Published in final edited form as:

Sleep Med Clin. 2016 June ; 11(2): 241–255. doi:10.1016/j.jsmc.2016.01.009.

The Role Of Big Data In The Management Of Sleep Disordered Breathing

Rohit Budhiraja, MD¹, Robert Thomas, M.D., M.M.Sc.², Matthew Kim, M.D.³, and Susan Redline, MD, MPH⁴

¹Department of Medicine, Division of Sleep and Circadian Disorders and Division of Pulmonary and Critical Care Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA

²Department of Medicine, Division of Pulmonary, Critical Care & Sleep Medicine, Beth Israel Deaconess Medical Center, 330 Brookline Avenue, Boston, MA

³Department of Medicine, Division of Endocrinology, Diabetes and Hypertension, Brigham and Women's Hospital, Boston, MA

⁴Brigham and Women's Hospital and Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA

Introduction

Despite considerable investments in health care, our health care systems often fall short in meeting quality metrics, patient satisfaction, and clinical outcomes goals.¹ Furthermore, despite large investments in clinical research, our research infrastructure often does not yield the needed data to support “evidence-based” health care. Deficiencies are partially due to inadequate representation of key stakeholders, including patients, in the design of health care research, as well as due to the overall paucity of valid and comprehensive data on sufficient numbers of patients. These gaps are especially evident in sleep medicine, where data are lacking that address questions such as which patients benefit from alternative treatments (e.g., CPAP, surgery, oral appliances, oxygen supplementation); the basis for inter-individual variation in cardiometabolic and cognitive effects of sleep apnea; and the individual determinants of CPAP adherence or residual sleepiness after CPAP treatment. These gaps are due to limitations of available data, which too often are inadequate to generate robust evidence for supporting health care decisions. There is simultaneously a richness of data (e.g., compliance tracking) and data waste (e.g., focusing on insurance payments requirements while discarding other available data). The traditional practice of generating data has been compared to working in silos. Data silos include those such as electronic medical records of clinical encounters, sleep laboratory databases, online positive airway pressure tracking databases that reside in independent non-communicating areas.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

While the strength of evidence is dependent on the validity and reliability of the data utilized to generate that evidence, with all other things being similar, a larger ‘n’ yields comparatively superior evidence with higher statistical power and lower likelihood of erroneous inferences. Small sized studies may be difficult to reproduce. Indeed, it has been suggested that a majority of published studies cannot be replicated.² The current healthcare model uses data from studies conducted on hundreds or possibly thousands of people to determine the recommendations for millions of people. Information is typically generated based on data from one or few clinical centers and is often limited in diversity of patient representation, environmental backgrounds, and practice variation. The understudied and underserved populations also are at a risk of inappropriate therapy based on using data from studies performed on majority cohorts. The use of data from single or a few centers also does not provide the statistical power to provide evidentiary information about individual variation in clinical response or risk specific to given individuals. Rather, inferences often only address “average” effects (in the sampled population) and the datasets are underpowered to detect differences that may be relevant to smaller subgroups of individuals.

The widespread use of electronic health record systems and advances in computational power and analytical methods provide new opportunities to amass data of unprecedented size, heterogeneity and complexity and analyze these with new analytical methods, thus overcoming the gaps described above. “Big data” refers to voluminous databases that allow exploration of the effects of individual differences and complex interactions with the goal of facilitating identification and management of individuals according to their unique characteristics (i.e. “personalizing” medicine or pursuing “precision” medicine). Big data can be defined by 3 ‘V’s - volume, velocity and variety. The ‘Volume’ or the size of the data underlines the appeal and the value of the database. Generally, the bigger the size of the dataset, the higher the statistical power for analyses and ability to drill down to analysis of subgroups and to reduce the likelihood of making spurious inferences. ‘Velocity’ is the rate at which the data are generated and analyzed. Modern devices such as smartphones generate data almost in real-time, which can lead to remarkably quick discoveries, quantifying dynamic changes in a population’s health, such as improving recognition of newly emergent food borne or infectious disease outbreaks. Finally, ‘Variety’ refers to different sources, types and formats of data. They may be obtained from social media, disease-oriented websites (such as MyApnea.org³), smartphones or medical records, or a combination of all these. Thus, in considering “big data,” one needs to consider the sample size of available data, the “dimensionality” of those data, the access to those data, and the analytical and computational resources used to maximize information from the data.

A major source of big data is from the electronic health record (EHR), which has the power to transform each patient encounter as a datum in a potentially massive longitudinal database. There are efforts to expand the EHR to contain more comprehensive and “searchable” fields of routinely collected clinical data, as well as to incorporate and link data from external sources, such as from validated patient surveys, health devices, and even genomic data. The ability to search multiple well-annotated medical notes, laboratory values, imaging studies, biomarkers, and data regarding progression of disease offers the prospect of utilizing knowledge from massive interconnected databases from large numbers of patients to guide therapy to an individual patient, improve public health and enhance

quality of care in general. Recognizing this potential, in addition to the challenges of accessing and integrating sleep data into the EHR, the American Academy of Sleep Medicine organized a Task Force to begin the process of specifying EHR needs in sleep medicine (Hwang, *in press*; see chapter x).

The widespread collection of data during routine daily activities through mobile phones, wearables, and other devices provides opportunities to non-intrusively extract real time data from multiple sources, integrating these with research and clinical datasets. These data are of high relevance to sleep medicine, where measures of sleep patterns, snoring, and PAP adherence are available through mobile devices.

“Big Data” has the potential to fast track scientific advancement in medicine by linking the ‘omics data (genomics, metabolomics, proteomics, expression, etc.), molecular biology, clinical data, population data and mobile data, among others, to understand the diverse aspects of disorders and to facilitate personalized management. These goals are tightly linked to establishing “precision medicine” as a goal for health care of the future. Supported by the National Institutes of Health (NIH), the U.S. Veteran’s Administration and other national and international organizations, Precision Medicine initiatives support activities leading to the use of multiple sources of patient-specific information-genomic, clinical, behavioral, environmental, and patient-preference- to identify and apply customized healthcare decisions, practices and products that meet individual patient needs. Moving away from models of care that focus on a population’s “average response” but often result in suboptimal treatment for many, precision medicine seeks to use “big data” to identify which treatments result in favorable responses, optimize safety and reduce harm, and improve long-term health outcomes for *specific* individuals. Although still nascent, notable progress has been made in tailoring therapies for patients with cancer, thrombotic diseases, immunological disorders, among others, and these examples provide models for sleep medicine.

It is useful to conceptualize levels or strata of Big Data in sleep sciences and sleep medicine. These include: 1) Population level information from multiple representative samples across the world comprised of individuals are not usually specifically selected for disease; 2) Disease-specific populations, comprised of data on patients ascertained on the basis of diseases (e.g., narcolepsy, sleep apnea, insomnia); 3) Individual level information – now increasingly enabled through Point-Of-Care longitudinal data tracking, and including information such as body weight, blood glucose, blood pressure, and from wearable devices.

The following sections discuss the potential of big data in providing insights into the pathophysiology, presentation and management of sleep-disordered breathing (SDB) and for approaches for Sleep Precision Medicine.

