Guided Self Testing: a quantitative framework for sleep-wake pattern discovery, intervention planning, and outcome assessment

Balaji Goparaju, MS, and Matt T. Bianchi, MD PhD

Neurology Department, Massachusetts General Hospital, Wang 720, Boston, MA 02114

Correspondence to: Matt Bianchi, thebianchi@gmail.com

Abstract

Self-tracking is a common approach to understanding and modifying aspects of mental and physical health. However, discovering patterns in a manner suitable to inform actions intended to improve some aspect of health or performance may be challenging especially if many potential factors are involved. Sleep-wake patterns represent a typical use case, and could potentially benefit from a framework to move beyond qualitative interpretation of patterns in tracking data. Different aspects of sleep quality and quantity can impact next-day function, and different aspects of daily experience and behaviors can impact the quality and quantity of next-night sleep, creating potentially complex feedback loops and embedded correlations. We introduce a quantitative framework for planning, diary tracking, basic and advanced analytics, including power calculations, designed to facilitate pattern discovery as illustrated in the use case of sleep-wake patterns. The statistical principles can generalize to any setting described by variability of input and outcome factors over time.

Introduction

Pattern recognition in our day-to-day lives can be an important aspect of optimizing certain aspects of health, wellness, and performance. This is particularly relevant for sleep, which can impact many aspects of daily life, and conversely, waking experiences and behaviors can impact subsequent sleep. Confident pattern discovery in a person’s sleep-wake experience is hindered by two main challenges: sleep quality and quantity can each be “cause” and “effect”, and relationships between factors are probabilistic rather than deterministic. Pattern recognition in what amounts to a noisy system of embedded cycles of influence is understandably difficult by qualitative inference, whether by diary or by recollection of patterns without explicit tracking efforts. These challenges are directly relevant for real-world investigations of sleep-wake patterns, whether in healthcare or consumer contexts. In clinical practice, for example, a typical sleep diary might be obtained for 1-2 weeks, and while some guidance is available on which variables the diary should contain, neither the American Academy of Sleep Medicine practice guidelines for chronic insomnia, nor the consensus statement on sleep diary tracking, suggests analytic methods for interpreting the data or even a suggested duration of tracking\textsuperscript{1-3}. In the consumer space, tracking devices include software intended to facilitate pattern recognition to improve sleep, but neither the methodology nor the performance in real-world settings is transparent.
In addition to the variability in sleep and wake factors that may occur for a given individual over time, inter-individual heterogeneity is also well described in sleep and circadian rhythms\(^4,5\). For example, heterogeneity has been shown in the vulnerability of workplace performance to sleep deprivation\(^6\) including in the medical profession\(^7,8\). Other examples include vulnerability to sleepiness in sleep apnea\(^9\), and sleep-wake perceptual disconnects in insomnia\(^10\), sensitivity to caffeine\(^11\), alerting effects of naps\(^12,13\), effect of exercise on sleep\(^14\), and the sedating effects of alcohol\(^15\).

Sleep is of course itself multi-dimensional, and different dimensions may have different importance for an individual's subjective experience, such as latency to sleep onset (SL), total sleep time (TST), number of awakenings (#W), etc. Although commonly assessed via self-report, these dimensions can also be objectively measured with sleep monitoring devices\(^16,17\). Within-individual differences between these objective and subjective measures may be of interest, especially in the context of insomnia\(^10,18,19\). Likewise, waking behaviors and experiences are also multi-dimensional. For example, exercise can be sub-categorized by time, duration, intensity, and type; naps may vary in timing and duration; caffeine and alcohol may vary in timing and quantity. These and myriad other sleep and wake factors exist in the context of individual heterogeneity in mental and physical health status, resilience to sleep deprivation, medication use (and vulnerabilities therein), stress, life obligations, and performance requirements.

Overall, this complexity creates a dimensionality problem: the more factors involved, the more difficult to tease apart patterns by qualitative review, and the longer tracking is required for quantitative review. Short term diary tracking in particular suffers from low power to detect correlations, multiple comparisons challenges in proportion to the number of factors being tracked, and embedded correlations among factors that could impact sleep (e.g., caffeine only consumed on weekdays, which are also the high stress days). To the extent that formal clinical trials mitigate some of these challenges by enrollment criteria and protocol requirements, the applicability to individuals in the real world is potentially reduced. To the extent that real-world practical information is desirable, motivated individuals might rationalize devoting effort toward actionable self-discovery. Thus, we introduce a framework, which we term "Guided Self-Testing" (GST). GST uses concepts of statistics commonly applied in randomized controlled trial (RCT) settings (power calculations, significance testing), but tailored to the individual discovery goals. To be maximally general, the system does not involve placebo comparisons, which require research protocols, and thus GST is distinct from so-called n-of-1 trial designs\(^20-26\).

