

DEVELOPMENT OF A DATA-MINING ALGORITHM TO IDENTIFY AGES AT REPRODUCTIVE MILESTONES IN ELECTRONIC MEDICAL RECORDS

JENNIFER MALINOWSKI

*Center for Human Genetics Research, Vanderbilt University, 2215 Garland Avenue, 519 Light Hall
Nashville, TN 37232, USA*

Email: jennifer.malinowski@vanderbilt.edu

ERIC FARBER-EGER

*Center for Human Genetics Research, Vanderbilt University, 1207 17th Avenue, Suite 300
Nashville, TN 37232, USA*

Email: eric.h.farber-eger@vanderbilt.edu

DANA C. CRAWFORD

*Department of Molecular Physiology and Biophysics, Center for Human Genetics Research,
2215 Garland Avenue, 519 Light Hall
Nashville, TN 37232, USA*

Email: crawford@chgr.mc.vanderbilt.edu

Electronic medical records (EMRs) are becoming more widely implemented following directives from the federal government and incentives for supplemental reimbursements for Medicare and Medicaid claims. Replete with rich phenotypic data, EMRs offer a unique opportunity for clinicians and researchers to identify potential research cohorts and perform epidemiologic studies. Notable limitations to the traditional epidemiologic study include cost, time to complete the study, and limited ancestral diversity; EMR-based epidemiologic studies offer an alternative. The Epidemiologic Architecture for Genes Linked to Environment (EAGLE) Study, as part of the Population Architecture using Genomics and Epidemiology (PAGE) I Study, has genotyped more than 15,000 patients of diverse ancestry in BioVU, the Vanderbilt University Medical Center's biorepository linked to the EMR (EAGLE BioVU). We report here the development and performance of data-mining techniques used to identify the age at menarche (AM) and age at menopause (AAM), important milestones in the reproductive lifespan, in women from EAGLE BioVU for genetic association studies. In addition, we demonstrate the ability to discriminate age at naturally-occurring menopause (ANM) from medically-induced menopause. Unusual timing of these events may indicate underlying pathologies and increased risk for some complex diseases and cancer; however, they are not consistently recorded in the EMR. Our algorithm offers a mechanism by which to extract these data for clinical and research goals.

1. Introduction

1.1 *Women's health and the reproductive lifespan*

Though women comprise more than 50% of the US population[1] and there are notable differences in the incidences and severity of diseases between men and women, from Alzheimer's disease[2] to inflammatory arthritis[3], only in the last few decades has the importance of women's health and physiologic differences between males and females in the research setting come to the forefront of researchers and government agencies[4]. Age at menarche (AM) and age at menopause (AAM) define the boundaries of the reproductive lifespan in women. The timing of these events has also been linked to numerous diseases and complex traits [5]. Fertility is directly impacted by the length of the reproductive lifespan. Earlier AM and later AAM have been associated with heightened risks for breast, ovarian, and endometrial cancers, elevated blood pressure, and increased glucose intolerance, driven by a significant extent by the additional exposure to circulating estrogens over an extended reproductive lifespan [6]. Early AAM has been associated with increased risk for cardiovascular disease [7]. More directly, extremely early or late attainment of these reproductive milestones can indicate underlying pathologies, such as pituitary diseases, hormone imbalances, and nutritional insufficiencies [5].

National surveys have calculated the average AM to be 12.4 years and age at natural menopause (ANM) at 51 years [8]. The genetic contribution to the timing of menarche and natural menopause is estimated to be approximately 0.50, however variants identified through numerous genome-wide association studies (GWAS) account for <10% of the observed variation in either AM or ANM [8]. Cross-sectional and longitudinal studies have shown recent secular trends in the earlier attainment of pubertal milestones (breast development, appearance of pubic hair, menarche) from the 1960s to present and later age at natural menopause [9]. The earlier AM is accelerated in girls of African American and Hispanic ancestry, a bias that remains after adjusting for socioeconomic variables and body mass index (BMI) [10]. The difference observed in the timing of reproductive events across ethnicities highlights the importance of conducting research in diverse populations—a challenging enterprise given the limited diversity in cohorts available for women's health outcomes research.

