Finding subtypes of obstructive sleep apnea using cluster analysis

Jason Li
M.S. Thesis

School of Medicine

Oregon Health & Science University

Certificate of Approval

This is to certify that the Master's Thesis of

Jason J. Li

"Finding subtypes of sleep apnea using cluster analysis"

has been approved

_Dr. Eilis Boudreau [redacted]
Thesis Advisor

_Dr. Guanming Wu [redacted]
Committee Member

__Dr. Melissa Haendel _____
Committee Member

_____
Committee Member

School of Medicine

Oregon Health & Science University

Certificate of Approval

This is to certify that the Master's Thesis of

Jason J. Li

"Finding subtypes of sleep apnea using cluster analysis"

has been approved

_Dr. Eilis Boudreau_____
Thesis Advisor

_Dr. Guanming Wu_____
Committee Member

__Dr. Melissa Haendel_[redacted]_____
Committee Member

_____
Committee Member

Chapter 1: Introduction

One hundred million people worldwide have obstructive sleep apnea (OSA), a disorder characterized by collapse of the upper airway and repeated pauses in breathing throughout sleep(1-5). Sleep fragmentation caused by sleep apnea can lead to daytime sleepiness(2, 4, 6, 7), increased risk for motor accidents(2, 6), and reduced quality of life. The physiological changes associated with these pauses are believed to contribute to the development of hypertension(2-4), diabetes mellitus(2), obesity(2), stroke(2, 3), and heart disease(2, 3).

Duration of the apneic events, age(3), degree of oxygen desaturation, and other features likely contribute to the severity of OSA. The severity of OSA is measured as the number of times a sleeper stops breathing per hour, which is called apnea-hypopnea index or AHI(5). However, it has become increasingly clear that this measure is insufficient and that additional measures are needed. Furthermore, it is likely that there are several different subtypes of OSA, each with different pathophysiological causalities, comorbidities, age range, and polysomnographic features(1, 3, 8-10). Heretofore, this heterogeneity has been poorly defined. By characterizing these subtypes and identifying comorbid medical conditions, we may be able to improve and personalize treatment.

Up until recently, tools for identifying OSA subtypes were limited. However, new computational approaches have made it more feasible. Separating and classifying entities such that entities in the same group are more similar than those in a different group, or cluster, has been a long-studied problem in the field of machine learning. Previous attempts to examine subtypes involve splitting the data along a predefined boundary (e.g., younger versus older age) and seeing how the rest of the variables differ(7, 9, 11). Clustering allows discovery of subgroups without any predefined parameters(12), and can be applied to many domains like cancer gene expression profiles(13), HIV/AIDS epidemiology(14), and ribosomal structure(15). A few groups have used cluster analysis to identify OSA subtypes in terms of clinical features, but these studies were limited due to their small sample sizes and homogeneous populations (16-18). These studies also had variable findings.

Another barrier is historical lack of access to large public datasets. Fortunately, there has been a shift in attitude, with more and more scientists recognizing that sharing their data is vital to good science. An example of an initiative for sharing data in the sleep community is the National Sleep Research Resource (NSRR), which will make 50,000 studies publically available for secondary analysis over the next few years(19, 20).

However, even when researchers are committed to data sharing, several issues arise. One of these key issues is trying to integrate data across studies. It is not uncommon for every study to use its own set of vocabulary terms. While this is not a large issue if only a few dozen data points have been collected for a limited number

of subjects, the problem rapidly becomes intractable when a study contain hundreds or more data points for thousands of subjects. Furthermore, the standardization of sleep terms and definitions across studies is not as well-developed as other fields. Over time, the standard definition of terms like apnea-hypopnea index (AHI) may change.(5) This makes comparisons across studies difficult. Resources such as Common Data Elements(21) and PhenX(22) are attempts to standardize vocabulary and protocols. Therefore, integrating heterogeneous datasets from different studies is a key informatics challenge.

Chapter 2: Background

Phenotype data in sleep

The traits of an organism are called phenotypes. Phenotypes result from interactions between an organism's genes and the environment over time. For example, obesity is a phenotype that is partially inherited, but also affected by diet. A disease will be associated with several phenotypes. Each phenotype is associated with one or more genes. The Monarch Initiative, a project developed for the integration of biological information using semantics from many sources, including gene, genotype, variant, and phenotype, has tools to investigate these associations.

Sleep studies collect a myriad of phenotypical data. There is the usual demographic and medical history data. The data specific to sleep studies comes in the form of subjective sleep questionnaires and objective polysomnograms (PSG). There are many different sleep questionnaires that ask the patient to self-report sleep symptoms, sleep-disordered breathing, insomnia symptoms, and family and social history. Typical sleep-related questions are "How many hours do you sleep a night?" and "How often do you fall asleep at work?" There are also psychiatric questions like "How many days do you feel nervous or anxious?" The answers are scored. This is typically how scores like the Epworth Sleepiness Scale are calculated.

PSGs are collected in an overnight patient sleep visit to a sleep clinic (they can also be collected in the home). The patient has electrodes and sensors applied all over their body, and then they will sleep in a private bed in the clinic while monitored by a sleep technologist. The signals recorded include EEG, EMG, ECG, EOG, airflow, breathing effort, oxygen saturation, and leg movements.

Ontologies

The formal naming and definition of entities and the interrelations between them is called an ontology(23-25). Popular websites using ontologies include Yahoo, which categorizes websites, and Amazon, which categorizes books and products for sale. Clinical studies often use a set of controlled vocabularies, called a data dictionary, in order to provide a standard naming convention for data capture and organization, but an ontology also defines the relationship between terms.

There are many reasons why ontologies are useful. One is sharing common understanding of the structure of information(26). In the case of having many different studies related to the sleep domain, if they used the same standardized vocabulary, then computer agents can extract and aggregate this information, which then could be used to respond to user queries or as input for other tools.

Enabling reuse of domain knowledge was one of the main drivers of ontology development(26). The Human Phenotype Ontology (HPO), for example, is a large-

scale ontology, containing over 12,000 terms, designed to support large-scale computational analysis of the entire human phenome(27). It is meant to be used by researchers, doctors, and patients for phenotype comparison and improved diagnosis. It merges many different ontologies together and invites contributors to extend its vocabulary.

An ontology makes domain assumptions explicit(26). What gives the HPO its power is that the curators have worked to develop computable logical definitions for the terms that relate phenotypic abnormalities to anatomy, pathology, physiology, biochemistry, and other areas. There is nothing inherent in the term "arthritis" that implicitly relates to "skeletal joint"; a human infers or learns this through education and experience, but it has to be explicit for machine computation. The structure of the ontology adds logic and makes functional comparison between entities possible.

Because the HPO incorporates a layer of 5,000 synonymous terms, it allows patients, basic science researchers, and machines to use it(28). For example, even if they may not know the term "hypopnea", they can search for "reduced breathing" and find it. In this way, it separates domain knowledge from operational knowledge.

Research Aim

Our aim is to improve treatment of obstructive sleep apnea through better subtyping. We mapped and integrated terms from a large, publically available sleep database, the Sleep Heart Health Study (SHHS)(29), to an ontology for phenotype comparisons, the Human Phenotype Ontology (HPO)(30), then used cluster analysis to find OSA subtypes.

To achieve the goals outlined above, we planned to use tools developed as part of the Monarch Initiative. They have created and collaborated on many computational tools and pipelines for ontology building and phenotype analysis, including the HPO(30), and OWLSim, an algorithm that compares phenotypes and calculates a similarity score based on information content(31-33). However, sleep terms are not well represented in the HPO. Therefore, novel discovery for sleep diseases through this methodology is limited. Conversely, integrating sleep terminology into the HPO will contribute to knowledge and discovery for other researchers.

The concept of semantic similarity clustering has been used for phenotype classification and disease gene discovery(34). Similar approaches have been done to study the genetic causes of bleeding and platelet disorders (35), drug associations in breast cancer tumor mutations (36). However, to our knowledge this has not been done for sleep apnea.

Precision medicine is a new approach for treatment of disease that accounts for difference in genes, environment, and phenotype for each patient. Different phenotypes respond differently to treatment, so understanding and characterizing those heterogeneous subtypes will allow us to more effectively treat each patient.

There are sleep components in rare phenotypes like Down syndrome which are heretofore unexplored and unquantified, and this work will help other researchers discover new relationships between these diseases. Lastly, OSA is a complex disease, and this approach can serve as a model for a workflow to use secondary data for data-driven subtyping of complex diseases.

Chapter 3: Aim 1

Aim 1: Determine which terms in SHHS, a large representative sleep dataset, are defined in the HPO, a tool for analyzing phenotypes.
Sub-Aim 1.1: Evaluate coverage of sleep-related phenotypes in HPO.
Sub-Aim 1.2: Enhance HPO with new terms covering sleep-related phenotypes.

Background

The Human Phenotype Ontology

The curators of the HPO have mapped nearly all the clinical descriptions in the OMIM (Online Mendelian Inheritance in Man) database, the largest database of human hereditary disorders(37). OMIM was developed 30 years ago and was not designed with a controlled vocabulary.

The HPO is structured as a directed acyclic graph (DAG) where the nodes are connected by **is.a** (class-subclass) transitions, which is similar to a hierarchy, but is different in that a child node can be related to more than one parent node (Figure 1). HPO annotations are represented in a simple tab-delimited format described at [http://human-phenotype-ontology.github.io/documentation.html](27).

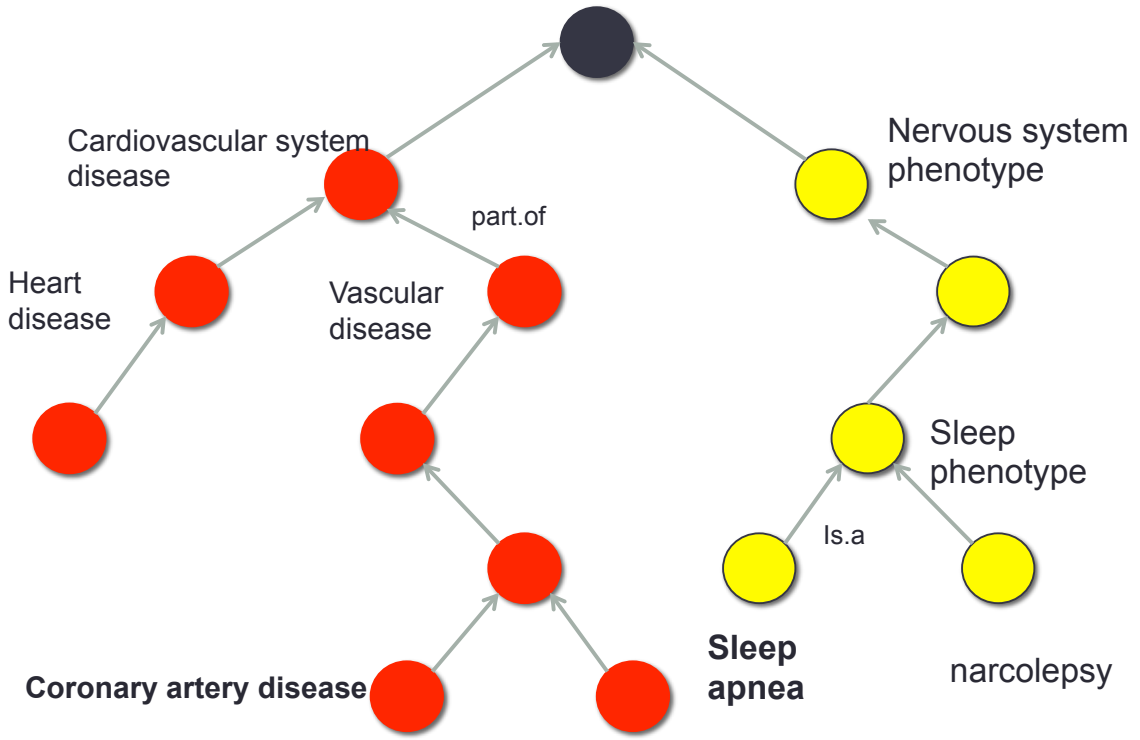


Figure 1. Schematic of a section of the HPO DAG. Not all nodes are labeled.

Sleep Heart Health Study

The data that is used in this study is from the Sleep Heart Health Study (SHHS), a multi-cohort study focused on sleep-disordered breathing and cardiovascular outcomes(29). It consists of 6,441 men and women aged 40 and older from six health centers. It is the largest widely cited study in the sleep field. There were two exam cycles, from 1995-1998 and 2001-2003. This dataset is now available for secondary analysis through the National Sleep Research Resource(20) (https://sleepdata.org), a comprehensive repository of de-identified sleep data, including bio-physiological signals, which are linked to risk factors and outcome data for participants in major National Institute of Health studies. Over 130 manuscripts have been published using this data. Other data sets will later become available, so our work in developing a pipeline for sleep phenotype discovery will be helpful for others wanting to use these data to their full advantage.

Data collected in the SHHS includes demographic information like age, ethnicity, and gender; anthropometric measurements like body mass index and neck circumference; physiological measurements like blood pressure and cholesterol; medications the patients are taking, self-reported health history like cigarette smoking and caffeine intake, sleep questionnaires that assess time and quality of sleep; and detailed polysomnographic measurements. Phenotypes collected include history of coronary artery disease, atrial fibrillation, asthma, stroke, diabetes, snoring, gasping during sleep, among others.

A request was submitted to the NSRR(19, 20), summarizing the goals of this research. The online NSRR Data Access and Use Agreement (DAUA) for this proposal was completed and executed. An Institutional Review Board/Ethics Committee review of this project was also performed by the NSRR Internal Review Committee to ensure the data will be used for appropriate research purposes and to ensure data security and appropriate handling of the data at the users site.

Methods

After acquisition of the SHHS data, the distribution and range of each variable was plotted. Obvious data entry errors were identified and removed. There are a total of 351 GB worth of 25,582 files, but the bulk of this data was EDF files containing the actual polysomography signals recorded during the sleep studies. This latter type of data was not needed for clustering. Also not relevant to our approach were EDF associated annotations, and EEG spectral analysis. Remaining were five main datasets: SHHS1, SHHS2, CVD (cardiovascular disease), HRV (heart rate variability), and interim, plus documentation, which were easily downloadable. The other important files were the data dictionary and domain information.