Etiology and Genetic predisposition

Developing pharmacological interventions for sleep apnea has been limited by insufficient understanding of its specific etiology and limitations of current phenotyping strategies. The pathogenesis of sleep apnea likely involves complex interactions between genetic factors

that influence anatomy and pharyngeal neuromuscular control, sleep fragmentation propensity, respiratory chemoreflex activation, and environmental exposures. For over 20 years it has been known that sleep apnea aggregates within families and that the Apnea Hypopnea Index (AHI) has significant heritability,^{4,5} defined as the portion of the trait variance attributable to familial, and typically genetic, influences. Although obesity is a significant risk factor for sleep apnea, approximately 60% of the genetic basis for sleep apnea is through body mass index-independent pathways,⁶ suggesting the potential for novel intervention targets. Early studies of the molecular genetics of sleep apnea were limited to modest-sized association studies of candidate genes which identified a number of genetic variants to be associated with the apnea hypopnea index (AHI),^{7,8} but results largely were not replicated in independent samples. Recent efforts to aggregate larger amounts of sleep phenotype and genotype data from NIH-funded cohorts have identified additional putative loci (Cade; in press), which also await independent replication. Even the latter efforts, representing the world's largest sleep apnea genetic data base and interactions among international centers (International Sleep Genetic Epidemiology Consortium; sleepgenetics.org), has analyzed genotype and phenotype data for less than 25,000 individuals. Although this sample is large for clinical trials or epidemiological studies, it is estimated to be only one-tenth the size needed to identify rare genetic variants. Furthermore, given that genetic architecture varies by ancestry and analysis of multiethnic samples can accelerate variant discovery, newer approaches are needed to achieve both the sample sizes as well as the racial/ethnic diversity to identify and validate functional variants. Due to the rapid reduction in genotyping costs, lack of genotypes is no longer a barrier to genetic discovery. Rather, there is a two-fold "big data" need: a) to link sleep apnea cases from the EHR to genetic samples; and 2) to identify sleep phenotypes that are both scalable across large samples and informative. The former need may be addressed through NIH and VA investments in initiatives such as the Million Veteran Program,⁹ eMerge (Electronic Medical Records and Genomics),¹⁰ and NIH Precision Medicine Cohorts, which each program aiming to deposit genotype data within the EHR. The challenge for Sleep Medicine is to ensure that the appropriate clinical data on sleep disorders are captured and are extractable within the EHR so that they can be linked to these genetic resources. The second need is to improve the sleep phenotype data within research cohorts with genetic data, including efforts to capture relevant and easily measured sleep phenotypes on large numbers of individuals. This may require moving beyond the AHI as the key metric for sleep apnea genetics. In addition to not providing information on attributes of sleep apnea that may reflect intermediate pathways, measurement of the AHI can be burdensome. Furthermore, recent research indicates that alternative sleep apnea measurements, such as the average apnea hypopnea duration and average nocturnal oxygen saturation have higher heritability than the AHI, with even greater heritability achieved by combining information from multiple sleep apnea traits as principle components (Wang, in press). Therefore, big data to promote the discovery of sleep apnea susceptibility genes and related molecular pathways require further integration of clinical sleep data within the EHR as well as collection of informative and scalable phenotypes in large research cohorts.

Genes likely influence not only the causation of sleep apnea, but also the complications thereof. While factors such as oxidative stress and endothelial dysfunction account for

pathogenesis of several complications of this disorder,¹¹ it is possible that a genetic susceptibility may be required for certain complications: a “two-hit” mechanism. For example, sleep apnea is associated with five times greater odds of risk of cognitive impairment in women with the apolipoprotein E (ApoE) epsilon4 allele, but no significant increase is encountered in women with sleep apnea lacking this allele.¹² Other studies confirm the worse cognitive function in APOE4-positive individuals with sleep apnea than those without this allele.^{13,14} It may be hypothesized that sleep apnea-related hypoxemia, inflammation and oxidative stress impacts brain function most severely in those with genetic susceptibility to cognitive disorders. A large repository of genetic and neurophysiologic markers of sleep apnea, can conceivably help predict the vulnerability to complications as well as the likelihood of benefit from sleep interventions.

Finally, there is shared heritability of sleep apnea with other traits, including obesity and diabetes. Big data efforts, which bring multiple phenotypes together can further identify shared genetic pathways as well as help identify additional predictors of health and disease. Identification of phenotypic and genetic markers will help understand the genetic basis and heritability of sleep apnea as well as related disorders, and might help identify novel therapeutic targets that address multiple disease pathways.

Diagnosis

Polysomnography (PSG) is the current gold standard for diagnosing sleep apnea, but suffers from several shortcomings. The current metric to quantify sleep apnea is the AHI. However, it is an imperfect predictor of symptoms and adverse health outcomes. Furthermore, apneas and hypopneas can be defined in many different ways (based on amplitude changes, desaturations and arousals), with limited data on the association between different definitions and symptoms and outcomes.¹⁵ Several different sensors are available for determining the magnitude of change in flow and effort, and may yield variable AHI values.^{15,16} Finally, phenomena like flow limitation and rapid eye movement (REM)-related sleep apnea may have symptomatic and outcome implications, but are not recognized by many health insurance providers due to lack of substantial data demonstrating their clinical significance.¹⁷

The traditional PSG interpretation does not routinely include quantitative analysis and reporting of potentially important metrics of physiological function, such as sleep assessment with electroencephalography (EEG) power spectral analyses and heart rate variability (HRV) analysis.¹⁸ Accumulation, aggregation and analysis of neurological and cardiorespiratory data could yield superior metrics for diagnosing the sleep apnea syndrome and improving the management, however a role for these more complex metrics requires systematic assessments of their independent prediction or ability to discriminate important subgroups of the population. While genetic and clinical specimen databases such as NHLBI's BioLINCC (Biologic Specimen and Data Repository Coordinating Center) allow linking biological and genetic specimen information with clinical data and promote research, there are few repositories for polysomnographic variables and sleep study data (see resources section).¹⁹ PhysioNet (physionet.org) is a web resource for several databases with diverse signals including ECG, continuous invasive blood pressure, respiration, oxygen

saturation, and EEG, interbeat (RR) interval databases, gait and balance databases, neuroelectric and myoelectric data and magnetic resonance angiography images.^{20,21} As described later, the National Sleep Research Resource (sleepdata.org) is a NIH-funded research repository that could be used to develop and test new quantitative sleep algorithms, which might then be applied to clinical records. To maximize the use of clinical PSG information, there needs to be further efforts at standardized not only the collection but the annotation and storage of the volumes of clinical PSGs performed each year. As discussed earlier, such data, when linked to other EHR data and potential deposits of genomic data, could provide fertile ground for developing better metrics for disease screening, diagnosis, and prediction.

Mobile fitness trackers promise another rich source of germane data. Sleep duration, physical activity and calorie intake are important factors that can lead to the genesis or progression of sleep apnea or the complications thereof. However, clinicians rarely query patients in details regarding these lifestyle variables during clinical encounters. Even when the patients are asked, their replies may not be sufficiently specific or may be limited by recall bias. The smartphones and wrist sensors now collect a bevy of health data, and the application designers continue to make the application software or ‘apps’ more sophisticated with increasingly valid and reliable data collection. Use of such data can provide objective information about factors including diet, physical activity and sleep duration, which can influence the severity of sleep apnea and symptoms including sleepiness. The apps can also help provide frequent visual feedback to the patient and promote healthy behaviors. They can significantly assist management of SDB, for education regarding these behaviors is an integral, albeit sometimes overlooked, part of therapy. With validation, these apps can play a role in screening for SDB. Wearable devices can also provide information on night-to-night variability of abnormality. Phenotyping of sleep quality during sleep apnea therapy and detecting persistence of periodic breathing and central apneas can be detected using wearable devices (e.g., M1²⁵, S+TM 26). There is ongoing interest in how such devices can be integrated into the EHR and used for clinical care, especially given that they are not currently designed to meet clear specifications (e.g., by FDA or professional societies).