1.2 *Research use of electronic medical records*

Electronic medical/health records (EMRs/EHRs) are becoming more widely used in clinical practice and hospital settings. Motivated in part by the 'meaningful use' requirement for supplemental reimbursements for Medicare and Medicaid claims through the Health Information Technology for Economic and Clinical Health (HITECH) Act, widespread adoption of EMR technology is expected to improve patient outcomes and streamline health care processes and may be helpful in the goal of "personalized medicine" [11-14]. A significant measure of 'meaningful use' is the recording of patient data (e.g., demographic, medication allergy, smoking status, vital signs) as structured data [12]. Additional measurements of 'meaningful use' include the dissemination of clinical quality measurements to states or other governmental oversight agencies.

Immunization and reportable disease statistics are two examples of EMR data that can be leveraged for public health research [15].

The rich phenotypic data existing in EMR systems allows clinicians and researchers to identify potential cohorts, while EMRs that are linked to biobanks extend this framework to genotype-phenotype association studies. Traditional epidemiologic studies are costly and require significant amounts of time to complete; furthermore, these studies may not include sufficient numbers of individuals from diverse ancestries. The Epidemiologic Architecture for Genes Linked to Environment (EAGLE) Study seeks to address these limitations by enabling high-throughput identification and generalization of genotype-phenotype associations in ethnically diverse research populations. Accessing data from EMRs for use in research may prove to be a cost effective alternative to traditional ascertainment and data collection. One challenge to research use of EMR-derived data is the lack of consistency in recording certain types of data in the EMR. Despite the obvious health implications, AM and AAM/ANM are not recorded consistently or in a standardized manner in the EMR. This presents a challenge for researchers and suggests algorithm development is a necessary first step in developing a resource for women's health studies in diverse populations.

1.3 BioVU

BioVU is the Vanderbilt University Medical Center (VUMC) biorepository linked to the EMR system. Beginning in 2007, discarded blood samples from routine clinical testing have the DNA extracted, stored, and linked to a de-identified version of the EMR termed the Synthetic Derivative (SD). As of mid-2012, more than 150,000 samples have been collected for BioVU, including more than 16,000 pediatric samples. Patients are given the opportunity to opt-out of BioVU at any time. Once a sample has been accepted into the system, a unique ID is generated through a one-way hash mechanism and linked to that patient's SD. The SD removes or de-identifies Health Insurance Portability and Accountability Act (HIPAA) information, such as names, geographical locations, and social security numbers, and replaces dates with dates that have been randomly shifted by up to six months. The date shifting is consistent within a single SD record. The SD enables researchers to examine genome-phenome associations and identify cohorts for research.

2. Methods

2.1. Population

As part of the Population Architecture using Genomics and Epidemiology (PAGE) I Study, EAGLE genotyped all non-European descent patients in BioVU (EAGLE BioVU, n=15,863) on the Metabochip, a custom genotyping array with an emphasis on cardiovascular disease and metabolic traits. This array also includes over 2200 SNPs associated at genome-wide significance to any trait published in the NHGRI GWAS catalog as of August 2009, with additional proxy SNPs chosen based on linkage disequilibrium (LD) in both CEU and YRI HapMap II datasets [16]. Overall, 11,521 African Americans, 1,714 Hispanics, 1,122 Asians and others were

genotyped on the MetaboChip by EAGLE. For the AM study, all females age >6 in EAGLE BioVU as of January 31, 2013 were eligible for inclusion. For the AAM study, all females >18 years were eligible for inclusion; for the ANM study, only women ages ≥ 41 were eligible for inclusion. All patients were of diverse ethnicity.