We have written code for the processing and manipulation of SHHS data, using R (38). This code is stored in an open Github repository (https://github.com/monarch-initiative/sleep-apnea-clustering). This code largely is built on top of the dplyr library(39), and a lot of functions are wrappers designed to work more closely with this particular dataset, breaking down into frequently used categories, removing unnecessary columns, and recoding fields and values into more usable formats.

Any field that consisted of more than 50% missing values was deleted and not considered for further analysis. Exploratory data analysis was performed to check for general trends and outliers.

There are 1,991 terms in the SHHS canonical data dictionary (version 0.11.0). Several of these differ only in category, degree, or visit (e.g. "REM power density at 8.0 Hertz" vs. "REM power density at 9.5 Hertz"). Duplicates were removed. We used a simple Python script that called the Monarch ontology as a Neo4j graph, designated Scigraph (https://github.com/SciGraph/SciGraph), and matched strings in an input text file ("traits") to terms in the ontology ("matches"). The script searched for up to three matches per trait. As a check, coverage of sleep terminology was assessed by using Monarch's built-in browser search function.

Extending and annotating the HPO is done by making term requests. A HPO curator is contacted through Github and the term, its definition, synonyms, and parent class are provided.

Case	Definition	SHHS	HPO
Match	SHHS term exists in HPO, same definition	hypertension	hypertension
Synonym	SHHS term does not exist in HPO, is synonymous with term that does	diabetes mellitus	Type II diabetes mellitus
Imperfect match	SHHS term exists in HPO, different definition	unrefreshed sleep	somnolence, drowsiness, sleep disturbance, etc.
No match	SHHS term does not exist in HPO, no synonyms	habitual snoring	
Category conversion	SHHS term can be a match to HPO if the data is put into categorical form	BMI	Overweight, obesity
Not appropriate	SHHS term does not exist in HPO, should not be put in categorical form, should be left out	NREM power density at 13.5 Hertz	

Table 1. The possible cases for term mapping from SHHS to HPO, with examples from preliminary search.

Table 1 describes the possible cases for term mapping. The easiest case is when there is a perfect match between the SHHS and HPO term. There is no need to do anything. Most synonyms will already be in the HPO.

An imperfect match where the definition is different in the HPO will require making term requests. We do not anticipate there will be that many instances where the definition needs drastic changing. Most likely there will just be some disambiguation

of terms. The same is true of synonyms. The synonyms are found through the search algorithm, then grouped all together, then term requests are made for each group.

For “No match” terms, a full term request was made for each one. Because our data dictionary is mostly in the sleep domain, which is not well-covered in the HPO, it may be more expedient because most of the terms will not have conflicts about where they fit in.

The domains of the terms were also considered for the “category conversion” case. There were variables such as “Type of stroke”, to which there were 8 separate types, such as cerebral hemorrhage and ischemic stroke. These would be considered separate terms in the HPO, and the coding of the dataset was altered to expand for these. In the same vein, domains that were not listed explicitly were created, for example, BMI can be converted from a number into categorical ranges, so that there are “underweight” (<18), “overweight” (25-29), and “obese” (>30) categories.

Results

By comparing the terms in the data dictionary of the Sleep Heart Health Study (SHHS) and those in the HPO using text mining, terms were identified that needed to be added to the HPO.

Deleting fields with more than 50% missing data eliminated columns related to HRV (heart rate variability) entirely. Because some patients were allowed to fill out questionnaires by mail, a lot of survey questions were left blank. Some variables were not collected by certain centers. The SHHS1 dataset was reduced down to 894 fields from 1,207, the SHHS2 dataset down to 804 fields from 1,213, and the CVD data down to 24 fields from 38. In the next step, any patient that was missing a value from any of the fields was eliminated. In the final clustering analysis, this reduced number of patients to 4,556.

Table 2 is a selection of the overall statistics of the SHHS dataset. The patient cohort, even though measures were taken to include as many minorities as possible, was overwhelmingly white (84.5%). Asians, Hispanics, and Native Americans were all put into the “other” category. There is evidence of Asians possibly requiring a different set of cutpoints for body mass index(40), but mixing them with other ethnicities makes it impossible to apply those separate thresholds. Our exploratory data analysis indicated that race did not have a significant effect on key variables, with the exception of hypertension (41). Some centers that recruited largely or exclusively from a minority patient base did not collect a lot of the variables, therefore, since our analysis method requires clustering over the full range of variables, those patients were not used.

	mean	SEM
BMI	28.16	0.06
Age	63.13	0.15
AHI 3%	14.65	0.2
AHI 4%	10.18	0.17
ESS	7.71	0.06
	n	**%**
sex(male)	2765	47.60%

Table 2. Overall statistics for SHHS dataset. BMI=body mass index, AHI 3% = apnea-hypopnea index (events/hour using 3% oxygen desaturation criteria), AHI 4% = (4% oxygen desaturation criteria), ESS= Epworth Sleepiness Scale

A preliminary assessment of sleep term coverage in Monarch showed that the "Sleep Phenotype" class is small, with three subclasses, and 16 phenotypes below that.

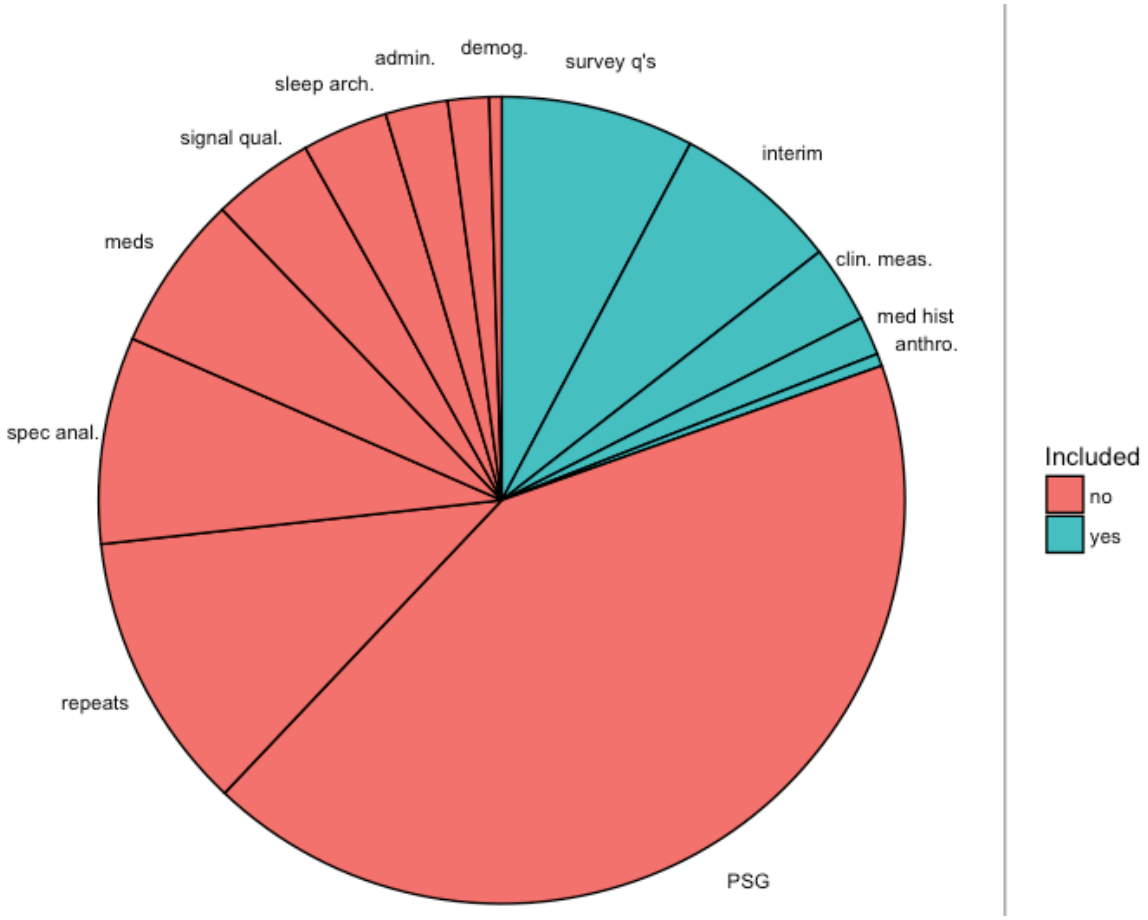


Figure 2. Overview of the data dictionary for the SHHS, divided into general categories. Of the 1991 terms, 392 were deemed possible candidates for phenotypes. 1599 were unsuitable.

From the NSRR canonical data dictionary, a list of 392 terms was drawn after removing inappropriate and duplicate terms, plus adding separate domains as terms (Figure 2). There are 340 primary matches out of the 392 terms, but it is clear that several of these are nonsensical or not applicable (e.g. “hypopnea events” is matched with “fewer egg laying events during active”). Scigraph’s search engine, although based on Lucene(42), a Java-based search engine library, is quite basic. It simply examined any word in the input string and found any Monarch term that also contained that word. When further restricted to HPO terms only and eliminating SHHS terms that point to the same HPO term, 79 primary matches were found.

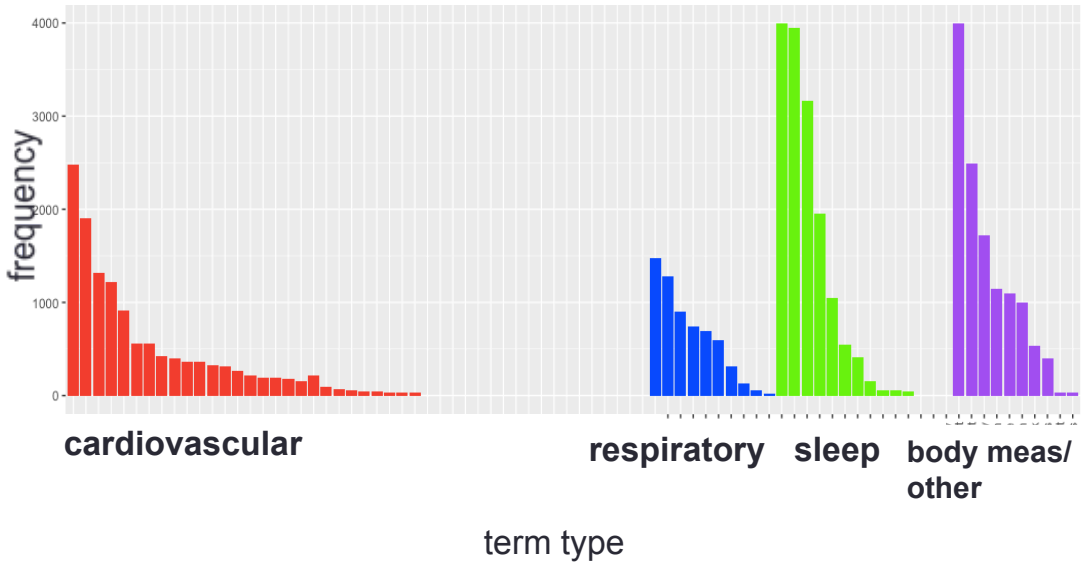


Figure 3. Frequency of occurrence in the SHHS cohort of the 79 HPO terms that could be mapped to the SHHS data dictionary. Whenever possible, matches were consolidated into as few terms as possible. For example, “heart attack” was consolidated under “myocardial infarction”. In the end, 79 terms were found. These fell under four general categories.

There were very few non-matches between the HPO and the SHHS, which was a pleasant surprise and a testament to the hard work of the curators. As a result, only three term requests were made to the Monarch HPO Github. These were for snoring (Figure 4), loud snoring, and wide hips. All requests were answered promptly, and were accepted with some revisions. “Wide hips” was already included under “large pelvis”.

We would like to submit a revised definition of snoring.

Preferred label: Snoring
Synonyms: snore, snores, snoring symptoms

Definition: Intense, noisy breathing during sleep(43) caused by air pressure and resistance changes in the upper airway(44).

References: UMLS C0037384, NCI thesaurus code C116315 [(https://ncimeta.nci.nih.gov/ncimbrowser/ConceptReport.jsp?dictionary=NCI%20Metathesaurus&code=C0037384)], Hoffstein, 1996; Levartovsky et al., 2016

Suggested Parent term: sleep phenotype, breathing dysregulation

A term for Snoring already existed (HP:0025267). I have created a hybrid definition from your suggestion
Deep, noisy breathing during sleep accompanied by hoarse or harsh sounds caused by the vibration of respiratory structures (especially the soft palate) resulting in sound due to obstructed air movement during breathing while sleeping.
and added the synonyms. I also agree with 'breathing dysregulation' as a parent class.

Figure 4. An example of a Github term request that was successful.

Discussion

The major drawback of using HPO and Monarch tools for clustering is that there is a lot of information lost when eliminating the polysomnography data such as arousals, oxygen desaturations, and apnea and hypopneas. At the same time, this data is too granular to effectively incorporate into the HPO. Because of this, we cannot leverage the semantic similarity calculator as a clustering metric. Perhaps the biggest loss of information is that there is no designation of the severity of the phenotype in many cases. A future direction would be adding the severity of the phenotype with an annotation (http://human-phenotype-ontology.github.io/documentation.html#annot).

There are also a lot of interesting subjective data that does not exactly qualify as a phenotype, such as the frequency of "falling asleep while watching TV" or "how blue do you feel." Grouping them under existing terms like "excessive daytime sleepiness" or "depression" would be implying causality.

The final 79 terms predominately consist of rare ECG anomalies. Many apply to as few as 0 or 1 patients out of the whole cohort. We do have information, but it might not be the right type of information needed to produce a quality semantic clustering.

What was accomplished was preparation and cleaning of the SHHS data for data manipulation, as well as some enhancement of the HPO with new sleep terms. There should be more work done in this area, and perhaps it could be guided by the results of Aim 2, which could indicate what terms to focus on.

Chapter 4: Aim 2

Aim 2: Identify subtypes of obstructive sleep apnea.

