Modeling the Association between Smoking and Sleep Fragmentation Using Log-Linear Models

by

Gina M. Norato

A thesis submitted to The Johns Hopkins University in conformity with the requirements for the degree of Master of Science

Baltimore, Maryland
April, 2016

Copyright 2016 by Gina Norato
All rights reserved
Abstract

Given that the etiology of poor sleep quality in regards to cigarette smoking is not well-defined, it is of interest to further investigate the risk of transition between various sleep stages for never, current, and former smokers. Polysonomography and clinical data from 5139 participants of the Sleep Heart Health Study were used for the current analysis. A log-linear generalized estimating equation (GEE) modeling approach was used, adjusted for demographic covariates and comorbidities. Previous literature has described the use of this approach as a computationally faster alternative to multi-state survival models.

Compared to never smokers, former smokers transition more often into wake, from non-REM and REM sleep (NW: 1.05 (CI 1.02, 1.09); RW: 1.06 (CI 1.02, 1.11)). Current smokers are more likely to transition out of non-REM into wake (1.07 (CI 1.02-1.13)), and correspondingly less likely to transition into REM sleep from non-REM sleep (0.91 (CI 0.87-0.96)). Moreover, current smokers have less risk for transition out of REM sleep (RN: 0.87 (CI 0.80-0.94); RW: 0.88 (CI 0.82-0.95)). In more detailed analyses that characterized sleep into five states, former smokers again show increased transitions to wakefulness, and also seem to have highly altered stage 2 sleep. For current smokers, transitions out of stages 1 and 2 sleep were reduced, and overall they are more likely to transition to and stay in lower levels of sleep (ie, wake or stage 1), are less likely to transition into deeper levels of sleep, and are less likely to transition out of REM, likely due to REM pressure. The results of the current study show that cigarette smoking does influence sleep stage transitioning, even for former smokers. Furthermore, using models that incorporate the granularity of sleep using five states are of more utility than a simpler 3-state model.
Acknowledgments

First, I would like to thank my advisor, Ciprian Crainiceanu, for his guidance and support throughout my time at Hopkins. I greatly appreciate your confidence in me and the many opportunities you’ve given me. I’d also like to thank my co-advisor Naresh Punjabi; your enthusiasm for research is contagious, and your expertise was invaluable and inspired the core of this work.

Thank you to Bruce Swihart for laying the framework for me with this analysis— and also for your saucy plots.

I would also like to thank my peers, especially Jisoo Kim and Ryan Andrews for being sounding boards when I was stuck, which was often. Thank you for believing in me and helping me take breaks when I needed them. Also thank you to the rest of my cohort for being a fantastic support system.

Finally, to my wonderful parents, thank you for always lending your ear and offering advice (even when unsolicited). Your support and love mean the world to me.
Table of Contents

List of Tables vi

List of Code Samples vii

List of Figures viii

1 Introduction 1
   1.1 Characterizing Sleep ............................. 1
      1.1.1 Sleep Stages .................................. 1
      1.1.2 Assessing Sleep ............................... 2
      1.1.3 Quantifying Sleep .............................. 3
   1.2 Nicotine and Sleep ................................. 4
   1.3 Analytical Approaches .............................. 6
   1.4 Dataset Selection ................................. 7
   1.5 Focus of this Analysis .............................. 8

2 Data Source and Processing 9
   2.1 Parent Cohorts .................................... 9
   2.2 Formation of SHHS Cohort .......................... 10
   2.3 Variable Selection ................................. 10
2.4 Polysomnography Data ........................................... 11
2.5 Merging Data ....................................................... 15

3 Methods ................................................................. 17
  3.1 Visualization ..................................................... 17
  3.2 Poisson Regression Analysis ................................. 20

4 Results ................................................................. 26
  4.1 Visualizations and Data Exploration ....................... 26
  4.2 3-state Modeling ............................................... 31
  4.3 5-state Modeling ............................................... 34

5 Conclusions .......................................................... 38

References ............................................................... 40

Appendix A Additional Figures ................................... 47
# List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>5-state Polysomnogram Data</td>
<td>12</td>
</tr>
<tr>
<td>2.2</td>
<td>3-state Polysomnogram Data</td>
<td>12</td>
</tr>
<tr>
<td>2.3</td>
<td>Sample 3-state Poisson Data for a Single Subject</td>
<td>14</td>
</tr>
<tr>
<td>3.1</td>
<td>Contrast Matrix</td>
<td>24</td>
</tr>
<tr>
<td>3.2</td>
<td>Relative Risk Estimates</td>
<td>25</td>
</tr>
<tr>
<td>4.1</td>
<td>Covariate Information</td>
<td>27</td>
</tr>
<tr>
<td>4.2</td>
<td>Sleep Information</td>
<td>31</td>
</tr>
<tr>
<td>4.3</td>
<td>3-state Model Estimates</td>
<td>33</td>
</tr>
</tbody>
</table>
List of Code Samples

3.1 Lasagna Plotting, Subject Information ......................... 18
3.2 Lasagna Plotting, Data Manipulation ............................. 18
3.3 Lasagna Plotting, Visualization ................................. 19
3.4 Poisson Regression ................................................. 21
3.5 Relative Risk Creation ............................................ 22
3.6 Poisson Regression, Adjusted ..................................... 23
3.7 Adjusted Poisson Output ......................................... 23
3.8 Linear Combination Function ..................................... 25
# List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Sample Hypnograms</td>
<td>3</td>
</tr>
<tr>
<td>2.1</td>
<td>Data Flow Diagram</td>
<td>16</td>
</tr>
<tr>
<td>3.1</td>
<td>Transition Relative Risks</td>
<td>21</td>
</tr>
<tr>
<td>4.1</td>
<td>Lasagna Plots</td>
<td>29</td>
</tr>
<tr>
<td>4.2</td>
<td>Proportion of Slow-Wave Sleep</td>
<td>30</td>
</tr>
<tr>
<td>4.3</td>
<td>Relative Risks, Fully Adjusted, 3-state</td>
<td>32</td>
</tr>
<tr>
<td>4.4</td>
<td>Relative Risk Diagrams</td>
<td>32</td>
</tr>
<tr>
<td>4.5</td>
<td>Relative Risks, Fully Adjusted, 5-state</td>
<td>34</td>
</tr>
<tr>
<td>4.6</td>
<td>3-state Relative Risks and 5-state Analogues</td>
<td>37</td>
</tr>
</tbody>
</table>
Introduction

Cigarette smoking is known to cause several serious chronic conditions, including cardiovascular disease, respiratory disease, and various malignancies [1]. While such effects are relatively well-known, the impact of smoking, even former smoking, on sleep quality is not well-defined. To investigate the effects of smoking on sleep, we must first characterize and visualize the evolution of sleep, understand how sleep quality affects health, and determine how smoking may influence the delicate processes involving sleep. We also must investigate the various statistical methodologies that have been applied in these areas.

1.1 Characterizing Sleep

1.1.1 Sleep Stages

The typical adult spends nearly one-third of their life asleep. Sleep and wake are regulated by neurotransmitters in the brain, namely serotonin and norepinephrine. These chemicals activate neural networks in different parts of the brain, and orchestrate whether the brain is “asleep” or “awake”. Once asleep, the brain cycles through five main stages of sleep, each of which have different characteristics. These are stages 1, 2, 3, 4, and rapid eye movement (REM). Stage 1 sleep is the lightest, and typically occurs just after falling asleep. Muscle
activity is slowed and the arousal threshold from this sleep stage is low. Stage 2 sleep typically follows stage 1 sleep, and consists of less eye movement and a higher threshold of arousal. Stages 3 and 4, together often called “slow-wave sleep”, are characterized by low-frequency oscillations in the electroencephalogram, called delta waves. Slow-wave sleep is much deeper than the other stages and it is harder to be awakened from this stage. REM sleep typically occurs 1-1.5 hours after falling asleep, and is characterized by more irregular breathing, relative muscle hypotonia, and, as the name suggests, rapid movement of the eyes. Also, vivid dreams are often reported during this stage. A complete sleep cycle lasts about 90-110 minutes. As the night progresses, REM sleep periods lengthen and slow-wave sleep decreases. Typically, a person will spend 50% in stage 2 sleep, 20% in REM, and 30% in the other stages [2].