The Guided Self Testing Platform

GST consists of a web-based software with diary and analytics components that can be viewed on any computer or smart phone (www.guidedselftesting.com). The user can interact with the software at different levels, tailored to the resource bandwidth and interest of the user, ranging from very limited tracking and simple analytics to detailed tracking and more advanced analytics. Basic statistical concepts of variability, power, and correlation are used, though with simplifications aimed to strike a balance between the qualitative views typical of personal tracking and clinical practice, and the statistically strict assessments typical of formal research studies.
Fundamentally, qualitative assessment of diary tracking data intends to identify associations or correlations. In this sense, it can be considered exploratory. We adopt a basic approach of computing pairwise correlation coefficients between tracked variables. Most real-world tracking will differ between individuals, not only in the number of variables and their potential relationships, but also in the statistical structure of any given variable. For example, some might be continuous variables like SL or TST (in minutes or hours), while others might be categorical (e.g., yes or no for exercise) or a limited integer value (for, say, number of drinks with caffeine or alcohol). Even the statistical distribution for a given variable may differ between individuals (normal versus skew, bimodal, etc). Strict application of statistical tests involves assumptions that may or may not be known for an individual, especially for short duration tracking (e.g., 1-2 weeks), including knowledge of the distribution, variance, collinearity, etc. In GST, for simplicity, we perform a basic correlation coefficient for all pairwise variables, regardless of their structure. This approach has risks: the failure of parametric or non-parametric correlation coefficients to detect complex or non-linear patterns is well described, as are other paradoxes involving outliers and heterogeneity violations (e.g. Simpson's Paradox). Thus, performing a simple Pearson correlation coefficient represents an approximation of possible relationships, and could suffer from false positives and false negatives. Ideally, sufficient longitudinal data would be collected to understand the distribution of each variable, and of collinearity and other features. Although this is unlikely in real-world settings, those particularly motivated to obtain extensive measurements could perform more advanced statistics.

Preparation: considerations before customizing the tracking plan

One of the core principles of quantifying patterns obtained through self-tracking is to understand the trade-offs associated with the number of factors tracked, how much variability occurs in each, the willingness to track for extended periods of time (weeks or longer), and the goals of tracking. That which is not tracked will not be discovered, so the more factors one chooses to track, the more possible patterns can be explored. However, tracking more factors may become tedious and reduces compliance, while also raising concerns about multiple comparisons as it pertains to the false positive risk. The duration of tracking impacts the power of inferences, such as the statistical confidence in the strength of a relationship. In the case of correlation analysis, this means the p-value associated with each correlation’s R-value. We can consider examples of individuals with different self-testing goals, each of which will engage in tracking in a different manner.

First, consider a person who is interested in short term tracking, who does not have many variables changing from day to day, who generally sleeps well, and the goal is simply the curiosity of exploration. This person might perform tracking for one week, with a simple diary that includes only two or three awake variables (say, exercise and caffeine), and a single sleep factor (say, TST). In general, if daytime factors are fairly constant day to day, and sleep quality can be tracked with one measure (e.g., TST), then if an intervention is planned, tracking is less involved and inferences can be drawn with shorter tracking durations.

Second, consider a person who is experienced with self-tracking in several areas like diet and exercise, who now wants to add a sleep component because they have intermittent difficulty with insomnia, and who has many ideas for possible factors with the motivation for
extended tracking periods. This person might design a more detailed diary, with 5 or 10 waking factors, as well as several sleep factors, and then begin a multi-week tracking period with weekly analysis and potential interventions along the way.

Third, consider a person with chronic insomnia who is looking to find patterns that predict good versus poor sleep from night to night, but also wants to try some herbal supplements with a structured plan in mind. This person might begin by performing some baseline tracking to understand candidate waking factors most relevant to their life (e.g., naps, diet, exercise, caffeine, alcohol). That baseline might suggest behavioral interventions to reduce variability before starting a trial of herbal supplements, or at least provide a picture of the variability of waking and sleep measures to plan the goals of the herbal trial, which informs how long it might be tried. Then, the baseline also sets the stage to evaluate the extent of benefit.

**Customizing the tracking plan**

After consideration of the resource bandwidth and tracking goals in the preparation phase, the next step is selecting the sleep and wake factors to track. For each factor, a unit of measurement is considered. Some metrics have a natural unit (like #W or minutes of SL), while others, especially awake factors, might have distinct sub-variables to consider if felt to be important. For example, the variable of exercise might be tracked as binary (present or absent) or could include features such as the time of day, duration, or intensity. The GST software offers examples of common sleep and awake factors from which to choose, as well as an option to create custom sleep and awake factors. As patterns are evaluated over time, the individual can change and adapt the sleep and awake factors, as described below. The software can also accommodate entry of data obtained from objective measurements of sleep or wake factors, through adding custom variables, but currently does not interface directly with other devices.

