Big data in sleep medicine: prospects and pitfalls in phenotyping

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters

Citation

Published Version
doi:10.2147/NSS.S130141

Citable link
http://nrs.harvard.edu/urn-3:HUL.InstRepos:31731751

Terms of Use
This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA
Big data in sleep medicine: prospects and pitfalls in phenotyping

Matt T Bianchi1,2, Kathryn Russo1, Harriett Gabbidon1, Tiaundra Smith1, Balaji Goparaju1, M Brandon Westover1
1Neurology Department, Massachusetts General Hospital, 2Division of Sleep Medicine, Harvard Medical School, Boston, MA, USA

Abstract: Clinical polysomnography (PSG) databases are a rich resource in the era of “big data” analytics. We explore the uses and potential pitfalls of clinical data mining of PSG using statistical principles and analysis of clinical data from our sleep center. We performed retrospective analysis of self-reported and objective PSG data from adults who underwent overnight PSG (diagnostic tests, n=1835). Self-reported symptoms overlapped markedly between the two most common categories, insomnia and sleep apnea, with the majority reporting symptoms of both disorders. Standard clinical metrics routinely reported on objective data were analyzed for basic properties (missing values, distributions), pairwise correlations, and descriptive phenotyping. Of 41 continuous variables, including clinical and PSG derived, none passed testing for normality. Objective findings of sleep apnea and periodic limb movements were common, with 51% having an apnea–hypopnea index (AHI)>5 per hour and 25% having a leg movement index >15 per hour. Different visualization methods are shown for common variables to explore population distributions. Phenotyping methods based on clinical databases are discussed for sleep architecture, sleep apnea, and insomnia. Inferential pitfalls are discussed using the current dataset and case examples from the literature. The increasing availability of clinical databases for large-scale analytics holds important promise in sleep medicine, especially as it becomes increasingly important to demonstrate the utility of clinical testing methods in management of sleep disorders. Awareness of the strengths, as well as caution regarding the limitations, will maximize the productive use of big data analytics in sleep medicine.

Keywords: polysomnography, sleep disorders, subjective symptoms, correlation, plotting, statistics

Introduction
Polysomnography (PSG) offers a wealth of physiological information, informing clinical decision-making and clinical research. Large sleep-related datasets are increasingly available for public analysis. For example, the National Sleep Research Resource (NSRR),1 PhysioNet (www.physionet.com), the Montreal Archive of Sleep Studies (MASS),2 and even consumer-facing efforts are underway.3 As “big data” analysis efforts gain momentum, it is increasingly important to understand not only the potential benefits but also the potential pitfalls of PSG phenotyping. In an era when in-laboratory PSG is increasingly restricted, enhancing signal processing and big data analytics could justify resource allocation to inform individual- and population-level insights.

The goals of sleep phenotyping span basic and clinical investigations as well as genotype–phenotype associations, especially as academic centers are increasingly banking bio-samples. Advanced knowledge about normal and pathologic sleep
physiology might be derived from studying the relationship between sleep-disordered breathing events and heart rate variability, or about how electromyography (EMG) dynamics in rapid eye movement (REM) sleep vary depending on the presence of different medications and disease states. Big data insights might link indices of fragmentation to comorbidities or predict response to treatment of obstructive sleep apnea (OSA).

The opportunity for advanced analysis and phenotyping from the rich data obtained in routine clinical practice cannot be overstated. However, the allure of big data should not distract from the potential risks associated with even basic statistics and inferential efforts. Numerous cautionary statistical articles4–12 and even entire monographs13–15 have been published in recent decades highlighting the existence (and persistence) of common statistical misconceptions and pitfalls in basic and clinical research contexts. Large datasets do not mitigate these risks and in fact may present further challenges. We explore various kinds of PSG data in this framework, including insights and pitfalls from the existing literature, and through empirical analysis of a large dataset of diagnostic PSGs from our center. With the growing capacity for large-scale analytics, recognizing the strengths and limitations of phenotyping will help maximize the utility of large database resources.

Methods

The Partners Institutional Review Board (IRB) approved retrospective analysis of our database without requiring additional consent for use of the clinically acquired data (IRB number: 2009P000758). We selected diagnostic PSGs from adults in our database from 2011 to 2015, yielding n=1835 studies in our dataset. We did not have any exclusion criteria.

PSG was performed and scored according to the American Academy of Sleep Medicine (AASM) practice standards. Channels included six electroencephalogram leads, bilateral electrooculogram, submentalis EMG, nasal thermistor, oronasal airflow, snore vibration sensor, single-lead electrocardiogram (ECG), chest and abdomen effort belts, pulse oximetry, and bilateral anterior tibialis EMG.

From presleep questionnaires, we analyzed self-reported symptoms associated with sleep apnea, insomnia, restless legs, and narcolepsy. OSA symptoms included checkboxes spanning symptoms of sleep apnea, such as snoring, gasping, and witnessed apnea. Insomnia symptoms included checkboxes for difficulty regarding sleep onset (30–60 minutes or >60 minutes of sleep-onset latency), sleep maintenance (>3 awakenings per night), and insomnia as the reason for the PSG. Although at the time our intake form included “waking earlier than desired”, we have found clinically that, for this question, many false positives were occurring (e.g., work requiring early waking, as not desirable), and thus we did not include in our current analysis. Restless leg syndrome (RLS) symptoms included checkboxes for legs feeling uncomfortable, feeling better with movement, and feeling worse at night. Narcolepsy symptoms included checkboxes for perisleep hallucinations, sleep paralysis, and cataplexy. The intake form was designed to provide basic language symptom inventories, but it has not been independently validated against clinical diagnosis of sleep physician evaluation. We did not include standardized questionnaires for each of the many subcomponents, to strike a balance between information that assists providers in interpreting PSG data and the burden on patients. As the majority (>70%) of patients undergoing PSG in our center are direct referrals (have not seen a sleep specialist in our division before the PSG), we did not have clinical interview-based validation of the symptom reporting.