Background

There are a couple of challenges that need to be addressed before the SHHS dataset can be used for clustering. The first issue is how to handle the different types of data. The second is dealing with the high dimensionality of the dataset.

Clustering

Clustering is the task of grouping a set of objects or events such that objects in the same group ("cluster") are more similar to each other than objects in other groups(12). Clustering is an unsupervised task, meaning that there is no "training" by a labeled sample set. This is why it is suitable for this problem. As stated before, there is no consensus solution for subtyping sleep apnea. Clustering is not a particular algorithm, rather, many different methods fall under the umbrella of clustering. It is assumed the reader has knowledge about clustering algorithms. The common algorithms that will be used in this work are Partitioning Around Medioids (PAM) and hierarchical clustering. A detailed discussion of these, as well as semantic clustering based on OWLSIM(31), is contained in the Appendix.

The pipeline for our analysis is described in Figure 5. The features to be used must be selected (dimensionality reduction). Then each patient record containing these features is compared to every other patient record, and the distance or dissimilarity between them is calculated according to some metric. These scores can be arranged in a matrix, which is used by a clustering algorithm to group similar records together. The clustering result is measured for quality and validated.

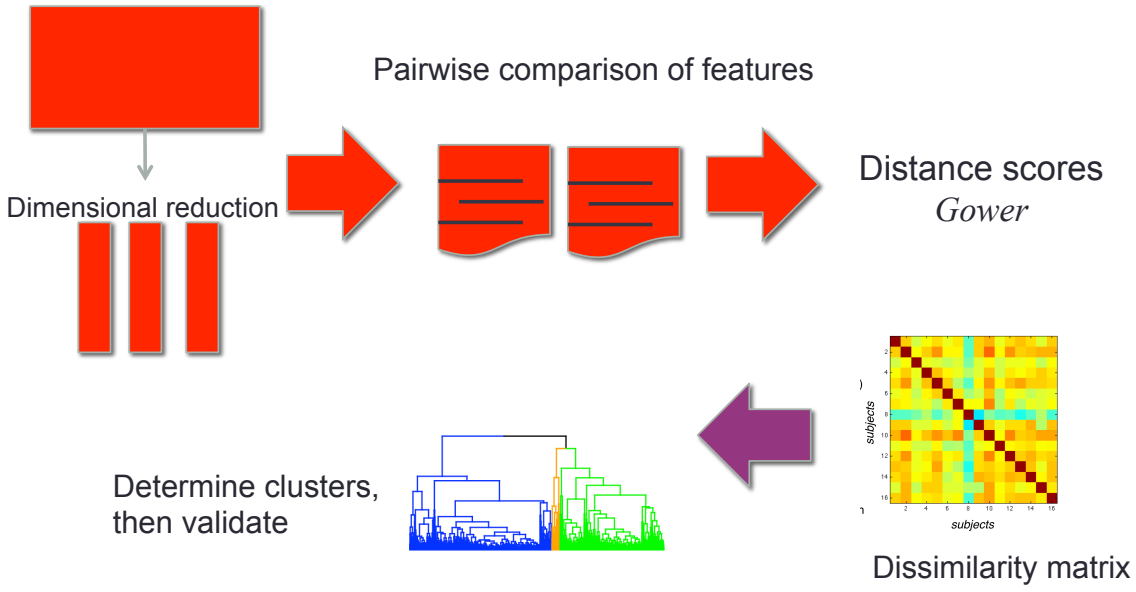


Figure 5. Pipeline of clustering analysis.

Handling mixed data

This dataset shares many characteristics with psychological research, in particular, having mixed data, including ordinally scaled data from survey questionnaires. Because of this, the R library `psych` (45), which is able to handle binary and ordinal data was used for analysis.

The SHHS dataset consists of a mix of continuous, dichotomous, and polytomous variables. Kolenikov and Angeles wrote a detailed discussion about ways to deal with the violations of the normality assumption when there is discrete data(46). We considered several approaches to handle this problem. One was coding our own functions to calculate chi-square statistic for comparisons between binary variables and ANOVA statistic for comparisons between binary and continuous variables, which can be used like Pearson or Spearman correlation statistics. Another approach was to use network analysis of mixed graphical models(47). The easiest approach was to use polychoric correlation(48).

The tetrachoric correlation(49) calculates the equivalent of Pearson correlation for binary data. The polychoric correlation(50) does the same for ordinal category data.

Binary data was coded as 0 (if a particular variable is absent) and 1 (if that variable is present). What complicated the SHHS data was that there were various ways of coding "don't know" or "not applicable" from survey to survey. We rectified this by treating all "don't know" responses as NA (missing data). In one variable that we thought was important to preserve, loud snoring, we imputed missing data by means of a random Gaussian function. Ultimately, this variable was not used in the final analysis. (Data showing imputation did not affect final distribution can be provided in the appendix.)

Dimensionality reduction

We used factor analysis to perform dimensionality reduction. The principle behind factor analysis is that multiple observed variables have similar response patterns because they are associated with a latent factor(51, 52). Essentially, a high-dimensional matrix can be represented as the product of two smaller matrices plus noise. This can be expressed in a form of an equation:

$$X_{n\times p} = Z_{n\times k}\Lambda_{k\times p} + \epsilon_{n\times p}$$

where

1. $X_{n\times p}$ is the observed data matrix consisting of n observations of p features;
2. $Z_{n\times k}$ is the factor matrix consisting of k factors for each of the n observations
3. $\Lambda_{k\times p}$ is the factor *loading* matrix which tells the contribution of each feature to each factor.
4. $\epsilon_{n\times p}$ is random Gaussian noise, $\epsilon_i \sim (0, \Psi)$.

The question of how many factors to retain has many conflicting answers, and a balance must be struck between them(53). The number of factors was estimated using a parallel analysis(54) on a scree plot. A scree plot is a plot of the eigenvalues of each factor. The scree test (55)is a subjective test where the cutoff for number of factors is determined by where there is a bend in the graph and the eigenvalues level off. Parallel analysis attempts to introduce objectivity. A random data set of the same size and the same range of values is simulated. The eigenvalues of this random data are extracted and plotted on the same scree plot as the real eigenvalues. The factors that should be retained are those with eigenvalues greater than the equivalent random eigenvalues.

An alternate analysis called Very Simple Structure (VSS) was also considered(33). VSS compares the original correlation matrix (R) to that produced by a simplified version (S) of the original factor matrix. If we consider that R is an n x n matrix that is a factor matrix of dimensions n x k multiplied by its transpose, plus a diagonal matrix of uniquenesses,

$$R_{nn} \sim F_{nk}F'_{kn} + U^2_{nn}$$
$$R \sim SS' + U^2$$

Then we can also have a simplified matrix S and its transpose S'. S is composed of the c largest loadings of each factor, where c (complexity) is a parameter from 1 to the number of factors. The VSS criterion is the fit of the simplified model to the original correlation matrix.

$$VSS = 1 - \frac{\sum r*^2}{\sum r^2}$$

where R* is residual matrix R*=R-SS' and r* and r are the elements of R* and R respectively.

Wayne Velicer's Minimum Average Partial (MAP) criterion(56) is a related test to VSS to decide the optimal number of factors to extract. In short, it involves finding the average squared correlation of the full correlation matrix, then subsequently partialing-out of the first factor and finding the average squared correlation of that matrix. The process repeats. The MAP is the step where the lowest average squared partial correlation was found, and represents the point at which variance in the matrix is still systemic, rather than random error.

Validation of clusters

Since unsupervised learning defines clusters that are not known *a priori*, there has to be some kind of evaluation of the clustering. There are two types of validation. One is biological validation, where one uses another sample or set of samples to confirm the results, using an alternate method(57). We could apply our algorithm to a previous paper's data to compare our clustering results to theirs, but we do not have easy access to their data. Thus we are limited to computational validation. If

we can make the case that our computational results hold true, then biological validation can be attempted by experimental scientists.

There is no generally agreed upon categorization of sleep apnea subtypes, and previous research disagrees on the number and features of the subtypes. Furthermore, as we only have access to this particular data set, we can only computationally validate based on internal criteria.

Homogeneity is the measure of the similarity of elements in a cluster to one another (Figure 6A). Alternately, it measures the amount of variation within a cluster. Formally, it is defined as the average distance of an observation to its assigned cluster center over all points. It is an unbounded measure and requires cluster centers as an input.

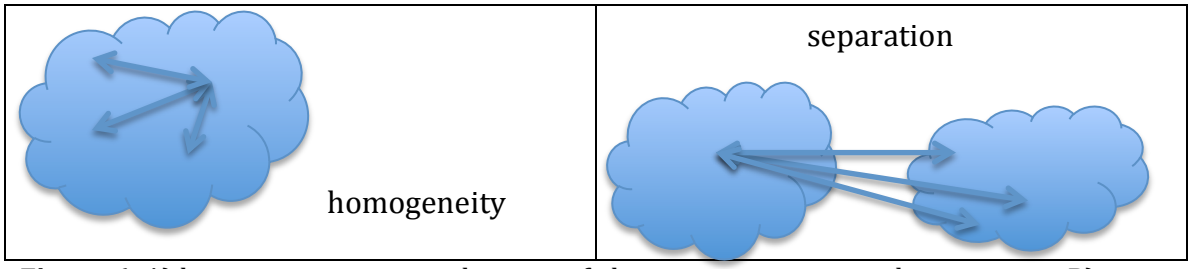


Figure 6. A) homogeneity is mean distance of observation to its own cluster center. B) separation is the average distance between the cluster centers, normalized for the number of cluster members. Silhouette width incorporates both the distance of observations to their own cluster center and to neighboring cluster centers.

Separation is the measure of the amount of variation between clusters (Figure 6b). The distance between each cluster center is calculated and normalized for the number of cluster members. The average distance is taken as the separation. As with homogeneity, separation is unbounded and requires knowledge of the cluster centers.

Silhouette width(58) can be thought of as a balance between homogeneity and separation. Let a be the average distance of an element to all other points in its own cluster. Then call average distance from the same point to points in the next closest cluster b. The silhouette width is then defined as:

$$s = \frac{b - a}{\max(a, b)}$$

We then divide by the number of total data points for the average silhouette width. This value is bounded by [0,1], and values closer to 1 are preferred.

The *Dunn index* is the shortest distance between all points not in the same cluster, divided by the longest distance between points in the same cluster (59). Mathematically,

$$D(\mathbb{C}) = \frac{\min_{C_j, C_k \in \mathbb{C}, C_j \neq C_k} \left(\min_{i \in C_j, j \in C_k} dist(i,j) \right)}{\max_{C_m \in \mathbb{C}} diam(C_m)}$$

where diam(C_m) is the maximum distance between observations in cluster m. Dunn index is bounded by [0, ∞) and larger values are interpreted as better(60).

Connectivity differs from silhouette and Dunn index, in that it is somewhat like the precision or recall statistic. For a given distance matrix, it determines how often the observations that are the nearest to each other are placed in the same cluster(60). If N is the number observations in a dataset, define $nn_i(j)$ as the jth nearest neighbor of observation I, and let x_i, $nn_i(j)$ be zero if I and j are in the same cluster and 1/j otherwise. Then, for a particular clustering C into K clusters,

$$Conn(C) = \sum_{i=1}^{N} \sum_{j=1}^{L} x_{i, nn_{i(j)}}$$

where L is a parameter giving the number of nearest neighbors to use. Connectivity is bounded by [0, ∞) and smaller values are preferred(60).

There is an R package designed for comparison of clustering quality called Consense(61, 62). While we did not directly use this package, many of the ideas from it were incorporated in this part of the analysis.

It should be stressed that these measures only serve as a guideline, and that ultimately, the clusters should also be assessed as to whether they make sense biologically.

Methods

The polychoric correlation matrix of the data was calculated using `heterochor` in the R library `polycor(63)`, with factoring method of Maximum Likelihood and the stipulation that the resulting matrix must be positive-definite (that is, no negative eigenvalues allowed and the matrix must be invertable). Factor analysis was performed using `psych` functions. The number of factors was determined by considering parallel analysis, the VSS, MAP, and finally, the individual eigenvalues and percentage of variance explained (based on cumulative eigenvalue) are considered. The highest load feature for each factor was taken as the representative of that factor.

Clustering

After dimensionality reduction, clustering was performed using the `daisy` function in the `cluster` library(64). The Gower distance was used as the distance metric. (65) Certain factors were marked as symmetrical binary (e.g. gender) and others were marked as asymmetrical binary (e.g. coronary artery disease) and scaled

ordinal factors were marked as such. All other factors were considered to be continuous.

The resultant dissimilarity matrix serves as the input to a number of clustering functions, including `pam` and `hclust`. The clustering algorithms chosen were PAM (partitioning around medioids), `diana` (divisive hierarchical clustering), and three methods of hierarchical agglomerative clustering: single-linkage, complete-linkage, and average-linkage. Other clustering methods were not considered as they either took unreasonably long to run or did not accept the distance matrix as an input. For a detailed explanation about clustering algorithms, including OWLSIM semantic clustering, see the appendix.

The proposed pipeline for OWLSim semantic clustering was as follows:
1. Each patient would be coded as a list of HPO phenotypes (the "patient record") in a csv file.
2. OWLSIM would be adapted to compare each patient record to each other, outputting the IC and simJ scores.
3. Choosing either IC or simJ across all pairs, each pairwise comparison score is placed in a similarity matrix. The scores are subtracted from 1 to make a dissimilarity matrix.
4. The dissimilarity matrix is then fed into one of the other clustering algorithms like `PAM` or `hclust`, and clusters are determined.

Validation

Silhouette width, Dunn index, and connectivity were calculated for clusterings with 2 to 10 clusters for each separate method. Clusters were set by cut point in the case of the hierarchical clustering, and specified beforehand in the case of PAM.

Comparisons between clusters were made by one-way ANOVA with post-hoc Bonferroni correction and chi-square tests for continuous and binary data respectively. Data were expressed as mean ± standard error of the mean when reporting continuous variables and counts or percentages when reporting categorical variables. For all analyses, a p-value of less than 0.05 was considered significant. All analysis was performed in R version 3.2.2.