1.1.2 Assessing Sleep

Sleep is assessed using polysomnography (PSG), which includes recording of certain physiological signals during sleep, and is commonly used to diagnose various sleep disorders [3]. A polysomnogram consists of three groups of recordings: electroencephalography (EEG), electrooculography, and surface electromyography [4]. Airflow is also monitored for detecting any breathing abnormalities during sleep. There are two major types of breathing abnormalities: apnea, where airflow is reduced almost completely, and hypopnea, where airflow is reduced with an associated arousal from sleep or a decrease in oxygen saturation. Often, these two measurements are combined into an apnea-hypopnea index (AHI) or the respiratory disturbance index (RDI), which represent the total number of apneas and hypopneas per hour of sleep [5].
1.1.3 Quantifying Sleep

The EEG recordings are used to detect wake, REM, stage 1, stage 2, and slow-wave sleep, using standardized criteria [6]. Once the stages are visually annotated, the data can be visualized using hypnograms (Figure 1.1). These hypnograms show the temporal evolution of sleep (in the figure, denoted as W, R, 1, 2, S for wake, REM, stage 1, stage 2, and slow-wave sleep (stages 3 and 4)) as a function of time. In these three sample hypnograms, the first person has relatively stable sleep, with few transitions between the stages. However, the second and third individuals experience more transitions between the stages, demonstrating increased transition frequency and sleep stage instability.

![Figure 1.1: Sample hypnograms in individuals with varying levels of sleep fragmentation. W=wake, R=REM, 1=stage 1, 2=stage 2, S=slow-wave](image)

Although an illustrative example, it is easy to see how hypnograms may be very informative for understanding how sleep is structured and how it may be disrupted. The number of shifts or time spent in each stage has been used for various analyses, but such summaries may be associated with significant loss of
information [7, 8]. More in-depth methods for data extraction and analysis will be described in Section 1.3.

1.2 Nicotine and Sleep

Sleep is a complex process which can be disrupted by numerous agents, including caffeine, antidepressants, and alcohol [2]. Sleep disruption can result in secondary sleep deprivation, which can be detrimental and lead to impairments in memory, physical performance problems, emotional well-being, and social functioning. As mentioned above, sleep is dictated by the activity of several neurotransmitters in the brain. If the function of these neurotransmitters is interfered with, sleep is disrupted.

Nicotine is a chemical which acts on the cholinergic receptors in the brain, due to its structural similarity to acetylcholine, which is produced by the brain [9, 10]. Nicotine in the brain stimulates several other neurotransmitters, including norepinepherine, which plays an important role in sleep regulation (Section 1.1.1). Because of the interference with normal neurotransmitter signaling, cigarette smoking has been found to be related to a higher prevalence of sleep-related respiratory disorders [11, 12]. Also, interestingly, across studies, there does not seem to be a strong indication that smoking cessation reduces these disorders.

In terms of sleep architecture, cigarette smoking can contribute to suppression of REM sleep and this suppression may be dose-dependent [13, 14, 15].
Moreover, smokers tend to spend more time awake and have a higher sleep latency [16]. More qualitatively and in self-reported surveys, smokers report having more trouble falling asleep and staying asleep [11, 17, 18], increased daytime sleepiness [18, 19], and short sleep duration [19]. Studies of nicotine withdrawal using PSG data indicate that smokers spend less total time awake and have reduced sleep latency [16]. Other studies reported increased arousal frequency and higher number of sleep stage changes for smokers [20]. In self-reports and survey studies, sleep was generally worse in those who were experiencing nicotine withdrawal (decreased sleep quality and frequent awakenings) [21, 22], but after the withdrawal period, some improvements were observed [23].

Previous analyses on the sleep-related effects of smoking have been conducted on the Sleep Hearth Health Study (SHHS), the current data of interest. One investigation of sleep architecture in smokers showed that current smokers, relative to never smokers, had more stage 1 sleep and less slow-wave sleep, and that never and former smokers had similar sleep architecture [24]. Interestingly, the amount of REM sleep was similar across groups. These aforementioned findings were independent of other covariates such as body mass index, RDI, cardiovascular disease, and race. Another study in a matched sample of smokers and non-smokers showed that differences in the EEG power spectrum between these groups were highest at the earlier parts of the night compared to the later parts [25]. Also, smokers were shown to have higher levels of EEG activity in the $\alpha$-bandwidth and lower $\delta$-bandwidth, which corresponds to less deep, restful sleep, and reported feeling less well-rested than never smokers in qualitative measures [25].
Taken together, these findings indicate that smoking impacts sleep architecture and quality of sleep. Using the largest cohort of subjects with PSG, we aim to investigate the details of the association between smoking and sleep architecture. In the next section, various analytical methods that have been used to address this problem are discussed.

1.3 Analytical Approaches

There have been several approaches for quantifying sleep architecture. Often, analyses focus on measures such as sleep stage distributions or EEG spectral analyses. However, the hypnogram provides complementary information that has rarely been explored beyond crude sleep stage summaries. The key to using the hypnogram to its full potential is to quantify both the number of transitions between stages and the time spent in each stage before transitioning, or the time at risk of transition [7]. This “event-history” approach is relatively novel, and has only recently been applied to as many states, transition-types, and subjects as the current analysis [26]. The two statistical approaches that can be used for this type of event-history data are multi-state survival models and log-linear regression models.

Multi-state survival models are models that describe a list of possible states, and all combinations of transitions between those states [27]. The movement between the states can be investigated through proportional hazards regression [28]. For group $g$ and transition type $h$, the survival model can be expressed as

$$\alpha_h(t) = \alpha_{h0}(t) + g : h + \text{covariates},$$

where $\alpha_h(t)$ is a transition-specific log-hazard, $g : h$ is the interaction term.
without the main effect, and $\alpha_{h0}(t)$ is the transition-specific baseline log-hazard.

There is an established equivalence between multi-state survival models and log-linear models, due to the equivalence of their likelihoods [29, 30, 31]. A log-linear model for a multi-state transition model can be written as:

$$\log \lambda^{(gh)} = g + h + g : h + \text{offset}\{\log(\text{tar}_h)\} + \text{covariates},$$

where $\lambda^{(gh)}$ is the rate of transition for group $g$ and transition type $h$, and $\text{tar}_h$ is the time at risk for transition type $h$. As log-linear models can be implemented using generalized estimating equation (GEE) approaches, they can be fit much faster than multi-state survival models. Indeed, in the GEE framework, the correlation structure is a nuisance parameter, and estimators based on within-subject independence assumption are sufficient to provide unbiased estimates [26].

Previous data from the SHHS have shown that the two approaches provide very similar results, while computation time for GEE log-linear models is an order of magnitude smaller [7, 31, 26]. An important technical component that helped reduce computation time is the fact that the Poisson regression uses a much smaller data set than the multi-state survival model, which will be discussed in more detail in Section 2.4.

1.4 Dataset Selection

The SHHS data are particularly well-suited to answering questions using PSG, especially in this particular analytic framework. Specific details about data collection and cohort formation are discussed in Chapter 2. Briefly, the large cohort
of subjects allows for well-populated, balanced subgroups for analysis, even after reducing to complete cases. Indeed, the SHHS cohort has been shown to be among the largest with PSG-based transition analyses as well [26]. Thorough collection of demographic and clinical characteristics allow for minimization of confounding in model construction. Specific choices of covariates and rationale will be discussed in Section 2.3.

1.5 Focus of this Analysis

In summary, the goal for the current analysis is to investigate and quantify the association between smoking and sleep fragmentation, while controlling for potentially confounding demographic variables and medical conditions. Log-linear (Poisson) models will be used to analyze the relative risk of transitioning between the sleep stages, for three categories of smokers: current, former, and never. We will investigate 5-state and 3-state models, where the five states are wake, stage 1, stage 2, slow-wave sleep (SWS), and REM sleep, and the three states are REM, wake, non-REM. Our goal is to quantify and contrast sleep fragmentation as a function of smoking status.
Data Source and Processing

Data collection and methods for the SHHS have been described previously [32]. Some of the key points will be summarized here.