Statistical analyses were performed using GraphPad Prism 6 software (GraphPad Software Inc., La Jolla, CA, USA).

Results and discussion

In the following sections, we combine the analysis of simulated data with the analysis of a large sample (n=1835) of diagnostic PSG data from our center to illustrate important considerations in analyzing big data in sleep medicine. We can consider four basic categories of information that support clinical phenotyping derived from PSG databases (Figure 1). In each category, methods of cleaning and analysis are implemented, which we discuss in the following sections. Inferential analysis and insights can also be obtained by combining information across categories. For example, correlation and regression analysis can be performed on variables within or between categories, as can more complex predictive analytics be performed using methods of supervised machine learning. Unsupervised learning, also known as clustering, can be applied as well for discovery of novel phenotypes.

Standard PSG metrics and data types

The standard metrics in most clinical PSG reports are readily accessible in sleep databases without requiring off-line extra processing. These include basic demographics and summary statistics of PSG scoring, such as stage percentages, total
sleep time (TST), efficiency, and apnea–hypopnea index (AHI). The importance of standardization in human scoring and basic metrics has been emphasized, especially for multicenter trials and data repositories involving PSG. Clinical metrics should be assessed in several steps to prepare for large-scale analytics.

**PSG scoring annotations**

PSG annotations include technician-scored labels for sleep–wake stage and various events (arousals, limb movements, breathing events) often with time stamps. These data can be exported for off-line processing and/or combining with other sources of clinical information. Aligning these files with exported time series data allows stage- or event-specific analysis of physiological signals.

Event label errors include errors of omission and of commission, and they are best assessed by manual rescoring. Some scoring errors may have indirect effects, such as failure to score an epoch of wake that interrupted a block of REM, which will have the dual effect of missing an awakening and resulting in a larger REM bout duration measurement.

Annotation data are also commonly used for inter-rater reliability analysis, or in groupwise comparisons of technician- or center-level differences in scoring. Because inter-rater reliability for various scoring tasks tends to be in the 80%–85% range, this sets a theoretical ceiling for performance of automated algorithms.

**PSG time series data**

Each channel of a standard PSG is a time series, to which a number of signal processing techniques can be applied to extract information. Initial preprocessing can involve detection and removal of periods with prominent muscle artifact, or removal of ECG signal contaminating electroencephalography (EEG) channels, or de-trending if slow drift is present.

Spectral analysis of the EEG is commonly performed using the fast Fourier transform (FFT) algorithm, which is applied using a moving window to produce an image of spectral characteristics as they change over time. However, the FFT alone provides noisy estimates of the underlying spectral characteristics of the data; thus, it is common to apply spectral and temporal smoothing to improve the estimates. The multi-taper spectral analysis method optimizes the trade-off between retaining fine details (spectral and temporal resolution) while still reducing noise (variance reduction). EEG time series analysis is common in research settings but has not enjoyed similar clinical applications in routine practice, although some clinical acquisition software includes basic frequency analysis options.

The ECG time series has been the subject of extensive analysis of heart rate variability, as well as point-process modeling variants. Another method of ECG analysis, known as cardiopulmonary coupling (CPC), has been suggested to provide an important window into sleep quality beyond that observed in EEG-defined states. Whereas stable non-REM (NREM) sleep is characterized by stable breathing and “high-frequency” coupling (HFC) at the respiratory frequency, processes that disrupt sleep tend to increase low-frequency coupling (LFC). Of particular interest is that treatment-emergent central apnea (and clinical failure of continuous positive airway pressure [CPAP]) was predicted by the degree of narrow band LFC. Whereas obstructive apnea is characterized by the broadband LFC phenotype, chemoreceptor-driven sleep apnea (e.g., central apnea) is associated with narrow band coupling.

**Self-reported clinical information**

Patients undergoing in-laboratory PSG are often asked to self-report symptoms, medical problems, and medications in the questionnaire form. Our center uses a custom form as a basic symptom and history screening tool, which includes the Epworth Sleepiness Scale (ESS), as well as checkboxes and boxes for free-text responses. When self-reporting methods are used, the data require manual or semiautomated review and cleaning before analysis is possible. If medications are listed as free text, spelling errors or nonstandard
terminology (e.g., sleeping pill) requires reconciliation. Internal inconsistencies require attention, such as listing multiple antihypertensive medications but not listing hypertension in the medical history. We found ~75% concordance between listing hypertension from a checkbox selection of medical problems and listing of antihypertensive agents (data not shown). Such a discrepancy could be a simple omission or could be that the patient is on treatment and thus no longer feels they have the disorder.

