Women's Health Patterns in Perimenopause

Roux Institute Data Analytics in collaboration with Hey Freya

Capstone Project, Summer 2022

Leah Arsenault, Eric Deerwester, Cole Guerin, Jeff Lavoie, Clare Ruhlman, Joe Reynolds

Project and Sponsor Information

Project Title: Women's Health Patterns in Perimenopause

Description: Women's bodies, and thus the conditions and symptoms that affect women (such as autoimmune disorders, insomnia, and mental health ailments that are overwhelmingly present in women) receive significantly less funding, attention, and research than men's bodies do. Some readers may even be shocked to learn that women were not allowed into clinical trials until 1993. The learning from this research is only beginning to be explored and understood.

Real-world consequences arise from the absence of data about women's health: women are misdiagnosed 30-40% more often than men with an average of 5 years before getting a diagnosis at all. Medications, supplements, and their respective dosing are administered without the ability to take women's unique physiology into account—despite evidence showing that women metabolize medication differently than men, have unique nutritional requirements, and might even benefit from dosing that modulates alongside menstrual cycles.

Hey Freya (HF) is an online community application for women to learn and share about women's health. To build and inform their membership, Hey Freya is seeking to make use of the first longitudinal study of women's health – the SWAN (Study of Women's Health across the Nation) study. In order to do so, they have engaged a group of Data Analytics graduate students at Northeastern's Roux Institute (RI) to investigate what the SWAN data might reveal about a series of core-questions and pain-points that Hey Freya is seeking to understand and explore for its members.

Deliverable: Hey Freya is interested in patterns and/or correlations between sleep quality & quantity, and/or self-reported lack of energy with nutritional supplementation patterns, sleep medication patterns, dietary estimates of nutrient consumption, and family lifestyle and support patterns in perimenopausal women.

Specifically, Hey Freya would like to see these in the context of age, race, socio-economic status, and disease status (any conditions diagnosed at the time of the interviews) and related to the day of their menstrual cycle (if they were still menstruating, Day 1 being the first day of bleeding/period).

Introduction

- The following report documents the SWAN data analysis work done in collaboration between Hey Freya and Northeastern University graduate students. The analysis centers on HF's mission to provide innovative and holistic care to perimenopausal women.
- The organization's work is motivated by the need for better research and evidence-based practices in women's healthcare.
- The following research questions guided the investigation:

1. Longitudinal patterns of lifestyle stress indicators with forms of medications, nutritional supplementation, or alternative medicines
2. Baseline, midline, end line patterns of nutritional markers among women in the study and if deficiencies were successfully addressed by supplementation
3. Correlations of lifestyle factors with dietary estimates of nutrient consumption
4. Assessing the indicators of lifestyle stress, hormone levels (done in the latter years) and sleep quality/quantity
5. The influence of indicators of community (sense of support and time with friends and partners) on energy, anxiety, stress indicators, sleep
6. The impact of motherhood shifts or career stage changes on stress, sleep and nutritional markers

- At the mid-point of the project, the progress into the questions was discussed, insights shared, and questions modulated. Importantly, terms like "nutritional supplementation," (question 2) "lifestyle stress," (question 4) and "indicators of community" were honed in relationship to available data in the SWAN research.
- The Roux Institute team addressed these questions using a range of statistical methods.
- In addition to this report, the RI team also developed an interactive dashboard for the HF team to disaggregate and view the data.

Project Data

Dataset

- "The Study of Women's Health Across the Nation (SWAN) is a first-in-kind, multi-site, longitudinal, epidemiologic study designed to examine the health of women during their perimenopausal and menopausal years. The study examines the physical, biological, psychological and social changes during this transitional period. The goal of SWAN's research is to help scientists, health care providers and women learn how mid-life experiences affect health and quality of life during aging."[1]
- This research is co-sponsored by the National Institute on Aging (NIA), the National Institute of Nursing Research (NINR), the National Institutes of Health (NIH), Office of Research on Women's Health, and the National Center for Complementary and Alternative Medicine.
- Although the study began in 1994, participants were recruited between 1996 and 1997 through seven designated research hospital systems in six? major U.S. cities.
- The dataset contains longitudinal data on 3,302 participants between the years 1997 and 2008 (over the course of 10 visits).
- The data collected pertains to various physiological, social and psychological variables.