**Inferences from first week of diary tracking**

We suggest beginning with a 7-day tracking period, as a minimum. This ensures sampling of both weekday and weekend time, and is roughly the minimum duration for statistical power to detect very large correlations (R values of ~0.9). For many individuals, one week of tracking will represent too under-sampled of a view. However, as a baseline minimum, one week of data can illustrate conceptual topics and inform decisions about continued monitoring. More than 7 days is encouraged, and we suggest analysis not even be attempted until at least 7 days, and then in 7-day increments to avoid responding to ultra-short term variation.

Analysis involves correlation calculations in four distinct categories: 1) between sleep factors, 2) between awake factors, 3) between wake factors and subsequent sleep, and 4) between sleep factors and subsequent wake. Most individuals engaging in diary tracking for sleep are trying to detect waking factors that impact subsequent sleep (i.e., category #3). This category of relationship is the most obvious that could be acted upon to improve sleep. However, the other three categories provide potentially important and practical information.

Analyzing the relationships among sleep factors can provide important context, as the goals of tracking typically involve improvement of one or more sleep factors. Sleep factors that are strongly related suggest redundancy, and individuals wishing to simplify their tracking might prioritize based on this information. For example, if the TST and sleep quality were strictly
correlated, one might simplify by only recording one of the two. Conversely, one could justify keeping multiple sleep factors in the diary if they varied at least somewhat independently.

Analyzing the relationships among awake factors offers a different set of advantages. In this category, if two factors correlate very strongly, then any potential impact on sleep cannot be statistically teased apart. In such cases, the individual would need to actively decide to interrupt this correlation, if it were feasible to do so. For example, if a person either had both or neither of the waking factors of nap and coffee on a given day, any potential relationship with sleep could not be uniquely attributed to one or the other no matter how long one tracked, unless the person decided to actively break apart this connection by choice in subsequent tracking.

Analyzing which sleep factors are related to the subsequent night’s sleep is the most common goal of diary tracking to identify potentially modifiable factors to improve sleep. In this category, each individual might have different levels of interest in correlations depending their strength. After 7 days, only very large correlation values will reach statistical significance (because of low power to detect anything but large R-value effect sizes). Such small samples are also at risk for false positives, because of the “law of small numbers”29. Large effects may be biologically unlikely. If present, the individual may have already recognized the relationship without formal tracking (e.g., someone who is exquisitely sensitive to caffeine).

Analyzing which awake factors are related to next-day awake factors can reveal potentially reactive behaviors triggered by better versus worse sleep. Correlations found in this category can also be acted upon, especially if they are part of a cycle of influence. Cycles of influence can be helpful (virtuous cycle) or harmful (vicious cycle) with respect to sleep outcomes of interest. An example of a vicious cycle: if poor sleep increases the use of next-day caffeine, which then negatively impacts subsequent sleep. An example of a virtuous cycle: if exercise improves sleep, resulting in better energy the next day and more likely to repeat the exercise.

For each of these categories, the strength of the correlation gives a sense of how confident to be in the finding. However, the assumption that larger effects are more likely to be true is balanced by the risk of false positives and the fact that even true positives from undersampled data tend to have over-estimated effect sizes29. Early data can be used to suggest how much additional tracking would be useful to confirm correlations, if they were in fact true. In this way, the first week becomes the basis for an interim power calculation to plan continued tracking. Under the simplifying assumptions typically implied by a simple correlation coefficient, only R-values of ~0.9 (absolute value) would be considered statistically significant for n=7 observations (for n=14 observations, ~0.7, and for n=28 observations, ~0.5; all assuming independent measures, which admittedly is not the case for repeated measures within an individual). In this way, some context can be provided for correlation values seen after one week, including how much longer one might wish to record to be powered for statistical significance of correlations lower than R~0.9, assuming the R value observed in week one is true (even if not p<0.05). Just as non-significant R-values can be re-assessed by additional tracking time, extending the observation period could result in lower R-values for some pairs that appeared significantly correlated after week one. This phenomenon can be viewed as a form of regression to the mean, a well-known phenomenon related to under-sampled views of any process. From the user standpoint, it is important to realize that changes in R-values over time can occur even if all other aspects of behavior and daily exposures remained constant.
Changes in R value from week to week can of course also occur if aspects of behavior and daily exposures actually have changed. This highlights a key challenge of pattern discovery: considering the possible explanations for the observed data, both false positive and false negative, and deciding whether to continue monitoring with or without added constraints or changes to tease apart the possibilities.