2.2. Algorithm development

We developed a flow chart to visualize the inclusion/exclusion processes for the algorithms (Fig. 1A (AM) and Fig. 1B/C (AAM/ANM)). AM and age at menopause or age at natural menopause (AAM/ANM) are not consistently recorded in the EMR system at VUMC; individuals may enter BioVU through numerous outpatient clinics. The lack of a pre-specified field for AM and AAM/ANM in the EMR necessitated a combination of free text data mining using regular expressions/pattern matching, billing (ICD-9) codes, and procedure (CPT) codes to identify AM and AAM/ANM in the subsequently generated SD. All analysis for this study was performed using the SD.

2.2.1 Age at menarche (AM)

Primary exclusion criteria for AM phenotype consisted of four components: age <7 years, male sex, ICD-9 codes for delayed puberty/sexual development (259.0) and precocious puberty/sexual development (259.1), and keywords (Figure 1A). Inclusion of any of the preceding criteria in the SD resulted in exclusion for the AM study. As part of the de-identification data scrubbing to convert a patient's EMR to the SD, ages and dates may be masked and listed as "birth-12" or "in teens." Dates and ages which are not masked were date shifted by up to six months forward or backward from the actual date.

To identify a listed AM for an individual, we utilized pattern matching to seek instances with menarche keyword phrases (Figure 1A). Numbers and dates were allowed to be included as numerals only. Instances where the AM was listed as a date used the subject's birthdate to calculate the age (in years) at menarche. In cases of ties, where more than one AM was identified and recorded an equal number of times in the SD, the AM was determined to be the one listed first in the SD. If the algorithm identified multiple versions of the AM (an exact age, an age calculated from a date, or a de-identified age), a hierarchy was used to determine the AM for the output, where an exact age or date was prioritized over de-identified age ranges. Instances where multiple different ages were listed in the SD as AM defaulted to the age listed most frequently. We considered situations where the algorithm identified an exact AAM and a de-identified AAM range containing the exact AAM to be the same for purpose of calculating sensitivity, specificity, and positive predictive value (PPV), but different for the purpose of calculating accuracy. The resulting output file contained the subject's unique research id (RUID), date of birth, and either an algorithm-generated AM or null value.