Data Collection - Risks

- Data is collected in annual intervals, which is not ideal for many features, especially concerning numeric medical data, which would benefit from multiple touchpoints throughout the year.

[1] (N.A.) 2022. Study of Women's Health Across the Nation. Retrieved from https://www.swanstudy.org/ on August 1st, 2022.

- In many cases, study questions ask participants to report on issues based on memory over a time-period. For example, “in the last month, have you experienced...”. These types of questions rely on participants to provide accurate reporting, and memory is notoriously unreliable.
- Many variables provide ordinal response options, which provide some level of understanding, but are not well-suited for eliciting subtleties.
- Some variance exists for attributes collected at each visit. For example, the “ALLARE” variable was only collected during the baseline, fifth and ninth patient visits.

Analysis

Data Exploration & Pre-processing

- The data were provided in eleven; separate datasets. Compilation of a comprehensive dataset was made challenging by the filename encoding.
    - **All** the variables had a suffix with the number visit they were recorded. Combining into one long data set required removing these suffixes and combining variables.
    - Some variables changed names over time or categories were completely changed for how values were recorded.
    - Common NLP practices such as removing punctuation and white space were used to standardize responses, as well as removing numerical indicators for categorical variables that also had text versions of those categories.
- A large portion of variables were missing over 50% of observations. Thus, for the purposes of analysis, features with significant missing values (50% or more) were eliminated.
- Not very much data imputation was done for missing values because imputing too many missing values would completely change the distributions of the variables.
- There were numerous variations in responses (e.g., “1-5 days” vs. “1: 1-5 days”) for close-ended survey items.
- There was evidence of both structurally missing data and data missing at random.
- An indexed response variable measuring sleep quality was created by combining variables: night sweats (NITESWE), trouble sleeping or latency (TRBSLEP), frequent waking throughout the night (WAKEUP), and early waking (WAKEEARL).

Data Analysis: Modeling and Dashboard

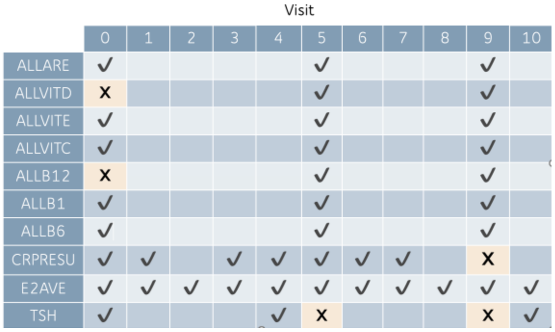
Baseline, midline, end line patterns of nutritional markers among women in the study and if deficiencies were successfully addressed by supplementation

Nutrition is a fundamental underlying factor for health. As such, HF was curious to investigate the nutritional data within SWAN. While there *is* data on Vitamins A, E, D, C, B12 and B6, they are collected only at the baseline, midline, and near the end of the study. HF was interested in modeling these against lab performance for C-reactive protein, Estradiol, and Thyroid Stimulating Hormone.

Unfortunately, the data contains some overt gaps—some vitamins do not have a baseline reading, and some of the labs do not have readings that map to visits where supplements are measured. Using one to predict the other would thus be fraught with issues when trying to draw conclusions about nutritional interventions.

Figure 1.

Table of nutritional and lab variables and the visits when recorded



	Visit										
	0	1	2	3	4	5	6	7	8	9	10
ALLARE	✓					✓				✓	
ALLVITD	X					✓				✓	
ALLVITE	✓					✓				✓	
ALLVITC	✓					✓				✓	
ALLB12	X					✓				✓	
ALLB1	✓					✓				✓	
ALLB6	✓					✓				✓	
CRPRESU	✓	✓		✓	✓	✓	✓	✓		X	
E2AVE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
TSH	✓				✓	X				X	✓

Some additional things to note about the nutritional data worth noting:

- Units of measurement vary between supplements.
- Unit of measurement for Vitamin A is unclear.
- The percentage of patients reporting supplement usage is oddly the *exact* same number across many supplements. We would usually expect a slight variance. Another way to say this is that every individual taking one supplement is taking them all, and these numbers have zero variance across the years they are recorded. It is so oddly the same that there could be an underlying data integrity issue that needs to be investigated further.

Understanding more about what constitutes a deficiency is key to addressing HF's question. If other studies have shown the threshold of vitamin deficiency, it stands to reason that the SWAN data could be sufficient as a comparison. Simple arithmetic could demonstrate intervention on the patient-level.