Results

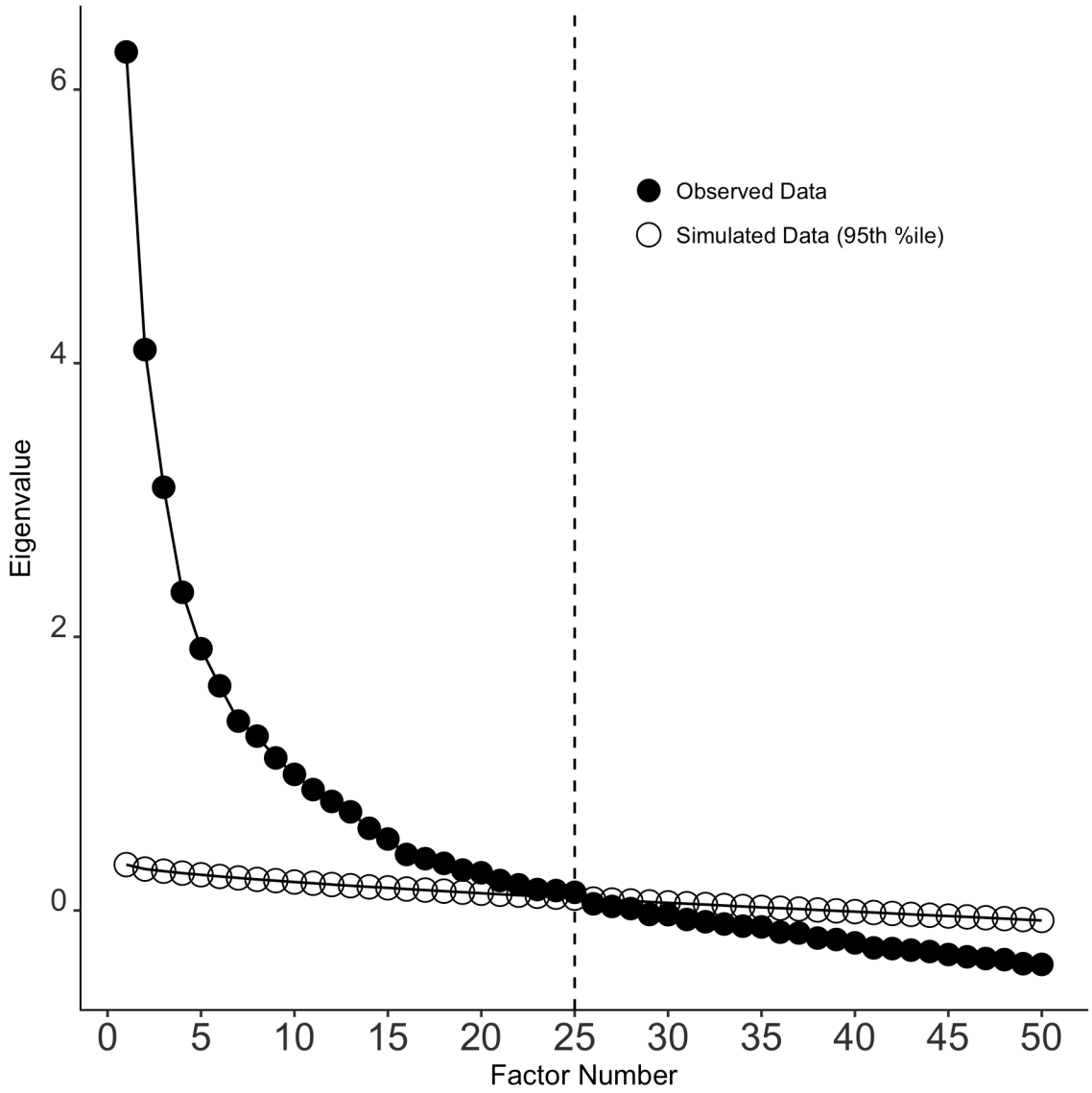


Figure 7. Scree plot and parallel analysis.

The scree plot (Figure 7) and table of factor loadings (Appendix, Table B) shows that the first five factors have relatively large eigenvalues compared to the rest and explain 24.2% of the variance. Parallel analysis indicates that the eigenvalues of the random junk data set begin to be larger than those of the real dataset at factor 25.

VSS and MAP analysis agreed with the assessment, with MAP reaching a minimum at 5 factors and VSS suggesting 4 factors may be the most appropriate.

Factor analysis suggests the dataset reduces down to five factors: (1) respiratory index (defined below), (2) presence of cardiovascular disease, (3) diagnosis of sleep apnea by a medical doctor, (4) presence of hypertension, and (5) SF-36 Physical Component Score (Figure 8, Appendix Figure C).

Respiratory index includes several correlated variables within the sleep architecture and polysomnographic categories, with high loadings for "percent of sleep time in apnea or hypopnea with greater than 3% desaturation", AHI, and average apnea length. The cardiovascular disease factor was most associated with coronary artery disease, but also associated medications and conditions like nitroglycerin and coronary artery bypass graft. There were two separate variables for diagnosis of sleep apnea by a medical doctor (abbreviated "SA15" and "MDSA02"). They differed with each other on some patients, which may indicate that some patients were diagnosed later. The link between hypertension and sleep apnea has been well-studied. SF-36 PC score is itself a factor score of how tired a person gets doing various tasks. PC score was correlated with several Quality of Life survey questions.

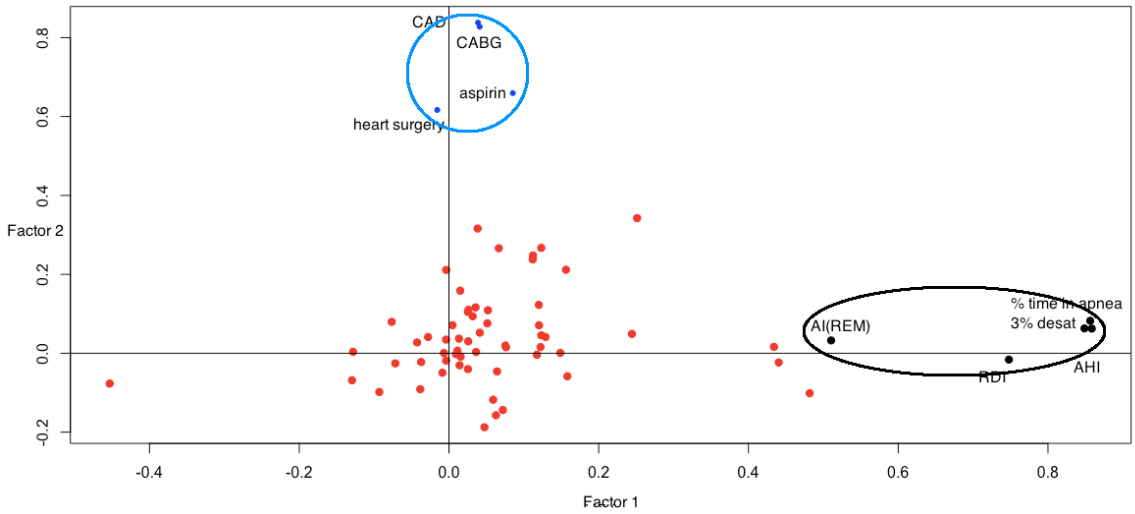


Figure 8. Factor analysis plot of first two factors against each other. All the respiratory index features line up along factor 1 axis. CAD (coronary artery disease) and CABG (coronary artery bypass graft) line up along another. Note that there are a lot of features that do not contribute much in any direction and congregate in the center (red points). Removing these reduces the random variation in the matrix. AI(REM) = Arousal index in REM sleep, AHI=apnea-hypopnea index, RDI=respiratory disturbance index, % time in apnea 3% desat = percentage of time spent in apnea over 3% oxygen desaturation.

Python code was written and adapted for the semantic clustering. Specifically, there was a function to input a csv file of patient records as lists of HPO terms, a function to split that data into separate lists, and a function to use OWLSIM to compare the lists and extract the IC and simJ scores from the JSON output.

However, the function to convert the SHHS features into HPO phenotype terms was not written, partially because most of the code for manipulating the SHHS was already written in R, and also because by this time we had found that much of the sleep architecture and respiratory data could not be easily adapted into the HPO.

Therefore, this part of the aim was abandoned, and the rest of the results will focus on the traditional quantitative clustering.

Because of the goal of the study was to find subtypes of sleep apnea, only clusterings that resulted in 3 or more clusters were considered (Figure 9). PAM performed very poorly in all three metrics. This may be because of the binary features in the model, which tend to dominate if PAM is applied.

Most of the hierarchical clusterings showed marked worsening after more than 6 clusters. At six clusters, the complete linkage clustering performed the best with respect to silhouette width and connectivity. Single linkage performed best in Dunn index, but its dendrogram is much lower in height than the complete linkage, meaning a lot of objects were grouped together very early and the clusters are quite diffuse. Upon checking the key variables in a single-linkage clustering, it was revealed that there were no significant differences in characteristics between clusters (i.e., ANOVA tests failed to return a p-value below 0.05 when testing a difference between mean AHI of different clusters).

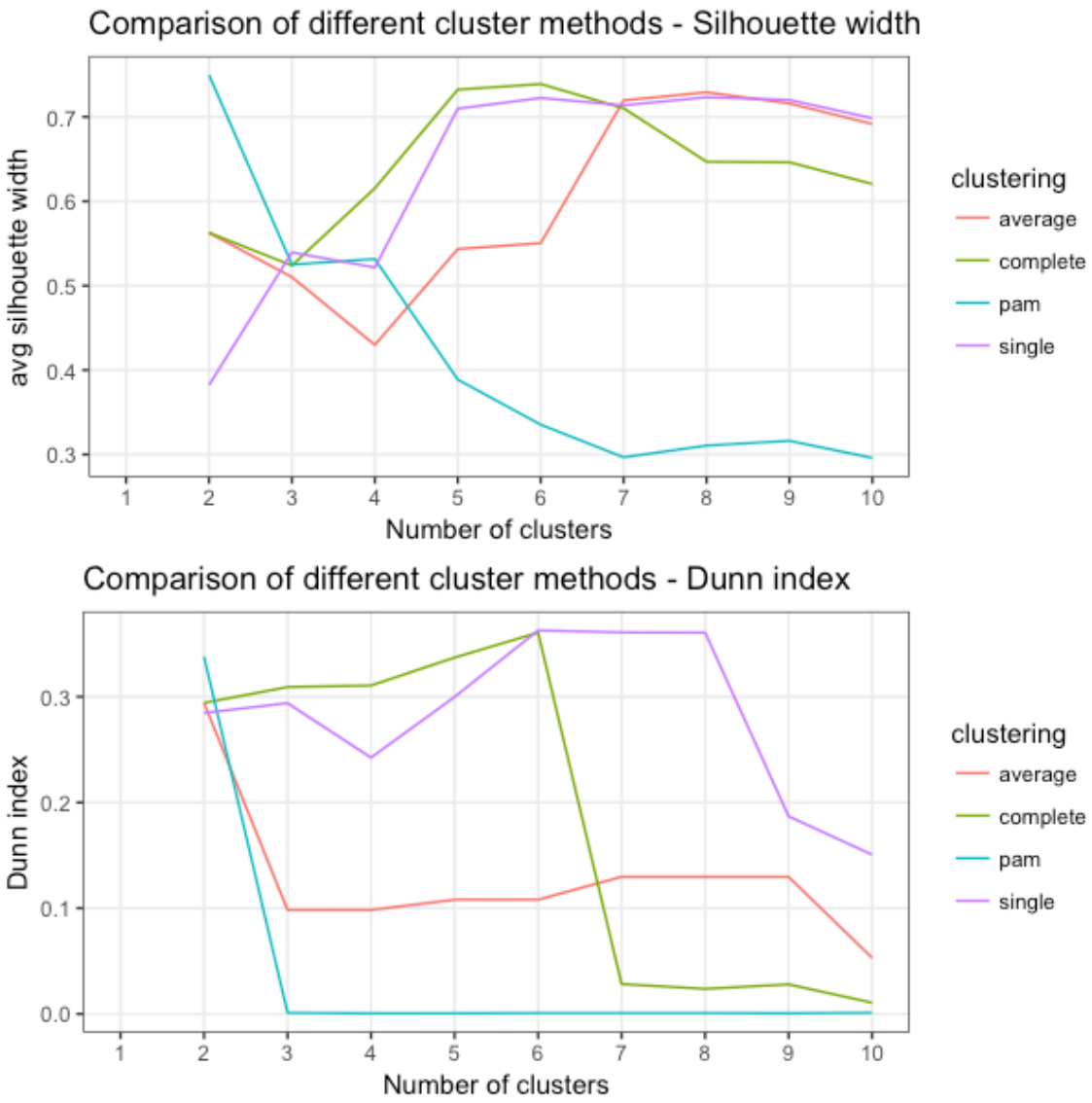


Figure 9. Comparison of connectivity (top), average silhouette width, and Dunn index for four different methods of clustering. Other hierarchical clustering methods were omitted for the sake of clarity.

It was decided to choose complete linkage clustering with six clusters, based on both internal validation metrics and the clinical interpretability of the clusters, which will be discussed in the next section. The average silhouette width was 0.72, which is considered fairly good separation (66).