Working example: basic inferences from a one week diary

Figure 1 is a typical one week diary with four sleep and four awake factors. A qualitative view reveals that the sleep latency (SL) is fairly stable and low (each was 5 minutes except one was 10 minutes), and that there are weekend versus weekday differences in behaviors: caffeine is only on weekdays, exercise is mainly on weekend days, naps are only on weekends, alcohol is highest on Friday and Saturday nights, and total sleep time (TST) is longest on Friday and Saturday nights. The sleep quality values peak on the two nights when TST also peaked. Looking at alcohol, it might appear that it predicts sleep quality, but the correlation could have been driven more perhaps by sleeping in on the weekend mornings. While there is a traditional place for such qualitative narratives surrounding diary tracking, we turn to quantifications for a more principled view.

**Figure 1:** Example one-week diary and correlation analysis

A. Seven days (rows) are shown for sleep factors (four columns; gray) and wake factors (4 columns, white). In this example, sleep quality is a score from 1-10 (10 is best), sleep latency (SL) in minutes, total sleep time (TST) in hours, number of awakenings as integer values, caffeine and etoh (alcohol) as the number of drinks, while exercise and nap are a binary (yes = 1). B-E: Correlation coefficients are calculated for four categories of possible correlations as indicated, and the possible pairs listed within each along with their R-values. For the sleep -> wake category, note that, after 7 days of tracking, only 6 of the nights have a subsequent day for comparison.
The main goal of tracking is typically to determine if an awake factor is influencing subsequent sleep. Figure 1B shows the R values between the awake factors and each of two sleep factors, TST and sleep quality. Both positive and negative correlations are found, as expected, and the range of values was large (0.19 to 0.93). For a sample diary of this duration, only R values of ~0.9 will reach significance. What if the individual strongly hypothesized that exercise would improve sleep? The R value of 0.79 with TST and 0.68 with sleep quality both seem large and thus compelling, but are not statistically significant. The individual with this data would be presented with the option of continuing tracking for additional time, to further evaluate potentially important correlations that are less than R~0.9. After 14 days of tracking, R values of ~0.7 can reach significance, so the recommendation in this case would be to extend tracking an additional week to further evaluate the two sleep correlations with exercise. After 28 days of tracking R values of ~0.5 can reach significance. For lower correlations, even longer tracking would be required. The individual can decide whether to extend and for how long, depending on their bandwidth and interest.

The reverse relationship may also provide insights: that some aspect of sleep influences next-day awake behaviors. In Figure 1C, the R-values are given for this kind of correlation. For a 7-day diary, only 6 such pairs are possible, since the final night does not have next-day awake values recorded. A similar range of R values, positive and negative, are observed in our example diary. Some may seem intuitive, such as longer TST being associated with less caffeine the next day, as if longer sleep results in more energy and less caffeine need. This may be the true narrative, but it could also be that this individual specifically avoids weekend caffeine for some other reason, and that sleeping in only happens on the weekend, so the apparent relation of TST and next-day caffeine intake could be spurious. We are more likely to conclude the TST and nap correlation to be spurious, intuitively, because of the different behaviors occurring on the weekends only (sleeping in, and time for a nap), rather than a mechanistic link of longer sleep causing a nap to occur (the opposite would be predicted by sleep homeostasis).

Correlations also are evident among the sleep factors (Figure 1D). As mentioned in the qualitative view, sleep latency had little relation with other sleep factors (being fairly constant), while TST was correlated with sleep quality. For individuals who wish to simplify their tracking, the within-sleep correlations can provide a rationale to reduce the number of factors tracked. For example, highly correlated sleep factors may present redundant information, and thus only one need be recorded. For constant or nearly constant factors, like SL, these could be removed from the diary if they are already within a normal range.

Finally, correlations may occur among the awake factors, as shown in Figure 1E. We can see that naps and caffeine are highly anti-correlated, as are caffeine and exercise, while exercise and naps are correlated with R= 0.73. All three could be explained by the weekend versus weekday issue. Consider the caffeine and nap question: if this anti-correlation is simply due to weekend versus weekday, then extra tracking will not be able to tease apart correlates of naps and caffeine, if these behaviors continued in this manner. Rather, the individual wishing to tease apart the roles would have to actively decide to change the weekend behavior to break up the observed correlation. In other words, increasing the sample size is not sufficient, one must actively choose to change behavior to answer the statistical question of correlations with these
factors individually. By contrast, the individual might observe an intermediate correlation between exercise and alcohol, and simply carry on their usual tracking, if they did not feel strongly about further separating these two factors in their role with sleep, or if they were willing to track long enough, they could evaluate correlations by sufficiently sampling days with one, the other, neither, or both, for effects on sleep.

