Development and Validation of a Machine Learning Algorithm for Problematic Menopause in the Canadian Primary Care Sentinel Surveillance Network (CPCSSN)

Anh N.Q. Pham (anh.pham@ualberta.ca)
University of Alberta
Michael Cummings
University of Calgary
Nese Yuksel
University of Alberta
Beate Sydora
University of Alberta
Tyler Williamson
University of Calgary
Stephanie Garies
University of Calgary
Russell Pilling
University of Calgary
Sue Ross
University of Alberta

Research Article

Keywords: Primary care data, case definition, EMR phenotyping, treated/problem menopause

Posted Date: January 3rd, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2403081/v1

Version of Record: A version of this preprint was published at BMC Medical Informatics and Decision Making on October 5th, 2023. See the published version at https://doi.org/10.1186/s12911-023-02298-x.

Abstract

Background

Menopause is a normal transition in a women's life. For some women, it is a stage without significant difficulties; for others, menopause symptoms can severely affect their quality of life. Identifying problematic menopause is essential to study the condition and to improve quality of care. This study developed and validated a case definition for problem menopause using Canadian primary care electronic medical records.

Methods

We used data from the Canadian Primary Care Sentinel Surveillance Network (CPCSSN). A case definition was developed using a reference set created by expert reviewers and a machine learning approach was applied to produce a case definition. Methods to select the most appropriate features and to re-balance our cohort were also applied.

Results

We randomly selected 2,776 women aged 45–60 for this analysis. An algorithm of two occurrences of ICD-9-CM code 627 in diagnosis fields within 24 months *OR* one occurrence of ATC code G03CA in medication fields detected problem menopause. This definition produced sensitivity 81.5% (95%CI 76.3%-85.9%), specificity of 93.5% (95%CI 91.9%-94.8%), positive predicted value 73.8% (95%CI 68.3%-78.6%), and negative predicted value 95.7% (95%CI 94.4%-96.8%).

Conclusion

Our case definition for problem menopause is useful for epidemiological study and demonstrated strong validity metrics. This case definition will help inform future studies exploring management of menopause in primary care settings.

Background

The menopause transition is a normal stage in a woman's life, generally occurring around the age of 51 in North America (with a typical range between 45 to 55 years), when the production of hormones (i.e., estrogen, progesterone) from the ovaries gradually decline and women cease menstruation [1]. Menopause can also occur as a result of surgery that removed both ovaries, for example for women who are at risk of ovarian or breast cancer [2]. Most women experience menopause without significant difficulty. For others, menopause symptoms can be severe and debilitating. Among women referred to a menopause clinic in a large urban centre in Alberta, Canada, self-reported moderate or severe menopause

symptoms included sleep disorders (76%), night sweats (51%), hot flashes (50%), mood swings (48%), and depression (38%) [3]. The menopause transition, and especially during perimenopause, is associated with increased risk of developing depressive episodes, particularly among women who previously suffered from depression [4]. Female suicide rates in western countries peak around the age of menopause [5], which is considered a vulnerable period for women with a history of depression [4].

Most women experiencing difficult menopausal symptoms will first consult their family physicians for help, but little is known about the epidemiology or clinical management of this condition in a primary care setting. A literature search for how menopause is defined in healthcare identified only three papers [6]–[8] but none were focused on primary care. It is therefore important to understand the actual burden of menopausal symptoms in Canadian primary care and consider whether service provision could be improved.

The aims of this project were to develop and validate a case definition of 'problematic menopause' in women who consult family physicians, in preparation to address questions about the clinical management of menopause in primary care in Canada and to enhance the care of women with troublesome symptoms. The case definition for problem menopausal symptoms will be useful to primary care clinicians who wish to identify and follow-up these women to refine their treatment for better care, for example by optimising menopausal hormone therapy (MHT), and to explore comorbidities such as depression and managing related risk factors.

Methods

Data source

CPCSSN is a collaboration of practice-based, primary care surveillance and research networks across Canada to better understand the epidemiology and management of chronic health conditions to improve patient care [9], [10]. CPCSSN collects and combines deidentified, patient level, primary care electronic medical record (EMR) data from participating practices across the country. Depending on the local implementation of electronic medical records, CPCSSN data can derive from as long ago as 1998. In general, patients with data recorded in CPCSSN are slightly older and more likely to be women than the Canadian general population [11].