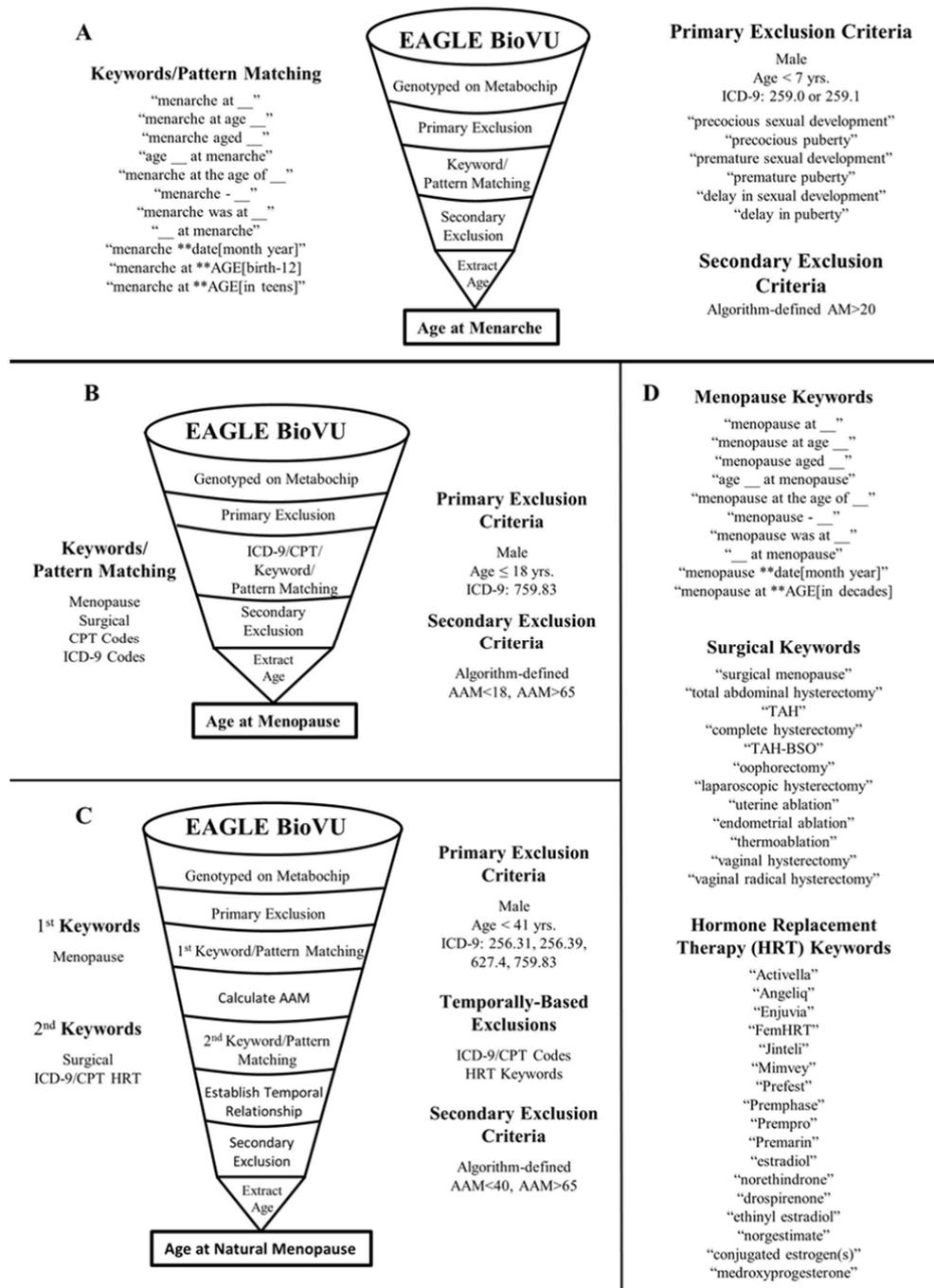


Figure 1. Flow chart for (A) age at menarche (AM), (B) age at menopause (AAM), (C) age at natural menopause (ANM), and (D) keywords for AAM and ANM algorithms.

2.2.2 Age at menopause (AAM)

For an algorithm to identify all post-menopausal women and their age at menopause (AAM), we initially excluded all males, set a minimum age of 18 years, and excluded patients with a Fragile X diagnosis (ICD-9 759.83) (Figure 1B). Pattern matching was utilized to find keyword phrases similar to those used in the menarche algorithm, substituting “menopause” for “menarche” (Figure 1D). Furthermore, we included keywords pertaining to surgical procedures that induce cessation of menses/menopause (Figure 1D). We excluded instances where the word “possible” immediately preceded a keyword. For instances where the SD had scrubbed the exact age, decade-specific results (e.g. “in 30s”, “in 50s”) were captured by our algorithm. CPT and ICD-9 (Table 1) codes were used to identify women with surgical menopause or menses-ceasing procedures.

Table 1. CPT and ICD-9 codes used for menopause (AAM/ANM) algorithm development.

| CPT codes | | | ICD-9 codes | | |
|-----------|-------|-------|-------------|-------|-------|
| 58150 | 58285 | 58548 | 65.5 | 68.3 | 68.69 |
| 58152 | 58290 | 58550 | 65.51 | 68.31 | 68.7 |
| 58180 | 58291 | 58552 | 65.52 | 68.39 | 68.71 |
| 58200 | 58292 | 58553 | 65.53 | 68.4 | 68.79 |
| 58260 | 58293 | 58554 | 65.64 | 68.41 | 68.9 |
| 58262 | 58294 | 58563 | 65.6 | 68.49 | |
| 58263 | 58353 | 58570 | 65.61 | 68.5 | |
| 58267 | 58541 | 58571 | 65.62 | 68.51 | |
| 58270 | 58542 | 58572 | 65.63 | 68.59 | |
| 58275 | 58543 | 58573 | 65.64 | 68.6 | |
| 58280 | 58544 | | 68.23 | 68.61 | |