**Other diary interpretation considerations**

**Handling outliers:** Abrupt changes in sleep may prompt the user to consider whether any external factors may have occurred to trigger an outlier event, such as unexpected stress, or acute illness. If an obvious trigger or cause is not apparent, then the individual would need to consider whether to stay the course with tracking or to adapt the plan (perhaps by adding more variables of potential sleep relevance).

**Coverage for distributions:** Statistical tests often include assumptions about the distribution or variance of the data, but it is challenging to understand distributions with only small datasets\(^29,30\). Understanding the distribution can even help interpret outliers: for example, under-sampled exponential distributions can be mistaken for normal distributions with outliers\(^30\). The distribution could be informative in a more practical way: consider the observation of a bimodal distribution in a variable expected to be normally distributed. This might imply an unmeasured factor contributing to the heterogeneity.

**Long-term patterns:** Variables that change over months, such as menstrual cycles or seasonal variability, require tracking for longer than the cycle length of interest. Similar power considerations can be applied, including variability and effect sizes of interest, to estimate how many cycles to examine. We previously described a similar challenge in epilepsy and the duration of observations required to evaluate the impact of anticonvulsant therapy on seizure frequency\(^31\).

**Working example: advanced inferences from a one week diary**

The correlation values in Figure 1B and 1C, when viewed together, can provide insights into potential cycles of influence. In other words, awake factors can influence subsequent sleep, and that sleep experience can influence next-day awake behaviors. Figure 2 illustrates examples of such cycles. An example of a virtuous cycle is shown: exercise correlates with longer subsequent TST, and longer TST is correlated with next-day exercise (Figure 2A). We can rephrase the narrative in more natural language: exercise improves my sleep, and improved sleep allows me the energy required to exercise again the next day. An example of a vicious cycle is also given: caffeine is negatively correlated with subsequent sleep quality, while poor sleep quality correlates with next day caffeine use (Figure 2A). We can rephrase this narrative as well: caffeine causes me to have sleep disturbance, and poor sleep results in low energy that I combat with caffeine.

More complicated cycles are possible (Figure 2B). The GST platform allows visualization of these potentially causal influence cycles, using the correlation analysis for all pairs of factors. The user can evaluate cycles for potential relevance, including motivation to prioritize interventions that either break vicious cycles or enhance virtuous cycles.
**Figure 2: Examples of factors that occur in sleep-wake cycles**

A. Viscous and virtuous cycles are shown by arrows between sleep factors (left column of circles) and wake factors (right column of circles), and the symbol (+ or -) and arrow indicate the kind of influence. In the virtuous cycle example, higher TST increases the chance of next-day exercise, and exercise increases next-night TST. In the vicious cycle example, worse sleep quality leads to next-day caffeine intake, and caffeine leads to worse sleep quality on the subsequent night. B. More complex cycles can be viewed in these kinds of diagrams, but interpreting causality remains challenging (for example increased TST appearing to increase nap probability, but this may have been driven by the confounding factor of weekend vs weekday (not shown here).

### Taking action to directly or indirectly improve sleep

Direct interventions refer to any externally introduced factor aimed at improving sleep, such as a prescription, an herbal supplement, cognitive behavioral therapy for insomnia (CBT-I), meditation, etc. Indirect interventions refer to minimizing a factor or behavior that might be interfering with sleep, such as reducing the intake of caffeine or alcohol.

A traditional approach to testing intervention within subjects would be a before-after design: obtain a baseline of tracking data, then make the change, and then continue tracking. Deciding how long to perform baseline tracking before the intervention may depend on the number of factors being tracked, how much variability exists in each, and the end-goal of the intervention. Tracking more factors improves understanding of potential influencers of sleep, but raises concerns for multiple comparisons. Variability allows for correlation analysis “naturally”, but also represents statistical noise that reduces power to detect meaningful changes post-intervention. Longer tracking provides better evaluation of variance, and increased power to detect differences statistically. Phrased as a power consideration, we can ask: how large of an effect is desired, and how confidently does one want to know?
The before-after strategy can be applied to an indirect or a direct intervention. Planning for both approaches is also reasonable, since correlations discovered through baseline tracking could inform indirect interventions that set the stage for increased likelihood of success during a subsequent direct intervention. Consider a person who sleeps poorly when they drink alcohol, have too much caffeine, or take a late nap – that person might wish to engage in reducing these behaviors to indirectly benefit sleep and reduce night-to-night variability before attempting a direct intervention, both of which will increase power to detect the effect of a direct intervention. Even if changing awake factors discovered during baseline is not feasible, the baseline tracking can provide context to interpret the success or failure of a direct intervention. Put another way, variable or even consistent daily presence of waking factors that negatively impact sleep (e.g. alcohol) can reduce the chances of success of a direct intervention to improve sleep both statistically (due to noise) and biologically (due to competing mechanisms). Consider planning a trial of an herbal supplement to achieve a goal of one hour of extra sleep. If two people have the same average TST of 5 hours per night, but have different range of hours night-by-night (i.e., differ in variability of the outcome variable), the less variable individual would require shorter tracking time to evaluate their success from a statistical standpoint.