The risk stratification of people who present with sleep complaints is based on identification of key symptoms or presenting signs, often collected by questionnaires. Existing screening questionnaires have been developed using data that reflect the “average” patient with sleep apnea—who often are middle-aged, male and overweight. These questions are likely to have more limited utility in identifying important subgroups with sleep apnea, including individuals who are normal or underweight, women, and young or older adults. Big volume data may allow better screening algorithms specific to individuals based on their gender, ethnicity, anatomy etc. Furthermore, while questionnaires like Epworth Sleepiness Scale (ESS)^{22,23} are widely used to assess daytime sleepiness, they only have a limited correlation with objective sleepiness measurements. Multiple sleep latency tests and multiple wakeful tests provide ‘objective’ assessments of daytime sleepiness but are expensive, burdensome and not commonly used. Use of “big data” to infer levels of alertness based on patterns of daily activities (e.g., activity levels, vigilance) may provide novel opportunities to assess patient function and response to treatment with minimal patient burden. Continuous real life

collection of these data provides can provide more representative information than a one-time questionnaire or a one-day ‘objective’ study in an artificial environment.

Management

Big data may facilitate dynamic, accurate and effective management of sleep disorders. Data from large EHR databases may help determine the best treatment for a patient by evaluating therapy outcomes over time, and comparing data in patients who are similar but who receive alternative treatments. This observational study design, although not as rigorous as a randomized controlled trial, can use techniques such as propensity score matching to provide valid inferences. Big data allows these propensity scores to be calculated reliably from large numbers of individuals. Use of multiple indices of disease risk and outcomes also may more accurately predict the chances of success of a therapy than the crude clinical estimates based on smaller numbers of variables. Effectively, millions of data points help create a multidimensional resource and allow more sophisticated pattern matching. Such an analysis would ask, “What was the outcome when a patient with the most similar characteristics and circumstances was managed in a particular manner?” These questions also can be framed to be “patient-centric”---providing patients with information on how patients most similar to them fared with specific therapies. The strict eligibility criteria of most clinical trials exclude many patients seen in routine clinical practice. Having access to large-scale EHRs can help the broadest group of patients make more informed decisions regarding their management.

Positive airway pressure (PAP) is currently considered the treatment of choice for most patients with obstructive sleep apnea (OSA). Alternative therapies, including oral appliances, surgery and positional therapies are usually reserved for patients who fail PAP therapy. However, is PAP therapy indeed the optimum first line therapy, no matter the unique characteristics and personal preferences of the patient? Availability of big data would afford more reliable subgroup analyses, including use of data on patient preference, and thence, facilitate a more individualized approach to managing OSA. Such insight will be instrumental in improving treatment adherence, effectiveness and appropriate use of resources. Future decisions informed by big data could be based on algorithms utilizing patient’s anatomy, presentation, traditional and novel polysomnographic variables, biomarkers, imaging, genetics and preferences. With better insight into what will likely work for a patient, and initiating the right therapy at the right time, has potentially huge advantages over current predominantly heuristic approaches.

Comprehensive use of data from the EHR could help identify which interventions may be most beneficial for a particular patient and patients most similar to each given patient (based on the multiple factors enumerated above). Did such patients have better adherence with spousal support or psychological support or peer support or other mediation? If so, similar interventions could be used to optimize CPAP therapy without having to embark on an ad hoc trial and error intervention. This could significantly diminish the wasted time and effort for both the provider and the patient, and alleviate the disease and debility and the financial burden on the healthcare system. What are the “non-traditional” factors that contribute to poor adherence? Are the people who are less able to articulate or express themselves – e.g., are those with limited English or health literacy or with limited social support at higher risk

for non-adherence? Are they the ones who will likely not contact the healthcare system if there are any issues with management? How can the provider minimize the risks in these groups and improve the adherence? Access to sufficient and relevant data could facilitate such analyses.

Big data may also shed further light on the substantial issue of PAP therapy non-adherence. Currently, several factors have been suggested to be associated with CPAP adherence, but the current models of non-adherence explain only a small degree of variation in outcomes. Analysis of large samples and daily trends in PAP use may elucidate patterns that relate to what limits or promotes CPAP adherence. Better understanding of factors influencing adherence can help decide which patients to start on PAP therapy. For example, knowing if a given patient was non-adherent to asthma inhalers or hypertension medications might predict adherence to CPAP. Big data analyses could predict the likelihood of non-adherence based on multiple inputs including such past adherence behavior as well as patient's age, gender, race, severity of sleep apnea, daytime symptoms, personality, and hundreds of other variables. This kind of analyses should be superior to a provider's empirical judgment, which usually takes into account only on a small subset of these factors, and is limited by one's individual ability to integrate large amounts of information and potential personal biases. Ideally, such information would not replace the provider judgment or insight, but enhance it and make it more precise and evidence-based.

Big data may also improve the understanding of PAP adherence by providing dynamic feedback on factors that can be leveraged to improve adherence, as well as offering the ability to individualize support. Smartphones provide continuous objective data over long periods rather than from a single point in time, often only at a clinic visit. For individuals, feedback on behaviors such as steps walked, weight lifted, nutritional information, subjective sleep, mood, emotions, etc. could be correlated with the CPAP use the night before, and help develop individual targets, and provide motivation for CPAP use. The feedback patients receive from their devices can also help them decide to seek help earlier help before a pattern of non-adherence has been established.²⁴ Visualization and analysis of temporal patterns of use from PAP machines, even including non-wear time, would better describe actual CPAP patterns and "residual" sleep apnea than current approaches that focus on simple summary measures (average hours per night). For example, a patient uses CPAP for 2 hours on one night and 6 hours the next. Based on the current payer requirements, a goal of 'an average of 4 hours a night' would have been met, and many providers may not further address important night-to-night variation in use. However, the patient might do much worse on the day following the 2-hour use night and may be vulnerable to not only to the side effects of poor sleep such as increased irritability, poor concentration, lower productivity at home, but also to increased traffic accidents. Despite an arguably arbitrary cutoff of 4 hours use for every patient, the definitions and consequences of underuse may vary significantly among different patients. Longitudinal estimation of total sleep time and residual apnea off positive airway pressure can enable calculation of the true residual untreated disease burden. The same principles are relevant to any sleep apnea treatment, including an oral appliance, weight loss, surgical approaches and body positioning. Models for tracking of sleep and breathing during sleep in home environments can be developed using approaches for blood pressure, weight, and home glucose monitoring.

Big data can be used to not only guide the management of sleep apnea, but also to understand the geographical, regional institutional or even provider-level variations in outcomes. This can be used for quality improvement purposes, and to direct educational resources towards specific providers to improve outcomes.

Complications of OSA

Sleep apnea is associated with several physiological alterations including hypoxia, inflammation, endothelial dysfunction, metabolic dysregulation and sympathetic activation.¹¹ These, along with genetic susceptibility, partly explain the association between OSA and diverse adverse consequences including hypertension,^{27–29} cardiovascular disease,^{30–32} metabolic derangements,^{28,33–35} neurological changes^{36,37} and increased mortality.³⁷ However, several issues pertaining to these associations remain uncertain. It is yet unclear which patients will suffer from which consequences. Studies with limited numbers of subjects have variably shown AHI, ODI, arousal index, time below 88% oxygen saturation and sleep architecture as polysomnographic predictors of consequences.³⁸ However, these metrics are not robust predictor of consequences. It also is unclear what demographic and clinical features suggest a worse outcome. Factors such as sleepiness and habitual short sleep duration may modify the risk of adverse outcomes with OSA and response to treatment.^{39,40} Several studies suggest racial differences in the outcomes,^{35,41} but whether race determines the consequences independent of multiple confounders such as socioeconomic and environmental factors is less clear. Similarly, there gender-based differences in outcomes which need further exploration, and which may be further influenced by age and cumulative years of sleep apnea exposure.⁴²

Improved prediction of outcomes may be anticipated through better access and use of Big Data. Analyses of EHRs from millions of patients can provide higher power to detect the risk factors independently associated with diverse outcomes in patients with sleep apnea. It may be particularly informative to access longitudinal data to estimate years of exposure to OSA and consider how age of onset of sleep apnea, years of exposure and timing of interventions all influence sleep apnea outcomes and treatment response. There is likely a large degree of individual variability in susceptibility to the various effects of sleep apnea, including differences in which organ systems may be most affected. It is unclear how such differences relate to differences in subtypes of sleep apnea (e.g., associated with different degrees of hypoxemia or sleep fragmentation), duration of untreated disease, or to other underlying co-morbidities including obesity and diabetes, or genetic variation. Only through access to very large datasets can patterns of risk be identified. This would require deep analysis of fully detailed sleep apnea data with data on demographics, clinical variables, and genetic factors.