After SD review of initial algorithms and subject matter knowledge, we implemented secondary exclusion criteria based on the algorithm-identified AAM and excluded subjects with a calculated $AAM < 18$ or $AAM > 65$ (Figure 1B). A hierarchy was used to determine the AAM for the output, with an exact age or date identified by keyword or pattern matching and ICD-9/CPT codes prioritized over de-identified age ranges. In rare instances where the algorithm identified more than one AAM for a subject, the age recorded most frequently was determined to be the AAM for that patient. In cases of ties, where more than one AAM was identified and recorded an equal number of times in the SD, the AAM was determined to be the one listed first in the SD. We considered situations where the algorithm identified an exact AAM and a de-identified AAM range containing the exact AAM to be the same for purpose of calculating sensitivity, specificity, and PPV, but different for the purpose of calculating accuracy. The resulting output file contained the subject’s unique research id (RUID), date of birth, race/ethnicity, either an algorithm-

generated AAM or null value, the method by which the AAM was calculated (e.g., from ICD-9 code, keyword), and the date in the SD corresponding to the AAM identification.

2.2.3 Age at natural menopause (ANM)

To discriminate age at natural menopause (ANM) from all instances of menopause (AAM), we extended the AAM algorithm to exclude women aged <41 years, men, and subjects with ICD-9 codes signifying premature ovarian failure/premature menopause (256.31), artificially induced menopause (627.4), ovarian failure (256.39), and Fragile X syndrome (759.83) (Figure 1C). We used pattern matching with the menopause keywords to identify an age at menopause (Figure 1D). We did not use ICD-9 codes, CPT codes, or keywords associated with procedures that induce menopause to identify subjects for the ANM cohort.

Medication delivery and prescriptions are captured by the EMR at VUMC and are included in the SD. To ascertain the temporal relationship between AAM and menopause-inducing/menses-ceasing surgery or hormone replacement therapy (HRT) use, we first calculated the AAM with the alternate algorithm (Figure 1C). Surgery-inducing menopause, determined through CPT and/or ICD-9 codes or keywords, and HRT were not exclusion criteria unless the first instance of surgery or HRT occurred prior to the extended algorithm-identified AAM. Keyword pattern matching was performed using surgical keywords (Figure 1D). We used a combination of brand-name and generic names for HRT identification (Figure 1D). If AAM was identified and no keywords or CPT/ICD-9 codes were found to indicate artificially induced menopause, the subject was deemed to have undergone natural menopause. If surgery or HRT occurred after the algorithm-determined ANM, the subject was also considered to have undergone natural menopause. If the subject had either surgery or used HRT prior to menopause, they were excluded from the cohort and the resulting output was a null value.

We implemented secondary exclusion criteria (Figure 1C) based on the algorithm-identified age at menopause and excluded subjects with a calculated $ANM < 18$ or $ANM > 65$ based on subject matter knowledge and review of early versions of our algorithms. A hierarchy was used to determine the ANM for the output. If the algorithm determined more than one ANM for a subject, we used the same procedure as described above to determine the final ANM generated by our query. We again considered situations where the algorithm identified an exact ANM and a de-identified ANM range containing the exact ANM to be the same for purpose of calculating sensitivity, specificity, and PPV, but different for the purpose of calculating accuracy. The resulting output file contained the subject's unique research id (RUID), date of birth, race/ethnicity, either an algorithm-generated ANM or null value, the method by which the ANM was calculated (e.g., from exact date, de-identified age), and the date in the SD corresponding to the ANM identification.

2.3. Manual review

To determine the sensitivity, specificity, PPV, and accuracy of the AM, AAM, and ANM algorithms, extensive manual chart review was performed by a single individual for consistency. Each algorithm output contained three types of values: exact ages, de-identified ages, and null values. For each algorithm, a random number generator was used to randomize RUIDs within each of the three types of output and the subjects were then sorted in ascending value by the random number. The first 50 subjects in the exact age and de-identified age categories and the first 100 subjects with a null value had their SD reviewed manually to determine the AM, AAM, or ANM. Sensitivity, specificity, PPV and accuracy were calculated by comparing the automated algorithm result to the manual review result for each subject.