For planning interventions, power can also be approximated by using traditional methods applied to groups (where “n” is here representing number of days, not number of subjects), if baseline data is obtained first: assume a paired statistical test, using observed baseline variability, with each individual choosing an expected goal effect size during post-intervention tracking. More advanced approaches can be used that incorporate the within-subject correlations inherent in repeated measures studies of course.

Despite the benefits of obtaining a baseline period of tracking, the GST platform does not require it. If an individual chooses to immediately begin an intervention, GST provides a path for this, assuming that they can conceptualize their sleep as a binary probability outcome (i.e., each night has a probability of “good” sleep), and that they have a goal probability in mind as the outcome. We have previously proposed such binary probabilistic goal-oriented testing strategy\textsuperscript{20}. This approach has several advantages beyond circumventing the need for baseline data. Sleep is considered probabilistically, which reminds the user not to react to individual nights but rather to focus on trends and the “big picture”. Further, the early trends can over-, under-, or accurately-estimate the true probability of good sleep. Figure 3 (next page) illustrates this by plotting the cumulative probability of a good night across 30 nights of simulated tracking, assuming in each case that $p$\text{(good)} = 0.5. Without knowing these patterns...
were drawn from the same binomial distribution, it would be very challenging to perform the reverse task, which is of course the real-world task, of observing a pattern of sleep and inferring the underlying probability. Displaying ones data with reference “funnels” of the target \( p(\text{good}) \) value can guide interpretation of tracking data, to answer the question, could my observations be consistent with my goal? The assumptions of independence from night to night and the governance by a simple binomial distribution may not hold in real-world tracking; nevertheless, the example illustrates important lessons about making inferences even under simplifying conditions, to caution against over-interpreting patterns under more complex circumstances.

Figure 3: Example patterns under binomial distribution assumption
Each panel represents the cumulative probability of a good night, based on repeated draws from a binomial distribution with \( p(\text{good}) = 0.5 \). The jagged line is the observed data from 30 draws, over 6 simulated experiments. The curved “funnel” (red lines) is the boundary probability over many draws (same in all cases, given the \( p(\text{good}) = 0.5 \)), and shows how wide the interval is initially and then narrows toward the true probability over time.
Discussion

Self-tracking as phenotyping

We describe a method that enables users to track health-, wellness-, and/or performance-related factors, combined with subjective or objective self-chosen outcomes. Clinical practice uses the history and qualitative diary evaluation to understand the sleep phenotype. The GST system helps to formalize these real-world topics to provide a deeper view of the phenotype, with emphasis on informing decisions. GST can be viewed as an extension of the clinical phenotyping goals of history-taking, and, like a clinical visit, the reason to emphasize characterization of the phenotype is the chance to inform decision making in ways that are sensible and individualized. The willingness to track, as well as the willingness to conceptualize sleep as a varying or probabilistic experience, are highly individual features of a phenotype as well, as these will shape the capacity to design and interpret interventions.

GST supports pattern detection, hypothesis testing, and decision making that is individualized along important axes such as time horizon, resources, goals of tracking, and type(s) of intervention. The iterative flow goes from tracking, to analysis, to actions; the actions could include a new choice about tracking, could involve direct or indirect influencers of sleep, and could be assessed in a before-after design (traditional) versus target-oriented (does not require baseline assessment). This flow is illustrated in Figure 4.

Figure 4. Schematic of the Guided Self Testing (GST) approach
Although the current GST software platform is implemented in the context of sleep, the general framework can be applied to any health, wellness, or performance factor that fluctuates over time scales of days or weeks and might benefit from principled pattern detection to find triggers or assess treatment response in a noisy background. Examples may include migraine, anxiety, pain, stress, or when chronic disorders exhibit shorter term lability (e.g., blood pressure, blood sugar). Custom guidance for individuals who perform self-tracking, from simple to advanced pattern detection, to principled assessment of success of interventions, can bring structure to the typically qualitative efforts used to improve health, wellness, and performance.