A complete exploration of each supplement is available in the appendix of the deliverables so that Hey Freya has an idea of what information exists in the data.

Longitudinal patterns of lifestyle stress indicators with forms of medications, nutritional supplementation, or alternative medicines

To address Hey Freya's question regarding longitudinal patterns between nutritional data and lifestyle stress indicators, the research team developed a multi-level, mixed effects, binomial regression model to explore the relationship between nutrition and sleep. Although this approach differs from traditional longitudinal studies, the lack of observations made it difficult to capture the effect of changing nutrition levels over time. Additionally, it was unclear from the data if these values were expected to increase or decrease as a result of some intervention or therapy.

A mixed effects model was chosen as the format of the data satisfied the three criteria for multi-level, mixed effects logistic regression. First, the sleep target variable was easily converted to a binary

Commented [CR1]: Are we using first-person?

Commented [JR2R1]: third

outcome i.e., the participant reported having at least one sleep issue three times a week or more. Second, it was deemed important to include all observations within each variable to estimate the possible range of values for the population. Lastly, there was no multicollinearity present in the independent variables (see **Appendix**). Covariance was uniformly low relative to the units of measurement between independent variables. Finally, due to the lack of available data and subject matter expertise, sleep quality was adopted as an indicator of lifestyle stress. However, identical methods can be applied to similar attributes for future studies.

Specifically, the model looked at both fixed effects (e.g., levels of vitamins and other nutrients and reproductive stage) and random effects (e.g., study participant) on sleep quality. Given the large number of items relating to nutrient consumption, only variables estimating the total amount of nutrient (through diet and supplement) were selected for regression (ALLARE, ALLVITC, ALLVITD, ALLVITE, ALLB1, ALLB6, ALLB12, ALLCALC, ALLFOL, ALLIRON, ALLZINC, E2AVE, SWANID, VISIT, STATUS).[2]

For the purposes of modeling, a new dichotomized target variable was created using similar variables to the sleep index (TRBLSLE, WAKEUP, and WAKEEARL). To maintain as much accuracy as possible, the new variable was binarized to represent general sleep disturbance (where 1 indicates the participant reported experiencing at least one of the above sleep disturbances three or more times per week). Finally, the data (except for visit number, participant ID, menopausal status, and the target variable) were scaled to account for differences in measurement units.

Commented [CR3]: I will update if we incorporate another var. to replace NITESWE.

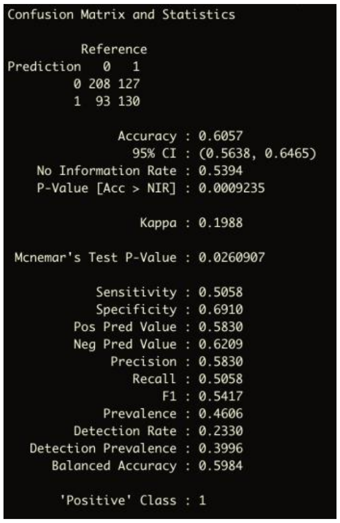
As the data was hierarchical in nature (i.e., the same participants across multiple visits), a multi-level model was constructed to cluster individual participants and visits (see **Figure 2**). Doing so instructed the model to account for any interactions between observations that would potentially influence the model. Prior to fitting, the data were split into training and test sets using an 80:20 ratio. Lastly, the BOBYQA optimizer was chosen using R's lme4 fit.All() function.

Figure 2.

Summary of the fitted classifier's performance on the test set

Commented [CR4]: To insert

[2] Because measurements were collected inconsistently across the dataset, the model only represents data from visits 5 & 9. Additionally, CPRESU and TSH were omitted from analysis as these measurements did not align with other supplement data collections.



```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 208 127
         1  93 130

               Accuracy : 0.6057
                 95% CI : (0.5638, 0.6465)
    No Information Rate : 0.5394
    P-Value [Acc > NIR] : 0.0009235

                  Kappa : 0.1988

 Mcnemar's Test P-Value : 0.0260907

            Sensitivity : 0.5058
            Specificity : 0.6910
         Pos Pred Value : 0.5830
         Neg Pred Value : 0.6209
              Precision : 0.5830
                 Recall : 0.5058
                     F1 : 0.5417
             Prevalence : 0.4606
         Detection Rate : 0.2330
   Detection Prevalence : 0.3996
      Balanced Accuracy : 0.5984

       'Positive' Class : 1
```

