019;30:S65-S75.
5. Feghali M, Venkataramanan R, Caritis S. Pharmacokinetics of drugs in pregnancy. Paper presented at: Seminars in perinatology2015.
6. Davidson L, Boland MR. Enabling pregnant women and their physicians to make informed medication decisions using artificial intelligence. *Journal of Pharmacokinetics and Pharmacodynamics.* 2020:1-14.
7. Boland MR, Polubriaginof F, Tatonetti NP. Development of a machine learning algorithm to classify drugs of unknown fetal effect. *Scientific reports.* 2017;7(1):1-15.
8. Duan R, Boland MR, Moore JH, Chen Y. ODAL: A one-shot distributed algorithm to perform logistic regressions on electronic health records data from multiple clinical sites. Paper presented at: PSB2019.
9. Maybin JA, Boswell L, Young VJ, Duncan WC, Critchley HO. Reduced transforming growth factor-β activity in the endometrium of women with heavy menstrual bleeding. *The Journal of Clinical Endocrinology & Metabolism.* 2017;102(4):1299-1308.
10. Husseini-Akram F, Haroun S, Altmäe S, et al. Hyaluronan–binding protein 2 (HABP2) gene variation in women with recurrent miscarriage. *BMC women's health.* 2018;18(1):143.
11. Hacaliefendioglu A, Ylmaz S, Koyuturk M, Karakurt G. Co-occurrence Patterns of Intimate Partner Violence. Paper presented at: PSB2021.
12. Chen IY, Alsentzer E, Park H, et al. Intimate Partner Violence and Injury Prediction From Radiology Reports. *PSB.* 2021.
13. Canelon S, Boland MR. Not All C-sections Are the Same: Investigating Emergency vs. Elective C-section deliveries as an Adverse Pregnancy Outcome. *PSB.* 2021.