Combining information categories to inform phenotyping

Using simple combinations of existing metrics, or more involved extractions from the clinical scoring (annotation files), additional data for phenotyping can be generated beyond that which may be available in the acquisition software system. For example, event-related signal analysis, manual scoring annotations, and temporally associated time series data can be combined to explore phenotypes. Several examples of event-specific metrics have been reported, with potential clinical relevance. Chervin et al.22,23 analyzed respiratory event-linked EEG changes to sub-phenotype OSA patients and found a stronger relation with sleepiness by this advanced analysis compared to the usual AHI value. EEG analysis of alpha power and spindle activity has been used to predict arousal response to auditory stimulation delivered during sleep,24,25 reflecting possible biomarkers of sleep fragility. Additional work investigating arousals and autonomic features highlights opportunities to stratify episodic physiological events during sleep that are not currently distinguished in routine scoring.26–32

Database-driven sleep phenotyping

Symptom heterogeneity

We used a convenience sample of n=1835 individuals who underwent diagnostic PSG in our laboratory. In this dataset, symptom combinations were common. Figure 2 illustrates the overlap between self-reported OSA symptoms, insomnia symptoms, and leg-related symptoms in the cohort. The majority of individuals reported more than one of these categories, with less than one-third reporting from only one category. Within the group reporting OSA symptoms, isolated snoring was present in over half, with nearly as many reporting a combination of snoring and either gasping arousals or witnessed apneas (Figure S1). Among those with insomnia symptoms, difficulties with sleep maintenance was the most common isolated symptom, while about half reported more than one insomnia symptom or checked “insomnia” from a list of reasons for the study, along with at least one insomnia symptom (Figure S1). In rare cases, insomnia was listed as the reason for study by the patient, but no insomnia symptoms were checked. Among those with leg symptoms, about one-quarter reported all three symptoms consistent with RLS (uncomfortable sensation while awake, worse at night, better with movement; Figure S2). Narcolepsy symptoms were the least common. The isolated reporting of only one of the three cardinal REM-related phenomena was more common than combinations of any two or all three (Figure S2).

Sleep–wake architecture and fragmentation

Sleep–wake stages are most commonly reported as the number of minutes, and relative percentage of wake, REM, and N1–3. Stage percentage during PSG may be noted in clinical interpretations, and there are normative data available across the lifetime.33 In some settings, such coarse descriptive metrics may be useful. For example, when commenting on the presence or severity of sleep apnea, one might consider the potential for underestimation if the night happened to contain little or no REM, as OSA is often more pronounced in REM sleep (i.e., REM dominant). A relative increase
in N3 percentage may suggest rebound sleep after recent deprivation. Certain medications may alter stage composition (e.g., commonly by reducing REM or N3). Thus, stage percentage might combine with other categories of information, such as the clinical history, rather than provide a basis for actionable clinical recommendations in isolation.

A somewhat more granular approach to sleep architecture is to quantify sleep fragmentation, for example, via sleep efficiency, or by increased time spent in N1 (often because of frequent arousals), or reduced time spent in REM or N3 that may indirectly occur. We can consider the use of sleep efficiency to describe two patients with very different hypnograms, but similar efficiency values. Because efficiency does not distinguish between different patterns of wake after sleep onset (WASO), it runs the risk of lumping together quite different patterns of fragmentation. Figure 3 shows two PSGs with similar sleep efficiency, but which differ by fivefold in terms of the number of transitions to the wake state. In fact, the PSG with greater frequency of wake transitions (Figure 3B) actually has a slightly higher efficiency than the PSG with fewer but longer wake bouts (Figure 3A; 82% versus 78%, respectively). The reasons behind these patterns, the potential clinical impact, and therapeutic considerations may be quite distinct. Figure 3C–G shows the distribution of several metrics in a cohort of n=100 individuals with sleep efficiency values of 79.5%–80.5%. These broadly distributed values are a reminder that a sleep efficiency of “80%” can not only be achieved with distinct patterns of wake but can also be associated with diverse patterns of other potential contributors to (or markers of) fragmentation.

Figure 3 Sleep efficiency: a limited view of heterogeneous sleep physiology.
Notes: (A and B) Hypnograms from different patients. (A) Sleep efficiency of 78.3%, associated with nine transitions to wake (age 23, male). (B) Sleep efficiency of 82%, associated with 46 transitions to wake (age 42, male). The Y-axis indicates scored stage; the time bar indicates 1 hour. (C–G) Distribution of n=100 individuals with essentially the same sleep efficiency values (79.5%–80.5%), but differ widely across other factors that potentially contribute to or signify sleep fragmentation (age, AHI, PLMI, # W, and N1%).
Abbreviations: R, rapid eye movement sleep; N1–3, non-rapid eye movement stages 1–3; AHI, apnea–hypopnea index; PLMI, periodic limb movement index; # W, number of wake transitions.
Stage percentage is also an insensitive measure of fragmentation. We can consider two patients each with 120 minutes of stage REM within an 8-hour TST: one patient could have four REM blocks each lasting 30 continuous minutes, and the other could have four blocks interrupted by frequent brief transitions to wake or N1. Each patient’s summary report could indicate REM as 25% of TST, yet the two patterns of consolidation versus fragmentation are quite different and might imply distinct pathophysiology. This example illustrates that stage percentage does not capture fragmentation phenotypes associated with a known cause of fragmentation (OSA). Several alternative methods to stage percentage have been proposed, including bout duration histograms,\(^{39}\) bout duration survival analysis,\(^{40}\) and others.\(^{41-44}\) Some have used multi-exponential transition models,\(^{45}\) while others have used power-law approaches\(^{46}\) to describe these skewed patterns. Even stage transition rates have proven useful where percentages have shown no discriminatory value.\(^{47,48}\) Which one of these is the “best” descriptor of the distribution of sleep–wake stages remains open to discussion, although even statistically principled model selection methods may not distinguish true from alternative functions in simulation studies.\(^{49}\)

If stage percentage cannot distinguish individuals with no OSA from those with severe OSA (despite the obvious fragmentation seen visually), then it might be even less sensitive for comparing groups or evaluating interventions expected to have less dramatic impact on sleep. For example, in a study of yoga in healthy adults, distribution analysis revealed stage differences not evident by percentage analysis.\(^{50}\) Likewise, stage percentage does not appear to distinguish patients with versus without misperception of TST, whereas differences were evident when stages were examined using bout distribution methods.\(^{51}\) One wonders how often initial analysis of stage percentage reveals little or no group differences, and deeper analysis of fragmentation is simply not pursued.