The resulting model achieved 60.57% accuracy (**Figure 1**). Notably, the model had considerably more trouble identifying participants with reported symptoms as opposed to those who did not. Based on the model's classification performance, the classifier had a sensitivity rate of 50.58% only correctly identifying 130/257 instances of the positive class (i.e., women with reported sleep issues) and a specificity rate of 69.10% (i.e., the model correctly identified 208/30 women without reported sleep issues). Moreover, analysis of the model's coefficients (**Figure 2**) revealed that the variables ALLVITC, ALLIRON, and early perimenopausal STATUS (the dummy-encoded variable for participants in early perimenopause) had statistically significant effects on sleep quality ($p < 0.1$, $p < 0.05$, and $p < 0.001$, respectively). More precisely ALLVITC and early perimenopausal STATUS had a negative effect on the likelihood of experiencing sleep disturbances, as indicated by their negative coefficients. These log-odds may be interpreted as follows:

- Each unit increase of vitamin C is associated with **a decrease of 12% in the odds of experiencing sleep disturbance**.
- The early perimenopausal stage is associated with **a 54% reduction in the relative risk of sleep disturbance**.

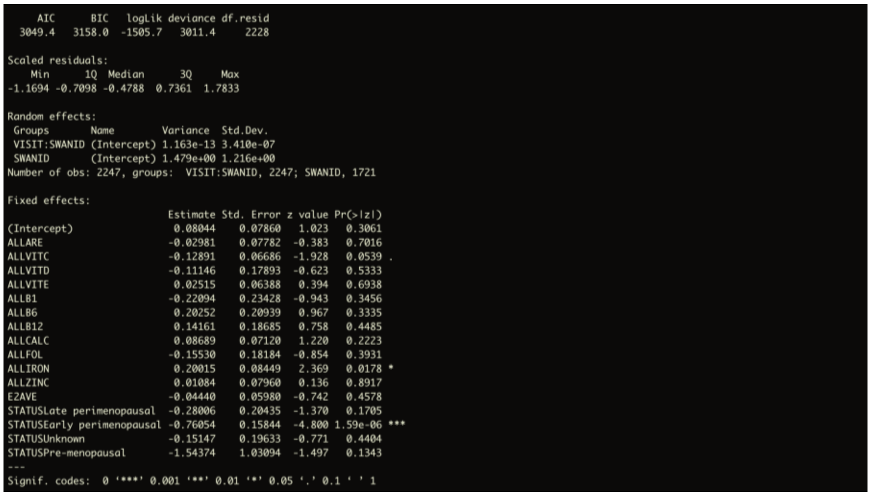
Iron, on the other hand, had a positive impact on sleep quality. That is, a unit increase in iron is associated with **an increase of 22% in the odds of experiencing sleep disturbance**. Finally, the model's output summary also showed that random effects did contribute largely to the variation in the dependent variable.

Thus, in addition to identifying vitamin C as a potential nutritional therapy, these findings suggest that iron levels should be monitored as a possible contributor to sleep quality issues. Moreover, these results support the observation that women are more likely to suffer from disordered sleep patterns during later stages of perimenopause.

Figure 3.