Finally, there is the opportunity to better analyze big data for the purpose of better defining outcomes using online data sharing and crowdsourcing-use of community wisdom to address a wide variety of challenges. A recent example was evident is a competition where algorithms designed by an audience registrant team predicted amyotrophic lateral sclerosis (ALS) disease progression better than a baseline algorithm as well as clinicians from top ALS clinics.⁴³ This competition also identified potential new predictors of disease

progression including creatinine and blood pressure. Using the National Sleep Research Resource, the community has interacted to help develop algorithms for detecting airflow limitation (sleepdata.org).

Associations with other disorders

Voluminous databases with data variables encompassing several disorders offer the potential for studying the prevalence, presentation and consequences of SDB in patients with diverse disorders. Several conditions including but not limited to heart failure, stroke, renal disorders, pulmonary hypertension, idiopathic pulmonary fibrosis, Cushing's syndrome, acromegaly and Down syndrome are associated with an increased prevalence of sleep apnea.⁴⁴⁻⁴⁸ The risk factors, features and consequences of sleep apnea in these conditions, especially the less common disorders, have not been adequately researched. The patients with some of these disorders may lack usual symptoms of sleep apnea, which may go undetected and untreated because of atypical presentations. Additionally, the direction of causality when sleep apnea is comorbid with diverse disorders is not always clear- SDB may result from pathophysiologic changes attributable to the primary disorders or it may contribute to their progression, and some of these associations may be bidirectional. However, limited data have impeded a better understanding of these relationships. Furthermore, the consequences of sleep apnea and its therapy on sleep quality, quality of life and clinical progression of the comorbid disorders are unclear. Conversely, the effect of treatment of these disorders on prevalence of sleep apnea requires better definition. Improved insights into the association of sleep apnea with these disorders, as feasible through big data, may not only facilitate improving outcomes in the primary disorders, but may shed more light on the pathophysiology of sleep apnea. One caveat is that while pooling of data will allow analyses of large numbers of patients even with rare disorders, several studies show that patients are rarely asked about their sleep symptoms unless they volunteer this information. Thus there is a need to ensure that the sleep history is included as a standard part of history-taking along with medical, social and family history.

Developing big data resources

Relevant data for research and clinical sleep medicine are diverse and include data from overnight sleep recordings containing neuro-cardio-respiratory signals, PAP machines, patient-report outcome surveys, other clinical information and biomarker and genetic data. Several steps can facilitate development of big data resources in the field of sleep that yield valid inferences (Table 1). Identifying the best sources of such data and defining those data using structured approaches, such as defining the meta-data associated with each set of terms and measurements, enhances the ability to aggregate and analyze data from many sources. When possible, data from different sources should be harmonized to minimize the differences in the property being measured. This serves to minimize random or misclassification and improve the ability to draw valid inferences.

In the absence of complete data standardization, several approaches to structuring and defining data can facilitate data curation, federation, integration and analysis. 'Metadata' is a standardized way of describing the data to facilitate location, retrieval and utilization

information. 'Provenance' describes the origin of the data- the when, where and how it was created. Provenance for sleep variables would specify the sensors used for data collection and how the data were scored and interpreted. Provenance allows the data to be tracked through various changes over time, such as occurs with sleep laboratory equipment replacements or new scoring standards. 'Ontology' refers to a classification scheme which describes the terms or concepts associated with specific data elements and shows how these elements relate to one another. It allows the implicit knowledge of the data to be characterized by specific logical properties, and thus, provides a framework for extracing and sharing data across users. 'Domain ontologies' formalize and standardize the data language to overcome the semantic heterogeneity, and create a machine-understandable specification of what exists in a given domain. Sleep Domain Ontology (SDO) contains standard terms representing sleep disorders, phenotypes, presentation and medications as well as sleep procedures and devices such as PSGs.⁴⁹ The concepts are stored using a formal logic language (such as the Web Ontology Language, OWL).⁵⁰

Of specific interest to sleep medicine is the information available from the PSG, which contains physiological and bioelectric signals such as airflow, respiratory effort, electroencephalogram (EEG), electromyogram (EMG), electrooculogram (EOG), electrocardiogram (ECG) and oxygen saturation. This rich source of information has enormous potential for characterizing dynamic changes in neuro-physiological and cardiorespiratory parameters over a period of hours. Curation of data from multiple physiological systems may allow a more improved characterization of sleep apnea related stresses as well as may provide a rich source of physiological data that describes multiple organ systems, including the interactions across those systems. Centralized and open-access libraries of well-defined PSGs from multiple studies and sites could enhance the objectives described in the preceding sections. Big data opportunities include those that ask questions such as, Does a combination of signals such as heart-rate variability and EEG changes or flow pattern define the outcomes better than traditionally used definitions of sleep apnea? Does a combination of signals provide a more reliable quantification of central versus obstructive origin of hypopneas? There is also opportunities to extract responses to PAP titration in systematic formats to identify subgroups of individuals, such as those with complex sleep apnea, or with phenotypes, such as high loop gain.

There is a strong push, based on economic consideration among others, to increase the use of portable sleep testing devices. Portable systems frequently record a subset of signals used in PSG (frequently including airflow, respiratory effort and oxygen saturation) and/or other signals such as peripheral arterial tonometry. This can introduce a heterogeneous set of data, and data collected in less controlled settings than laboratory-based assessments. Establishing standards for data quality for data derived from these portable systems is needed as those data are curated into larger data repositories.

A major concern when aggregating data relates to the comparability of the source or derived data. When harmonizing data from PSG or home sleep tests, it is important to recognize that the sensors used are variable in their material, mechanism of acquiring signal (pressure changes, temperature changes), validation process and reliability. Adherence to a common set of procedures for measurement and scoring improves the comparability of data from

different sources and laboratories.^{15,51} However, often data were collected when no or different data standards were in place, or were collected with equipment that uses different collection and processing algorithms despite meeting minimal published standards. In those cases when data were collected differently, it is critical that each measurement is defined with sufficient specificity to allow the user to make appropriate analysis decisions, and the provenance of those measurements is provided.