2.1 Parent Cohorts

The SHHS is a prospective cohort study with the primary goal of investigating the impact of sleep-disordered breathing on the development of cardiovascular and cerebrovascular disease. It is unique in the sense that it did not involve recruiting new participants, but instead, individuals were enrolled into SHHS from other cardiovascular study cohorts. These studies are the Atherosclerosis Risk in Communities Study (ARIC), the Cardiovascular Health Study (CHS), the Framingham Heart Study (FHS), the Strong Heart Study (SHS), the Tucson Epidemiologic Study of Airway Obstructive Diseases (TES), the Tucson Health and Environment Study (H&E), and three New York cohorts enrolled as part of the “Psychosocial Factors and Cardiovascular Disease” program. Each of the institutional review boards associated with these studies provided approval for the SHHS study protocol, and informed consent was obtained from all subjects.
2.2 Formation of SHHS Cohort

Protocols and data availability within each cohort were compared, and any variables not collected by a parent cohort that were deemed necessary for SHHS objectives were collected additionally. All participants recruited into the SHHS cohort had in-home PSGs. Detailed information about the study protocol has been published [33]. SHHS had specific inclusion criteria, including only individuals who were either 40 or older. Those treated with continuous positive airway pressure were excluded. Recruitment was balanced generally across genders, and emphasis was placed on enrolling minorities [32].

The baseline visit data, our visit of interest, included questionnaires to assess important co-morbidities such as hypertension and cardiovascular disease, smoking history, race, sex, and age. Additional variables including measures of sleep disordered breathing were obtained from the polysomnogram [33].

2.3 Variable Selection

Data were provided in two parts- the clinical data from the first visit (SHHS1) and the PSG data. Data processing and formatting needed to be performed on both sets. In addition to smoking status, the clinical covariates of interest are gender, race, age, body mass index, hypertension, cardiovascular disease, RDI based on a 4% oxygen desaturation threshold, and FEV$_1$ (forced expiratory volume in 1 second). These characteristics are known to influence sleep quality and/or breathing characteristics during sleep. Lighter sleep as been shown to occur more often in men, as well as in African-Americans and American Indians—older populations experience more disrupted sleep as well [34].
Sleep-disordered breathing which is clinically manifested as snorting/gasping and snoring have been shown to be more common in males and Latino populations, and therefore have the potential to alter sleep architecture and confound the analysis [35]. Sleep-disordered breathing is also associated with incident cardiovascular disease [36]. Additionally, it has been shown that short and long sleep duration is positively associated with obesity, hypertension, and cardiovascular disease, which may also indicate important influences on the structure of sleep [37]. Thus, all of the aforementioned factors are important to control for in the current analysis, since they may confound the association of smoking and sleep architecture.

Each of these variables were ensured to be of the proper type (binary, categorical, continuous) before data exploration and analysis. Complete-case data was used throughout the analysis [26].

2.4 Polysomnography Data

The largest data processing step was structuring the PSG data. As described in the introduction, PSG data as obtained in the SHHS cohort (discrete-time, discrete-state) consists of a string of symbols indicating the current sleep stage, in 30-second epochs. These “symbols” will be used throughout this paper, for simplicity: R for REM, S for slow-wave sleep, W for wake, 1 for stage 1, and 2 for stage 2. We will use a portion of a random subject’s data for illustration. For this subject, the initial 10 epochs/5 minutes of recording were denoted as 1222W12222. The temporal array of visually annotated sleep stage data in this temporal order are as follows: stage 1 for 30 seconds, 1.5 minutes of stage 2
sleep, 30 seconds of wake, 30 seconds of stage 1 sleep, and 4 minutes of stage 2 sleep. A transition occurs when there is a change between stages. Reorganizing these data allows it to take matrix form, as seen in Table 2.1, where each transition that occurs is given its own row. Using the first row as an example, `obsno` is a simple subject counter, and `shiftno` counts the transitions within a subject. Columns `startepoc` and `endepoc` indicate the start and end epochs of the recording where the individual was in stage 1, prior to transitioning to stage 2. Time prior to transition is column `t`, or put another way, the time since the last transition. Recall that an epoch is a 30-second span of time.

<table>
<thead>
<tr>
<th>pptid</th>
<th>obsno</th>
<th>shiftno</th>
<th>shift</th>
<th>startepoc</th>
<th>endepoc</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>0007</td>
<td>1</td>
<td>1</td>
<td>12</td>
<td>28</td>
<td>29</td>
<td>0.5</td>
</tr>
<tr>
<td>0007</td>
<td>1</td>
<td>2</td>
<td>2W</td>
<td>29</td>
<td>32</td>
<td>1.5</td>
</tr>
<tr>
<td>0007</td>
<td>1</td>
<td>3</td>
<td>W1</td>
<td>32</td>
<td>33</td>
<td>0.5</td>
</tr>
<tr>
<td>0007</td>
<td>1</td>
<td>4</td>
<td>12</td>
<td>33</td>
<td>34</td>
<td>0.5</td>
</tr>
<tr>
<td>0007</td>
<td>1</td>
<td>5</td>
<td>2S</td>
<td>34</td>
<td>43</td>
<td>4.5</td>
</tr>
<tr>
<td>0007</td>
<td>1</td>
<td>6</td>
<td>S2</td>
<td>43</td>
<td>47</td>
<td>2</td>
</tr>
<tr>
<td>0007</td>
<td>1</td>
<td>7</td>
<td>2S</td>
<td>47</td>
<td>48</td>
<td>0.5</td>
</tr>
<tr>
<td>0007</td>
<td>1</td>
<td>8</td>
<td>SW</td>
<td>48</td>
<td>52</td>
<td>2</td>
</tr>
</tbody>
</table>

**Table 2.1:** 5-state Polysomnogram Data. A segment of sleep for a single subject from onset of sleep, represented in 5 states. Transitions between stages are denoted as “2W” for a stage 2 to wake transition, for example. Column `t` denotes time in the first stage (2) before transitioning to the second stage (W).

A similar format can be used for 3-state data, where a distinction is not made between subsets of non-REM sleep (slow-wave sleep and stages 1 and 2; Table 2.2). It is rather straightforward to see how the rows are collapsed in this case.

To perform multi-state survival analysis on these data, the dataset would need to be augmented by creating one additional row for each shift, where
Table 2.2: 3-state Polysomnogram Data. Data for a single subject represented in 3-state form, where a distinction is not made between subsets of non-REM sleep (slow-wave sleep and stages 1 and 2).

<table>
<thead>
<tr>
<th>pptid</th>
<th>obsno</th>
<th>shiftno</th>
<th>shift</th>
<th>startepoc</th>
<th>endepoc</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>0007</td>
<td>1</td>
<td>1</td>
<td>NW</td>
<td>28</td>
<td>32</td>
<td>2</td>
</tr>
<tr>
<td>0007</td>
<td>1</td>
<td>2</td>
<td>WN</td>
<td>32</td>
<td>33</td>
<td>0.5</td>
</tr>
<tr>
<td>0007</td>
<td>1</td>
<td>3</td>
<td>NW</td>
<td>33</td>
<td>52</td>
<td>9.5</td>
</tr>
</tbody>
</table>

Each added row is an unobserved (censored) transition. For example, if the observed transition was non-REM to wake (NW), the censored (unobserved) transition would be non-REM to REM (NR). For 3-state modeling, this doubles the size of the dataset, since for each observed transition, there is only one other possible unobserved transition. For 5-state modeling, this would increase the dataset by 300%—indeed, for each observed transition, there are 3 unobserved transitions. For example if the observed transition is 1R then transitions 12, 1W, and 1S are unobserved. Thus, the data structure for a multi-state survival model is 1 observed transition and either 1 or 3 censored transitions depending on the number of states considered. Although we are not performing multi-state survival modeling in this analysis, these data arrangements are useful for visualizations after subsetting to only the observed cases, or as shown prior to augmentation in Tables 2.1 and 2.2. We will refer to these data arrangements as a “survival” arrangements in this paper.