Summary of model’s coefficients

Commented [CR5]: To insert



```
     AIC      BIC   logLik deviance df.resid
  3049.4   3158.0  -1505.7   3011.4     2228

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.1694 -0.7098 -0.4788  0.7361  1.7833

Random effects:
 Groups       Name        Variance  Std.Dev.
 VISIT:SWANID (Intercept) 1.163e-13 3.410e-07
 SWANID       (Intercept) 1.479e+00 1.216e+00
Number of obs: 2247, groups:  VISIT:SWANID, 2247; SWANID, 1721

Fixed effects:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)                 0.08044    0.07860   1.023   0.3061
ALLARE                     -0.02981    0.07782  -0.383   0.7016
ALLVITC                    -0.12891    0.06686  -1.928   0.0539 .
ALLVITD                    -0.11146    0.17893  -0.623   0.5333
ALLVITE                     0.02515    0.06388   0.394   0.6938
ALLB1                      -0.22094    0.23428  -0.943   0.3456
ALLB6                       0.20252    0.20939   0.967   0.3335
ALLB12                      0.14161    0.18685   0.758   0.4485
ALLCALC                     0.08689    0.07120   1.220   0.2223
ALLFOL                     -0.15530    0.18184  -0.854   0.3931
ALLIRON                     0.20015    0.08449   2.369   0.0178 *
ALLZINC                     0.01084    0.07960   0.136   0.8917
E2AVE                      -0.04440    0.05980  -0.742   0.4578
STATUSLate perimenopausal  -0.28006    0.20435  -1.370   0.1705
STATUSEarly perimenopausal -0.76054    0.15844  -4.800 1.59e-06 ***
STATUSUnknown              -0.15147    0.19633  -0.771   0.4404
STATUSPre-menopausal       -1.54374    1.03094  -1.497   0.1343
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

While helpful in responding to Hey Freya’s research questions, the capstone team would like to acknowledge a couple of limitations that may affect the validity of these conclusions. These considerations may also prove helpful in determining directions for future research.

1. Additional data may increase the model’s explanatory power. That is, the model was limited to observations across two visits. Perhaps more data would have revealed additional trends.
2. Additional information around why nutritional levels changed over time (i.e., if intervention occurred and/or if deficiencies were present), may compromise these findings.
3. Multinomial modeling of the various target variable levels (i.e., the number of symptoms participants experience) may allow for more symptom-specific observations.

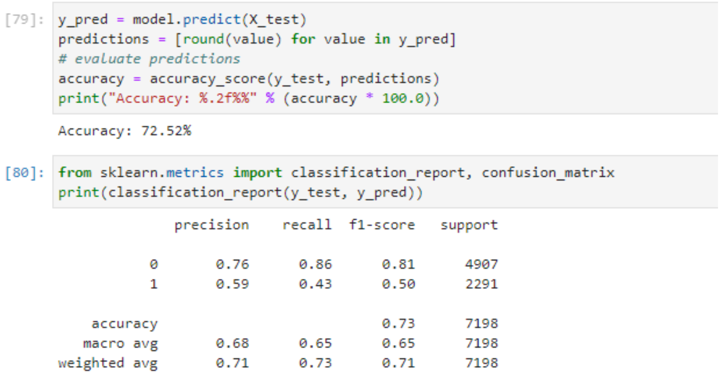
Assessing sleep quality in relationship to indicators of lifestyle stress, community support, energy, and hormone levels.

XG Boost modeling was used to provide insight on the target sleep variable. The target sleep variable consisted of Overall Sleep Quality, Trouble Sleeping, Wake up, and Wakeup Early. These four variables were altered into binary responses so that they were one if the patient responded negatively (for example, they had trouble falling asleep often) and zero if the patient responded positively. A new sleep index was then created to average the patient’s response to those four questions, so the index was a number between zero and one. This index was then created into a sleep target variable by making the response one if the index was greater than or equal to 0.75, and zero if the index was lower than 0.75. This resulted in a binary variable to describe those having three of more reported sleep problems. This model assists in data that contains missing values and is more powerful than a basic decision tree model. The model outcome had a weighted accuracy of 72.52%. The features that heavily contributed to the outcome of the model were features capturing the patient’s mood (frequent mood changes and feeling blue), the patient's stress levels (upsetting responsibilities, legal problems, relationship) and the patient’s physical state (cold sweats, back pain, headaches).

Figure 4.

Summary of model performance: The model has a higher precision predicting patients who don’t have sleep problems. This is likely because the model was trained on a smaller population of people with sleep problems than the number of patients who don’t have sleep problems that the model was trained on.

Commented [CR6]: To insert

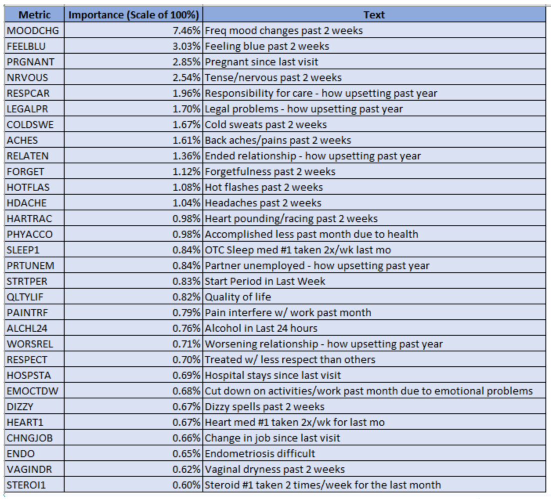