**Comparison to qualitative decision making**

How do we find correlations and patterns in regular life? It may seem easy for high-valence isolated experiences (e.g., food allergy), or from big-picture trends when trial and error is occurring (e.g., how much exercise is tolerable). Most people draw inferences about associations in daily life without rigorous assessments of tracked variables or systematic adjustments of key variables one at a time to test hypotheses. One interpretation is that we are all equipped with pattern detection skills and causal inference intuitions, even if we can’t or don’t articulate it as such. However, even under simplifying assumptions, we can see how challenging such tasks are. Some combination of the relative importance of valence regarding the pattern in question, the benefits of being correct, the risks of being incorrect, and the magnitude of effect we consider important - all figure into how much energy we invest in formalizing the process. A few examples are illustrative.

Consider an individual who becomes sick with nausea and vomiting, and naturally considers if any new kinds of food were recently eaten, and recalls having a dragon fruit for the first time a few hours earlier. The person might decide never to eat dragon fruit again, reinforced by the fact that they didn’t care for the taste anyway (i.e., no downside or risk to being wrong about the fruit being the cause). If the illness were either random or related to sporadic contamination of something else eaten, there is little way to figure this out, so there is not much else to act upon, except to re-challenge with consuming dragonfruit.

Next consider a person who enjoys coffee, but also has insomnia and heard that it was bad for sleep quality. That person might periodically take a coffee late in the day, and when asked, reports that they do not notice any consistent effect on their sleep. Without some tracking data, it could be challenging to tease apart an effect. Unless coffee notably worsened sleep every time it was consumed, which is possible but unlikely, a partial effect might easily be overlooked. Perhaps the person is only interested in factors that have a dramatic or near-certain relationship with sleep, in which case their personal experience with recognizing non-large effects would be limited.

**Real-world variability: friend and foe**

The very reason RCTs attempt to increase homogeneity is a reminder of the reality that each individual brings a different set of factors in daily life that contribute to variability. For self-testing, within-individual variability manifests as temporal variability. This means noise and reduced power in traditional clinical trial design, and inclusion/exclusion criteria and other
protocol constraints are designed to reduce these confounding elements. However, these elements cannot be assumed to be under similar control in the real world. For the individual willing to engage in self-tracking and potentially modify behaviors, the variability can be understood, can inform decisions, and can even potentially be controlled. One way to conceptualize the use of baseline tracking variability is that correlation analysis provides some answers (or at least hypotheses) regarding the variance in one factor explained by another (e.g., see\textsuperscript{32}). Each individual may differ in the number of factors of potential influence, the relative importance or vulnerability to each factor, and the extent to which behavioral correlations are discoverable or amenable to change. In GST, variability is specifically anticipated, it provides insights, and can even be harnessed for intervention planning, as described above.

Consider the patient who tries a sleeping pill one night, and feels no improvement, and discontinues it. This would be a sensible interpretation if the relationship between sleeping pill ingestion and sleep were deterministic and binary (it either works every time or doesn’t ever work), or if the person has low tolerance for partial response. Many aspects go into these kinds of seemingly simple decisions – some of which are malleable via frameworks such as GST. Consider the person who, like in the diary example (Figure 1) has some variability in TST from night to night, and TST happens to be the main determinant of their sleep quality. If such a person started a new intervention like an herbal supplement to help sleep on a night which, by chance, was “bad” (say, it would have been a 5-hour night), even if the herbal was mildly helpful, the TST after taking the herbal could still be “bad” (say, 5.5 hours, a gain of 30 minutes over the already bad value of 5.0). The individual experiencing this 5.5 hour TST as still being in the “bad” range for them might (falsely) conclude no effect of the herbal, or, worse still, even conclude (again, falsely) that the herbal remedy was causally responsible for the bad night. Neither of these interpretations would be true in this example. Only with a framework to think about variability more formally, and to the extent the person is willing to track sleep-wake data over multiple nights, could the effect of the herbal be ascertained.

Identifying natural sources of variability can inform the process of self-discovery. Potential sources include those which are measured (i.e., awake factors from the diary), those which are unmeasured but are in principle measurable (e.g., room temperature, diet), and those which are not typically feasible to measure outside of research settings (e.g., core body temperature). Anything not measured (regardless of feasibility) will contribute noise to the observations and thus reduce the chances of finding correlations among factors that are measured. Although variation over time is the basis for discovering correlations statistically, even “static” consistent observations can point to important associations (e.g., if daily caffeine is constant, and so are nocturnal awakenings caused by the caffeine).

Some forms of variability are more difficult to predict or analyze, due to their sparse occurrence rates, such as acute illness, or possibly travel. Some baseline phenotypes are so extreme that any variation takes on a distinct context. Consider the athlete who trains to achieve consistently high performance measures for competition. Were such an individual to engage in self testing, searching for correlations of performance that is already near its statistical ceiling is challenging from a power perspective because of small the feasible change in absolute effect size is. The very consistency of their performance may lead to rationalizing
what others might call superstition\textsuperscript{33-35}: reacting to details that were temporally associated with important (albeit small) fluctuations in performance. In this setting, experimental proof of an effect might take non-feasible durations of self-testing, because of the small variance of high performance settings. On the other hand, such high-performance individuals may be motivated to engage in extensive tracking to experimentally evaluate direct or indirect measures that could increase performance.