**Phenotyping sleep apnea**

The summary metric most commonly acted upon in clinical practice is the AHI, which is used to define the presence and severity of OSA. This event rate has become the cornerstone of diagnosis, a threshold index for insurance coverage of therapy, and a metric for inclusion and outcome of research trials. However, the OSA phenotype is much more heterogeneous than differences in AHI values might suggest, even if we put aside desaturation thresholds for scoring hypopneas\(^{52-55}\) and the potential for night-to-night variability.\(^{56-61}\) and other anatomical and physiological contributors.\(^{62-64}\) OSA phenotypes can be described by extracting further details from routine PSG. For example, the severity of OSA often depends on sleep stage, and on body position, although a single night of PSG recording may not contain sufficient time in the different combinations of stage and position to make this determination.\(^{65}\) Figure 4A illustrates an example of severe hypoxia to <60% during REM in a highly REM-dominant case, despite categorization as normal (AHI, 4.7) when the event rate is calculated over the full night. Figure 4B illustrates a strongly supine-dominant case, with normal AHI while sleeping in the lateral position, and very severe AHI while sleeping supine. The full-night AHI is the weighted average of these extremes, which happened to be 19.2 on this night. Had the person slept supine the whole time, or lateral the whole time (or if positions were not recorded), then very different conclusions about the presence and severity of OSA would likely be drawn. In this case, it is also interesting that REM dominance could not be assessed as only lateral REM was seen, and no apnea was present while lateral.

Which AHI is most relevant depends on several factors. For example, in a study of airway anatomy while supine in a scanner, the supine AHI might be most informative even if the individual never sleeps supine in the home. By contrast, in a study of clinical outcomes, the real-world AHI experienced by the patient is paramount: if the patient sleeps exclusively non-supine (and this can be demonstrated), then the lateral AHI is the relevant “phenotype” for that individual.

Characterizing supine dominance also has direct implications for clinical care. Patients with strong supine dominance may benefit by pursuing positional therapy. Much work exists in this area,\(^{66}\) and devices to assist in position therapy exist in the consumer and prescription\(^{67}\) spaces. Device-assisted therapy is important, especially because patients’ self-report of body position during sleep carries substantial uncertainty.\(^{58}\) By contrast, REM dominance does not as easily translate into clinical care recommendations for therapy, but REM-dominant OSA has been increasingly linked to hypertension,\(^{69}\) and thus might impact treatment motivation. Insufficient evidence exists regarding REM-suppressing agents as pharmacological therapy for OSA.\(^{70}\)

The heterogeneity in clinical features is apparent by examining a distinct set from our database with AHI in a very small range, 30–35 (n=100). In this group of very tightly clustered “severe” AHI cases, the age, body mass index (BMI), supine AHI, periodic limb movement index (PLMI), and central apnea index (Figure 4C–G) are each quite broadly dispersed. In addition, the distributions do not visually suggest obvious cutoffs or subgroups. In each case, clinical decisions might be distinct depending on where in...
the dispersion an individual resides in each category, and across categories; similar clinical sub-phenotypes have been discussed recently. For example, an AHI of 30 in a slender 25-year-old with no periodic limb movements of sleep (PLMS) and a high supine AHI might have different treatment options or preferences (not to mention risks and outcomes) than an older obese patient with comorbid PLMS and increased central component. Clinically and in many research settings, severity categories span much larger AHI ranges, and are thus likely to have even more heterogeneity across these and other potentially important phenotypic axes (medications, alcohol, airway anatomy, etc).

**Phenotyping insomnia**

Insomnia is clinically defined entirely by self-reported symptoms. While research efforts impose cutoffs for sleep latency or WASO as inclusion criteria, in clinical practice the emphasis is on the severity of the complaint and the self-reported impact on daytime function rather than on numerical requirements of sleep parameters. Even in research...
settings, it can be challenging to demonstrate objective impact on daytime function, a reminder that chronic insomnia is not phenomenologically equivalent to experimental sleep restriction in healthy adults and the ensuing performance decrements. There is growing interest in using objective measures to study insomnia, with respect to the hyperarousal pathophysiology, as well as recent work indicating that it is the combination of insomnia and objective short sleep duration on PSG that is specifically associated with incident medical and psychiatric risk.

Despite the clinical reliance on self-report, extensive work highlights the challenges associated with the subjective experience of insomnia. As an example, the seemingly simple question of sleep duration, which forms the basis of nearly all epidemiological sleep research, depends on the demographics, comorbid psychiatric disorders, and comorbid sleep disorders. In addition, within-individual analysis reveals some striking observations that when and how sleep–wake durations are queried impacts patient responses. The prospect of internal inconsistency across query methods remains an important yet unresolved issue.