Data from sleep tests also provide opportunities to generate numerous quantitative measures that can be used in big data analyses to enhance discovery of new patterns and predictors of disease. The various signals on the polysomnogram provide risk signal analysis opportunities, at the level of individual signals, coupled signals (e.g., cardiopulmonary coupling and cardiorespiratory synchronization), and network analysis of multiple signals. Traditionally the 30-second sleep epoch is considered the basic unit of sleep data and is assigned a sleep stage irrespective of the significant heterogeneity that may exist in EEG within that epoch (alpha and delta rhythms, K-complexes, sleep spindles etc.). Information from each epoch is often collapsed and summarized rather than analyzed using quantitative methods that characterize all of the information within each epoch and the relationship of data in surrounding epochs. Spectral and non-linear analyses of these variables (see the section below) might further enhance the information content of PSG data often presented in simpler “summary” or binary format. Linear methods like Fast Fourier transform (FFT) of EEG data reformat EEG into component frequency spectral patterns can help elucidate pathological EEG patterns and their relationship to outcomes.¹⁸ These analyses need to appropriately consider challenges in analyzing such signals, such as assumptions of stationary and impact of artifact. The human brain is neither just deterministic nor fully stochastic, and its properties and signal generation are nonlinear.⁵² Newer non-linear machine learning techniques such as support vector machines (SVMs) for EEG signal classification or multi-scale entropy analysis can be informative regarding EEG changes in sleep apnea and provide more insight into diagnosis and correlates.⁵³ Unsupervised and supervised machine learning algorithms can be employed to understand the pattern differences as well as correlate them with presence or absence of significant disease.^{54–56}

Electronic Health Record Data

EHRs constitute a valuable collection of patient health information and can be transformative for individual patient as well as public health. A systematic collection of patient records helps track care, assess quality and safety of care and trigger alerts, warning and reminders. The Health Information Technology for Economic and Clinical Health (HITECH) Act was enacted as part of the American Recovery and Reinvestment Act of 2009 to promote the health information technology adoption, including certified EHR adoption, formation of patient registries and achievement of ‘meaningful use’ objectives. Use of electronic prescribing and cross-referencing with diagnoses can help keep track of use of medications such as hypnotics or stimulants in patients with SDB and promote safety. Such data help understand which patients need these medications, and how these medications impact diverse outcomes such as sleep quality, quality of life, daytime sleepiness, traffic accidents etc. Given the variation in SDB phenotypes across the population, the development of SDB registries could create resources to better understand the

differences in risk factors, outcomes and treatment responses across patients, and thus inform more individualized screening, risk stratification and treatment strategies. For example, registries can be used to identify patient with developmental disorders associated with OSA, patients with complex sleep apnea, patients with different levels of adherence, and patients with residual sleepiness despite sleep apnea treatment⁵⁷. The American Academy of Sleep Medicine is in the process of publishing recommendations for sleep data to be included in the EHR, highlight “end to end” needs—from patient screening, diagnostic information, and population management, and are discussed in chapter x. The Academy also reviewed and has recommended a list of recommended patient-reported surveys to be included as well as data fields from sleep studies to be included in the EHR.

Sleep Medicine Big Data Resources

The National Institutes of Health (NIH) Big Data to Knowledge (BD2K) initiative was developed to support the development of state-of-the-art big data access and analysis methods. Its specific goals are to: a) Facilitate broad use of biomedical digital assets by making them discoverable, accessible, and citable. b) Conduct research and develop the methods, software, and tools needed to analyze biomedical Big Data. c) Enhance training in the development and use of methods and tools necessary for biomedical Big Data science, and d) Support a data ecosystem that accelerates discovery as part of a digital enterprise. BD2K initiatives include the development of Centers of Excellence for big data computing, Data Discovery Coordination consortia; and big data courses and open educational resources. Although many of these initiatives are generic, at least one BD2K funded project identifies sleep-related data as a “use case” for developing more advanced approaches for data annotation and visualization. The White House Brain Research through Advancing Innovative Neurotechnologies (BRAIN) initiative⁵⁸ defines a national research agenda to uncover mechanisms involved in neurological diseases such as Alzheimer’s Disease, Parkinson’s Disease and Depression, which are topics of relevance to the sleep medicine community.

In past few years, there has been a concerted effort towards development of repositories of sleep data that are more comprehensive, easily accessible and well-annotated. These resources are aimed at satisfying the informatics and data resource needs in the sleep community for clinical and research purposes. Large diverse collections of data can help elucidate the effect of age, gender and demographic variables on the clinical and polysomnographic presentation of sleep apnea, identify predictors that may optimize management of sleep apnea, and provide novel characterizations of the associations between SDB and adverse health effects such as cardiovascular disorders. Such large data sources, especially when combined with genetic data, can help improve sleep apnea phenotyping and identify genetic fingerprints of sleep apnea and the associated consequences.

The National Sleep Research Resource (NSRR), established through a National Heart Lung Blood Institute resource grant, supports the goals of the NIH Big Data to Knowledge (BD2K) initiative by providing researchers and trainees with access to de-identified well-annotated physiological signals from curated sleep studies linked to data aggregated from large cohort studies and clinical trials.⁵⁹ Datasets that have been made available through this

resource include sleep studies and covariate data from the Sleep Heart Health Study (SHHS), the Cleveland Family Study, the Heart Biomarker Evaluation in Apnea, the Study of Osteoporotic Fractures, the MrOS Sleep Study, the Cleveland Children's Sleep and Health Study, and the Childhood Adenotonsillectomy Trial. The NSRR implements a scalable architecture designed to accommodate expansion to incorporate data abstracted from ongoing cohort studies, planned clinical trials, and clinical resources including current generation EHRs. Available covariate data includes demographic information, anthropometric parameters, physiologic measurements, medical history elements, and neurocognitive testing results. Files that can be downloaded from the NSRR include raw signal data, covariate data, and a suite of modifiable tools that can be used for signal processing and analysis. Measurements of key parameters from EEG, EOG, EMG, and ECG using open-source codes are available, ready for use by investigators for biostatistical analysis. A cross-dataset query interface has been developed to facilitate the collection, management and sharing of data (Figure 1).^{60,61} A canonical data dictionary that maps study specific data variables to a standardized set of variable definitions is used to standardize terminology across studies. This resource incorporates an expanding set of core terms drawn from a range of domains prioritized on the basis of their anticipated utility to facilitate intra-cohort exploratory data analysis and cross-cohort query generation. As each new dataset is added to the NSRR, an external data file is generated that maps core terms listed in the canonical data dictionary to corresponding study variables (Figure 2). As each study variable is mapped, it is annotated with provenance attributes that specify the source of the data, the time point of collection, the method of collection, and the specifics of any equipment used to acquire the data.^{62, 63}

Social networking websites that contain personal and social, and occasionally medical information as well as websites designed specifically to disseminate medical information or allow active patient engagement are other valuable sources for big data.^{3,64} The emerging big data models have allowed patients to generate and share their data rather than just relying on healthcare institutions for data accumulation. Crowdsourcing has recently been used to help build prediction, detection and therapeutic models. The current push is to allow open-access model for such data, allowing democratization of the data.⁴³ Patients are now playing a role not only in generating the data, but even analyzing and publishing it.^{3,65} In this regard, a sleep apnea resource is the Sleep Apnea Patient Centered Outcomes Network, and its public facing portal, MyApnea.Org.³ This is a virtual network of patients, family members, clinicians and researchers who interact through a forum and series of interactive tools, including tools for nominating and voting on research priorities and sleep apnea screening tools. In addition, patients provide self-reported survey information through a series of validated and new tools that assess a range of information from experience with diagnosis, treatments, symptoms, quality of life and health conditions.³ As of January 2016, over 6,300 individuals have joined the network, providing large amounts of data to help the community understand, from a patient's perspective, issues related to satisfaction with care, treatment outcomes, and co-morbidities. Data are collected using a common data model, allowing integration with other national efforts. Such portals represent a significant evolution in patient engagement from being a mere participant to being the designers and drivers of the research. Over time, portal such as this may be enhanced by integrating information from

wearables and PAP devices, EHR data, and biospecimens, and thus truly achieving “big data” goals of multi-dimensional data.