```
[79]: y_pred = model.predict(X_test)
      predictions = [round(value) for value in y_pred]
      # evaluate predictions
      accuracy = accuracy_score(y_test, predictions)
      print("Accuracy: %.2f%%" % (accuracy * 100.0))

      Accuracy: 72.52%

[80]: from sklearn.metrics import classification_report, confusion_matrix
      print(classification_report(y_test, y_pred))

                    precision    recall  f1-score   support

                 0       0.76      0.86      0.81      4907
                 1       0.59      0.43      0.50      2291

          accuracy                           0.73      7198
         macro avg       0.68      0.65      0.65      7198
      weighted avg       0.71      0.73      0.71      7198
```

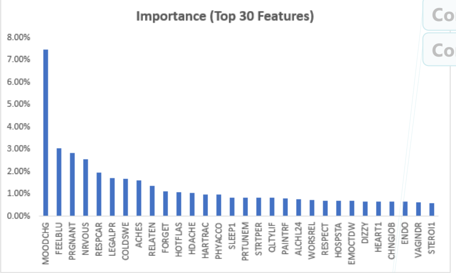
Figure 5.

Summary of model’s weighted variable importance:

Commented [CR7]: To insert



Metric	Importance (Scale of 100%)	Text
MOODCHG	7.46%	Freq mood changes past 2 weeks
FEELBLU	3.03%	Feeling blue past 2 weeks
PRGNANT	2.85%	Pregnant since last visit
NRVOUS	2.54%	Tense/nervous past 2 weeks
RESPCAR	1.96%	Responsibility for care - how upsetting past year
LEGALPR	1.70%	Legal problems - how upsetting past year
COLDSWE	1.67%	Cold sweats past 2 weeks
ACHES	1.61%	Back aches/pains past 2 weeks
RELATEN	1.36%	Ended relationship - how upsetting past year
FORGET	1.12%	Forgetfulness past 2 weeks
HOTFLAS	1.08%	Hot flashes past 2 weeks
HDACHE	1.04%	Headaches past 2 weeks
HARTRAC	0.98%	Heart pounding/racing past 2 weeks
PHYACCO	0.98%	Accomplished less past month due to health
SLEEP1	0.84%	OTC Sleep med #1 taken 2x/wk last mo
PRTUNEM	0.84%	Partner unemployed - how upsetting past year
STRTPER	0.83%	Start Period in Last Week
QLTYLIF	0.82%	Quality of life
PAINTRF	0.79%	Pain interfere w/ work past month
ALCHL24	0.76%	Alcohol in Last 24 hours
WORSREL	0.71%	Worsening relationship - how upsetting past year
RESPECT	0.70%	Treated w/ less respect than others
HOSPSTA	0.69%	Hospital stays since last visit
EMOCTDW	0.68%	Cut down on activities/work past month due to emotional problems
DIZZY	0.67%	Dizzy spells past 2 weeks
HEART1	0.67%	Heart med #1 taken 2x/wk for last mo
CHNGJOB	0.66%	Change in job since last visit
ENDO	0.65%	Endometriosis difficult
VAGINDR	0.62%	Vaginal dryness past 2 weeks
STEROI1	0.60%	Steroid #1 taken 2 times/week for the last month



Commented [JR8]: comment on this

Commented [JR9R8]: and figure numbers

Relating to the sleep quality index, those who showed higher prevalence of sleep issues were also the likely to indicate high levels of lifestyle stressors. The subset of respondents showing the most prominent sleep issues were also likely to report recent feelings of less accomplishment and upset about money or work-related issues.

Additionally, those who showed higher prevalence of sleep issues were also the most likely to report a lack of feeling supported (Ex: never feeling like they had a listener or confidant available to them when needed). Those who were separated from their spouse, divorced, or widowed were also more likely to suffer from sleep issues than those who were married and still together, or those who were single/never married.

The impact of motherhood shifts or career stage changes on stress, sleep and nutritional markers

The data set did not contain any indicators about motherhood shifts or career changes, such as whether the patient had become the mother of a new child, had become newly responsible for another family member or friend, or had started a new job. Analysis was done on a set of variables identified by Hey Freya relating to stress from the patient's career, relationship, role as a mother, and role as a caregiver. These variables did not disclose anything other than what one would intuitively expect. A variable for perceived stress (P_STRESS) was also in the cross-sectional codebook and created for the other visits by combining responses to four different variables (CONTROL, CONFIDE, YOURWAY, PILING).