For Poisson analyses, the data structure is altered from this arrangement and requires much smaller data sizes. The analysis requires a summary of each type of transition and the time “at-risk” for the respective shift. Thus, the transition dataset for a Poisson log-linear model is much smaller in size. Indeed, for each
subject, the PSG dataset is a matrix with the number of rows equal to $K(K-1)$, where $K$ is the number of states, and 4 is the number of columns. Thus, $K(K-1)=6$ for $K=3$ stages and 20 for $K=5$ stages. The number of columns of the subject-specific data matrix is 4, where the first and second columns correspond to subject ID and type of transition. The third column is the number of observed transitions of that type. The fourth column contains time at risk for individual transitions and is calculated by adding up the time that is spent in each stage for each person. For example, for NW and NR transitions, the time at risk is the total time spent in non-REM for that person. A complete set of 3-state Poisson data for subject 0007 is provided in Table 2.3.

<table>
<thead>
<tr>
<th>pptid</th>
<th>shift</th>
<th>count</th>
<th>tar</th>
</tr>
</thead>
<tbody>
<tr>
<td>0007</td>
<td>NR</td>
<td>6</td>
<td>324.5</td>
</tr>
<tr>
<td>0007</td>
<td>NW</td>
<td>40</td>
<td>324.5</td>
</tr>
<tr>
<td>0007</td>
<td>RN</td>
<td>0</td>
<td>51.0</td>
</tr>
<tr>
<td>0007</td>
<td>RW</td>
<td>7</td>
<td>51.0</td>
</tr>
<tr>
<td>0007</td>
<td>WN</td>
<td>45</td>
<td>65.0</td>
</tr>
<tr>
<td>0007</td>
<td>WR</td>
<td>1</td>
<td>65.0</td>
</tr>
</tbody>
</table>

Table 2.3: Sample 3-state Poisson Data for a Single Subject. A data format such as this results in 6 x $N$ rows of data, where $N$ is the number of subjects.

This table also contains information about the length of sleep which can be obtained as $324.5 + 51 + 65 = 440.5$ minutes $= 7.34$ hours, and total number of transitions, which is the sum of entries in the third column (99 transitions). Thus, this subject has an average number of transitions of 13.5 transitions per hour.

The process is similar for creating the 5-state Poisson dataset. The only difference is that each subject will have a $20 \times 4$ matrix. Because the dataset
is much smaller and simpler for the Poisson log-linear model, computations and data handling are much easier.

Another important aspect of the PSG data is data quality. In particular, we are interested in incorporating information about quality of visual scoring obtained during REM stage scoring. The SHHS dataset contains information on whether or not classification of stages went beyond REM, non-REM, and wake, whether or not classification of stages went beyond sleep and wake, and whether or not there were issues with these classifications. In the current analysis, we only used recordings that had none of these quality problems.

2.5 Merging Data

Once the PSG data is organized for use with the Poisson model as described in Section 2.4 and Table 2.3, it can be merged with the clinical variables available. In Figure 2.1, we present a flowchart of data processing indicating at every step the number and reason for dropping specific subjects. The 3-state data will take a final form similar to that in Table 2.3, but with additional columns describing clinical characteristics for each subject. The 5-state Poisson data and the 3- and 5-state “survival” datasets can be supplemented in a similar way.
Figure 2.1: Data Flow Diagram. After eliminating subjects without polysomnography (PSG) or with poor quality recordings, and without complete covariate information, the sample eligible for analysis is 5139.
Methods

3.1 Visualization

Once the data are formatted, it is useful to perform visualizations. A helpful visualization approach of hypnogram data is to display subject-level hypnograms and the associated lasagna plots [38]. Lasagna plots were introduced to improve visualization of longitudinal, continuous, or categorical measurements. Indeed, the traditional “spaghetti” plots are not very useful in this context due to overplotting and complexity of individual trajectories. These plots use color to denote changes in status (or values in a continuous outcome), and use each row to denote an individual under observation. Rather than plotting subjects on the same scatterplot as a function of time, lasagna plots display data on separate rows for each subject and use color to provide an overview of population trends, while preserving the subject-specific complexity. Color selection is an important consideration, so as to encourage proper interpretation and reduce misinterpretation [38].

Readily available R-packages for lasagna plots can be used to explore the data, once the data are in the proper format [39]. We will be using the data in “survival” format, with clinical covariates added onto each row for each subject. We will also add a new column called curr, which denotes the current
state. Additionally, we will save smoking group information and total number of epochs for each subject as follows:

```
# create current state
dat$curr[dat$shift %in% c("12", "1R", "1S", "1W")] <- 1
dat$curr[dat$shift %in% c("21", "2R", "2S", "2W")] <- 2
dat$curr[dat$shift %in% c("R1", "R2", "RS", "RW")] <- 0
dat$curr[dat$shift %in% c("S1", "S2", "SR", "SW")] <- 3
dat$curr[dat$shift %in% c("W1", "W2", "WR", "WS")] <- -1

# smoking group info
group <- by(as.numeric(dat$smokstat_s1), dat$pptid, max)

# get total number of epochs for each subject
totalepochs <- by(dat$t, dat$pptid, function(W){2*sum(W)})

# calculate max number of epochs across subjects
max.totalepochs <- max(totalepochs)
```

**Code Sample 3.1: Lasagna Plotting, Subject Information**

This creates two `by` objects, one for smoking group, and the other for the length of epochs for each subject. For smoking, the `max` argument is the simplest way to get a unique identifier for each person. For total epochs, each person’s total sleep time (achieved by the `sum` function), can be multiplied by 2 to get the number of epochs. For example, 1000 total epochs = 500 minutes. To set up the PSG data, we use the following code:

```
# create empty N x E matrix, where E is the maximum sleep length across the cohort
hm <- matrix(NA, nrow=5139, ncol=max.totalepochs)

# arranges data from survival-formatted data frame into "by" object of length N with each patient having a vector denoting their sleep stage in each epoch
hyp5 <- by(dat[, c("t","curr")], dat$pptid,
	function(W){
		holder=NULL
		for(i in 1:nrow(W)){
			holder<- c(holder,kronecker(W[i,"curr"],rep(1,2+W[i,"t"])))
		}
		holder
	}
)

# fill hyp5 into empty hm
for(i in 1:5139) { hm[i,1:length(hyp5[[i]])] <- hyp5[[i]]}
```

**Code Sample 3.2: Lasagna Plotting, Data Manipulation**

First we create an empty matrix `hm` to store the data, where the number of
rows is equal to the number of subjects, and the number of columns is equal to the maximum number of epochs across the cohort. Then, we take the data and arrange it into a by object of length 5139. Each subject will have a vector that contains their sleep stage in each epoch. Each vector will have a different length due to the different sleep durations across subjects. This is done by employing a user-created function that, for each subject, sequentially reads the current state, which is a numeric code, and time-in-state (t), and strings it out for the appropriate number of epochs. The kronecker multiplication achieves this by multiplying the current state numeric code across a vector of length t filled with ones. For example, suppose for a subject we have 0 as the current stage (REM) and this stage lasted for t=1.5. We know that they were in REM for 3 epochs (2*1.5=3). Thus, kronecker(0,rep(1,3)) would result in the series [0,0,0], which indicates 3 epochs of REM sleep. We would then proceed to the next row for this subject (ie, the next sleep stage that occurs) and concatenate that resulting string of numbers onto the previous. Once all subjects’ vectors are joined into a by object, then using a for loop, we can easily fill the empty matrix with this data. Plotting is straightforward from this point:

```r
library(RColorBrewer)
palette <- brewer.pal(11, "PuOr")[c(1,8,3,5,10)]

# plot unsorted lasagna plot
lasagna <- function(X, col=palette, axes=F, ...){
  image(t(X)[,(nrow(X):1)], col=col, axes=axes, ...
}

lasagna(hm, col=palette)

# plot sorted lasagna plot
lasagna(hm[order(group,totalepochs),], col=palette)
```

**Code Sample 3.3: Lasagna Plotting, Visualization**

The unsorted lasagna plot is created with no additional specifications needed, and the sorted lasagna plot requires a simple sorting of the hm matrix using
Lasagna plots were obtained for the 5-, 3-, and 2-state categorizations of sleep, modeled after Swihart et al [26]. Density plots showing the proportion of sleep were created to explore the association between amount of time in a sleep stage and smoking status. The density plot for slow-wave sleep was quantified using a test of equal proportions. Results are provided and discussed in Section 4.1.

The data were also explored quantitatively, across clinical covariates and various sleep parameters, by smoking status. Continuous clinical/demographic variables were analyzed using F-tests, and categorical variables employed $\chi^2$ tests. Post-hoc comparisons between smoking categories were performed, using never smokers as the reference group. Sleep parameters were described using medians and interquartile ranges, as well as F-tests. P-values were classified as being significant ($p < 0.05$) or highly significant ($p < 0.001$) between groups.