Prolonged sleep latency is a common complaint, and although it may seem a straightforward metric, it carries special challenges when understanding insomnia and specifically the misperception phenotype. Objective sleep latency measurement requires an operational definition, for which there is no gold standard. Although prior literature considered behavioral (non-EEG) approaches to identify sleep onset, clinical reporting usually involves defining sleep onset by either the first epoch of any sleep or the first instance of a consolidated bout (e.g., 10 epochs) of uninterrupted sleep. Different definitions impact calculations and therefore experimental results. We can consider a patient with delayed sleep phase syndrome, who exhibits a 2-hour latency, but subsequent sleep was well consolidated, compared to an individual who spends the first 2 hours with fragmented brief transitions between wake and sleep, perhaps due to pain, and also has a 2-hour onset latency to persistent sleep. It is difficult to rationalize lumping these together under a definition of latency to persistent sleep (both are 2 hours).

For studies of misperception, the subjective sleep latency is compared to some definition of objective sleep onset; clearly, the definition of objective onset may impact the resulting calculation. We recently introduced a novel metric of sleep onset misperception that obviates the need to define objective sleep onset. The fundamental goal of quantifying sleep onset misperception is to capture how much sleep was misinterpreted as wake, and thus we calculated the total sleep duration occurring during the time between lights out and subjective onset. This also addresses a potential confound of assuming that onset misperception and TST misperception are independent. We can consider patients with objective sleep of 8-hour duration, with a 1-minute onset latency, who report subjectively a 4-hour onset latency and 4 hours of TST. Typically, these persons would be labeled as having both onset and total sleep misperception (4 hours each). However, if they had anchored their total sleep estimate to their own sleep latency estimate of 4 hours, then their 4-hour total sleep guess is an accurate estimate of TST occurring since they believed that they fell asleep. We recently showed that a substantial portion of patients would be reclassified if their misperception phenotype is based on the sleep during subjective latency, and the “corrected” total sleep misperception. Big datasets may allow further evaluation of misperception phenotype(s), which have not enjoyed consistent predictors in the prior literature.

**Analysis and inference**

**Missing and erroneous database entries**

Routine clinical data can be easily arranged in tabular format to facilitate an initial data evaluation. When these metrics are exported into spreadsheets with columns of features (and each row is one patient’s data), some straightforward cleaning methods can be implemented (Figure S3). Minimum, maximum, and counting commands can identify columns with missing data (e.g., count if empty), improperly formatted data (e.g., count if text is present), or implausible values (e.g., count if outside limit value). In some cases, such outliers or errors would be missed in routine plotting such as bar plots with standard deviations (SDs) or even box and whisker plots depending on whether outliers are plotted and how the axis ranges are chosen.

Several reasons for missing values are possible, including corrupted data (data were collected but were no longer accessible), collected but not recorded (paper copies fail to transfer to electronic database), and not collected. Some qualitative assessment of the distribution of variables from individuals who are missing at least one other entry in a data matrix can be useful. Specific decisions regarding how to handle missing data points or error values are best handled by prespecifying a plan, which could involve excluding subjects or imputing missing values; more advanced discussions are available.
In some cases, missing values occur for appropriate reasons and the absence can be informative. For example, the REM AHI cannot be calculated if no REM sleep was observed during a PSG. Likewise, position dependence of OSA cannot be calculated for PSGs in which only supine position was observed. In these cases, removing such subjects might be favored over imputation.

Determining whether a given value is erroneous (versus a biologically plausible outlier) may depend on certain clues such as “impossible” values or known placeholders for missing data used by an acquisition software system. For example, negative values where only positive values are possible (e.g., age) are easily identified as both errors and outliers. Likewise, when letters are present instead of numbers, or where the value is out of range (e.g., ESS value of >24), these are also easily identified. In some examples (Figure S3), simply plotting the data identifies outliers. Database software such as Excel can easily show the maximum and minimum values for inspection of implausible values. As an example of errors not readily detected by the abovementioned methods, in our database the BMI and ESS are manually entered in adjacent fields, such that an out-of-range ESS value prompted inspection of the BMI as well, and in some cases it was shown by viewing the original data that these two values were interchanged (in this instance, the BMI value of 18 is plausible, and so it would not have been flagged as an outlier).

In some cases, we may still wish to exclude plausible data from analysis. Examples are related to stage- and position-specific metrics, wherein the amount of time spent in the condition of interest is the “sampling” problem, rather than the number of subjects. We may wish, for instance, to exclude people with minimal time in REM or minimal time spent sleeping supine, not just the zero time individuals, when calculating OSA dominance ratios or oxygen nadirs in REM. For AHI, the values could be artificially high (one apnea in one epoch), or artificially low (insufficient time spent in REM to manifest obstructions).

**Distributions and plotting**

Evaluating the distribution of individual variables can inform multiple aspects of analysis and inference. The most basic reason to understand the distribution is to decide what kind of statistical approach is most appropriate, such as whether continuous data are normally distributed or skewed in some manner, in which case data transformations to make the data distribution approximately normal (e.g., logarithmic transformation of positive-valued data) or nonparametric analysis methods may be preferred. Moreover, like plotting the raw data, evaluating the distribution using one of several techniques can also inform the approach to outliers, or the possibility of interesting biological heterogeneity. For example, multimodal distributions may imply that the population contains different sub-phenotypes that might be worthy of further investigation.

In our cohort, none of the variables passed statistical testing for normality, similar to prior work using the Sleep Heart Health Study database. Of note, large samples may be highly “powered” to reject the null hypothesis of a normal distribution, even when the distribution appears nearly normal. Conversely, small samples are more likely to pass tests of normality, even if known to be non-normal. Indeed, when we under-sampled the current dataset, there was increasing probability of passing tests for normality (Figure S4). Non-normal data can be handled by either nonparametric methods or transformation that render the data approximately normal. The challenge is as much statistical as biological: non-normal distributions may have phenotyping implications.