BioPortal provides a comprehensive repository of biomedical ontologies, including sleep ontologies.⁶⁶ PatientsLikeMe is another online data-sharing platform millions of unique data points from patients with diverse disorders including sleep apnea.⁶⁴

Limitations and challenges

Big Data holds enormous promise but can be associated with significant challenges. Since data can come from several different sources, in different formats, with different annotations, harmonizing and standardizing can be a major undertaking. Although there exist some standards through the International Classification of Sleep Disorders and scoring manuals, and the NSRR has begun development of ontologies and meta-data for sleep terms, there is not yet a nationally endorsed Sleep Domain Ontology. The disparate sources of sleep data across the medical system—from sleep laboratories to data collected by durable medical equipment companies to use of wearable medical devices and PAP machines present special challenges for data capture and integration. New efforts by the American Academy of Sleep Medicine promise to begin the process for defining and prioritizing those items most relevant for sleep medicine. However, full participation by EHR vendors and healthcare systems will be required to realize the potential of these data. From the clinician’s perspective, there will be further practical challenges. Data collection, access and support need to improve both the efficiency and effectiveness of the clinical outcome to be of value and pragmatic. The role of regulatory and professional organizations in defining benchmarks for electronic data maintenance needs clarification and should enhance clinical care.

In establishing centralized data repositories or enhancing access to data, there are important privacy, security and confidentiality concerns that need to be carefully and continuously addressed. However, there is also a need for the sleep medicine community to enhance its culture of collaboration where the value of sharing deidentified data and collaborative analysis is embraced. The public, funders, and industry all recognize that greater data sharing can improve transparency, enhance discovery, and accelerate dissemination of findings.⁶⁷ To achieve this goal there is a need for the field to revisit traditional attitudes towards data “ownership.”

Volumes of data cannot compensate for poor quality data, and defining and implementing data verification steps are critical. Big data also can lead to spurious outcomes if appropriate adjustment for multiple comparisons are not applied. Achieving a balance between heuristics and data-driven decision-making will need to be achieved until we have evidence of what is the most efficient and effective process to improve patient outcomes.^{68,69}

Finally, big data can be extremely expensive to create and maintain. Collaborations between healthcare providers, information technologists, patient representatives, mathematicians, computer scientists and statisticians among others; as well as financial and regulatory support from Federal, State and private supporters will be required for maintenance and implementation of effective and viable clinical and research big data resources. It is

imperative that the procurement, storage, and analytical methods be cost-efficient to be widely accepted and adopted and sustainable.

Conclusions

Analysis of large volume data hold the promise for improving the application of precision medicine to sleep, including improving the identification of patient subgroups who may specifically benefit from alternative therapies. Big data used within the health care system also promises to facilitate “end to end” screening, diagnosis and management of sleep disorders, improved the recognition of differences in presentation and susceptibility to sleep apnea; and lead to improved management and outcomes. To meet the vision of personalized, precision therapeutics and diagnostics and improving the efficiency and quality of sleep medicine will require ongoing efforts, investments and some change in our current medical and research cultures. Expanding the EHR with well defined sleep data and further development and growth of national research repositories and patient networks and registries are important strategies for achieving these goals. Any successful strategy will need to integrate patient collected data through the rapidly evolving wearable device market.

Acknowledgments

Support in part by NIH HL114473 and PCORI PPRN-1306-04344

References

1. Tricoci P, Allen JM, Kramer JM, Califf RM, Smith SC. Scientific evidence underlying the ACC/AHA clinical practice guidelines. *JAMA*. 2009; 301(8):831–41. [PubMed: 19244190]
2. Prinz F, Schlange T, Asadullah K. Believe it or not: how much can we rely on published data on potential drug targets? *Nat Rev Drug Discov*. 2011; 10(9):712. [PubMed: 21892149]
3. [accessed 12/25/15] <https://myapnea.org>
4. Mathur R, Douglas NJ. Family studies in patients with the sleep apnea-hypopnea syndrome. *Ann Intern Med*. 1995; 122(3):174–8. [PubMed: 7810934]
5. Liu Y, Patel S, Nibbe R, Maxwell S, Chowdhury SA, Koyuturk M, Zhu X, Larkin EK, Buxbaum SG, Punjabi NM, et al. Systems biology analyses of gene expression and genome wide association study data in obstructive sleep apnea. *Pac Symp Biocomput*. 2011:14–25. [PubMed: 21121029]
6. Patel SR, Larkin EK, Redline S. Shared genetic basis for obstructive sleep apnea and adiposity measures. *Int J Obes (Lond)*. 2008; 32(5):795–800. [PubMed: 18209735]
7. Larkin EK, Patel SR, Goodloe RJ, Li Y, Zhu X, Gray-McGuire C, Adams MD, Redline S. A candidate gene study of obstructive sleep apnea in European Americans and African Americans. *Am J Respir Crit Care Med*. 2010; 182(7):947–53. [PubMed: 20538960]
8. Patel SR, Goodloe R, De G, Kowgier M, Weng J, Buxbaum SG, Cade B, Fulop T, Gharib SA, Gottlieb DJ, et al. Association of genetic loci with sleep apnea in European Americans and African-Americans: the Candidate Gene Association Resource (CARE). *PLoS One*. 2012; 7(11):e48836. [PubMed: 23155414]
9. Gaziano JM, Concato J, Brophy M, Fiore L, Pyarajan S, Breeling J, Whitbourne S, Deen J, Shannon C, Humphries D, et al. Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *J Clin Epidemiol*. 2015
10. McCarty CA, Chisholm RL, Chute CG, Kullo IJ, Jarvik GP, Larson EB, Li R, Masys DR, Ritchie MD, Roden DM. The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC medical genomics*. 2011; 4(1):13. [PubMed: 21269473]