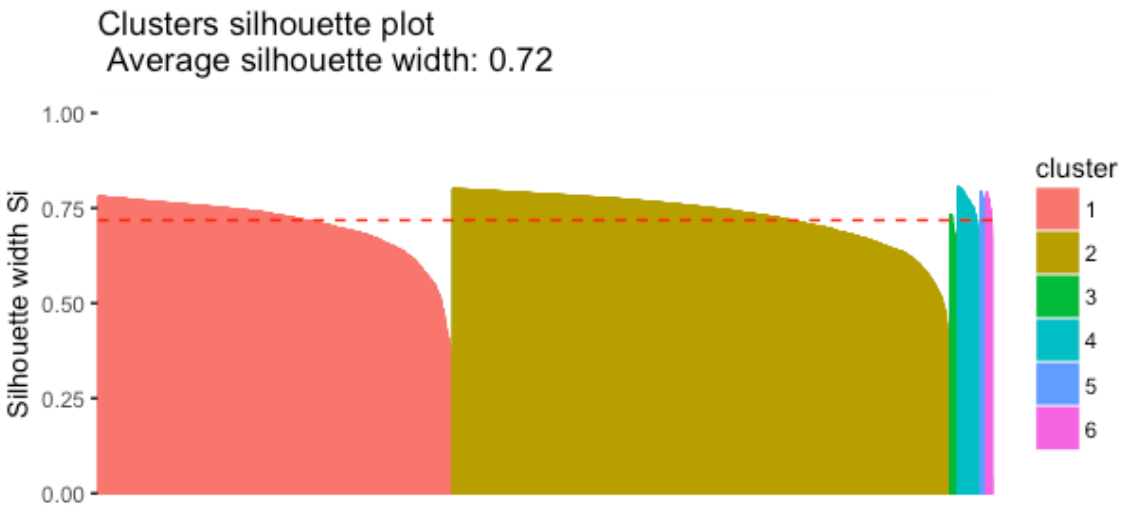
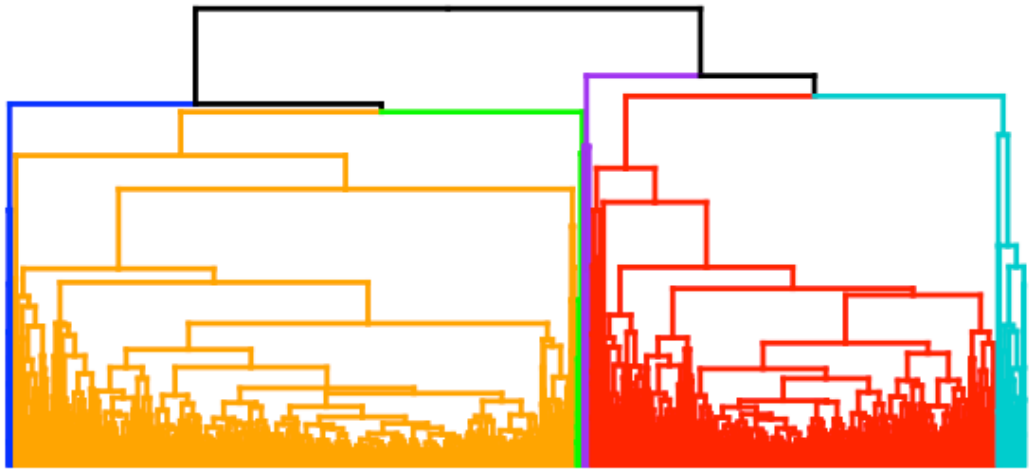


Figure 10. Silhouette width plot for complete linkage hierarchical clustering with six clusters. Average silhouette width of 0.75 is considered fairly good in terms of homogeneity and separation.

The six clusters can be characterized as follows (Tables 3 and 4). Cluster 3 is confirmed sleep apnea cases with the highest AHI (33.90 ± 4.51) out of all the clusters. Related measures like percent of sleep time under 90% oxygen saturation (T90) and percent of sleep time spent in apnea/hypopnea with desaturation greater than 3% were also the greatest in this cluster. Subjects in this cluster also had the highest BMIs (31.90 ± 0.90), highest systolic and diastolic blood pressure (139/80), and highest FVC (3.90 ± 0.15). This cluster also had the lowest minimum SaO2 (81.7 ± 1.87). This cluster also had the highest percentage of males (83%), and subjects take significantly more medication related to hypertension.



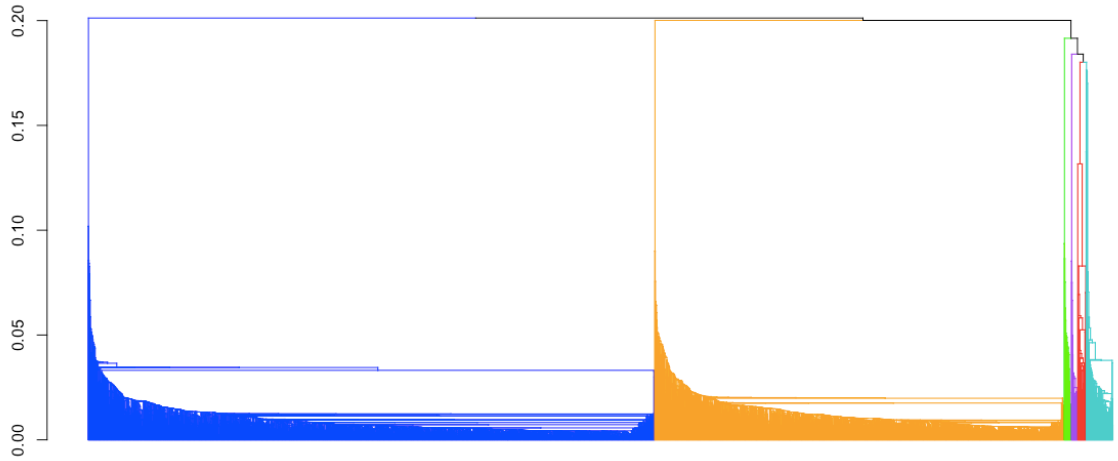


Figure 11. (Top) Cluster dendrogram for complete linkage hierarchical clustering with six clusters. (Bottom) Dendrogram for single linkage clustering with six clusters. Single-linkage or nearest neighbor method tends to "chain" or add the nearest point one at a time, resulting in a spread-out cluster.

The other confirmed sleep apnea subgroup is cluster 5. The major difference is that subjects in this cluster have not been diagnosed with hypertension. However, they do have a higher proportion of respiratory diseases, stroke, diabetes, and smoking than all the other clusters. They have the highest cholesterol levels out of all the groups (214.8 ± 7.9), as well as the highest ESS (11.9 ± 1.1). Their AHIs are not as severe (25.5 ± 3.9) but they have similar mean BMI as cluster 3. This group has a smaller percentage of males and is on average the youngest (56.7 ± 1.8).

Cluster	1	2	3	4	5	6	p-value
Age	67.15 ± 0.25	61.02 ± 0.22	63.46 ± 1.64	69.55 ± 0.88	56.65 ± 1.78	67.38 ± 1.5	<0.0001
BMI	28.79 ± 0.13	27.52 ± 0.09	31.9 ± 0.9	28.31 ± 0.47	31.03 ± 1.01	27.15 ± 0.66	<0.0001
AHI 3 %	17.08 ± 0.41	12.45 ± 0.27	33.9 ± 4.51	19.23 ± 1.51	25.49 ± 3.87	18.89 ± 3.51	<0.0001
AHI 4 %	12.18 ± 0.37	8.31 ± 0.22	28.02 ± 4.37	13.51 ± 1.39	18.97 ± 3.53	13.71 ± 3.2	<0.0001
T90	4.45 ± 0.27	2.45 ± 0.16	12.05 ± 3.97	7.86 ± 1.88	6.92 ± 3.09	3.73 ± 1.49	<0.0001
FEV1	2.45 ± 0.02	2.81 ± 0.02	2.94 ± 0.13	2.58 ± 0.08	2.95 ± 0.15	2.74 ± 0.15	<0.0001
FVC	3.25 ± 0.02	3.72 ± 0.02	3.9 ± 0.15	3.43 ± 0.08	3.87 ± 0.17	3.69 ± 0.17	<0.0001
Cholesterol	209.76 ± 0.91	206.93 ± 0.77	199.22 ± 6.52	198.79 ± 3.22	214.81 ± 7.89	208.19 ± 6.55	0.0105
Diastolic BP	75.29 ± 0.32	72.09 ± 0.21	80 ± 1.96	71.11 ± 1.25	74.87 ± 2.33	67.44 ± 1.94	<0.0001
Systolic BP	135.05 ± 0.48	121.52 ± 0.34	138.58 ± 3.05	132.12 ± 1.72	122.45 ± 3.13	124.96 ± 3.36	<0.0001
ESS	7.77 ± 0.11	7.64 ± 0.09	11.22 ± 0.98	9.07 ± 0.42	11.94 ± 1.14	7.3 ± 0.71	<0.0001
Minimum SaO2	86.07 ± 0.15	87.543 ± 0.114	81.69 ± 1.87	85.26 ± 0.66	83.64 ± 1.26	86.48 ± 1.17	<0.0001
SF-36 Mental Comp	53.25 ± 0.2	53.83 ± 0.15	51.63 ± 1.44	53.58 ± 0.71	53.56 ± 1.57	54.51 ± 1.12	0.149
SF-36 Physical Comp	45.11 ± 0.24	49.7 ± 0.17	47.13 ± 1.96	42.79 ± 0.98	44.98 ± 2.14	46.27 ± 1.86	<0.0001

Table 3. One-way analysis of variance (ANOVA) for continuous variables

Of the three non-sleep apnea-diagnosed clusters, cluster 2 can be thought of as healthy subjects. They have low proportions of all the diseases tested for, no hypertension, and the lowest AHI (12.45 ± 0.27), highest minimum SaO2, and highest SF-36 physical component score, meaning they have the highest endurance in performing tasks during the day.

Cluster	1	2	3	4	5	6	p-value
n	1808	2533	37	116	31	31	
gender	0.465	0.463	0.838	0.698	0.677	0.774	<0.0001
CAD	0	0	0.081	1	0	1	0
Sleep apnea MD	0	0	1	0	1	0.065	0
Hypertension	1	0	1	1	0	0	0
Smoking	0.518	0.52	0.595	0.621	0.71	0.677	0.0262
Alpha-blockers	0.077	0.029	0.189	0.086	0.032	0.065	<0.0001
Asthma	0.084	0.079	0.065	0.079	0.12	0.071	0.964
Runny nose	0.287	0.231	0.351	0.302	0.452	0.452	<0.0001
COPD	0.014	0.006	0	0.04	0.08	0.036	<0.0001
Diabetes	0.116	0.041	0.065	0.208	0.08	0.071	0.0157
Heart failure	0.03	0.004	0.065	0.109	0	0	<0.0001
Stroke	0.055	0.014	0	0.119	0.04	0.036	<0.0001
Angina	0.101	0.019	0.129	0.604	0.08	0.464	<0.0001
CABG	0.051	0.014	0.032	0.277	0.04	0.357	<0.0001
MI	0.092	0.017	0.065	0.505	0.08	0.429	<0.0001
Pacemaker	0.015	0.003	0.065	0.04	0	0.036	<0.0001

Table 4. Cluster characteristics

Cluster 4 has no sleep apnea cases but 100% of subjects have hypertension and coronary artery disease. As such, there are a high proportion of related cardiovascular disorders, stroke, and diabetes. This is the oldest group (69.6 ± 0.9). their AHI is the highest among the clusters with no diagnosed sleep apnea cases, and the ESS is the highest in any category.

Cluster 6 consists of subjects with coronary artery disease cases, but with no hypertension. Two sleep apnea patients were grouped in this cluster, which indicates sleep apnea diagnosis is not the strongest factor in the hierarchical clustering, and there may be some undiagnosed sleep apnea cases in all the clusters. There is a high proportion of angina, coronary artery bypass graft, and myocardial infarction.

The two sleep apnea groups can be characterized as severe sleep apnea with hypertension (cluster 3) and moderate sleep apnea without hypertension, but with higher likelihood of respiratory comorbidities (cluster 5).

It may seem like it is a foregone conclusion that using sleep apnea diagnosis as one of the factors will result in a solution that sorts the most severe sleep apnea cases into neat clusters. It should be noted that by the criterion of mean AHI, all the clusters except for group 2 would have moderate to severe sleep apnea, so there are probably several undiagnosed cases. Also, we repeated the analysis using four factors, removing the factor of sleep apnea diagnosis (MDSA02). Similar clusters were found, except now the “healthy subjects” cluster includes the healthier MDSA02 subjects.

An independent analysis on the SHHS2 dataset gave similar results. Respiratory index (AHI) was by far the most important factor, then physical limitations on activity, aspirin (related to cardiovascular phenotypes), sleep efficiency, percentage of time with oxygen saturation is below 75%, and then hypertension.

An analysis of the "interim" dataset, without PSG data or medication data, came up with blood pressure (hypertension) as the most important factor, followed by "falling asleep during some sort of activity" (related to physical limitations), "sleep apnea symptoms", and "awakening during sleep."

We saw similarities between our clusters and those of the Vavougious *et al.* study, particularly in the existence of severe obstructive sleep apnea syndrome (OSAS) phenotypes with and without comorbidities, as well as moderate OSAS phenotypes with and without comorbidities. In both our studies, the phenotype with the most severe AHI had fewer concomitant diseases, other than obesity, while a more moderate phenotype was associated with a lot of comorbidities.

Age did not seem particularly linked to OSA severity in our study. Age has been associated with an increased rate of OSAS but reduced severity in men (67). The group with the second highest AHI was also collectively the youngest, but they also had the highest rate of smoking and respiratory disorders.

The link between obesity and OSA has been well documented(68). It is a complicated interaction in which obesity can lead to the development of sleep apnea, which leads to increasing severity of sleep apnea. Indeed, the two clusters with 100% sleep apnea diagnosed subjects also had the highest mean BMI.

Hypertension is a well-known comorbid condition with OSA(68). It is itself associated with diabetes, COPD, and cardiovascular disease. While in the most severe OSA phenotype, there were low proportions of concomitant disorders, in moderate OSA, when hypertension was also present, there were the highest proportion of comorbid disorders.

Chapter 5: Discussion of findings

In general, our study seems to confirm the results of the Vavougious *et al.* study(16) and addresses some of the shortcomings of that paper. Namely, it was centered around only a single center, whereas the SHHS sourced its participants from several centers and was more than four times bigger. In addition, the SHHS was more balanced with regards to gender.

In addition, a newer study by Lacedonia *et al*. found that the main differences in their three clusters of sleep apnea patients was caused by differences in AHI, BMI, and ESS(17). They theorized that that a more moderate form of sleep apnea could be comorbid with a disease that is far more threatening to health. This seems to be the case in our cluster 4, which has the third highest AHI, but very high proportion of cardiovascular disease and diabetes.