From the GST standpoint, changing performance that is already near its ceiling represents a virtually untestable circumstance: powering for small effect sizes requires very lengthy observation periods, which is often impractical. Absent feasible testing, the narrative of potential variables takes on necessarily greater importance, which may explain so-called superstition behaviors so common in elite performance\textsuperscript{33-35}. Since mental focus and discipline is critical, including lack of distraction, it is plausible that superstitions represent a combination of little downside, potential upside, and that failure to comply with them could cause reduced focus or doubt and thus potential regret if performance dips occur.

**Democratizing longitudinal data analysis**

Various statistical approaches exist for analysis of longitudinal data streams. These approaches include strategies for within and between subject factors, and evaluation of the impact of an intervention\textsuperscript{36,37}. Certainly, any individual with sufficient data could subject their tracking experience to more advanced analytics than presented here. We propose that GST falls into a middle ground, between purely qualitative inferences, and formal experimental studies with sophisticated statistical designs. Even though most diary tracking data is likely to violate the assumptions of a Pearson correlation test, using this information as an approximation can still provide actionable context to routinely acquired data in four important ways: 1) moving from qualitative to quantitative assessment of relationships; 2) incorporating concepts of power and significance to interpret the strength of relationships; 3) motivate additional tracking time or re-design of tracking goals; 4) to plan and interpret interventions.

**Incorporation of objective data into tracking**

In many aspects of health care, the introduction of objective testing to the evaluation and management of individuals, especially in the home, is on the rise. For years, self-monitoring devices like scales, thermometers, and blood pressure cuffs, have been available for home use via over-the-counter purchase. In sleep medicine, consumer-facing trackers continue to expand, with or without concurrent fitness-aimed measures like activity or heart rate\textsuperscript{16,17}. Sleep monitoring devices can be used to supplement to subjective measures of a sleep diary, especially when the individual is unwilling or unable to record sleep factors on a regular basis. The discrepancy between diary and objective measures of sleep may be of interest as a part of the phenotype. For example, the perception of sleep may be at least in part malleable in healthy adults\textsuperscript{38} as well as those with insomnia\textsuperscript{39,40}. Objective device monitoring can provide outcome metrics themselves, particularly if the metrics do not have a natural self-reported correlate, such as heart rate variability or sleep stage or electroencephalogram power.
Comparison with randomized controlled trials

RCTs are the benchmark in modern evidence based medicine. Randomization and blinded placebo comparisons are considered strong techniques to isolate causal relationships between health interventions and health outcomes. We have an extensive framework for judging evidence, we have criteria for systematic reviews, and we have clinical guidelines with formal levels of confidence in evidence. Despite this rigor, in clinical practice decisions about an individual patient often requires extrapolation from group- or population studies, which may or may not clearly apply at the individual level. Much of this concern relates to the question of external validity: can I apply the results of this evidence if my patient would have been excluded from the trial, perhaps due to comorbidities? Put another way, we may ask: is my patient similar enough to the population(s) of the RCTs? If not, do I have any next-level evidence to sway decisions one way or the other? These and other critiques of RCTs (and evidence based medicine in general, and null-hypothesis testing that much of it is based on) have been described elsewhere41-45.

In some settings, RCTs are not feasible or unlikely to be completed in a timeframe relevant to the individual’s decision making. It is possible that waiting for an RCT could be viewed by the individual as a risk, to be weighed against risks associated with making decisions on lesser forms of evidence. In fact, many questions in the clinic are answered in the absence of directly relevant RCT-level evidence, and the reign of RCTs in clinical practice has been scrutinized43,44.

It is not unreasonable to ask: under what circumstances should we seek or at least allow alternatives to RCT evidence? Potential considerations that are relevant directly to investigation of sleep include: a) When symptom heterogeneity is prominent between individuals; b) when symptom expression is variable over time within individuals; c) when vulnerability to modifying factors shows within- and between-individual heterogeneity; d) when many low-cost and low-risk interventions exist; e) when the placebo effect is not crucial to isolate (and may even be welcome).

We propose that ultimately individuals might place a premium on information relevant to them, especially if population RCTs are either unavailable or not applicable to their specific circumstance. For sleep, one of the main reasons to perform RCTs (demonstrate causal efficacy relative to a placebo effect) may not be important to an individual. Put another way, a reasonable person might rather anchor the desired outcome in their own context, rather than to answer question: “what portion of the benefit was attributed to placebo”.

References