3. Results

3.1 Population characteristics

A total of 10,051 females were genotyped on the MetaboChip in BioVU by EAGLE for various studies. We identified an age for menarche (exact or de-identified) in 1,618 individuals. For the AAM algorithm, we identified an AAM (exact age or de-identified decade) for 1281 individuals. We identified 83 individuals with an ANM (exact or de-identified decade) (Table 2). The algorithm-extracted mean AM in our population was 12.7 (+/- 2.1) yrs. The mean AAM in our population was 44.6 (+/- 9.8) yrs. and the mean ANM was 49.7 (+/- 5.6) yrs. (Table 2). Approximately half of the algorithm extracted AM (54.7%) and ANM (47.0%) were exact ages, while the majority of AAM (92.5%) were exact ages (Table 2).

Table 2. Population characteristics for women with algorithm-identified age at menarche (AM), age at menopause (AAM), and age at natural menopause (ANM) from EAGLE BioVU. Abbreviations: standard deviation (sd), years (yrs).

| | AM | AAM | ANM |
|---------------------------------|------------|------------|------------|
| N, total | 1618 | 1281 | 83 |
| exact age (n) | 885 | 1185 | 39 |
| de-identified age (n) | 733 | 96 | 44 |
| Age at event, mean +/- sd (yrs) | 12.7 (2.1) | 44.6 (9.8) | 49.7 (5.6) |
| Age range at event (yrs) | 8-20 | 18-65 | 40-65 |
| Race/ethnicity (n) | | | |
| African American | 1232 | 1112 | 62 |
| Hispanic | 120 | 45 | 4 |
| Asian | 115 | 66 | 11 |
| Other | 151 | 58 | 6 |

3.2 AM algorithm performance

We manually reviewed 200 SD entries for the AM algorithm to determine sensitivity, specificity, PPV, and accuracy. Of the 100 subjects with an algorithm-specified AM, 94 were confirmed by manual review. For the 100 subjects without an AM captured by the algorithm, 99 were not found to have an identifiable AM upon manual review. The AM algorithm had a sensitivity and specificity of 99.0% and 94.3%, respectively, and a PPV of 94.0% (Table 3). We calculated the accuracy of the algorithm by comparing the results for the 94 subjects with both manually identified and algorithm identified AMs, requiring identical results for concordance. Of these 94 subjects, we found 87 where the AM matched in both manual and algorithm identification for an accuracy of 92.6% (Table 4). We observed instances where the algorithm calculated an exact AM (e.g., 8) and manual review found a de-identified AM (e.g., birth-12), or vice-versa. If we allow these to be concordant, accuracy increases to 94.7%.

Table 3. Performance of the age at menarche (AM), age at menopause (AAM), and age at natural menopause (ANM) algorithms in women from EAGLE BioVU.

Abbreviations: positive predictive value (PPV).

| | Sensitivity | Specificity | Accuracy | PPV |
|--------------------|--------------------|--------------------|-----------------|------------|
| AM (n=200) | 99.0% | 94.3% | 92.6% | 94.0% |
| AAM (n=200) | 94.4% | 85.6% | 52.4% | 84.0% |
| ANM (n=183) | 89.8% | 75.8% | 75.5% | 63.9% |

3.3 AAM algorithm performance

For the AAM algorithm, we manually reviewed 200 SD entries to determine sensitivity, specificity, PPV, and accuracy. Of the 100 subjects with an algorithm-identified AAM, we identified 82 with AAM via manual review. Only five of the 100 subjects without an algorithm-identified AAM were found to have an identifiable AAM with manual review. Overall, our algorithm was found to have 94.4% sensitivity, 85.6% specificity, and a PPV of 84.0% (Table 3). We also calculated the accuracy of our AAM algorithm by comparing the algorithm-obtained AAM to the manual review-obtained AAM. We observed a 52.4% exact concordance within our 82 subjects with AAMs calculated from both manual review and the algorithm. If we allowed a de-identified age range encompassing an exact age to be considered concordant with the exact age obtained from the other method, our accuracy improved to 61.9%.