### 3.2 Poisson Regression Analysis

Poisson regression models are appropriate for the SHHS sleep data, and provide a good approximation to the multi-state survival model approach [26]. The Poisson regression used in our analyses is log-linear on the rate of transitions:

$$\log\{\lambda^{(gh)}\} = g + h + g : h + \text{offset}\{\log(tar_h)\} + \text{covariates}$$

The above model describes the transition rate $\lambda^{(gh)}$ for smoking group $g$ and transition type $h$ and indicates how the information on the time at risk for transition type $h$ ($\text{tar}_h$) is incorporated into the model.
Figure 3.1: Transition Relative Risks. $\lambda_{XY}$ represents the relative risk of transition from stage $X$ to stage $Y$, compared to never smokers. This figure will be used to visualize risk of transitioning for both former and current smokers.

Exponentiating linear combinations of coefficients (which is explained in more detail below) will result in relative rates of the counts between smoking group $g$ and the reference group for transition type $h$. Each of these sets of rates can be visualized graphically (Figure 3.1) [26]. The regression is executed in R as follows:

```r
library(geepack)
geeglm(count ~ offset(I(log(tar+1))) + age_s1 + gender + rdi4p + BMI_s1 + race + HTNDerv_s1 + cvd + FEV1 + shift*smokstat_s1, id=pptid, data=dat, family="poisson", corstr="independence", scale.fix=TRUE, wave=f.wav, control=geese.control(epsilon=1e-4, maxit=as.integer(10), trace=TRUE, scale.fix=TRUE))
```

Code Sample 3.4: Poisson Regression

The specification of this model follows most basic specifications of Poisson regressions, with a few points to note. The first is that we must add 1 to the time at risk, to avoid issues with $\log(0)$, ie, a case where an individual was not at risk for a certain transition due to not having been in one of the stages. This
occurred in the dataset with individuals not entering REM sleep. The second is that we assume an independent correlation structure since the main motivation is to obtain point estimates and confidence intervals. Correlation in GEE is a nuisance, and consistent estimates are obtained regardless of correlation structure [26]. These regressions were fit for 3- and 5-state models, adjusting for age, gender, RDI (4%), body mass index, race, hypertension, cardiovascular disease, and FEV$_1$, as discussed in Section 2.3. To obtain relative risks, the output of the Poisson model is processed to focus on the linear combination of parameters of interest. In order to perform this linear combination, we must first create the appropriate contrast matrix. Using a function for this contrast matrix creation simplifies the process (Code Sample 3.5).

```r
longtab.func <- function(n, model , t, shifttype){
  # contrast for current smokers
  con.current<- rbind (c(rep (0 ,n) ,1,0,rep(0,t-1),rep(0,t-1)),
    cbind(matrix(0,ncol=n,nrow=t-1),rep(1,t-1),rep(0,t-1),
      diag(t-1)*1, diag(t-1)*0))
  # contrast for former smokers
  con.former<- rbind(c(rep (0 ,n) ,0,1,rep(0,t-1),rep(0,t-1)),
    cbind(matrix(0,ncol=n,nrow=t-1),rep(1,t-1),rep(0,t-1),
      diag(t-1)*0, diag(t-1)*1))
  cont.all<-rbind(con.current , con.former)
  # use user-defined function "linear.comb" to perform the linear combination on the model of interest, using the contrast matrix, and clean up the output
  results<-linear.comb(model , cont.all)
  longtab <- cbind(smokstatus=rep(c("Current","Former"),each=t),shifttype,results)
  names(longtab)[1:2]<- c("Smoking Status", "Shift")
  return(longtab)
}
```

**Code Sample 3.5:** Relative Risk Creation

This function requires that the model is specified with the interaction of shift and smoking status coming last in the model. For an illustrative example, we will show the results of an adjusted Poisson regression:
```r
library(geepack)
geeglm(count ~ offset(I(log(tar+1))) + age_s1 + gender + rdi4p + BMI_s1 +
    race + shift*smokstat_s1,
    id=pptid,
    data=dat,
    family="poisson",
    corstr="independence",
    scale.fix=TRUE,
    wave=f.wav,
    control=geese.control(epsilon=1e-4, maxit=as.integer(10),
     trace=TRUE, scale.fix=TRUE))
```

**Code Sample 3.6:** Poisson Regression, Adjusted

The output from this model appears as follows:

| Coefficients | Estimate | Std.err | Wald | Pr(>|W|) |
|--------------|----------|---------|------|---------|
| (Intercept)  | -3.58472 | 0.05061 | 5015.69 < 2e-16 *** |
| age_s1       | -0.00249 | 0.00056 | 19.79 8.6e-06 *** |
| genderM      | 0.13154  | 0.01107 | 141.17 < 2e-16 *** |
| rdi4p        | 0.00518  | 0.00058 | 79.20  < 2e-16 *** |
| BMI_s1       | -0.00491 | 0.00121 | 16.57  4.7e-05 *** |
| as.factor(race)2 | 0.00848   | 0.02491 | 0.12  0.73346 |
| as.factor(race)3 | -0.04957  | 0.02242 | 4.89   0.02702 * |
| as.factor(race)4 | 0.08823   | 0.04698 | 3.53   0.06041 . |
| as.factor(race)5 | -0.02479  | 0.02702 | 0.84   0.38587 |
| shiftNW      | 1.182532  | 0.01627 | 5278.48 < 2e-16 *** |
| shiftRN      | 0.663821  | 0.01413 | 2206.76 < 2e-16 *** |
| shiftRW      | 1.197207  | 0.01661 | 5193.80 < 2e-16 *** |
| shiftWN      | 2.863970  | 0.01737 | 27189.25 < 2e-16 *** |
| shiftWR      | 0.514429  | 0.03021 | 289.87  < 2e-16 *** |
| smokstat_s1Current | -0.09499 | 0.02523 | 14.18  0.00017 *** |
| smokstat_s1Former | -0.00614 | 0.01567 | 0.15   0.69483 |
| shiftNW:smokstat_s1Current | 0.164094 | 0.03623 | 20.51  5.9e-06 *** |
| shiftRN:smokstat_s1Current | -0.042286 | 0.03518 | 1.44   0.22943 |
| shiftRW:smokstat_s1Current | -0.035461 | 0.03894 | 0.83   0.36245 |
| shiftWN:smokstat_s1Current | 0.148250 | 0.04262 | 12.09  0.00051 *** |
| shiftWR:smokstat_s1Current | 0.003397 | 0.07709 | 0.00   0.96485 |
| shiftNW:smokstat_s1Former | 0.059392 | 0.02335 | 6.37   0.01563 * |
| shiftRN:smokstat_s1Former | -0.036454 | 0.02141 | 2.90   0.08865 . |
| shiftRW:smokstat_s1Former | 0.068352 | 0.02379 | 8.25   0.00407 ** |
| shiftWN:smokstat_s1Former | -0.027488 | 0.02501 | 1.21   0.27182 |
| shiftWR:smokstat_s1Former | 0.009099 | 0.04338 | 0.04   0.83386 |

Signif. codes:  0 ***  0.001 **  0.01 *  0.05 .  0.1

| Scale is fixed. |
| Correlation: Structure = independence |
| Number of clusters: 5139 Maximum cluster size: 6 |

**Code Sample 3.7:** Adjusted Poisson Output

We can see that the main effects of smoking and the interaction terms all appear at the end of the output as needed. Now let’s step through the parameters
needed to input into the \texttt{longtab.func} function. For \( n \), we need to know how many lines of the output will appear before these two estimates. In this case, it is 14 – this will be our specification of \( n \) for the function. The parameter \texttt{model} is simply the saved \texttt{geeglm} object from the regression, in this case \texttt{model2}. The parameter \( t \) describes the number of possible transitions (ie, 6 or 20, in the case of 3-state or 5-state analysis, respectively). Here, \( t \) will be 6. The final specification is that of \texttt{shifttype}, which is a vector of names of the transitions, in this case NR, NW, RN, RW, WN, WR. In this example, the contrast matrix will appear as in Table 3.1. However, for simplicity, the first 12x14 section of the matrix has been omitted. This section of the matrix would correspond to the covariates of the model that are not smoking status or transition type/smoking status interactions, and would be entirely zero.