The method of plotting can impact the viewer’s impression of the data. Bar graphs with mean and SD or standard error of the mean (SEM) are commonly used, but these routine methods risk inadvertently concealing potentially important information. Figure 5 illustrates different plotting methods for groups of simulated data from known distributions. In the case of bar plots with SEM, casual inspection might give the false sense of reduced variance of the actual observations in the dataset (Figure 5A and B). This happens because the SEM is obtained by dividing the SD by the square root of the sample size, which makes error bars smaller. The SEM thus does not reflect variance in the data, but rather reflects the precision of the estimate of the mean value – one should not conflate the two.

The SD, in contrast, reflects the dispersion in the data, and does not diminish with increasing sample size like the SEM. However, the SD may still be misleading in a bar graph when it is constructed from data with a non-normal distribution. Because the SD is by convention shown as symmetric bars around the mean (regardless of the actual underlying data distribution), viewers may be left with the potentially false impression of symmetric spread around the mean simply because of the display convention (Figure 5C). Sometimes the only clue in a bar graph that the population is skewed is that the SD value is greater than the mean value for a dataset that cannot take on negative values, which implies a long tail (i.e., non-normal). This is common, for example, in known
skewed distributions such as AHI or sleep latency, where a value of, say, 15±18 would be interpreted as evidence of a long-tail non-normal distribution. Asymmetries and skew are visually evident in box and whisker plots (Figure 5D). However, even the box and whisker method can “hide” the distribution for unusually structured data, such as bimodal distributions, which would be phenotypically important to recognize.

Other techniques for visually assessing structure in populations include frequency histograms and cumulative
distribution functions (CDFs). Histograms are used in Figure 6 to illustrate the distribution of TST, sleep efficiency, and the number of ≥30-second awakenings. The choice of bin size for histogram plots should consider the trade-off between granularity of the variable of interest, and sample size per bin. Too many bins cause the variable values to be either 0 or 1 in each bin, or to vary randomly from bin to bin due to sampling noise, and therefore offer little visual insight. Too few bins cause the underlying distribution to be overly smoothed. Histogram views can reveal outliers, suggest heterogeneity of the population, or inform selection of cutoff values (e.g., if a “valley” was seen between two modes within the data, suggests two populations). In contrast, the histograms shown in Figure 6 do not have clear “valleys” on visual inspection.

CDF plots can also be informative, especially when comparing groups, or when the metric of interest is a threshold imposed upon a continuous variable. Unlike histograms, CDFs do not require specification of bin size; however, their visualization may be less intuitive. Figure 7 shows CDF plots for different sleep apnea metrics, such as position dependence of the AHI (Figure 7A) and of the central apnea index (Figure 7B). Figure 7C shows the distribution of time spent in different body positions during sleep. Threshold values can be evaluated visually, such as the portion of the population with at least 50% of the night spent supine (Figure 7C; ~60%), or who had a supine AHI value >30 (Figure 7A; ~20%).

Correlation analysis
One of the powerful approaches enabled by large datasets is investigating correlations between variables. Nonparametric (Spearman) correlation was performed between AHI and BMI, which are well known to be positively correlated. Taking the full cohort, the unadjusted Spearman’s $R$-value is $\sim 0.25$. Figure 8A shows the distribution of $R$-values for AHI versus BMI obtained when repeatedly analyzing randomly selected smaller subsets of the cohort. For the subgroups of the cohort, the range of $R$-values is larger for smaller sample sizes. In other words, smaller samples of the large cohort (n=1835) show much larger range of correlation values than the value of the whole set ($\sim 0.25$). This variation includes more extreme $R$-values such as actually negative correlations in some cases (for the subsets of size n=10, n=20). Similar patterns are observed with another pair of parameters that showed a positive correlation in the large cohort (age and PLMI; Figure 8B).
To emphasize the risks of under-sampled (small) data resulting in spurious findings using correlation analysis, we also demonstrate the $R$-values obtained when pairing REM% from the cohort with a vector of random numbers (Figure 8C). This plot clearly shows that significant correlations can occur, even with convincingly large $R$-values, with random data. These plots illustrate the concept that extreme values for statistical estimates (such as correlation coefficient or mean) are more common in under-sampled data. Most investigators reflexively think of “power” in the sense that lack of statistical significance when a true difference exists (type 2 error) could be a symptom of insufficient sample size. However, small sample sizes also harbor false-positive risk (type 1 error).

Large datasets can mitigate false-positive risks associated with the issue of small numbers mentioned earlier, except when the dataset is parsed into smaller and smaller subsets in data-mining queries of ever more specific subsets. Across all 41 continuous variables in the database, the threshold $R$-value for meeting significance was quite small when calculated from the entire cohort. The threshold for obtaining a significant $R$-value increases as progressively smaller subsets of the cohort are considered. Figure 8D illustrates how the “power” of the Spearman correlation calculated between any two pairs of variables decreases as the sample size decreases. In other words, when the entire cohort is considered, even quite small $R$-values in pairwise correlations meet significance criteria, because the large size essentially provides power to detect small correlations as significant. By contrast, small samples provide insufficient power to detect small correlations as statistically significant, and thus only large $R$-values meet significance criteria. This latter issue creates an interesting conundrum: because only large $R$-values can be significant when small datasets are considered, any significant correlation (whether true or false in reality) will necessarily have compelling-appearing $R$-values, which may overestimate the true $R$-value of that pair of variables, had a larger sample been utilized.

