A Text Mining Approach for Obstetric History Identification in Electronic Health Records

Cosmin A. Bejan PhD[1], Joshua C. Denny MD, MS[1,2]

[1]Department of Biomedical Informatics, [2]Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA

Introduction

Analyzing obstetric histories of female patients is highly relevant for population health studies. However, the women's obstetric history is not typically available in a computable form in Electronic Health Records (EHRs), but rather is often recorded in clinical notes. The *gravida/para/abortus* (GPA) system is the most common recording system to express the gravidity, parity, and abortion counts in clinical text. For instance, "$G_2P_2A_0$" encodes the obstetric history of a woman who has had two pregnancies that resulted in live births and no abortions. An extension of GPA is the GTPAL system, which makes the distinction between premature (P) and on term (T) births, and additionally records the number of living (L) children. Based on the specifications of the above-mentioned recording systems, we developed a large-scale text mining approach to automatically extract gravidity, parity, term, preterm, abortion, and living children counts corresponding to female patients admitted at Vanderbilt University Medical Center (VUMC).

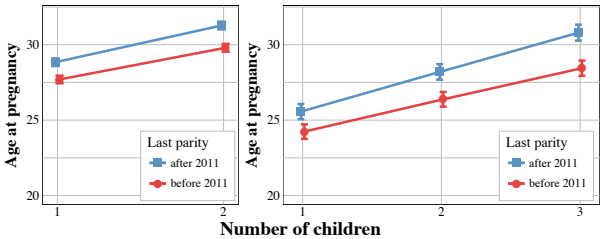
Materials and Methods

The dataset used in our study included all the clinical notes of the female patients from the Synthetic Derivative (SD), a de-identified version of the Vanderbilt EHR. We performed manual annotations on 1,000 randomly sampled notes with types either from the general clinical domain (e.g., 'discharge summary', 'history and physical examination') or from the obstetrics and gynecology domain (e.g., 'gynecology clinic visit', 'gynecology annual clinic visit'). Common proportions of 70%–30% were used to split the annotated notes into training and test sets, respectively. We analyzed the notes from the training set to capture predefined patterns of the annotated obstetric histories and to implement these patterns with regular expressions. We compared the manual annotated mentions with the ones extracted by the regular expressions and computed the precision, recall, and F1-measure. During training, we iteratively refined the regular expressions with the purpose of maximizing the F1-measure. For evaluation, we reported the results from applying the regular expressions derived from the above-mentioned optimization process on the test set. The extraction of all female patients with obstetric histories in the SD allowed us to test the hypothesis that, more recently, women defer their pregnancies to later ages. For this, we equally divided the mothers of the same parity by the time of their last pregnancy and compared the mean ages at parity event from each subgroup.

Results

The system evaluation revealed a precision, recall, and F1-measure of 98.9, 87.7, and 93.0, respectively. The error analysis indicated that most of the false negative obstetric histories were expressed in text using different or modified recording systems. When running the system on >50 million notes corresponding to >1.1 million women from the SD, >750,00 obstetric history mentions of >150,00 distinct patients were identified. From the mothers with all their parity events mentioned in notes, there were 5,683 and 737 of them with 2 and 3 children, respectively. For these two cohorts, Figure 1 confirms the hypothesis that in recent years women give birth to their children at later ages. For instance, the difference in age at the second pregnancy between the two subgroups of mothers with 2 children is approximately one year (29.35±0.25 vs. 30.27±0.23; $P<0.001$). We also noticed a tendency for women who had their first child earlier in their life to have more pregnancies than those having their first pregnancy at a later age. For instance, the mean age for the parity one of women with four pregnancies is 22.76±0.62 years whereas the mean age for the parity one of women with two pregnancies is 27.63±0.16 ($P<0.001$).

Figure 1 Temporal trends in mean age at the time of parity event for mothers of 2 and 3 children.



Conclusion

Our text mining approach is a scalable solution for high-precision extraction of obstetric histories from the EHR. Future work could enhance the identification of large-scale patient cohorts to investigate clinical and genetic associations with pregnancy events.