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|c|c|c|c|c|c|c|c|c|c|c|c|}
\hline
\multicolumn{2}{|c|}{Linear Comb Estimate} & \multicolumn{8}{|c|}{Model Coefficient Estimate} \\
\hline
 & C & F & C*NW & C*RN & C*RW & C*WN & C*WR & F*NW & F*RN & F*RW & F*WN & F*WR \\
\hline
C & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
& NR & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
& RW & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
& RN & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
& WR & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\hline
F & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
& NR & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
& RW & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
& RN & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
& WR & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\hline
\end{tabular}
\caption{Contrast Matrix $\dagger$}
\end{table}

$\dagger$ This matrix would be bound on the left with a 12 x 14 matrix of zeros, where 14 (\( n \)) is the number of lines in the model designating adjusted covariates.  
$\dagger$ $C = $ current smoker; $F = $ former smoker; $C^{\ast}NW = $ model coefficient for smoking status and transition type interaction
Code Sample 3.5 employs another user-defined function, `linear.comb` which helps calculate and clean up the display of the linear combination, using the created contrast matrix. This is shown below:

```r
library(esticon)

# input model fit and contrast matrix
linear.comb <- function(fit, cntrst.mtrx1){
  hold <- cbind(exp(esticon(fit, cntrst.mtrx1)[,c("Estimate", "Lower", "Upper")]),
                esticon(fit, cntrst.mtrx1)[,c("Pr(|X^2|)")])
  hold <- round(hold,2)
  # label columns
  colnames(hold) <- c("RR est", "95% L", "95% U", "p-value")
  hold
}
```

**Code Sample 3.8: Linear Combination Function**

Thus, the final matrix calculated relative risk dataframe would be as shown in Table 3.2. These together allow us to create relative risk plots along with the associated confidence interval, for each transition, for both 3- and 5-state models. We will also visualize entry and exit plots, which illustrate the risk of transitioning in and out of a particular stage. Plots comparing 3- and 5-state risk are used to illustrate the loss of information associated with data coarsening [26].

**Table 3.2: Relative Risk Estimates.** Data is stored in this format after analysis, and is manipulated from this arrangement for plotting and other visualizations.

<table>
<thead>
<tr>
<th>Smoking Status</th>
<th>Shift</th>
<th>RR est</th>
<th>95% L</th>
<th>95% U</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Current</td>
<td>NR</td>
<td>0.91</td>
<td>0.87</td>
<td>0.96</td>
<td>0.00</td>
</tr>
<tr>
<td>Current</td>
<td>NW</td>
<td>1.07</td>
<td>1.02</td>
<td>1.13</td>
<td>0.01</td>
</tr>
<tr>
<td>Current</td>
<td>RN</td>
<td>0.87</td>
<td>0.80</td>
<td>0.95</td>
<td>0.00</td>
</tr>
<tr>
<td>Current</td>
<td>RW</td>
<td>0.88</td>
<td>0.82</td>
<td>0.94</td>
<td>0.00</td>
</tr>
<tr>
<td>Current</td>
<td>WN</td>
<td>1.05</td>
<td>0.98</td>
<td>1.13</td>
<td>0.14</td>
</tr>
<tr>
<td>Current</td>
<td>WR</td>
<td>0.91</td>
<td>0.79</td>
<td>1.06</td>
<td>0.23</td>
</tr>
<tr>
<td>Former</td>
<td>NR</td>
<td>0.99</td>
<td>0.96</td>
<td>1.02</td>
<td>0.69</td>
</tr>
<tr>
<td>Former</td>
<td>NW</td>
<td>1.05</td>
<td>1.02</td>
<td>1.09</td>
<td>0.00</td>
</tr>
<tr>
<td>Former</td>
<td>RN</td>
<td>0.96</td>
<td>0.91</td>
<td>1.01</td>
<td>0.12</td>
</tr>
<tr>
<td>Former</td>
<td>RW</td>
<td>1.06</td>
<td>1.02</td>
<td>1.11</td>
<td>0.01</td>
</tr>
<tr>
<td>Former</td>
<td>WN</td>
<td>0.97</td>
<td>0.93</td>
<td>1.01</td>
<td>0.10</td>
</tr>
<tr>
<td>Former</td>
<td>WR</td>
<td>1.00</td>
<td>0.92</td>
<td>1.09</td>
<td>0.94</td>
</tr>
</tbody>
</table>
Results

4.1 Visualizations and Data Exploration

Out of the 5139 subjects for this analysis, 11% of them were current smokers, 42% of them were former smokers, and 47% of them were never smokers (Table 4.1). Former smokers were significantly older (mean age = 64.1) than never smokers (63.2), although this difference is likely not clinically relevant. Current smokers (59.2) were significantly younger than never smokers. There were significant differences across groups in gender composition, with former smokers having the highest proportion of male subjects. The distribution of race was different across smoking groups as well ($p < 0.001$). Body mass index and FEV were similar across groups, with former smokers having highly significant $p$-values due to large sample size. Both groups show differences in RDI, with current smokers having significantly lower median RDI than never smokers (2.6 vs 3.7; $p < 0.05$), and former smokers having significantly higher RDI (5.1 vs 3.7; $p < 0.001$). Cardiovascular disease was significantly higher in former than never smokers (19.9% vs 13.0%; $p < 0.001$).

Lasagna plot visualization of the data can provide useful insights into data patterns. Figure 4.1 provides the lasagna plots with each column of panels corresponding to 5-, 3-, and 2-state data. Each lasagna plot displayed in one of
### Table 4.1: Covariate Information

<table>
<thead>
<tr>
<th>Covariate</th>
<th>Smoking Status</th>
<th>Never Smoker n = 2406</th>
<th>Current Smoker n = 550</th>
<th>Former Smoker n = 2183</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age in years</td>
<td></td>
<td>63.2 (11.3)</td>
<td>59.2 (9.2)*</td>
<td>64.1 (10.2)**</td>
</tr>
<tr>
<td>Male gender (%)</td>
<td></td>
<td>35.8</td>
<td>49.6**</td>
<td>57.9**</td>
</tr>
<tr>
<td>Race (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td></td>
<td>79.3</td>
<td>65.1**</td>
<td>84.3**</td>
</tr>
<tr>
<td>African American</td>
<td></td>
<td>6.1</td>
<td>5.5**</td>
<td>4.9**</td>
</tr>
<tr>
<td>American Indian</td>
<td></td>
<td>7.2</td>
<td>25.1**</td>
<td>8.0**</td>
</tr>
<tr>
<td>Other</td>
<td></td>
<td>7.2</td>
<td>4.4**</td>
<td>2.8**</td>
</tr>
<tr>
<td>Body mass index</td>
<td></td>
<td>28.1 (5.2)</td>
<td>27.9 (5.2)</td>
<td>28.8 (5.1)**</td>
</tr>
<tr>
<td>Respiratory disturbance index–median (IQR)</td>
<td></td>
<td>3.7 (1.2-10.2)</td>
<td>2.6 (0.8-7.9)*</td>
<td>5.1 (1.6-12.0)**</td>
</tr>
<tr>
<td>Forced expiratory volume</td>
<td></td>
<td>2.6 (0.8)</td>
<td>2.7 (0.8)</td>
<td>2.7 (0.8)**</td>
</tr>
<tr>
<td>Cardiovascular disease (%)</td>
<td></td>
<td>13.0</td>
<td>14.5</td>
<td>19.9**</td>
</tr>
</tbody>
</table>