Women in this study going through perimenopause are slightly more stressed by work than they are by their role as a mother or personal relationships. Not many patients in the data (around 15% of people who responded) indicated that they were responsible for another family member or friend, and very few were upset by this responsibility. This is in keeping with HF's understanding of social behavior reported by perimenopausal women—they value serving their primary relationships, and do not find them bothersome.

There were very marginal differences between the distributions of these variables for perimenopausal women vs. post-menopausal women.

The research team also looked at this data to see how this information on motherhood and careers impacts stress. Creation of a new XGBoost model for the previously created sleep index, but only using the variables of interest and a variable for race as predictors. The model was 78.7% accurate, and identified the responsibility, race, perceived stress, and indicator for whether the patient had children as the most important features in the model. When perceived stress wasn't used in the model, the variable for the stress associated with the role of the mother was found to be very important, which implies that this variable and the perceived stress variable are correlated.

Impacts from nutritional markers and supplements are hard to visualize in the data with respect to motherhood and career stress because most variables of interest were only tracked at baseline and mid-line visits that nutritional markers and supplements were recorded (visits 0 and 5, not visit 9). It's hard to attribute any changes over time in these areas of stress for individuals directly to any supplements they are taking, considering there are only two different time measurements for this in the data. For the future, Hey Freya would want to take more supplemental measurements, be more consistent about measuring variables of interest each visit, and track indicators for if motherhood shifts or career changes occurred between visits.

Additional areas of interest: Exercise

Using the XGBoost model, with mild, moderate, and strenuous exercise as the included features, the model predicted with 71% accuracy whether a participant had "wake up" problems. When the model was used for the combination sleep target of the previous model, the accuracy was far lower, around 50%. But using the wake-up variable specifically provided better results. This data was only from the 4th visit, as this was the only visit where participants were asked about each of the three levels of exercise. Mild exercise was the most important feature in the model.

Client Deliverables

Interactive Excel Dashboard

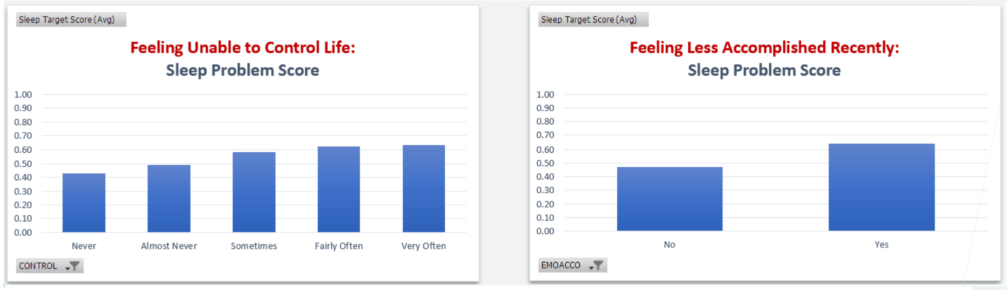
Hey Freya wants to use SWAN research to create meaningful and influential narratives for their user base. With that goal in mind, the research team created dashboards that provide HF with end-user control. Interactive tooling enables data analysis related to variables their users care about, while filters and slicers allow HF to run different scenarios with at-a-glance outputs, making the data come to life.

Subsets based on ethnicity, hormonal status, and/or the presence of children in the household, enable comparison of responses pertaining to specific groups, and the resulting visualizations focus on the frequency and distribution of responses pertaining to lifestyle, family, support network, stress, and sleep quality indicators. Users can drill down into specific categories of interest, allowing for observation of patterns which may be uncovered from the source data. More specifically, one can observe how certain health conditions, such as the prevalence of sleep problems, may be correlated with strong responses to questions revolving around lifestyle, family stress, or lack of a support network.