Ye *et al*. published the first paper to attempt a cluster analysis approach for subtyping OSA(18). Concerns about that study were the fact that it was only on a small homogeneous population, they did not perform dimensionality reduction on the survey data, and that they lumped cardiovascular disease into one variable.

We still believe semantic clustering could be done, if further work was done to incorporate different severity sleep apnea phenotypes into the HPO. The stratification that came from the respiratory index factor was key, and that information was excluded in the semantic analysis.

The reason for the exclusion was that we hypothesized that AHI was 1) not a sufficient classifier for sleep apnea subtypes, and 2) it would be possible to construct better classifiers from the component variables of AHI. Hypothesis 1 turned out to be correct: other factors are important in describing a full sleep apnea phenotype. However, while there are indices and measurements that are slightly better predictors of OSA severity, they are highly correlated with AHI. The widespread clinical use of AHI makes it easy to understand.

Additionally, the findings from the traditional clustering make the case that moderate AHI and severe AHI are distinct phenotypes that interact with other phenotypes differently, and thus should be added to the HPO.

Chapter 6: Summary and conclusions

Different OSA phenotypes may respond differently to treatment, thus understanding and better characterizing the subgroups is expected to allow us to assign more effective precision treatment to the right patients. In terms of our results, patients in category 4 are at much higher risk for heart attacks compared to others. Therefore, these patients' cardiovascular health should be monitored more closely.

Viewing a disease as a collection of individual but interacting phenotypes may be the way diagnosis is headed in the future. It may be that we do not need to categorize someone with moderately severe OSA, but rather they have "moderately high AHI phenotype", "increased diastolic blood pressure phenotype", "coronary disease phenotype", etc.

In addition, incorporating sleep data into the HPO enhances it as a resource for future users. The code we wrote to work with the SHHS dataset is an open resource for users. We also demonstrated the power of using shared data from several centers, showing that we can extend the analysis to a bigger and more diverse data set.

There are sleep components in rare phenotypes which are heretofore unexplored and unquantified, and this work will help researchers discover new relationships between these diseases. Conversely, the additional phenotype data will potentially allow us to discover new aspects of OSA. The complex interaction of different stratifications of AHI with comorbid conditions is worth exploring further. OSA is a complex disease, and this approach can serve as a model for a pipeline to use public data to subtype complex diseases.

Acknowledgments

I would like to thank my advisor, Dr. Eilis Boudreau, for her constant support and advice. This thesis could not have been completed without her. Additionally, the suggestions and knowledge provided by the rest of my thesis committee, Dr. Guanming Wu and Dr. Melissa Haendel, was invaluable.

I would also like to thank the entire staff of the Department of Medical Informatics and Clinical Epidemiology office, especially Diane Doctor, who has guided me through my OHSU career, and Lynne Schwabe, who helped schedule the dates and deadlines.

To Prof. Haendel's Ontology Development Group, thank you for welcoming me warmly and teaching me about ontologies. Special thanks go to Matthew Brush and Kent Shefcheck for their pointed questions and help on OWLSIM.

I would like to acknowledge Dr. Shannon McWeeney and all the Bioinformatics and Computational Biology program professors who taught me. The skills I learned here will be useful for the rest of my career.

I would like to thank my BCB colleagues, especially my cohort. It was hard work, but I had fun with all of you.

Lastly, to M, seeing your smile heals me.

APPENDIX

The workflow for cluster analysis is as follows and corresponds to Figure 5 in the main text:

-Feature selection: We must decide which traits or features to perform clustering on. A balance must be struck between using as much information as possible while reducing computational load by eliminating unnecessary, redundant, low-information features.

-Clustering algorithm selection: The method of actually dividing the data points into clusters is selected in this step. Generally, each algorithm is defined by a proximity measure and clustering criterion. The proximity (or distance) measure is how similar two data points are. The distance between a point and itself must equal zero. The clustering criterion can be expressed as a cost function or some other kind of mathematical rule that defines how the data set is to be partitioned.

-Measure of clustering quality: Since unsupervised learning defines clusters are not known a priori, there has to be some kind of evaluation of the clustering. There are two types of validation. One is biological validation, where one uses another sample or set of samples to confirm the results, using an alternate method. We could apply our algorithm to a previous paper's data to compare our clustering results to theirs, but we do not have easy access to their data. Thus we are limited to computational validation. If we can make the case that our computational results hold true, then biological validation can be attempted by experimental scientists.

-Interpretation of the results: The final step is to analyze the clusters to see if they make biological sense. This will draw on our domain expertise in the field of sleep.

Clustering
1 Partitioning
The k-means algorithm(69) is a well-known, simple, clustering method. The general k-means clustering procedure is as follows:

1. Choose k cluster centers (centroids) inside the space containing the observation set, where k is the number of clusters desired.
2. Assign each observation to the closest centroid (based on Euclidean distance or squared error criterion).
3. Recompute the centroids using the current cluster memberships (i.e., the new centroids must be the center of the clusters).
4. If the convergence criterion is not met, repeat step 2. Typical convergence criteria are no reassignment of points to new cluster centers, or minimal decrease in squared error.

The squared error criterion is also known as the sum of the squared errors. The squared error for clustering L of a pattern set X containing K clusters is

$$e^2(X, L) = \sum_{j=1}^{K} \sum_{i=1}^{n_j} \left| x_i^{(j)} - c_j \right|^2$$

where c_j is the centroid of the jth cluster and x_{ij} is the ith pattern in the jth cluster. In essence, the squared error is the square of the Euclidean distance of points to the centroids of the clusters to which they belong.

The potential problems of k-means clustering include never reaching convergence, converging on a local minimum due to a poor initial partition (choice of centroids), and unusually shaped clusters. Domain-specific knowledge can help with choosing a good initial partition. Setting up a good convergence criterion can solve the second problem (e.g., if the algorithm is forever switching one pattern back and forth between a cluster, redefine what happens in case of a tie or terminate when only one pattern is switching back and forth).

Unusually-shaped clusters, such as elongated clusters, are not reproduced by k-means clustering because of the squared error/Euclidean distance criterion. Because it seeks to minimize the squared error, it tends to make round clusters.

K-means is an O(n) algorithm, which is very efficient.

Another partitioning method is Partitioning Around Medioids. The objective is to find a central data point (medioid) within each cluster. Initially, c number of medioids are chosen. Objects are then grouped with the medioid they are closest to (most similar to). Medoids are swapped with non-selected objects until all objects qualify as medioid. PAM is a quite computationally expensive algorithm, because it has to compare an object with the entire dataset.

2. Hierarchical

Within this category are two general types: agglomerative, where the clusters closest to each other are merged together at each step; and divisive, where clusters are split apart at each step.

In R, hierarchical clustering can be handled by `hclust()` or `agnes()` for agglomerative hierarchical clustering, and function `diana()` does divisive hierarchical clustering. The method used by Vavougios *et al.* was based on SPSS (IBM Analytics, Almaden, CA) Two-Step Clustering, a proprietary agglomerative hierarchical method which can combine both continuous and categorical variables and is good for large datasets.(16) It defines distance between two clusters as the log-likelihood decrease from combining them together. The first step of the cluster analysis is sorting the points into pre-clusters. This reduces the size of the matrix

that contains the distance between all pre-clusters. The next step is to combine the closest clusters.

The linkage is a function of dissimilarity between groups. Agglomerative hierarchical clustering starts with each object as its own cluster, then repeatedly merging the two groups that have the smallest dissimilarity, or linkage(70).

In single linkage (or nearest-neighbor linkage), the dissimilarity between clusters G and H is the smallest dissimilarity between a point in G and a point in H.

$$d_{single}(G,H) = \min_{i\in G, j\in H} d_{ij}$$

Therefore, for a cut at a certain height (e.g., 0.9), for each point Xi in a cluster, there will be a point Xj with dij<=0.9.

Complete linkage (or farthest-neighbor linkage) is when the dissimilarity between G and H is taken to be the largest dissimilarity between a point in G and a point in H.

$$d_{complete}(G,H) = \max_{i\in G, j\in H} d_{ij}$$

The cut interpretation is that if there is a point Xi in a certain cluster cut at height=0.9, then every other point Xj in that cluster will have dij<=0.9.

The average linkage between G and H is the average dissimilarity between all points in both groups.

$$d_{average}(G,H) = \frac{1}{n_G \cdot n_H} \sum_{i\in G, j\in H} d_{ij}$$

There is not a very good interpretation of cut height in average linkage.

Single, complete, and average linkage do not need to have dissimilarities in Euclidean space, which means the Gower distance matrix can be used. Agglomerative clustering with any of these linkages results in a dendrogram with no inversions, i.e., the height of a parent node will always be higher than its children nodes.

Single linkage tends to suffer from chaining, or adding one close point at a time. Therefore, clusters can be too spread out and not compact. Complete linkage suffers from the opposite problem, crowding. A point could be closer to points in other clusters than its own cluster. Clusters tend to be compact, but not separated enough. As might be expected, average linkage strikes a balance between the two, but has the problem of not being very interpretable.

There are many more linkage functions, as well as divisive clustering, which works from the top down instead of bottom up.

3. OWLSIM

A publicly available tool from Monarch is OWLSim, which is used to compare phenotype profiles(31, 32). The two primary uses are 1) disease knowledge base: that is, what phenotypes describe a disease, and 2) in a clinical setting, annotating each patient with a set of phenotypes. OWLSim can compare the patient phenotype profile (a set of phenotypes associated with a patient) to all the disease phenotype profiles in the HPO and output a list of diseases that are most similar. The structure of the HPO enables fuzzy matching when phenotypes are annotated at different levels of granularity.

We intended to use OWLSim in a novel way, to find clusters of patients that are similar to each other. Essentially, we will use OWLSim's semantic similarity score as the distance function. Pairwise comparisons of all patient profiles will be made to generate a similarity matrix between all possible pairs of data points.

OWLSim compares a list of HPO terms against another list of HPO terms. The comparison is then scored using either Information Content (IC) or Jaccard similarity (simJ) as metrics. SimJ is the ratio of shared attributes to total attributes:

$$simJ(p,q) = \frac{|a^p \cap a^q|}{|a^p \cup a^q|}$$

where a^p means the inferred attributes of phenotype p and a^q means the inferred attributes of phenotype q. Jaccard similarity is explained in Figure A.

The IC of a description is the negative logarithm of the ratio of the number of features annotated with that description over the total number of annotations.

$$IC(description) = -log_2(\frac{|annot_{description}|}{|annot|})$$

IC is calculated for the Least Common Subsuming (LCS) phenotype of the pairwise comparison, which is the most specific set of all shared attributes. The IC provides a measure of how unusual the set of attributes in common are. A match where the terms in common are rare will score higher than when the common terms are less specific (i.e., "stratum corneum of the epidermis" will score higher than "anatomical structure").

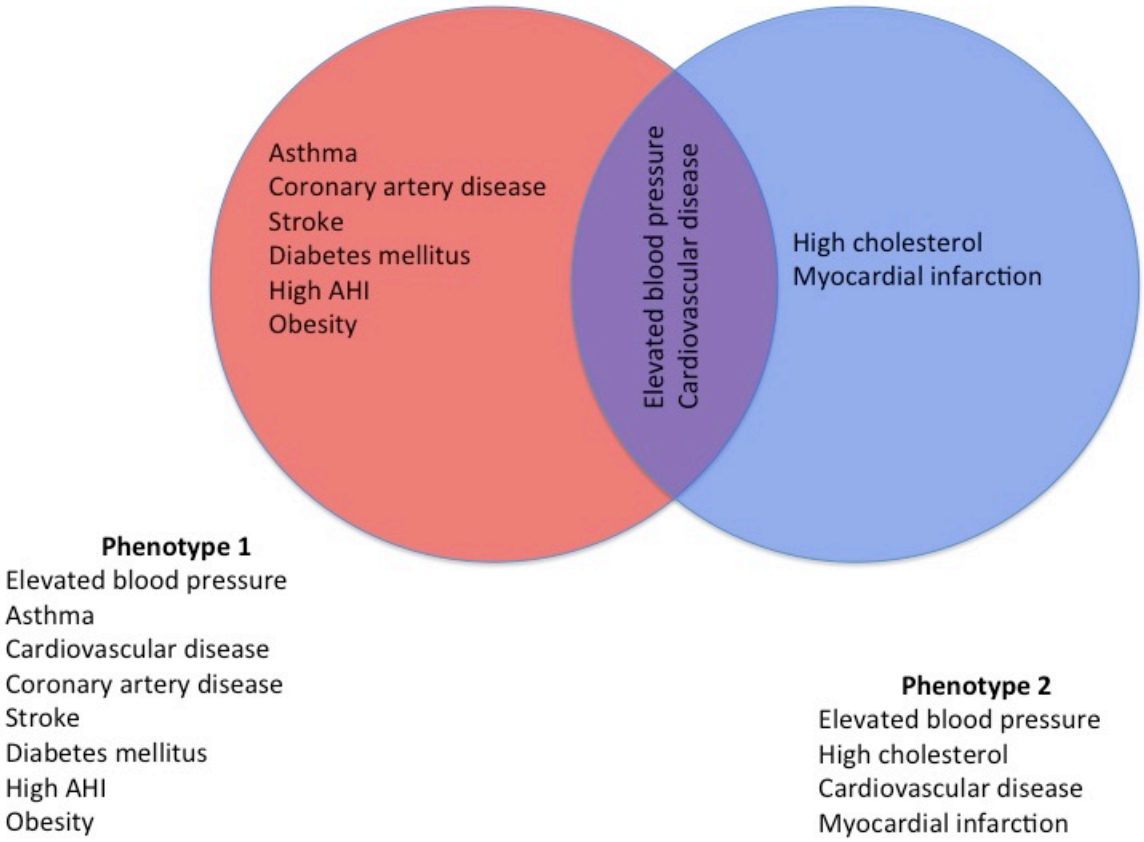


Figure A. How Jaccard similarity is calculated. In this example, the common terms, $a^p \cap a^q = 2$, while the total number of terms in the set, $a^p \cup a^q = 10$, therefore, simJ=0.2.