3.4 ANM algorithm performance

The ANM algorithm identified 83 individuals with an ANM; therefore, we manually reviewed 183 SD entries to determine the specificity, sensitivity, PPV, and accuracy of our ANM algorithm. Of the 100 individuals with no algorithm-identified ANM, manual review of the SD found 6 instances with an identifiable ANM (Table 3). Of the 83 individuals with an algorithm-specified ANM,

manual review confirmed 53. Overall, the sensitivity and specificity of the ANM algorithm were 89.8% and 75.8%, respectively, and the PPV was 63.9%. Of the 53 subjects with both algorithm- and manually-identified ANM, 40 were an exact match, yielding an accuracy of 75.5%. We again observed instances where the algorithm yielded an exact age, but manual review of the SD obtained only a de-identified ANM range that encompassed the exact age, and vice-versa; if we considered these as concordant, our accuracy increased to 81.1%.

4. Conclusion

Menarche and menopause are the bookends of the reproductive lifespan in women. The timing of these events may increase risk for various complex disorders and cancers, such as osteoporosis and breast cancer [5]. Precocious or delayed menarche may signal the occurrence of hormonal imbalance, inadequate nutrition or caloric intake, or pituitary diseases [5]. The timing of menopause directly affects reproductive capabilities. In addition, premature menopause may result from hormonal imbalances, genetic disorders such as Fragile X Syndrome, metabolic disorders, or autoimmune diseases such as thyroid disease or rheumatoid arthritis [17]. Though the timing of menarche and menopause may increase risk for disease or indicate underlying pathologies, this information is not consistently included in electronic health records, leading to missed opportunities to inform clinical care and represents a challenge to clinicians and researchers alike.

Data-mining EMRs has been used to identify cohorts for research studies [18-21], determine smoking status [22], and predict disease, such as sepsis [23]. Our development of algorithms to extract these important data is notable for the emphasis on diverse populations and attention to women's health, both historically underrepresented in health outcomes research. The menarche (AM) and menopause (AAM) algorithms have PPV>80% and high specificity and sensitivity, though accuracy of the AAM algorithm was just over 50%. The age at natural menopause (ANM) algorithm had moderately high (>75%) sensitivity and specificity but the lowest PPV, at 63.9%. However, the accuracy of the ANM algorithm bested that of the AAM (75.5% vs. 52.4%, respectively). In addition, the algorithm-extracted ages at menarche, menopause, and natural menopause are consistent with published research, validating our methodology.

Several factors may have reduced the performance of our menopause algorithms. We observed many instances where the ages calculated by the algorithm and by manual review differed by one year. This may have been the result of the date-shifting done within each individual's SD for de-identification purposes. If the method for calculating the age differed between the methods, it is possible this could result in the observed one-year difference. When we allowed a +/- 1 year difference in the algorithm and manual identified AAM and ANM, the accuracy of our algorithms improved to 70.2% and 90.6%, respectively. The timing of menopause is challenging to identify, as the menstrual cycle becomes more erratic as a woman moves through perimenopause into menopause. Months may lapse between cycles; hormone levels may change substantially. In addition, the normal menopausal age range is quite large, taking place between the ages of 40 and 60. These factors challenge the accurate dating of the onset of menopause.

Furthermore, an algorithm designed to identify the age at menopause may not accurately reconcile multiple mentions in an EMR of menopause. Discerning between natural menopause and medically/surgically induced menopause is an additional challenge. Our extensive list of time-dependent exclusions for HRT and surgical procedures was not exhaustive and may have led to the algorithm identifying an ANM where manual review identified HRT and/or a procedure artificially inducing menopause. Correctly identifying the temporal relationship between attainment of natural menopause and surgical procedures that result in menopause may perform inconsistently in the absence of these data in structured fields in an EMR. Addressing some of these issues by including structured fields for age at menarche, age at menopause, and type of menopause (natural/medical), and standardizing the reporting of these data could greatly improve the performance of our algorithms.