To date, CPCSSN case definitions have been validated for chronic obstructive pulmonary disease (COPD), dementia, depression, diabetes, type 1 diabetes, diabetic nephropathy, epilepsy, hypertension, osteoarthritis, parkinsonism, speech disorders, herpes zoster, pelvic floor disorders, adult asthma, and child asthma with validation metrics ranging from good to excellent [12], [13]. No previous case definition existed for menopause and therefore we designed this development and validation study.

Population

There were 211,103 women aged 45y-60y in the national CPCSSN database at the time of the December 2020 data extraction.

The sample size for the validation set was determined using the Wald 95% confidence interval formula:

$$\mathrm{Numberofchartsrequired} = 1.96^2 \frac{S_n(1-S_n)}{\left(\frac{c}{2}\right)^2 p} \text{ [14]}$$

where S_n is the expected sensitivity of the case definition, c is the full width of the confidence interval for the validation metrics, and p is the expected prevalence of the disease within the sample.

A preliminary search for problematic menopause in the dataset using a combination of an ICD-9 code for menopause (627*) and menopausal hormone therapy (i.e., estrogen and progesterone) yielded an estimated prevalence of 19%. To provide validation metrics of at least 80% sensitivity and the widest range of 95% confidence interval of 10%, a sample of 1,388 patients was required for the validation set. We decided to get the same number of patients for the training dataset. Therefore, a sample of 2,776 patients for a training set and a validation set was randomly selected for this analysis. The distributions of sex, regions, and EMRs have been checked to ascertain the representativeness of our cohort.

Variables

Development and initial validation of a case definition was modeled on the methods used to define other chronic conditions within the dataset [12], [15], [16]. A working case definition was used as guidance to support chart reviewers while labeling charts. It incorporated four features: International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) diagnostic codes, textual descriptors of menopausal disorders, relevant drugs, and relevant referrals.

ICD-9-CM diagnostic and procedure code 627 and relevant sub-classifications for menopausal and postmenopausal disorders.

Textual descriptors of menopausal symptoms were identified in collaboration with family physicians, a gynecologist and other clinical expert (NY). The text descriptors included: menopaus*, climacteric, plus abbreviations, alternative spellings and synonyms and other relevant variations.

Relevant drugs used for problem menopause were identified using e-CPS, the online Compendium of Pharmaceuticals and Specialities [17]; e-CPS was used to confirm the specific indications and availability of drugs in Canada for the management of menopause. The drug descriptions included current and recently discontinued generic and proprietary names and were coded to Anatomical Therapeutic Classification (ATC) codes.

Referral for problem menopause to specialist menopause clinic or gynecology.

Development of a problem menopause case definition

A woman was classified as having a problematic menopause when she satisfied criteria of the case definition. The process of developing a case definition in CPCSSN includes:

Step 1: Review of CPCSSN records to create a reference set

CPCSSN-trained clinical chart reviewers reviewed the full CPCSSN records of the sample to create the reference set by differentiating true cases from true non-cases. If a chart reviewer found an instance where 'caseness' was uncertain, the opinion of an independent expert (NY) was sought for adjudication. Fleiss's Kappa was used to measure agreement among the five EMR record reviewers (inter-rater reliability) [18].

Step 2: Machine learning-derived case criteria for problematic and/or treated menopause

A Classification and Regression Tree (CaRT) machine learning approach was applied to determine what data elements among the women in the sample were most relevant to a case definition for problematic or treated menopause. The calculated validation metrics indicate their epidemiological, and potentially clinical, utility. All steps of this analysis were coded in Python 3.7

Feature engineering

We extracted all possible features from multiple fields of patients' EMR data, including billing, encounter diagnosis, health condition (patient problem list), referral, medical procedure, and prescribed medications. Both diagnosis and medication codes, as well as data recorded as free text were used for the feature generation process. Free text was extracted using a "bag-of-words" approach, one feature being generated for each possible word, while special characters and numbers were removed; though bigrams with negated terms (e.g., 'not', '?') were used to avoid falsely diagnosis. To reduce the number of features generated, only features existing in at least 10 charts were selected. We also identified the frequency of features as one instance in the whole chart, two instances in 12 consecutive months, in 24 consecutive months, and in the entire patients' chart. All features were treated as binary variables (i.e., patient has or does not have the feature).