Sometimes, we may have prior information to help mitigate false inferences. Given the strong known relation between BMI and AHI, insignificant or paradoxical (negative associations) can be interpreted as false findings.
However, in other cases, we may not have the benefit of strong prior knowledge, so assessing new data becomes more challenging.

**Inferential pitfalls of small and large sample sizes**

We have seen through empirical analyses and simulations that a spectrum of information and pitfalls are possible when working with large datasets. We now turn to illustrative examples from the literature in which large datasets may not be as explanatory as they appear. While there are many examples from which to choose, these example situations are representative of some key challenges.

**Situation 1: when big data are still under-sampled**

A recent study of more than 50 million pregnancies in the US sought to correlate adverse maternal and baby outcomes with OSA. The study used billing codes from a massive registry to assign case labels for OSA. By this method, the prevalence of OSA in pregnancy was 3 per 10,000, approximately 100-fold lower than expected in this demographic. The discrepancy raises the possibility that the OSA coding is not just underestimating prevalence, but may also be biased, for example, toward the most severe or most symptomatic cases. If so, implications of any results based on these data greatly shrink in scope, as it they would apply only to the most severe cases of OSA or the most vulnerable or symptomatic individuals.
Situation 2: when big data explain little
While large sample sizes increase power, an associated risk involves the potential for being “overpowered” to detect very small correlations or group differences. We can observe this in the effect of modafinil on sleep latency of shift workers, or in the relation of OSA severity with ESS. A striking example of a very large sample supporting very small effects can be found in the analysis of mood rhythms detected in word analysis of more than 500 million tweets. Relative change in day length was significantly related to positive affect, with an R-value of 1.2×10⁻³, suggesting that the rhythm explained a fraction of a percent of affect fluctuations. These observations highlight the well-known but under-practiced mandate to focus scientific investigations on determining effect sizes, causal relationships, and establishing practical or clinical relevance, rather than focusing on simplistic binary questions at the heart of statistical significance testing.

Situation 3: when big data are misinterpreted
The largest published study of home sleep testing was recently published, with the stated goal of determining if home testing was being used clinically in accordance with AASM standards. The sample size of nearly 200,000 home tests is orders of magnitude larger than any prior home testing report. The authors concluded based on a high posttest probability (~80%) that indeed testing was in line with AASM guidelines. However, the AASM recommends that pretest probability (not posttest probability) should be >80% for at least moderate OSA (AHI >15). Bayes theorem tells us that the pretest probability of AHI >15 was <10% in the published cohort (and 50% if AHI >5 threshold is used), and thus the data actually support the opposite conclusion to that reached by the authors: home testing for OSA is being used too liberally, and not in line with the AASM guidance.

Another recent article using administrative data from more than 2000 patients to derive a screening algorithm for OSA cases failed to recognize, by Bayes theorem, that their algorithm’s sensitivity and specificity were indicative of chance performance. One always needs to consider both sensitivity and specificity when evaluating any test. We use a simple calculation, the “rule of 100”, which can avoid this statistical fallacy: if the sensitivity and specificity of a test add to 100%, the probability of disease is unchanged by the result of the test (i.e., chance performance).

Conclusion
Clinical databases have important strengths that can support big data research goals. Clinical data contain diversity and heterogeneity that may be specifically excluded in clinical trial databases, which are often designed to reduce sources of variability that can be detrimental to power calculations and outcome testing. Clinical databases are more likely to reflect “real-world” variation in clinical phenotypes. This can be important for testing whether predictive algorithms can generalize across a diversity of clinical phenotypes. In addition, heterogeneous sets may be more amenable to clustering and other exploratory methods that allow discovery of new phenotypes that can be explored in subsequent prospective studies. From a resource utilization standpoint, clinical databases are a natural extension of already acquired data supporting patient care, which allows valuable and limited resources to be applied at the analysis phase.

Despite these advantages, certain limitations must be recognized. Academic centers may have different referral biases, for example, being enriched for complicated cases. Although most clinical laboratories have standardized physiological recording protocols, the collection of self-reported clinical information may not be standardized. Variation across recording and scoring technologists may contribute heterogeneity despite quality efforts required in accredited laboratories. Centralized scoring common to large clinical trials may not be practical for clinical databases.

Large sleep datasets offer the opportunity to pursue complex phenotyping exploration, and to detect scientifically or clinically interesting differences or patterns in health and disease. Despite the clear advantages, analysis of big data in sleep medicine also carries risks. Understanding common pitfalls can help mitigate the risks, whether one is conducting the analysis or reviewing publications involving big data. Ideally, what is learned from population-level big data efforts can then inform individual clinical care decisions. In an era when insurance restrictions are driving at-home limited channel alternatives, these efforts will be critical to elaborate and justify the current and possibly more advanced future use of PSG for clinical care. The era of big data in sleep medicine is poised to provide unprecedented insights, especially as it coincides with massive shifts in reimbursement and availability of laboratory-based PSG.

Disclosure
Dr Matt T Bianchi has received funding from the Department of Neurology, Massachusetts General Hospital, the Center for Integration of Medicine and Innovative Technology, the Milton Family Foundation, the MGH-MIT Grand Challenge, and the American Sleep Medicine Foundation. He has a pending patent on a sleep wearable device, received research
funding from MC10 Inc and Insomniosolv Inc, has a consulting agreement with McKesson Health and International Flavors and Fragrances, serves as a medical monitor for Pfizer, and has provided expert testimony in sleep medicine. Dr M Brandon Westover receives funding from NIH-NINDS (1K23NS099000), the Rappaport Foundation, and the Andrew David Heitman Neuroendovascular Research Fund. The authors report no other conflicts of interest in this work.