* $p < 0.05$ compared to never smokers
** $p < 0.001$ compared to never smokers

The 3 panels in each column has 5139 rows (subjects) and 1218 columns (epochs), which corresponds to 10.15 hours, the maximum observed sleep time. The top panel in each column displays the color-coded hypnogram for each subject in the order subjects are organized in the data. The left panel, corresponding to 5-state data, contains more colors because there are more states, while the middle and right panels have fewer colors, indicating a coarser classification of the sleep hypnogram at the subject level. The middle panels in every column display the same data but sorted by the length of sleep and partitioned (from top to bottom) into three groups: never, current, and former smokers. These panels provide additional organization of the original data and indicate that transitions and stages are relatively stable in the first part of the night, with many subjects in slow-wave sleep (SWS) 20-40 minutes after falling asleep. The bottom panels contain the same data, but now sorted by sleep stage within
column and groups. In these panels, rows are no longer subjects, since sorting by sleep stage scrambles subjects. Instead, for each time point after sleep onset, the graph indicates how many subjects are in various stages of sleep. For example, one hour after falling asleep there are 313 (13%) non-smokers who are awake, 404 (16.8%) who are in REM, and 648 (26.9%) who are in SWS. At the same time, there are 68 (12%) smokers awake, 85 (15%) smokers in REM, and 126 (23%) in SWS. This type of data visualization is losing some information, especially about transitions, but provides useful insights in terms of marginal comparisons between groups. The population level distribution of sleep stages over time is consistent with the literature. Indeed, SWS is more prevalent during early stages of sleep, while the proportion of subjects in REM increases throughout the night. Figure 4.1 suggests that the proportion of slow-wave sleep across smoking categories varies. However, such visualizations should be taken only as exploratory analyses, especially because plots provide the number of subjects and not the proportion of subjects in various stages of sleep. Thus, some of the imbalance may be due to the different number of smokers and non-smokers in the sample (550 current versus 2406 never/2183 former). To account for the imbalance we have also produced plots where we up-sample the current smokers to obtain comparable graphs with the same number of smokers and non-smokers. Up-sampling is a procedure where we sampled with replacement the smokers, with the number of samples being equal to the number of non-smokers (Figure A.1).

Indeed, it may be useful to compare the proportion of subjects in specific states across groups over time. Figure 4.2 displays the proportion of slow-wave sleep for never (black line), current (red), and former (blue) smokers. The plot
Figure 4.1: Lasagna plots for 5-state, 3-state, and 2-state sleep (from left to right). Each panel is 5139 rows (subject) by 1218 columns (epochs). Top panels: color-coded hypnograms for each subject as organized in data. Middle panels: subjects sorted by sleep length and partitioned from top to bottom into never, current, former smokers. Bottom panels: sorted additionally by sleep stage within column and group.
suggests that current smokers have overall less slow-wave sleep, while former smokers have a proportion of slow-wave sleep relatively between current and non-smokers. (Figure 4.2). For example, about 2 hours after falling asleep, ∼22% of current smokers, 25% of former smokers and ∼28% of non-smokers are in slow-wave sleep. The absolute difference between the proportion of smokers and non-smokers in slow-wave sleep (∼6%) may appear small, but corresponds to a ∼20% reduction in the proportion of smokers in slow-wave sleep compared to non-smokers. We will further investigate these differences in a statistical framework, but it is important to understand the magnitude of differences in easy to understand plots. At the first peak, specifically, these differences in proportions are significantly different (never vs former, p=0.04; never vs current, p=0.013).

![Figure 4.2](image)

**Figure 4.2**: Proportion of Slow-Wave Sleep. Proportions were calculated as the proportion of individuals from each subgroup that were in slow-wave sleep during that epoch (x-axis converted to hours for interpretability).

We can also look at further quantitative differences between the groups in sleep variables (Table 4.2). Current smokers have a higher sleep latency than
never smokers (22 minutes vs 16 minutes; \( p < 0.001 \)). Former smokers have a shorter REM latency than never smokers, and this difference was not seen in current smokers (71 minutes vs 75 minutes; \( p < 0.05 \)). Current and former smokers also have shorter sleep time, more stage 1 sleep, more stage 2 sleep, and less slow-wave sleep. Former smokers have slightly less sleep efficiency. No differences were seen in REM sleep percentages.

Table 4.2: Sleep Information

<table>
<thead>
<tr>
<th>Sleep parameter</th>
<th>Smoking Status</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Never Smoker ( n = 2406 )</td>
</tr>
<tr>
<td>Sleep latency (minutes)</td>
<td>16 (9.5-28)</td>
</tr>
<tr>
<td>REM latency (minutes)</td>
<td>75 (56.5-108.4)</td>
</tr>
<tr>
<td>Total sleep time (hours)</td>
<td>6.2 (5.5-6.8)</td>
</tr>
<tr>
<td>Sleep efficiency (%)</td>
<td>85.7 (78.5-90.9)</td>
</tr>
<tr>
<td>Stage 1 sleep (%)</td>
<td>4.2 (2.6-6.4)</td>
</tr>
<tr>
<td>Stage 2 sleep (%)</td>
<td>56.2 (48.3-64.2)</td>
</tr>
<tr>
<td>Slow-wave sleep (%)</td>
<td>18.6 (10.1-26.6)</td>
</tr>
<tr>
<td>REM sleep (%)</td>
<td>20.2 (16.0-24.0)</td>
</tr>
</tbody>
</table>

* \( p < 0.05 \) compared to never smokers
** \( p < 0.001 \) compared to never smokers

4.2 3-state Modeling

Using the Poisson GEE models described in Section 3.2, we conduct statistical analyses of sleep stage transition data. We will start with the 3-state data and present results for the fully-adjusted model visually. We focus on estimating the relative risk for each transition type and smoking status (Code Samples for calculating the relative risks are shown in Sections 3.4 and 3.8). Figure 4.3 presents the results in the form of point estimates and confidence intervals with red intervals corresponding to current smokers and blue corresponding to former smokers.
smokers. If confidence intervals cross 1 then there is no evidence of a different transition risk at the alpha level of 0.05. Intervals that contain 1 are shown as faded (or more transparent) to further emphasize statistically significant results.

Figure 4.4 presents the same results from Figure 4.3 in a triangular diagram. Each type of transition is displayed as an arrow and the associated number is the relative risk. A star indicates statistical significance at the 0.05 level.

**Figure 4.3:** Relative Risks, Fully Adjusted, 3-state. Model adjusted for race, age, sex, RDI (4%), BMI, hypertension, cardiovascular disease, and FEV₁. Transparent intervals (ie those not containing 1) are statistically significant

**Figure 4.4:** Relative Risk Diagrams. Data from Figure 4.3 represented in flowcharts. A star indicates statistical significance at the 0.05 level.
Table 4.3: 3-state Model Estimates

<table>
<thead>
<tr>
<th>Shift</th>
<th>Current Smoker</th>
<th>Former Smoker</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Model 1&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Model 2&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>NR</td>
<td>0.92 [0.88, 0.97]</td>
<td>0.91 [0.87, 0.96]</td>
</tr>
<tr>
<td>NW</td>
<td>1.08 [1.03, 1.14]</td>
<td>1.07 [1.02, 1.13]</td>
</tr>
<tr>
<td>RN</td>
<td>0.88 [0.81, 0.96]</td>
<td>0.87 [0.80, 0.95]</td>
</tr>
<tr>
<td>RW</td>
<td>0.89 [0.82, 0.95]</td>
<td>0.88 [0.82, 0.94]</td>
</tr>
<tr>
<td>WN</td>
<td>1.07 [0.99, 1.14]</td>
<td>1.05 [0.98, 1.13]</td>
</tr>
<tr>
<td>WR</td>
<td>0.92 [0.79, 1.07]</td>
<td>0.91 [0.79, 1.06]</td>
</tr>
</tbody>
</table>

<sup>a</sup>Model 1: Unadjusted
<sup>b</sup>Model 2: Adjusted for age, gender, race, BMI, RDI (4%)
<sup>c</sup>Model 3: Adjusted for age, gender, race, BMI, RDI (4%), CVD, HTN, FEV<sub>1</sub>

The above results indicate that former smokers are more similar in their sleep transitions to never smokers than current smokers. Former smokers transition more often into wake from either non-REM or REM sleep. Current smokers are more likely to transition out of non-REM sleep into wakefulness and less likely to transition into REM sleep. Smokers have a lower risk of transitioning out of REM sleep, making REM sleep a “stickier state” for smokers relative to non-smokers.

We can also investigate the utility of adjusting the models for covariates (Table 4.3). We compare an unadjusted model (only smoking status and sleep stage), to a model adjusted additionally for age, gender, race, BMI, and RDI (4%), and to a third model adjusted for cardiovascular disease, hypertension, and FEV<sub>1</sub>. Model adjustment does not substantially change the relative risk estimates for current or former smokers.
4.3 5-state Modeling

We have conducted a similar analysis for the 5-state data. Figure 4.5 displays the results in a different format to account for the much larger number of possible transitions.