Initially, over 86,000 unique features were created. Their relative usefulness was calculated as the ratio of the difference between the number of patients who have the condition and have the feature (True Positive) and the number of patients who do not have the condition and have the feature (False Positive), over the total number of patients who have the condition. The 300 features with the highest ratios were selected from the training sample that was used by in the supervised machine learning algorithm to develop a case definition.

Imbalanced class distribution

Given the approximate prevalence for patients with problem menopause in our dataset of 19%, we expected an imbalanced class distribution with the ratio of cases over non-cases of about 1:4.

Imbalanced class ratios can lead to biased machine learning models [19], since trained models have more information about the majority class (here, non-cases) than the minority class (cases), which can lead to an increase in the misclassification rate. Therefore, techniques to create an artificial balance between classes, include Synthetic Minority Over-sampling Technique for Nominal (SMOTE-N) data [20], random over-sampling, and random under-sampling [21] were applied.

Supervised machine learning

Training of the machine learning models was completed using nested k-fold cross validation (with k = 10). Hyperparameter tuning was completed for each step of the k-fold cross validation and will be discussed in more detail in the next section. To reduce the effect of imbalanced class distribution, the data for training and testing at this step were split using stratified sampling, to ensure a consistent distribution of cases and non-cases in each fold.

Model development had three steps, (1) feature selection using either recursive feature elimination (RFE) [22] or selecting the k best features [23], (2) re-sampling to balance classes ratio using SMOTE-N, random over-sampling, random under-sampling, and combinations of them, and (3) classification using CaRT. Figure 1 presents the pipeline used. We also swapped step 1 and 2 (re-sampling first, then selecting features) to check for the effectiveness of each method by comparing validation metrics.

Fine-tuning

The optimal hyperparameters (criterion, splitter, max-depth, min_sample_split, and class_weight) of the built-in Python implementation of the CaRT algorithm was selected using k-fold cross-validation coupled with a grid search. The best model and algorithms were chosen by optimizing the F1-score ($\frac{sensitivity PPV}{sensitivity+PPV}$) [24] to balance recall and precision of the chosen model.

Translating to human readable algorithms and validation

We found that algorithms developed by the CaRT for problem menopause using this dataset, though valid, are not easily readable by people for having a lot of branches. To adjust for this, after obtaining the best performance CaRT algorithm, we selected the most important features to create a set of probable case definitions by ranking features by their importance scores. We then re-applied each of these definitions to the training dataset to identify the best and final algorithm; a validation dataset (n = 1,388) was kept separate from the training dataset (n = 1,388) and was used to check for the algorithm's generalisability.

Results

From 190,687 women aged 45–60 with at least one record in CPCSSN as of 2020, we selected a random subset of 2,776 EMR records for chart review. A set of 2,776 EMR charts have been reviewed individually by five reviewers; charts that reviewers could not classify were labelled uncertain and were re-reviewed independently by two reviewers (SR and AP); the two of them did not agree on ten charts (discrepancies);

those were then sent to a pharmacist and certified menopause practitioner (NY) for a final review. One record did not include sufficient information to be classified and was excluded from the analysis. Figure 2 presents the steps and number of charts used during this process.

Sample characteristics

Table 1 presents characteristics of the 2,775 patients whose CPCSSN records underwent expert review.

Table 1
Sample characteristics

Baseline characteristics	**Chart review cohort**	**CPCSSN 2020 dataset**
N	2,775	190,392
	n (%)	n (%)
Age		
45–49	802 (28.9)	58,636 (30.8)
50–54	806 (29.0)	57,630 (30.3)
55–60	1167 (42.1)	74,126 (38.9)
Urban residence	2,341 (84.4)	150,915 (79.3)
Number of chronic conditions (of 11 CPCSSN conditions*)		
0	965 (34.7)	72,509 (38.0)
1	765 (27.6)	43,743 (23.0)
2	421 (15.2)	28,097 (14.8)
3+	624 (22.5)	46,043 (24.2)

** As of 2019. CPCSSN has developed and validated case definitions for chronic kidney disease, COPD, dementia, depression, diabetes mellitus, dyslipidemia, epilepsy, Herpes Zoster, hypertension, osteoarthritis, Parkinson's disease* [25].