References


43. Klerman EB, Davis JB, Duffy JF, Dijk DJ, Kronauer RE. Older people awaken more frequently but fall back asleep at the same rate as younger people. Sleep. 2004;27(4):793–798.


Supplementary materials

**Figure S1** Overlap of symptoms associated with sleep apnea and insomnia.

**Notes:** (A) Venn diagram of symptoms related to sleep apnea: snoring (solid line, blue fill), gasping arousals (dashed line, red fill), and witnessed apnea (dotted line, yellow fill). The n-value (sample size) for each category: snoring only = 656; snoring and witnessed apnea = 337; gasping and snoring and witnessed apnea = 166; snoring and gasping = 70; gasping only = 34; witnessed apnea and gasping = 13; witnessed apnea only = 61. (B) Venn diagram of insomnia symptoms: onset (dashed line, blue fill), maintenance (solid line, yellow fill), and listing insomnia as the reason for PSG (dotted line, red fill). The n-values for each category: onset only = 148; onset and maintenance = 331; maintenance only = 521; maintenance and listing insomnia as the reason for PSG = 121; listing insomnia as the reason for PSG but no other symptoms were indicated = 10; onset and maintenance insomnia and listing insomnia as the reason for PSG = 288; onset and listing insomnia as the reason for PSG = 38.

**Abbreviation:** PSG, polysomnography.

**Figure S2** Overlap of symptoms associated with restless legs and with narcolepsy.

**Notes:** (A) Venn diagram of symptoms related to restless legs: uncomfortable sensation in the legs (solid line, blue fill), better with movement (w/mov’t) (dotted line, red fill), and worse at night (dashed line, yellow fill). The n-values (sample sizes) for each category: uncomfortable sensation alone = 94; uncomfortable and better with movement = 72; better with movement alone = 26; uncomfortable and worse at night = 33; better with movement and worse at night = 90; worse at night alone = 42; uncomfortable and worse at night = 49. (B) Venn diagram of narcolepsy symptoms: peri-sleep hallucinations (dashed line, blue fill), sleep paralysis (dotted line, red fill), and cataplexy (solid line, yellow fill). The n-values for each category: hallucinations alone = 77; hallucinations and cataplexy = 11; cataplexy alone = 82; hallucinations and sleep paralysis = 21; hallucinations and cataplexy and sleep paralysis = 17; cataplexy and sleep paralysis = 30; sleep paralysis alone = 59.
Figure S3 Assessing missing data.

Notes: (A) An example of missing data, outliers, and data reversals (subject code MGH24: the BMI and ESS scores are switched, but the error codes are only implausible for ESS), indicated by gray shading. Column statistics (maximum, minimum, count if text, and missing cell entries) can be helpful to alert potential anomalous data. (B) The age variable from (A) represented as a bar plot with SD, a box and whisker plot, and a dot plot; the outlier is not evident in the bar with SD. (C) The BMI variable from (A); similarly, the presence of an outlier is not evident in the bar with SD, and none hint at the switch with ESS because the erroneous value was plausible. In the Sex column in (A), 0 = female and 1 = male.

Abbreviations: BMI, body mass index; ESS, Epworth Sleepiness Scale; m, male; Max, maximum; Min, minimum; SD, standard deviation; Subj, subject; TST, total sleep time.
<table>
<thead>
<tr>
<th>Variable</th>
<th>n=1800</th>
<th>n=600</th>
<th>n=200</th>
<th>n=100</th>
<th>n=50</th>
<th>n=20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age, years</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>BMI</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>ESS</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>TST (min)</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Efficiency</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>LPS</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>+</td>
</tr>
<tr>
<td>N1%</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>N2%</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>+</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>N3%</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>REM%</td>
<td>–</td>
<td>+</td>
<td>+</td>
<td>–</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>#W ≥30s</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Sup%</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>AHI</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>AHI NonSup</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>AHI Sup</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>AHI REM</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>CAI</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>CAI Sup</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>CAI NonSup</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Min O₂ NR</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Min O₂ REM</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>+</td>
</tr>
<tr>
<td>Mean HR</td>
<td>–</td>
<td>–</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>PLMI</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Spont AI</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>+</td>
</tr>
</tbody>
</table>

**Figure S4** Normality testing results vary by sample size.

**Notes:** The variables listed were tested for normality (D’Agostino–Pearson test), with “–” indicating failed testing for normality and “+” indicating passed testing for normality. The columns indicate the sample size of random subsets of the full dataset, with none passing when the sample size was 1800.

**Abbreviations:** AHI, apnea–hypopnea index; AHI NonSup, apnea–hypopnea index in non-supine; AHI Sup, apnea–hypopnea index in supine; BMI, body mass index; CAI, central apnea index; CAI NonSup, central apnea index in non-supine; CAI Sup, central apnea index in supine; ESS, Epworth Sleepiness Scale; LPS, latency to persistent sleep; mean HR, mean heart rate; Min O₂ NR, minimum oxygen in non-REM; N1–3, non-REM stages 1–3; Min O₂ REM, minimum oxygen in REM; PLMI, periodic limb movement index; REM, rapid eye movement; Spont AI, spontaneous apnea index; Sup%, supine percentage; TST, total sleep time; #W ≥30s, number of wakess >30 seconds.