OWLSim outputs four metrics for each pairwise comparison into a JSON object:

avgIC = average IC score across all pairs
maxIC = maximum IC score across all the pairs
avgsimJ = average simJ score across all the pairs
maxsimJ = maximum simJ score across all the phenotype pairs

4. Other methods
Density-based algorithms such as DBSCAN think of clusters as dense regions of objects separated by regions of low density(71). The general idea is that for each point in a cluster, within a given radius must be a minimum number of points. DBSCAN is good at handling noisy outliers and discovering clusters of unusual shape (not round or ellipsoidal). Vavougios *et al.* used a variation of this method, clustering on data set values in order to determine groups of subjects in a correlation network, represented as a graph of nodes (subjects) and edges (connections between subjects)(16). Subjects that meet a minimum threshold of correlation are connected by edges.

Assorted tables and figures

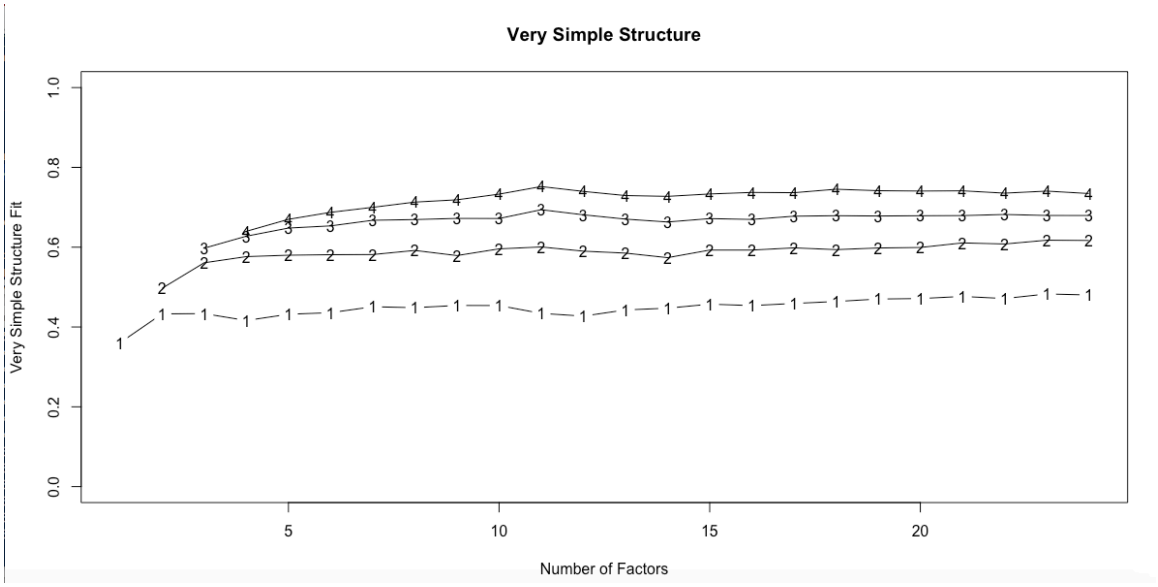
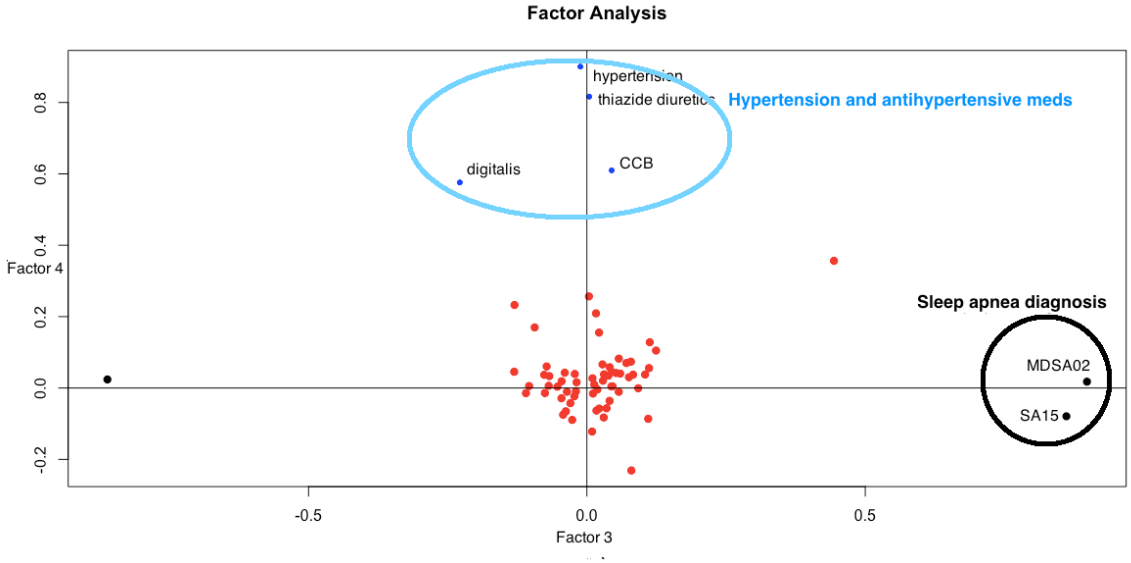


Figure B. VSS plot. With complexity of 1, the maximum VSS is 22 factors (we limited our analysis to up to 24 factors), but the output prompt suggested it is "probably more reasonable to think about 5 factors." Additionally, the Velicer MAP achieves a minimum of 0.015 at 4 factors.

Table 1. Factor analysis eigenvalues for SHHS1

	MR1	MR2	MR4	MR3	MR20	MR7	MR6	MR10	MR11	MR5	MR8
SS loadings	7.096	4.474	3.096	3.088	3.002	2.73	2.467	2.424	2.38	2.117	2.103
Proportion Var	0.091	0.057	0.04	0.04	0.038	0.035	0.032	0.031	0.031	0.027	0.027
Cumulative Var	0.091	0.148	0.188	0.228	0.266	0.301	0.333	0.364	0.394	0.421	0.448
	MR17	MR14	MR9	MR13	MR19	MR15	MR18	MR21	MR16	MR12	MR22
SS loadings	2.072	1.972	1.955	1.91	1.881	1.804	1.79	1.732	1.712	1.701	1.478
Proportion Var	0.027	0.025	0.025	0.024	0.024	0.023	0.023	0.022	0.022	0.022	0.019
Cumulative Var	0.475	0.5	0.525	0.55	0.574	0.597	0.62	0.642	0.664	0.686	0.705

Table A. Eigenvalues of the factor analysis.



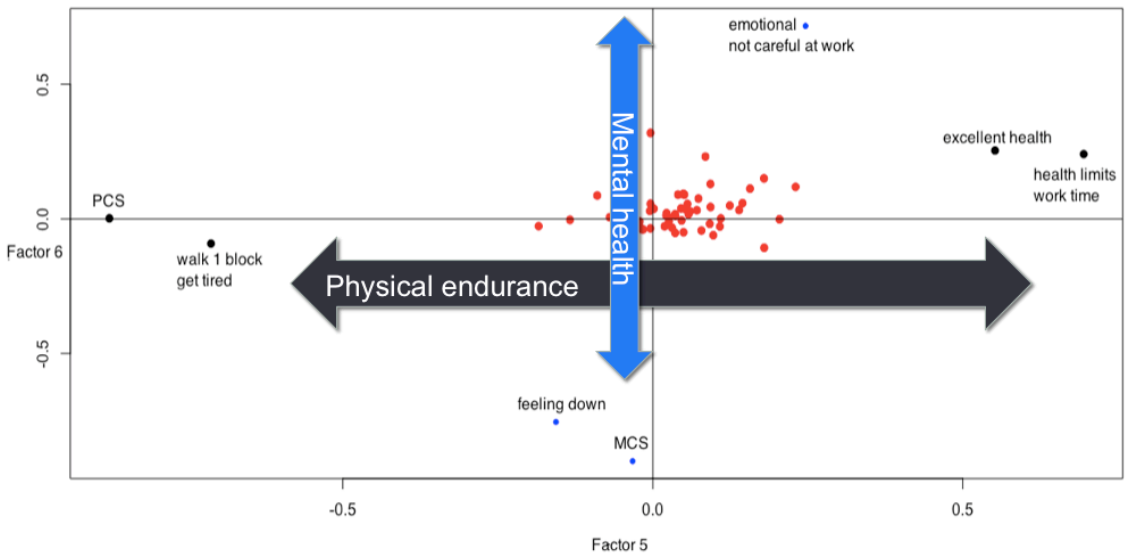


Figure C. (top) Factor analysis plot of factors 3 vs. 4 (hypertension, sleep apnea diagnosis). (bottom) Factor analysis plot of factors 5 vs. 6 (mental health vs. physical endurance). The red points are features that have loadings less than 0.5 on either axis. Blue dots are features that have loadings greater than 0.5 on the Y-axis, and black dots have loadings greater than 0.5 on the X-axis. CCB=calcium channel blocker, MDSA02 = medical diagnosis of sleep apnea, SA15 = sleep apnea (presence), MCS = SF-36 mental component score, PCS = SF36 physical component score.

Table B. Factor loadings for final dimensionality reduction. MR=minimum residual, which is the particular criterion for this factor analysis (maximum likelihood is an alternate criterion). The factor columns are arranged in order of largest to smallest eigenvalue.

Loadings:

	MR2	MR1	MR3	MR4	MR6	MR10	MR14	MR16	MR5	MR13
age_s1	0.349	0.263	-0.195	-0.527	0.160	0.156			0.121	
gender	0.334	0.265	0.228	0.647		0.290			0.122	0.233
bmi_s1	-0.111	0.179	0.208					0.198		-0.144
avsao2nh		-0.198	-0.138			-0.201		-0.759		
mcs_s1										
pcs_s1	-0.209	-0.106		0.325	-0.138		-0.196			
ai_nrem		0.737							0.202	
ai_rem		0.604		0.153			0.164			
arrembp		0.152								
fvc			0.188	0.801	-0.197		-0.207			0.111
chol		-0.132		-0.130				0.193		-0.128
ace1					0.794					
alpha1	0.146	0.136	0.100		0.120					
hctzk1				-0.186	0.712		-0.242			-0.116
sympth1	0.158				-0.198		0.890			
ohga1	0.207		-0.122	-0.253	0.113		-0.103	0.104		
ntca1		-0.148					0.200		0.144	0.125
thry1		-0.106	0.112	-0.594					-0.163	
nsaid1	-0.106									
ca15	0.864				0.120					
asthma15							0.863			
pacem15	0.249		0.291		0.249			-0.100		
htnderv_s1	0.307	0.102		-0.118	0.897			0.129		
runny15			0.259				0.155			
cgpkyr						0.910				
cabg15	0.832									
alcoh				0.338	0.108	0.447				
sa15			0.909							
othrcs15	0.711				0.119	0.110				
wrface10										
ltdp10									-0.104	

nitro15	0.789							0.115		
evsmok15			0.111			0.927				
carful25										
loudsn02		0.143	0.437			0.138				
sob02	0.150		0.234				0.400			
hrswd02				-0.139			-0.124		0.238	0.102
wu2em02										
tfa02										
sleepy02			0.381							
mdsa02			0.912							
surgtr02	-0.544		0.285	0.165	-0.152		-0.297	0.118	0.113	0.151
time_bed		-0.189	0.142	-0.255		-0.134			0.578	0.432
nsupinep										
pcstahda		0.833	0.166					0.322		
pctsa75h		0.116		0.116		-0.119		0.684		
tmstg2p		0.305		0.183		0.105				0.768
remlaip										
pctsa90h		0.257				0.116		0.779		
slpeffp	-0.106	-0.238							-0.255	0.188
hremt1p		0.197						-0.102		0.103
hremt2p				-0.139					-0.490	0.151
hremt34p		0.138	-0.193						-0.141	
slplatp										
tmremp		-0.309							0.120	-0.262
pcstah3d		0.770	0.184					0.378		
scstg1p		0.121		0.125					0.747	
scstg2p									0.147	0.920
slptawp		0.420							0.705	

	MR9	MR12	MR15	MR7	MR11	MR8	MR20	MR18	MR19	MR21
age_s1		-0.122					-0.196	-0.121		0.190
gender	-0.103		-0.132				-0.218			
bmi_s1							-0.139	0.577		-0.241
avsao2nh										
mcs_s1	-0.110	-0.861	-0.102							
pcs_s1	-0.116					-0.142		-0.507		-0.112
ai_nrem			0.155						0.127	
ai_rem				0.168	0.267					
arrembp			-0.112		0.900					
fvc										
chol			0.177		0.107					
ace1		0.119				0.145	-0.233			0.122
alpha1				0.858			-0.171			
hctzk1	0.112		0.116	-0.165		-0.113	0.265			-0.240
sympth1										
ohga1		0.159	-0.157				-0.601	0.184		
ntca1	0.103	0.362	0.587			-0.117	0.146		-0.235	
thry1		0.133	0.110			0.127				
nsaid1								0.773	0.107	
ca15										
asthma15										
pacem15	0.172			-0.809		0.129	-0.122			
htnderv_s1										
runny15							0.132			0.717
cgpkyr										
cabg15							-0.223	-0.116		
alcoh	-0.100					0.127			0.138	
sa15									0.129	0.194
othrcs15		-0.105		-0.253		0.148	0.110			-0.297
wrface10									0.759	
ltdp10	-0.231			0.114					-0.673	
nitro15	0.156	0.105		0.184				0.147		0.236
evsmok15										
carful25		0.827								
loudsn02									-0.132	-0.109
sob02	0.183	0.154						0.170	0.114	
hrswd02	-0.638						0.143			
wu2em02	0.758									
tfa02	0.729		0.126						0.112	
sleepy02	0.364	0.335	0.119							
mdsa02									0.139	0.139
surgtr02				0.285			-0.166	0.137		0.415
time_bed						0.188	0.228			-0.156
nsupinep					-0.860					-0.109

pcstahda										
pctsa75h										0.141
tmstg2p			0.234				-0.182			
remlaip			0.764							
pctsa90h										
slpeffp			-0.196			-0.690	0.143		-0.168	
hremt1p			-0.143		-0.116					
hremt2p			-0.364				0.217			-0.158
hremt34p			-0.108	-0.121			0.645	0.112		0.185
slplatp						0.836				
tmremp			-0.645		0.111	-0.104			-0.116	-0.162
pcstah3d										
scstg1p							-0.101			
scstg2p						-0.203				
slptawp										