We found that resampling methods (SMOTE-N, random under and over sampling) negatively affected the performance of the applied machine learning techniques; the best performing methods were produced by pipelines without resampling methods. Appendix 1 presents metrics obtained by different machine learning techniques.

We tuned hyperparameters of the CaRT algorithm for features generated by either RFE or select k-best features and produce the final case definition. After fine-tuning, a pipeline of the best five features produced the highest validation metrics for all four components. The complete decision tree had lower

sensitivity than the metrics obtained by the simplified case definition, while the other validation metrics are roughly similar between the two methods. Metrics for both definitions are included in Table 2. Table 3 presents the case definition selected to classify problem menopause cases in our sample of records.

Table 2
Validation metrics for the final case definition

	Sensitivity % (95%CI)	Specificity % (95%CI)	PPV % (95%CI)	NPV % (95%CI)
CaRT	79.2 (73.8–83.8)	92.7 (91.0-94.1)	70.9 (65.3–75.9)	95.2 (93.8–96.3)
Simplified case definition	81.5 (76.3–85.9)	93.5 (91.9–94.8)	73.8 (68.3–78.6)	95.7 (94.4–96.8)

Table 3
Case definition of problem menopause in CPCSSN

2 occurrences of ICD-9-CM code 627 in diagnosis fields within 24 months
OR
1 occurrence of ATC code G03CA in medication fields

Discussion

Few studies have focused on the distribution of menopausal symptoms in the community or management of menopausal symptoms in primary care settings. Among published studies, most tend to concentrate on particular aspects of menopause management, such as vasomotor symptoms and menopausal hormone therapy (e.g., Kiran, 2022 [17]) rather than describing primary care management more generally.

We believe this is the first such study in Canada, and certainly the first to apply population health analytic methods to explore a case definition for problematic or treated menopause. We found that re-balancing the dataset did not improve the validity of applied algorithms. It might reflect that the imbalance was not too severe to effect learning ability. Our findings are also aligned with some evidence that suggest decision trees do not suffer from the same degree of bias from imbalanced data sets as other methods [26].

However, reducing the number of features provided to the CaRT algorithm produced good validation metrics, with all aspects of sensitivity, specificity, PPV, and NPV being substantially greater than 70%. However, the tree itself was complicated with a maximum number of connections from the root to the furthest leaves (maximum depth) equals to five, which made the case definition rules challenging to understand. Therefore, we extracted results from the CaRT to reduce the number of criteria to simplify the case definition. This simplified definition also has good validation metrics and is clinically and

epidemiologically relevant, with slightly higher sensitivity than the CaRT-based case definition (89.9% vs 81.5%); although PPV was reduced from 80.3–73.9%.

The value of this research lies in developing and validating case definition algorithms that provide the foundation for quality improvement initiatives in the primary care treatment of menopause in Canadian women and better knowledge about the burden of menopausal symptoms in the community. Using the CPCSSN case definition, our next study will explore whether the treatment of Canadian women is optimal as recommended by the Society of Obstetricians and Gynaecologists of Canada and/or North American Menopause Society guidelines for the management of menopausal symptoms [27]. If treatments meet this ideal for most women, then we shall celebrate that outcome. If not, then we shall develop educational materials for distribution by the CPCSSN organisation to the individual primary care practice-based networks and enter into discussion with medical schools, professional organizations, and the College of Family Physicians of Canada to enhance physician education about menopause management.

This study was limited by the quality of the data collection as it depended on the data as recorded in family physicians' offices, and there is no opportunity to check the veracity of the anonymized data provided for our analysis. This limitation applies to the data used for the case definition algorithms as well as data abstracted from the EMR charts. The database does not include women who have not accessed family physician care: this ensures that women who are identified as “cases” will be those with menopausal problems sufficiently troublesome to seek family medicine health care, a more rigorous test of prevalence than self-report in a population survey. However, it may be that these patients do not represent all Canadian women; women who self-manage their symptoms will not be identified as cases of problem menopause.

The development of this case definition offers future opportunities to better understand the primary care management of menopause. For example, it can be used to estimate and predict the workload associated with providing care for patients who have problem menopause by family physicians and other primary care team members. As a result, primary care teams may be able to streamline their work, perhaps providing additional resources for menopausal women, such as introducing interdisciplinary menopause clinics that might off-load physician workload [3].