For example, the below charts compare the average “Sleep Problem Score” (indicating highest prevalence of sleep issues) to survey responses revolving around stress. The results indicate that the more one feels out of control in life, the more likely they will suffer from sleep issues. Similarly, those feeling less accomplished recently were also more likely to suffer from sleep issues.

Figure 6.

Sample of the Sleep & Stress Interactive Excel Dashboard



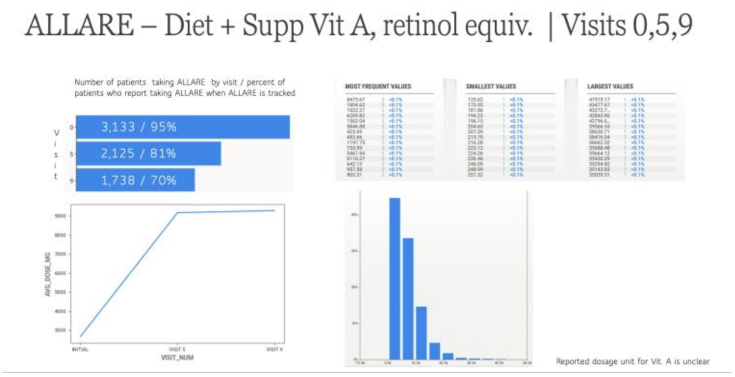
Commented [JR10]: figure #

Nutritional Data Exploration

Nutritional supplement and lab data are visualized in a PowerPoint deck in the deliverables. The below sample shows some of the features, such as the percentage of respondents per visit, total number reporting, a distribution of the dosage, and a time series of the mean dosage. Additionally, range and frequency data are highlighted.

Figure 7.

Sample of the nutritional supplement exploratory data analysis



ALLARE – Diet + Supp Vit A, retinol equiv. | Visits 0,5,9

Number of patients taking ALLARE by visit / percent of patients who report taking ALLARE when ALLARE is tracked

MOST FREQUENT VALUES			SMALLEST VALUES			LARGEST VALUES		
8473.67	2	<0.1%	125.62	1	<0.1%	47919.17	1	<0.1%
1904.62	2	<0.1%	173.05	1	<0.1%	45477.67	1	<0.1%
1022.27	2	<0.1%	181.86	1	<0.1%	43272.7...	1	<0.1%
6209.82	2	<0.1%	194.23	1	<0.1%	42845.65	1	<0.1%
1562.04	2	<0.1%	196.15	1	<0.1%	42796.6...	1	<0.1%
5846.88	2	<0.1%	204.63	1	<0.1%	39564.53	1	<0.1%
423.69	2	<0.1%	207.39	1	<0.1%	38630.71	1	<0.1%
453.66	2	<0.1%	213.79	1	<0.1%	38476.04	1	<0.1%
1197.75	2	<0.1%	216.28	1	<0.1%	36642.32	1	<0.1%
733.59	2	<0.1%	223.13	1	<0.1%	35688.48	1	<0.1%
5467.84	2	<0.1%	224.26	1	<0.1%	35664.12	1	<0.1%
6116.27	2	<0.1%	236.46	1	<0.1%	35430.29	1	<0.1%
642.13	2	<0.1%	246.09	1	<0.1%	35294.82	1	<0.1%
957.58	2	<0.1%	248.99	1	<0.1%	35143.83	1	<0.1%
803.31	2	<0.1%	251.32	1	<0.1%	35039.51	1	<0.1%

Reported dosage unit for Vit. A is unclear

Conclusion

Recommendations

- Additional biological marker data is available from the SWAN study through the NIA Biobank and can be obtained by requesting permission from the research administrators at this link.
- Having access to the NIA Biobank blood and urine samples would enable another layer of depth for findings and could be used to explore additional models.

Limitations

- As discussed in the Exploration and Pre-Processing section above (and at length during meetings with Hey Freya), the SWAN data is very much a real-world data set—it is big, and it can be quite "messy." Variables were often added and recorded differently throughout the study. In many cases, data is largely missing.
- The Northeastern research team saw considerable challenges and limitations as the result of these data governance issues. Nonetheless, insights can be derived.
- Challenges become inspiration for future studies!

Potential Directions for Future Research

- HF plans to conduct its own longitudinal data collection using an online survey.
- Use geospatial data to investigate impact of food deserts on disease in perimenopausal women.
- Design research to map world and cultural events to clinical reporting in order to better understand and control for macro environmental stressors in the data.
- Investigate the language of self-reporting and the variance between microoperations and cultures in order to establish better control.
- Design experiment to measure the impact of career stage on perimenopausal symptoms.
- Improve the research that measures perimenopausal women's coping strategies in order to move away from the dialogue about "resilience" that exists in current literature.

Commented [JR11]: add the individual projects