We have demonstrated the performance of algorithms designed to extract the age at menarche and age at menopause from the Synthetic Derivative, a de-identified version of the electronic medical record at Vanderbilt University Medical Center. Furthermore, we have developed an algorithm to discriminate naturally occurring menopause from artificially-induced menopause. Our method combining text-mining for regular expressions and pattern matching, and structured data derived from the EMR to obtain the age at menarche and the age at menopause is likely to be easily transferable to other institutions, given the simplicity of the approach. Overall, these algorithms provide an opportunity for researchers and clinicians to obtain these valuable, though inconsistently reported data.

Acknowledgments

This work was supported by NIH U01 HG004798 and its ARRA supplements. The dataset(s) used for the analyses described were obtained from Vanderbilt University Medical Center's BioVU which is supported by institutional funding and by the Vanderbilt CTSA grant UL1 TR000445 from NCATS/NIH. The Vanderbilt University Center for Human Genetics Research, Computational Genomics Core provided computational and/or analytical support for this work.

References

1. Bureau of Census, C2010BR-03 (2011).
2. K. Irvine, K.R. Laws, T.M. Gale, and T.K. Kondel, *J Clin Exp Neuropsychol*. doi:10.1080/13803395.2012.712676 (2012).
3. C. Barnabe, L. Bessette, C. Flanagan, *et al.*, *J Rheumatol* **39**, 1221 (2012).
4. C. Taylor, *J Womens Health* **3**, 143 (1994).
5. P. Hartge, *Nat Genet.* **41**, 637 (2009).
6. V. Dvornyk and U.H. Waqar, *Hum Reprod Update* doi:10.1093/humupd/dmr050 (2012).

7. W.A. Rocca, B. R. Grossardt, V.M. Miller, *et al.*, *Menopause* **19**, 272 (2012).
8. C. He and J.M. Murabito, *Mol Cell Endocrinol*, doi:10.1016/j.mce.2012.05.003 (2013).
9. P. Kaplowitz, *Curr Opin Obstet Gynecol* **18**, 487 (2006).
10. T. Wu, P. Mendola, and G.M. Buck, *Pediatrics* **110**, 752 (2002).
11. D. Blumenthal and M. Tavenner, *N Engl J Med.* **363**, 501 (2010).
12. D. Blumenthal, *N Engl J Med.* **362**, 382 (2010).
13. D. Blumenthal, *N Engl J Med.* **365**, 2426 (2011).
14. A.K. Jha, C.M. DesRoches, E.G. Campbell, *et al.*, *N Engl J Med.* **360**, 1628 (2009).
15. R. Kukafka, J.S. Ancker, C. Chan, *et al.*, *J Biomed Inform.* **40**, 398 (2007).
16. B.F. Voight, H.M. Kang, J. Ding, *et al.*, *PLoS Genet.* doi:10.1371/journal.pgen.1002793 (2012).
17. T.C. Okeke, U.B. Anyaehie, and C.C. Ezenyeaku, *Ann Med Health Sci Res.* **3**, 90 (2013).
18. K.M. Newton, P.L. Peissig, A.N. Kho, *et al.*, *J. Am. Med. Inform. Assoc.* doi: 10.1136/amiajnl-2012-000896 (2012).
19. M. J. Stratton-Loeffler, J. C. Lo, R.L. Hui, *et al.*, *J. Manag. Care Pharm.*, **18**, 497 (2012).
20. J. Warren, J. Kennelly, D. Warren, *et al.*, *Stud. Health Technol. Inform.* **178**, 228 (2012).
21. J. S. Brownstein, S.N. Murphy, A.B. Goldfine, *et al.*, *Diabetes Care* **33**, 526 (2010).
22. L.K. Wiley, A. Shah, H. Xu, W.S. Bush. *J. Am. Med. Inform. Assoc.* doi: 10.1136/amiajnl-2012-001557 (2013).
23. S. Mani, A. Ozdas, C. Aliferis, *et al.*, *J. Am. Med. Inform. Assoc.* 10.1136/amiajnl-2013-001854 (2013).