Population access to menopause treatment has not been previously studied in Canada. Thus, it is difficult to assess the demand for services to primary and higher-level care for patients who have problem menopause and to estimate the need for health care workforce assets, such as menopause specialists, and other additional services. Further, in planning services, it is also important to try to predict the demand for the increasingly aging population and to take account of the impact of changes in symptoms over a woman’s lifetime. The production of this problem menopause case definition could be used in conjunction with provincial and national population estimates to plan workload over time as the population ages.

Conclusion

This study developed and validated a case definition for problem menopause using EMR data recorded in primary care settings. The findings from this study will lead to further work about the epidemiology of problem menopause and its symptom management in primary care settings to provide better information for primary care providers and improve the care of their menopause patients.

Abbreviations

ATC	Anatomical Therapeutic Classification
CaRT	Classification and Regression Trees
COPD	Chronic obstructive pulmonary disease
CPCSSN	Canadian Primary Care Sentinel Surveillance Network
EMR	Electronic medical record
ICD-9-CM	International Classification of Diseases– 9th version (Clinical Modification)
MHT	Menopausal hormone therapy
NPV	Negative predicted value
PPV	Positive predicted value
RFE	Recursive feature elimination
SMOTE-N	Synthetic Minority Over-sampling Technique for Nominal data

Declarations

Ethics approval and consent to participate

This study has received approval from the Health Research Ethics Board at the University of Alberta (Pro00105443) and the University of Calgary (RES0051425) and adheres to all relevant guidelines and regulations for research involving de-identified health data (e.g. CHREB, Tri-Council Policy Statement on the Ethical Conduct for Research Involving Humans [TCPS2]).

The CPCSSN database has received ethics approval, including waivers of individual patient informed consent for their de-identified data to be used for surveillance and research, from each contributing network's local Research Ethics Board (Queen's University, Memorial University of Newfoundland, University of Ottawa, University of Calgary, Dalhouse University, University of Toronto, Western University, Bruyère Research Institute, University of British Columbia, University of Alberta, and University of Manitoba).

Consent for publication

Not applicable

Availability of data and materials

The data used in this analysis are available from the Canadian Primary Care Sentinel Surveillance Study (CPCSSN) upon reasonable request to SR, with permission from CPCSSN.

Competing interests

SR and NY received a research grant to fund this study (see below). NY has been on the advisory board and/or speaker for Biosyent, Bayer, Amgen, Organon and Duchesnay. All the remaining authors do not have any competing interest.

Funding

This study was funded by a Pfizer Quality Improvement Grant (Pfizer Grant ID 60293977). Pfizer had no influence on the study design, data analysis, findings interpretation, or any content of this manuscript.

Authors' contributions

SR, NY, BS: conceptualization, methodology, manuscript reviewing and editing; AP: methodology, data analysis, original draft preparation; MC, TW: methodology, supervision; SG: conceptualization, methodology; RP: chart reviewing. All authors have read, revised, and approved of the final manuscript.

Acknowledgements

Thanks Matt Taylor (data manager at the CPCSSN) for creating and managing a Data Presentation Tool, which was used to review EMRs.

Thanks Rebecca Miyagishima and Amanda Larocque for reviewing EMR charts.

References

1. N. Yuksel *et al.*, “Guideline No. 422a: Menopause: Vasomotor Symptoms, Prescription Therapeutic Agents, Complementary and Alternative Medicine, Nutrition, and Lifestyle,” *J. Obstet. Gynaecol. Can. JOGC J. Obstet. Gynecol. Can. JOGC*, vol. 43, no. 10, pp. 1188-1204.e1, Oct. 2021, doi: 10.1016/j.jogc.2021.08.003.
2. T. Yasui *et al.*, “Factors associated with premature ovarian failure, early menopause and earlier onset of menopause in Japanese women,” *Maturitas*, vol. 72, no. 3, pp. 249–255, Jul. 2012, doi: 10.1016/j.maturitas.2012.04.002.
3. B. Sydora *et al.*, “Patient characteristics, menopause symptoms, and care provided at an interdisciplinary menopause clinic: retrospective chart review,” *Menopause*, Jan. 2018.