11. Budhiraja R, Parthasarathy S, Quan SF. Endothelial dysfunction in obstructive sleep apnea. *J Clin Sleep Med*. 2007; 3(4):409–15. [PubMed: 17694731]
12. Spira AP, Blackwell T, Stone KL, Redline S, Cauley JA, Ancoli-Israel S, Yaffe K. Sleep-disordered breathing and cognition in older women. *J Am Geriatr Soc*. 2008; 56(1):45–50. [PubMed: 18047498]
13. Gozal D, Capdevila OS, Kheirandish-Gozal L, Crabtree VM. APOE epsilon 4 allele, cognitive dysfunction, and obstructive sleep apnea in children. *Neurology*. 2007; 69(3):243–9. [PubMed: 17636061]
14. Nikodemova M, Finn L, Mignot E, Salzieder N, Peppard PE. Association of sleep disordered breathing and cognitive deficit in APOE epsilon4 carriers. *Sleep*. 2013; 36(6):873–80. [PubMed: 23729930]
15. Berry RB, Budhiraja R, Gottlieb DJ, Gozal D, Iber C, Kapur VK, Marcus CL, Mehra R, Parthasarathy S, Quan SF, et al. Rules for scoring respiratory events in sleep: update of the 2007 AASM Manual for the Scoring of Sleep and Associated Events. Deliberations of the Sleep Apnea Definitions Task Force of the American Academy of Sleep Medicine. *J Clin Sleep Med*. 2012; 8(5):597–619. [PubMed: 23066376]
16. Budhiraja R, Goodwin JL, Parthasarathy S, Quan SF. Comparison of nasal pressure transducer and thermistor for detection of respiratory events during polysomnography in children. *Sleep*. 2005; 28(9):1117–21. [PubMed: 16268381]
17. Mokhlesi B, Finn LA, Hagen EW, Young T, Hla KM, Van Cauter E, Peppard PE. Obstructive sleep apnea during REM sleep and hypertension. Results of the Wisconsin Sleep Cohort. *American journal of respiratory and critical care medicine*. 2014; 190(10):1158–1167. [PubMed: 25295854]
18. Budhiraja R, Quan SF, Punjabi NM, Drake CL, Dickman R, Fass R. Power spectral analysis of the sleep electroencephalogram in heartburn patients with or without gastroesophageal reflux disease: a feasibility study. *J Clin Gastroenterol*. 2010; 44(2):91–6. [PubMed: 19661813]
19. Giffen CA, Carroll LE, Adams JT, Brennan SP, Coady SA, Wagner EL. Providing Contemporary Access to Historical Biospecimen Collections: Development of the NHLBI Biologic Specimen and Data Repository Information Coordinating Center (BioLINCC). *Biopreservation and biobanking*. 2015; 13(4):271–279. [PubMed: 26186276]
20. Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, Mietus JE, Moody GB, Peng CK, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation*. 2000; 101(23):E215–20. [PubMed: 10851218]
21. Moody GB, Mark RG, Goldberger AL. PhysioNet: physiologic signals, time series and related open source software for basic, clinical, and applied research. *Conf Proc IEEE Eng Med Biol Soc*. 2011; 2011:8327–30. [PubMed: 22256277]
22. Johns MW. A new method for measuring daytime sleepiness: the Epworth sleepiness scale. *Sleep*. 1991; 14(6):540–5. [PubMed: 1798888]
23. Johns MW. Sensitivity and specificity of the multiple sleep latency test (MSLT), the maintenance of wakefulness test and the epworth sleepiness scale: failure of the MSLT as a gold standard. *J Sleep Res*. 2000; 9(1):5–11. [PubMed: 10733683]
24. Budhiraja R, Parthasarathy S, Drake CL, Roth T, Sharief I, Budhiraja P, Saunders V, Hudgel DW. Early CPAP use identifies subsequent adherence to CPAP therapy. *Sleep*. 2007; 30(3):320–4. [PubMed: 17425228]
25. [accessed 1/19/2016] <http://www.sleepimage.com>
26. [accessed 1/19/2026] <http://www.resmed.com/us/en/consumer/s-plus.html>
27. Budhiraja R, Sharief I, Quan SF. Sleep disordered breathing and hypertension. *J Clin Sleep Med*. 2005; 1(4):401–4. [PubMed: 17564410]
28. Dean DA, Wang R, Jacobs DR, Duprez D, Punjabi NM, Zee PC, Shea S, Watson K, Redline S. A systematic assessment of the association of polysomnographic indices with blood pressure: the Multi-Ethnic Study of Atherosclerosis (MESA). *Sleep*. 2015; 38(4):587–96. [PubMed: 25348124]
29. Redline S, Sotres-Alvarez D, Loreda J, Hall M, Patel SR, Ramos A, Shah N, Ries A, Arens R, Barnhart J, et al. Sleep-disordered breathing in Hispanic/Latino individuals of diverse

- backgrounds. The Hispanic Community Health Study/Study of Latinos. *Am J Respir Crit Care Med*. 2014; 189(3):335–44. [PubMed: 24392863]
30. Budhiraja R, Quan SF. Sleep-disordered breathing and cardiovascular health. *Curr Opin Pulm Med*. 2005; 11(6):501–6. [PubMed: 16217175]
 31. Budhiraja R, Budhiraja P, Quan SF. Sleep-disordered breathing and cardiovascular disorders. *Respir Care*. 2010; 55(10):1322–32. discussion 1330–2. [PubMed: 20875159]
 32. Lin GM, Colangelo LA, Lloyd-Jones DM, Redline S, Yeboah J, Heckbert SR, Nazarian S, Alonso A, Bluemke DA, Punjabi NM, et al. Association of Sleep Apnea and Snoring With Incident Atrial Fibrillation in the Multi-Ethnic Study of Atherosclerosis. *Am J Epidemiol*. 2015; 182(1):49–57. [PubMed: 25977516]
 33. Seicean S, Kirchner HL, Gottlieb DJ, Punjabi NM, Resnick H, Sanders M, Budhiraja R, Singer M, Redline S. Sleep-disordered breathing and impaired glucose metabolism in normal-weight and overweight/obese individuals: the Sleep Heart Health Study. *Diabetes Care*. 2008; 31(5):1001–6. [PubMed: 18268072]
 34. Strand LB, Carnethon M, Biggs ML, Djoussé L, Kaplan RC, Siscovick DS, Robbins JA, Redline S, Patel SR, Janszky I, et al. Sleep Disturbances and Glucose Metabolism in Older Adults: The Cardiovascular Health Study. *Diabetes Care*. 2015; 38(11):2050–8. [PubMed: 26384390]
 35. Bakker JP, Weng J, Wang R, Redline S, Punjabi NM, Patel SR. Associations between Obstructive Sleep Apnea, Sleep Duration, and Abnormal Fasting Glucose. The Multi-Ethnic Study of Atherosclerosis. *Am J Respir Crit Care Med*. 2015; 192(6):745–53. [PubMed: 26084035]
 36. Gelber RP, Redline S, Ross GW, Petrovitch H, Sonnen JA, Zarow C, Uyehara-Lock JH, Masaki KH, Launer LJ, White LR. Associations of brain lesions at autopsy with polysomnography features before death. *Neurology*. 2015; 84(3):296–303. [PubMed: 25503626]
 37. Molnar MZ, Mucsi I, Novak M, Szabo Z, Freire AX, Huch KM, Arah OA, Ma JZ, Lu JL, Sim JJ, et al. Association of incident obstructive sleep apnoea with outcomes in a large cohort of US veterans. *Thorax*. 2015; 70(9):888–95. [PubMed: 26038534]
 38. Lutsey PL, McClelland RL, Duprez D, Shea S, Shahar E, Nagayoshi M, Budoff M, Kaufman JD, Redline S. Objectively measured sleep characteristics and prevalence of coronary artery calcification: the Multi-Ethnic Study of Atherosclerosis Sleep study. *Thorax*. 2015; 70(9):880–7. [PubMed: 26156526]
 39. Priou P, Le Vaillant M, Meslier N, Paris A, Pigeanne T, Nguyen XL, Alizon C, Bizieux-Thaminy A, Leclair-Visonneau L, Humeau MP, et al. Cumulative association of obstructive sleep apnea severity and short sleep duration with the risk for hypertension. *PLoS One*. 2014; 9(12):e115666. [PubMed: 25531468]
 40. Phillips B, Shafazand S. CPAP and hypertension in nonsleepy patients. *J Clin Sleep Med*. 2013; 9(2):181–2. [PubMed: 23372475]
 41. Shah N, Allison M, Teng Y, Wassertheil-Smoller S, Sotres-Alvarez D, Ramos AR, Zee PC, Criqui MH, Yaggi HK, Gallo LC, et al. Sleep apnea is independently associated with peripheral arterial disease in the Hispanic Community Health Study/Study of Latinos. *Arterioscler Thromb Vasc Biol*. 2015; 35(3):710–5. [PubMed: 25657310]
 42. Roca GQ, Redline S, Claggett B, Bello N, Ballantyne CM, Solomon SD, Shah AM. Sex-Specific Association of Sleep Apnea Severity With Subclinical Myocardial Injury, Ventricular Hypertrophy, and Heart Failure Risk in a Community-Dwelling Cohort: The Atherosclerosis Risk in Communities-Sleep Heart Health Study. *Circulation*. 2015; 132(14):1329–37. [PubMed: 26316620]
 43. Küffner R, Zach N, Norel R, Hawe J, Schoenfeld D, Wang L, Li G, Fang L, Mackey L, Hardiman O, et al. Crowdsourced analysis of clinical trial data to predict amyotrophic lateral sclerosis progression. *Nat Biotechnol*. 2015; 33(1):51–7. [PubMed: 25362243]
 44. Shipley JE, Scheingart DE, Tandon R, Starkman MN. Sleep architecture and sleep apnea in patients with Cushing's disease. *Sleep*. 1992; 15(6):514–8. [PubMed: 1335612]
 45. Minic M, Granton JT, Ryan CM. Sleep disordered breathing in group 1 pulmonary arterial hypertension. *J Clin Sleep Med*. 2014; 10(3):277–83. [PubMed: 24634625]
 46. Mermigkis C, Stagaki E, Tryfon S, Schiza S, Amfilochiou A, Polychronopoulos V, Panagou P, Galanis N, Kallianos A, Mermigkis D, et al. How common is sleep-disordered breathing in