	MR17
age_s1	
gender	-0.123
bmi_s1	
avsao2nh	
mcs_s1	
pcs_s1	
ai_nrem	
ai_rem	0.313
arrembp	0.115
fvc	
chol	0.408
ace1	
alpha1	
hctzk1	
sympth1	
ohga1	-0.135
ntca1	
thry1	0.126
nsaid1	
ca15	
asthma15	-0.148
pacem15	
htnderv_s1	
runny15	
cgpkyr	
cabg15	
alcoh	0.133
sa15	
othrcs15	
wrface10	
ltdp10	
nitro15	
evsmok15	
carful25	
loudsn02	
sob02	
hrswd02	0.169
wu2em02	
tfa02	
sleepy02	
mdsa02	
surgtr02	-0.188
time_bed	
nsupinep	0.147
pcstahda	
pctsa75h	
tmstg2p	
remlaip	
pctsa90h	
slpeffp	0.175
hremt1p	0.753
hremt2p	
hremt34p	-0.145
slplatp	
tmremp	
pcstah3d	
scstg1p	

```
scstg2p
slptawp    -0.165
```

Bibliography

1. Eckert DJ, White DP, Jordan AS, Malhotra A, Wellman A. Defining Phenotypic Causes of Obstructive Sleep Apnea: Identification of Novel Therapeutic Targets. Am J Respir Crit Care Med. 2013;188(8):996-1004.
2. Balachandran JS, Patel SR. Obstructive Sleep Apnea. Ann Intern Med. 2014;161(9):ITC1.
3. Edwards BA, Wellman A, Sands SA, Owens RL, Eckert DJ, White DP, et al. Obstructive Sleep Apnea in Older Adults is a Distinctly Different Physiological Phenotype. Sleep. 2014;37(7):1227-36A.
4. Wang Q, Zhang C, Jia P, Zhang J, Feng L, Wei S, et al. The association between the phenotype of excessive daytime sleepiness and blood pressure in patients with obstructive sleep apnea-hypopnea syndrome. International Journal of Medical Sciences. 2014;11(7):713-20.
5. BaHammam AS, Obeidat A, Barataman K, Bahammam SA, Olaish AH, Sharif MM. A comparison between the AASM 2012 and 2007 definitions for detecting hypopnea. Sleep Breath. 2014;18:767-73.
6. Teran-Santos J, Jimenez-Gomez A, Cordero-Guevara J. The association between sleep apnea and the risk of traffic accidents. New England Journal of Medicine. 1999;340(11):847-51.
7. Sforza E, Pichot V, Saint Martin M, Barthelemy JC, Roche F. Prevalence and determinants of subjective sleepiness in healthy elderly with unrecognized sleep apnea. Sleep Medicine. 2015;16:981-6.
8. Joosten SA, Hamza K, Sands S, Turton A, Berger P, Hamilton G. Phenotypes of patients with mild to moderate obstructive sleep apnea as confirmed by cluster analysis. Respirology. 2012;17:99-107.
9. Gabbay IE, Lavie P. Age- and gender-related characteristics of obstructive sleep apnea. Sleep Breath. 2011;16:453-60.
10. Luyster FS, Buysse DJ, Strollo PJ. Comorbid Insomnia and Obstructive Sleep Apnea: Challenges for Clinical Practice and Research. Journal of Clinical Sleep Medicine. 2010;6(2):196-204.
11. Kim KT, Cho YW, Kim DE, Hwang SH, Song ML, Motamedi GK. Two subtypes of positional obstructive sleep apnea: Supine-predominant and supine-isolated. Clinical Neurophysiology. 2016;127:565-70.
12. Xu R, Wunsch DC, II. Clustering Algorithms in Biomedical Research: A Review. IEEE Reviews In Biomedical Engineering. 2010;3:120-54.
13. Yu Z, Chen H, You J, Liu J, Wong H, Han G, et al. Adaptive Fuzzy Consensus Clustering Framework for Clustering Analysis of Cancer. IEEE/ACM Trans Comput Biol Bioinform. 2015;12(4):887-901.

14. Qian S, Guo W, Xing J, Qin Q, Ding Z, Chen F, et al. Diversity of HIV/AIDS epidemic in China: a result from hierarchical clustering analysis and spatial autocorrelation analysis. AIDS. 2014;28(12):1805-13.
15. Wolf A, Kirschner K. Principal component and clustering analysis on molecular dynamics data of the ribosomal L11-23S subdomain. J Mol Model. 2013;19(2):539-49.
16. Vavougios GD, Natsios G, Pastaka C, Zarogiannis SG, Gourgoulianis KI. Phenotypes of comorbidity in OSAS patients: combining categorical principal component analysis with cluster analysis. J Sleep Res. 2016;25:31-8.
17. Lacedonia D, Carpagnano GE, Sabato R, Lo Storto MM, Palmiotti GA, Capozzi V, et al. Characterization of obstructive sleep apnea-hypopnea syndrome population by means of cluster analysis. J Sleep Res. 2016.
18. Ye L, Pien GW, Ratcliffe SJ, Bjornsdottir E, Arnardottir ES, Pack AI, et al. The different clinical faces of obstructive sleep apnoea: a cluster analysis. Eur Respir J. 2014;44:1600-7.
19. National Sleep Research Resource 2017 [Available from: https://sleepdata.org.
20. Dean Dn, Goldberger A, Mueller R, Kim M, Rueschman M, Mobley D, et al. Scaling Up Scientific Discovery in Sleep Medicine: The National Sleep Research Resource. Sleep. 2016;39(5):1151-64.
21. Grinnon S, Miller K, Marler J, Lu Y, Stout A, Odenkirchen J, et al. National Institute of Neurological Disorders and Stroke Common Data Element Project - approach and methods. Clin Trials. 2012;9(3):322-9.
22. Hamilton C, Strader L, Pratt J, Malese D, Hendershot T, Kwok R, et al. The PhenX Toolkit: get the most from your measures. Am J Epidemiol. 2011;174(3):253-60.
23. Pesquita C, Faria D, Falcao AO, Lord P, Couto FM. Semantic Similarity in Biomedical Ontologies. PLoS Computational Biology. 2009;5(7):1-12.
24. Schuurman N, Leszszynski A. Ontologies for Bioinformatics. Bioinformatics and Biology Insights. 2008;2:187-200.
25. Gruber TR. A Translation Approach to Portable Ontology Specification. Knowledge Acquisition. 1993;5:199-220.
26. Noy NF, McGuinness DL. Ontology Development 101: A Guide to Creating Your First Ontology http://protege.stanford.edu/publications/ontology_development/ontology101-noy-mcguinness.html [
27. Kohler S, Doelken SC, Mungall CJ, Bauer S, Firth HV, Bailleul-Forestier I, et al. The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. Nucleic Acids Res. 2014;42(Database issue):D966-74.
28. Vasilevsky N, Engelstad M, Foster E, McMurry J, Mungall CJ, Robinson PN, et al. Monarch Initiative [Internet]. http://monarch-initiative.blogspot.com/2016/03/finally-medical-terminology-that.html2016. [cited 2016].
29. Quan SF, Howard BV, Iber C, Kiley JP, Nieto FJ, O'Connor GT, et al. The Sleep Heart Health Study: Design, Rationale, and Methods. Sleep. 1997;20(12):1077-85.

30. Robinson P, Kohler S, Bauer S, Seelow D, Horn D, Mundlos S. The Human Phenotype Ontology: A Tool for Annotating and Analyzing Human Hereditary Disease. American Journal of Human Genetics. 2008;83:610-5.
31. Chen C-K, Mungall CJ, Gkoutos GV, Doelken SC, Kohler S, Ruef B, et al. MouseFinder: candidate disease genes from mouse phenotype data. Hum Mutat. 2012;33(5):858-66.
32. Washington NL, Haendel MA, Mungall CJ, Ashburner M, Westerfield M, Lewis SE. Linking Human Diseases to Animal Models Using Ontology-based Phenotype Annotation. PLoS Biology. 2009;7(11):1-20.
33. Kohler S, Schulz MH, Krawitz P, Bauer S, Doelken SC, Ott CE, et al. Clinical Diagnostics in human genetics with semantic similarity. Am J Human Genetics. 2009;85:457-64.
34. Hwang T, Atluri G, Xie M, Dey S, Hong C, Kumar V, et al. Co-clustering phenome-genome for phenotype classification and disease gene discovery. Nucleic Acids Res. 2012;40(19):e146.
35. Westbury SK, Turro E, Greene D, Lentaigne C, Kelly AM, Bariana TK, et al. Human phenotype ontology annotation and cluster analysis to unravel genetic defects in 707 cases with unexplained bleeding and platelet disorders. Genome Med. 2015;7(1):36.
36. Wang C, Zimmerman MT, Prodduturi N, Chute CG, Jiang G. Adverse drug event-based stratification of tumor mutations: a case study of breast cancer patients receiving aromatase inhibitors. AMIA Annu Symp Proc. 2014;2014:1160-9.
37. Kibbe WA, Arze C, Felix V, Mitraka E, Bolton E, Fu G, et al. Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. Nucleic Acids Res. 2015;43(Database issue):D1071-8.
38. Team RC. R: A language and environment for statistical computing. 3.3 ed. Vienna, Austria: R Foundation for Statistical Computing; 2017.
39. Wickham H, Francois R. dplyr: A Grammar of Data Manipulation. R package version 0.5.0. ed2016.
40. Jih J, Mukherjea A, Vittinghoff E, Nguyen T, Tsoh J, Fukuoka Y, et al. Using appropriate body mass index cut points for overweight and obesity among Asian Americans. Prev Med. 2014;65:1-6.
41. Bentley-Lewis R, Powe C, Ankers E, Wenger J, Ecker J, Thadhani R. Effect of race/ethnicity on hypertension risk subsequent to gestational diabetes. Am J Cardiol. 2014;113(8):1364-70.
42. Pang C, Sollie A, Sijtsma A, Hendriksen D, Charbon B, de Haan M, et al. SORTA: a system for ontology-based re-coding and technical annotation of biomedical phenotype data. Database (Oxford). 2015;2015.
43. Levartovsky A, Dafna E, Zigel Y, Tarasiuk A. Breathing and Snoring Sound Characteristics during Sleep. Journal of Clinical Sleep Medicine. 2016;12(3):375-84.
44. Hoffstein V. Snoring. Chest. 1996;109:201-22.
45. Revelle W. psych: Procedures for Personality and Psychological Research. 1.73 ed. Evanston, Ill.: Northwestern University; 2017. p. R package.
46. Kolenikov S, Angeles G. The use of discrete data in PCA: Theory, simulations, and applications to socioeconomic indices. Measure Evaluation. 2004.

47. Haslbeck J, Waldorp LJ. mgm: Estimating Time-Varying Mixed Graphical Models in High-Dimensional Data2015.
48. Uebersax JS. Introduction to the Tetrachoric and Polychoric Correlation Coefficients, Statistical Methods for Rater Agreement. 2006 [updated 2015. Available from: http://www.john-uebersax.com/stat/tetra.htm.
49. Harris B. Tetrachoric correlation coefficient. In: Kotz L, Johnson N, editors. Encyclopedia of Statistical Sciences. New York: Wiley; 1988. p. 223-5.
50. Drasgow F. Polychoric and polyserial correlations. In: Kotz L, Johnson N, editors. Encyclopedia of Statistical Sciences. New York: Wiley; 1988. p. 69-74.
51. Costello AB, Osborne JW. Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis. Practical Assessment, Research & Evaluation. 2005;10(7).
52. Gorsuch RL. Factor Analysis. 2nd ed: Lawrence Erlbaum Associates; 1983. 425 p.
53. Courtney MGR. Determining the number of factors to retain in EFA: using the SPSS R-Menu v2.0 to make more judicious estimations. Practical Assessment, Research & Evaluation. 2013;18(8):1-13.
54. Horn JL. A rationale and test for the number of factors in factor analysis. Psychometrika. 1965;30:179-85.
55. Catell RB. The scree test for the number of factors. Multivariate Behav Res. 1966;1:245-76.
56. Velicer WF. Determining the number of components from the matrix of partial correlations. Psychometrika. 1976;41:321-7.
57. Brock G, Pihur V, Datta S, Datta S. clValid: An R package for cluster validation. Journal of Statistical Software. 2008;25(4).
58. Rousseeuw P. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics. 1987;20:53-65.
59. Dunn J. A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. J Cybernetics. 1973;3(3):32-57.
60. Brock G. clValid. 0.6-6 ed: CRAN. p. Validation of Clustering Results.
61. Laderas T. Consense. 2015.
62. Laderas T, McWeeney S. A consensus framework for clustering microarray data. OMICS: A Journal of Integrative Biology. 2007:116-28.
63. Fox J. polycor. 0.7-9 ed: CRAN.
64. Rousseeuw P, Struyf A, Hubert M. cluster: "Finding Groups in Data": Cluster Analysis Extended. 2.06 ed: CRAN.
65. van den Hoven J. Clustering with optimised weights for Gower's metric. Netherlands: University of Amsterdam; 2015.
66. Dimitriadou E, Dolnicar S, Weingessel A. An examination of indexes for determining the number of cluster in binary data sets. Psychometrika. 2002;67(1):137-60.
67. Bixler EO, Vgontzas AN, Ten Have T, Tyson K, Kales A. Effects of age on sleep apnea in men. Am J Respir Crit Care Med. 1988;157:144-8.
68. Wolk R, Shamsuzzaman ASM, Somers VK. Obesity, Sleep apnea, and hypertension. Hypertension. 2003;42(6):1067-74.

69. Jain AK, Murty MN, Flynn PJ. Data Clustering: A Review. ACM Comput Surv. 1999;31(3):264-323.
70. Everett B. Cluster Analysis. 2nd ed: Heinemann; 1980.
71. Ester M, Kriegel H-P, Sander J, Xu X, editors. A density-based algorithm for discovering clusters in large spatial databases with noise. Second International Conference on Knowledge Discovery and Data Mining; 1996: AAAI Press.