4. P. M. Maki *et al.*, “Guidelines for the evaluation and treatment of perimenopausal depression: summary and recommendations,” vol. 25, no. 10, pp. 1069–85, 2018.
5. P. Mendez-Bustos, J. Lopez-Castroman, E. Baca-García, and A. Ceverino, “Life cycle and suicidal behavior among women,” *Sci. World J.*, vol. 2013, no. 485851, 2013.
6. A. Morabia and P. Flandre, “Misclassification Bias Related to Definition of Menopausal Status in Case-Control Studies of Breast Cancer,” *Int J Epidemiol.*, vol. 21, no. 2, pp. 222–228, 1992.
7. W. Utian, “Ovarian function, therapy-oriented definition of menopause and climacteric,” *Exp. Gerontol.*, vol. 29, no. 3/4, pp. 245–251, 1994.
8. A. Phipps *et al.*, “Defining Menopausal status in epidemiologic studies: a comparison of multiple approaches and their effects on breast cancer rates,” *Maturitas*, vol. 67, no. 1, pp. 60–66, 2010.
9. M. Greiver, T. Williamson, T.-L. Bennett, N. Drummond, C. Savage, and B. Aliarzadeh, “Developing a method to estimate practice denominators for a national Canadian electronic medical record database,” *Fam Pr.*, vol. 30, no. 3, pp. 347–54, 2013.
10. T. Williamson, N. Natarajan, D. Barber, D. Jackson, and M. Greiver, “Caring for the whole practice: the future of primary care,” *Can Fam Physician*, Jul. 2013.
11. S. Garies, R. Birtwhistle, N. Drummond, J. Queenan, and T. Williamson, “Data Resource Profile: National electronic medical record data from the Canadian Primary Care Sentinel Surveillance Network (CPCSSN,” *Int J Epidemiol*, vol. 46, no. 4, pp. 1091–2, 2017.
12. A. Kadhim-Saleh, M. Green, T. Williamson, D. Hunter, and R. Birtwhistle, “Validation of the diagnostic algorithms for 5 chronic conditions in the Canadian Primary Care Sentinel Surveillance Network (CPCSSN): a Kingston Practice-based Research Network (PBRN) report,” *J. Am. Board Fam. Med. JABFM*, vol. 26, no. 2, pp. 159–67, 2013.
13. T. Williamson, M. E. Green, R. Birtwhistle, S. Khan, S. Garies, and S. T. Wong, “Validating the 8 CPCSSN case definitions for chronic disease surveillance in a primary care database of electronic health records,” *Ann Fam Med*, vol. 12, no. 4, pp. 367–72, 2014.
14. S. E. Vollset, “Confidence intervals for a binomial proportion,” *Stat. Med.*, vol. 12, no. 9, pp. 809–824, May 1993, doi: 10.1002/sim.4780120902.
15. M. Shaw and N. Drummond, “Developing a case definition for type 1 diabetes mellitus in a primary care electronic medical record database: an exploratory study,” *CMAJOpen*, vol. 7, pp. 246–251, 2019.
16. S. Ross *et al.*, “Developing and testing the validity of case definitions for pelvic floor disorders in women who consult primary care clinics: retrospective study using primary care electronic medical records,” *CMAJOpen*, May 2020.
17. A. Kiran *et al.*, “Epidemiology and treatment patterns of UK women diagnosed with vasomotor symptoms: Findings from the Clinical Practice Research Datalink GOLD database,” *Maturitas*, vol. 164, pp. 1–8, Jun. 2022, doi: 10.1016/j.maturitas.2022.05.013.
18. J. L. Fleiss, “Measuring nominal scale agreement among many raters.,” *Psychol. Bull.*, vol. 76, no. 5, pp. 378–382, 1971.