- patients with idiopathic pulmonary fibrosis? *Sleep Breath*. 2010; 14(4):387–90. [PubMed: 20232261]
47. Ezzat H, Mohab A. Prevalence of sleep disorders among ESRD patients. *Ren Fail*. 2015; 37(6): 1013–9. [PubMed: 25959021]
 48. Maris M, Verhulst S, Wojciechowski M, Van de Heyning P, Boudewyns A. Prevalence of Obstructive Sleep Apnea in Children With Down Syndrome. *Sleep*. 2015
 49. Arabandi, S.; Ogbuji, C.; Redline, S.; Chervin, R.; Boero, J.; Benca, R.; Zhang, GQ. Developing a Sleep Domain Ontology. Vol. 2010. San Francisco, CA: Mar 12–13. 2010
 50. Hitzler, P.; Krötzsch, M.; Parsia, B.; Patel-Schneider, PF.; Rudolph, S. W3C. 2009. OWL 2 Web Ontology Language Primer.
 51. Redline S, Budhiraja R, Kapur V, Marcus CL, Mateika JH, Mehra R, Parthasarthy S, Somers VK, Strohl KP, Sulit LG, et al. The scoring of respiratory events in sleep: reliability and validity. *J Clin Sleep Med*. 2007; 3(2):169–200. [PubMed: 17557426]
 52. Klonowski W. Everything you wanted to ask about EEG but were afraid to get the right answer. *Nonlinear Biomed Phys*. 2009; 3(1):2. [PubMed: 19470156]
 53. Khandoker AH, Palaniswami M, Karmakar CK. Support vector machines for automated recognition of obstructive sleep apnea syndrome from ECG recordings. *IEEE Trans Inf Technol Biomed*. 2009; 13(1):37–48. [PubMed: 19129022]
 54. Gentleman, R.; Carey, V. *Bioconductor Case Studies*. Springer; 2008. Unsupervised machine learning; p. 137-157.
 55. Kotsiantis, SB.; Zaharakis, I.; Pintelas, P. *Supervised machine learning: A review of classification techniques*. 2007.
 56. Alpaydin, E. *Introduction to machine learning*. MIT press; 2014.
 57. Strollo PJ Jr, Badr MS, Coppola MP, Fleishman SA, Jacobowitz O, Kushida CA. The future of sleep medicine. *Sleep*. 2011; 34(12):1613. [PubMed: 22131593]
 58. Bargmann CI, Newsome WT. The Brain Research Through Advancing Innovative Neurotechnologies (BRAIN) Initiative and Neurology. *JAMA neurology*. 2014; 71(6):675–676. [PubMed: 24711071]
 59. [accessed 12/25/15] <https://sleepdata.org/>
 60. Services USDoHaH, editor. Three New Informatics Pilot Projects to Aid Clinical and Translational Scientists Nationwide. National Center for Research Resources (NCRR); 2009.
 61. [accessed 12/25/15] Physio-MIMI. <https://sleepdata.org/tools/physiomimi>
 62. Zhang, GQ.; Siegler, T.; Saxman, P.; Sandberg, N.; Mueller, R.; Johnson, N.; Hunscher, D.; Arabandi, S. VISAGE: A Query Interface for Clinical Research. Vol. 2010. San Francisco, CA: Mar 12–13. 2011 p. 76-80.
 63. Mueller, R.; Sahoo, S.; Dong, X.; Redline, S.; Arabandi, S.; Luo, L.; Zhang, GQ. Mapping multi-institution data sources to domain ontology for data federation: the Physio-MIMI approach. 2011.
 64. Wicks P, Massagli M, Frost J, Brownstein C, Okun S, Vaughan T, Bradley R, Heywood J. Sharing health data for better outcomes on PatientsLikeMe. *Journal of medical Internet research*. 2010; 12(2)
 65. Wicks P, Vaughan TE, Massagli MP, Heywood J. Accelerated clinical discovery using self-reported patient data collected online and a patient-matching algorithm. *Nat Biotechnol*. 2011; 29(5):411–4. [PubMed: 21516084]
 66. Musen MA, Musen M, Shah NH, Shah N, Noy NF, Noy N, Dai BY, Dai B, Dorf M, Griffith N, et al. BioPortal: ontologies and data resources with the click of a mouse. *AMIA Annu Symp Proc*. 2008:1223–4. [PubMed: 18999306]
 67. Drazen JM. Sharing individual patient data from clinical trials. *N Engl J Med*. 2015; 372(3):201–2. [PubMed: 25587944]
 68. Cresswell K, Majeed A, Bates DW, Sheikh A. Computerised decision support systems for healthcare professionals: an interpretative review. *Inform Prim Care*. 2012; 20(2):115–28. [PubMed: 23710776]
 69. Marewski JN, Gigerenzer G. Heuristic decision making in medicine. *Dialogues Clin Neurosci*. 2012; 14(1):77–89. [PubMed: 22577307]

- Analysis of large volume data hold the promise for improving the application of precision medicine to sleep, including improving the identification of patient subgroups who may specifically benefit from alternative therapies.
- Big data used within the health care system also promises to facilitate “end to end” screening, diagnosis and management of sleep disorders, improved the recognition of differences in presentation and susceptibility to sleep apnea; and lead to improved management and outcomes.
- To meet the vision of personalized, precision therapeutics and diagnostics and improving the efficiency and quality of sleep medicine will require ongoing efforts, investments and some change in our current medical and research cultures.
- Expanding the EHR with well defined sleep data and further development and growth of national research repositories and patient networks and registries are important strategies for achieving these goals.
- Any successful strategy will need to integrate patient collected data through the rapidly evolving wearable device market.

Figure 1. NSRR cross-dataset query interface. Construction of a cross-dataset query involves (1) selection of datasets, (2) specification of core query terms, and (3) refinement of selectable parameters. A query returns counts of results annotated with attributes detailing the provenance of each included study variable.

Figure 2. NSRR cross-dataset query processing. Standardized core query terms listed in a canonical data dictionary are directly mapped to corresponding study variables in each dataset.

Table 1

Establishing big data resources in Sleep Medicine

Creation of comprehensive archives of sleep study data
The collection of data-the sensors, the scoring criteria and the recording should be standardized. However, technology is rapidly improving and evolving; avoiding getting locked down to an inferior option demands flexibility not generally considered in scientific investigations
A structured format for healthcare data-notes, documentation of lab results, imaging and genetic and other studies will promote homogeneity of such databases.
A normalized vocabulary of sleep terms is required for cross-referencing and more valid comparisons.
The data must be stored on widely accessible platforms and retrievable through user-friendly interfaces to be of maximum value to the researchers.
Open access or nominal cost will help widespread utilization.
Frameworks need to be in place to better define data access from different sources to help share and communicate information.
Settings up secure and largely automated data pathways to merge data from EHR, sleep laboratory polysomnograms, therapeutic-device tracking systems, and data from wearables