19. A. Luque, A. Carrasco, A. Martín, and A. de las Heras, “The impact of class imbalance in classification performance metrics based on the binary confusion matrix,” *Pattern Recognit.*, vol. 91, pp. 216–231, Jul. 2019, doi: 10.1016/j.patcog.2019.02.023.
20. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique,” *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.
21. K. Fujiwara *et al.*, “Over- and Under-sampling Approach for Extremely Imbalanced and Small Minority Data Problem in Health Record Analysis,” *Front. Public Health*, vol. 8, p. 178, May 2020, doi: 10.3389/fpubh.2020.00178.
22. I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, “Gene selection for cancer classification using support vector machines,” *Mach. Learn.*, vol. 46, no. 1, pp. 389–422, 2002.
23. F. Pedregosa *et al.*, “Scikit-learn: Machine Learning in Python,” *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.
24. S. A. Hicks *et al.*, “On evaluation metrics for medical applications of artificial intelligence.” medRxiv, p. 2021.04.07.21254975, Apr. 09, 2021. doi: 10.1101/2021.04.07.21254975.
25. CPCSSN, “CPCSSN Case Definition Version 2.” 2019. Accessed: Jul. 27, 2021. [Online]. Available: http://cpcssn.ca/wp-content/uploads/2019/05/CPCSSN-Case-Definitions-v2.pdf
26. C.-O. Truică and C. Leordeanu, “Classication of an Imbalanced Data Set using Decision Tree Algorithms,” *Univ. Politeh. Buchar. Sci. Bull. Ser. C - Electr. Eng. Comput. Sci.*, vol. 79, p. 69, Dec. 2017.
27. “The 2022 Hormone Therapy Position Statement of The North American Menopause Society” Advisory Panel, “The 2022 hormone therapy position statement of The North American Menopause Society,” *Menopause N. Y. N*, vol. 29, no. 7, pp. 767–794, Jul. 2022, doi: 10.1097/GME.0000000000002028.

Figures

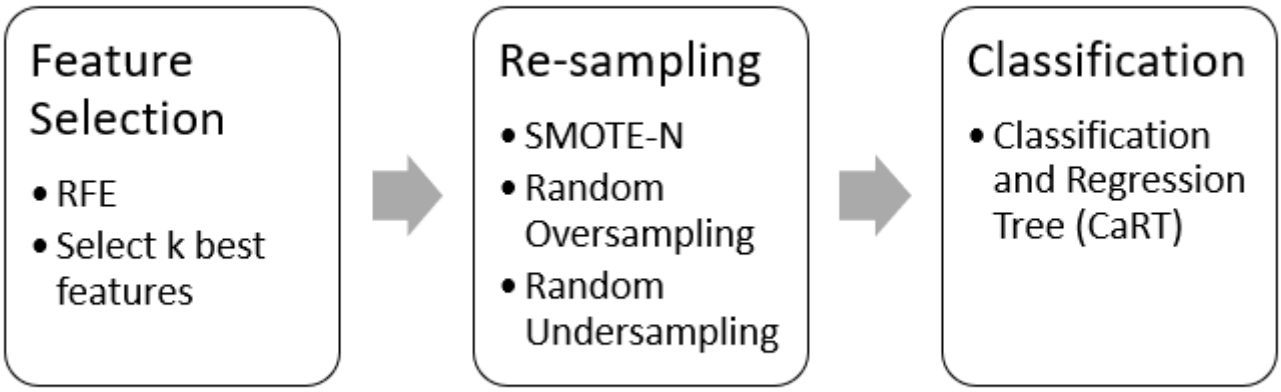

Figure 1

Machine learning pipeline used for case definition creation.

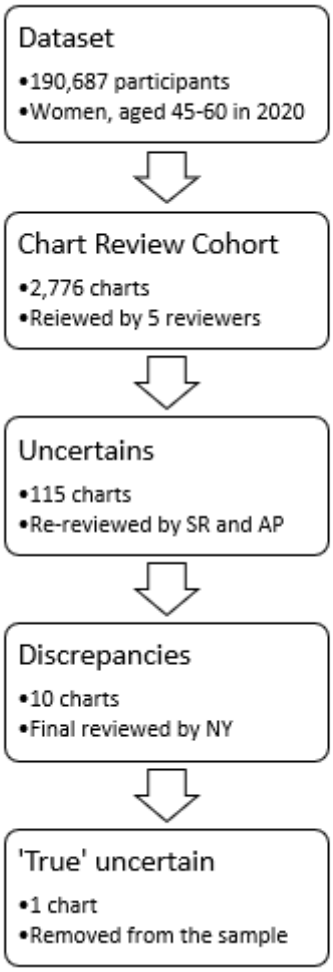

Figure 2

Flowchart for chart review's process

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- PhametalSupplementary.docx