**Figure 4.5**: Relative Risks, Fully Adjusted, 5-state. Model adjusted for race, age, sex, RDI (4%), BMI, hypertension, cardiovascular disease, and FEV$_1$. Results are organized by significance first, then direction of significance, then by stage, based on current smokers’ results. Former smokers used the same labeling scheme. Transparent intervals (ie those not containing 1) are statistically significant.

For plotting purposes we organized results by significance first, then by direction of significance, then by stage, based on the results for smokers. Results for former smokers use the same labeling scheme to make results for current and
former smokers easier to compare. This figure provides a more refined view of transitions than Figures 4.3 and 4.4, especially in the transitions in and out of non-REM states. In former smokers, stage 2 sleep seems to be most affected, with disruptions in the rates both into and out of this stage. Also, transitions to wakefulness are higher for nearly all states in former smokers. For current smokers, transitions out of stage 1 and 2 are reduced. Smokers are more likely to transition to and stay in wake and stage 1 (ie, 2W, 21, W1). Smokers are less likely than non-smokers (1S,1R, 12, 2S, 2R) to transition into deeper levels of sleep, and are less likely to transition out of REM (R2, RW). Some estimates (1S, WS, S1, and RS) are highly variable. This is due to the small number of this type of transitions in the population.

We can compare directly the results for the 3-state and 5-state transition models. Figure 4.6 provides a side-by-side comparison of results. The black vertical line in each of the four panels delineates results for the 3-stage and 5-stage models. For example, panel one (top/left) shows the non-REM to REM sleep transition relative risk for smokers (red) and former smokers (blue). To the left of the black vertical line the 3-state model indicates that smokers, but not former smokers, have a lower risk of transition into REM (R) from non-REM (N) sleep. To the right of the black vertical line the results for the corresponding 5-state model states are shown. More precisely, the NR transitions are broken down into transitions into REM from stage 1, 2 and SWS into R. Confidence intervals for the 5-state model are larger because the number of transitions is smaller (NR is broken down into 3 different types of transitions). The results indicate that there is a penalty for going to a more refined definition of stages, while when data are available a more granular view can be obtained. In this
case we obtain that the NR differences between smokers and non-smokers are due both to lower risk of transition between Stages 1 and 2 and REM. In this instance, the SWS to REM transitions do not appear to be statistically different between smokers and non-smokers, and indeed when looking at the 5-state analysis, there are no more finely visible differences. However, in the lower left panel, we see that non-significance in former smokers in the 3-state model was obscuring a more detailed picture in the 5-state model. As reported before, the confidence intervals for transitions related to SWS are wider and sometimes extreme because the number of transitions in and out of SWS is much smaller.

Furthermore, we can look at entry and exit plots for another view of the data (Figures A.2-A.6). These plots further indicate the preference for light sleep in both current and former smokers. Both seem to have reduced transitions from stages 1 and 2 into deeper stages of sleep.
Figure 4.6: 3-state Relative Risks and 5-state Analogues. The vertical black line delineates the results from the 3- and 5-state models.
Conclusions

The goal of the current analysis was to understand and quantify the relative risk for transitioning between different stages of sleep as a function of smoking status. The previous literature indicates that smoking impacts sleep architecture and quality of sleep [13, 14, 15, 16, 11, 17, 18, 19]. However, analyses have not investigated the effect of adjusting for event history. Using the Poisson (log-linear) model in a GEE framework, we provide a detailed description of differences in sleep architecture between groups. This approach is more comprehensive than using transition counts or number of arousals. Both Poisson models and survival models can be used to describe relative risks of transition between sleep stages. Results were reported to not differ significantly between these approaches, while Poisson models require a much simpler data structure and are much faster to implement [7, 26]. Thus, more complex models can be run very quickly and efficiently, which allows us to explore a larger number of models, run simulations, and obtain bootstrap estimators if desired [26]. We provide R code that describes both how to organize and fit the data using these Poisson models.

Results for the 3-state model indicate that former smokers transition more often into wake. This seems to be consistent with the previously published literature. Indeed, it has been shown that in nicotine withdrawal, there is an
increase in awakenings during sleep [20]. The 5-state analysis indicates that there is reduction in the probability of transitions in and out of stage 2 for former smokers. Current smokers are more likely to transition out of non-REM into wake and are less likely to transition from non-REM into REM. These findings complement previous literature showing that current smokers spend more time awake [16]. Our results also indicate that smokers are at reduced risk for transition out of REM from either stage, which may indicate less time spent in REM or more pressure to stay in REM sleep. Based on the 5-state model, current smokers have a reduced transition probability out of stages 1 and 2, and are more likely to transition to and stay in wake and stage 1 (ie, 2W, 21, W1). Current smokers are also less likely to transition into SWS and REM, and are less likely to transition out of REM.

Our results indicate that both 5-state and 3-state models can be useful in practice, especially when datasets are relatively large. When datasets are smaller or the number of stages increases, estimation uncertainty grows. Therefore, the two models are complementary and support each other. For example, the 5-state model provides a way to analyze and quantify disruptions in sleep transitions between non-REM sleep stages. However, the number of transitions (20) in a 5-state model makes it challenging to follow and translate findings in a way that is easily discernible by the scientific community. Thus, using a 3-state model may be helpful to present some of the findings in a more compact format.

One limitation of the data is the visual scoring of sleep stages done during pre-processing. Indeed, the stages in the 3-state models are probably less prone to misclassification. Another limitation is that sleep stages are defined in 30-second epochs, making transitions between stages that happen within 30-second
interval impossible to detect. A third limitation is that the former smoking
group is likely a group with highly heterogeneous smoking history. Indeed,
some individuals may be in earlier stages of cessation, while others may have
quit smoking a long time before the beginning of the study. Thus, even if
a dose/response effect exist as a function of time-from-quitting, it cannot be
estimated using this dataset.

Our analyses raise several important follow-up questions. For example, it is
important to quantify the association between smoking status and self-reported
sleep quality. Also it may be interesting to evaluate whether transition risks
mediate the effect of smoking on restless or light sleep. A previous paper has
laid the groundwork for this analysis [8]. Another technical problem is to further
explore data using lasagna plots and explore whether effects are relatively stable
as a function of time from falling asleep.
References


[34] Susan Redline, H. Lester Kirchner, Stuart F. Quan, Daniel J. Gottlieb, Vishesh Kapur, and Anne Newman. The effects of age, sex, ethnicity,


Additional Figures

Figure A.1: Up-sampled 5-stage Lasagna Plot. Same as the bottom left panel in Figure 4.1, but with the current smokers data plotted after upsampling (550*4=2200 subjects) to make visualization across groups more comparable.
Figure A.2: Entry/Exit for Stage 1. The four sets of estimates to the left indicate relative risk of transitioning into stage 1, either for current or former smokers, and the four sets to the right indicate transitions out of stage 1.

Figure A.3: Entry/Exit for Stage 2. The four sets of estimates to the left indicate relative risk of transitioning into stage 2, either for current or former smokers, and the four sets to the right indicate transitions out of stage 2.
Figure A.4: Entry/Exit for REM. The four sets of estimates to the left indicate relative risk of transitioning into REM, either for current or former smokers, and the four sets to the right indicate transitions out of REM.

Figure A.5: Entry/Exit for Slow-Wave Sleep. The four sets of estimates to the left indicate relative risk of transitioning into slow-wave sleep, either for current or former smokers, and the four sets to the right indicate transitions out of slow-wave sleep.
Figure A.6: Entry/Exit for Wake. The four sets of estimates to the left indicate relative risk of transitioning into wake, either for current or former smokers, and the four sets to the right indicate transitions out of wake.
Biographical Statement

The author was born in Trenton, NJ on June 20, 1990. She attended Dickin-
son College and graduated with a bachelors degree in Neuroscience in 2012, with
minors in Psychology and Mathematics. After graduation, she conducted data
collection and data management for various stroke trials, first at Georgetown
University and then at the University of Arizona.

She began her masters degree in Biostatistics at Johns Hopkins Bloomberg
School of Public Health in 2014. She conducted research in various areas includ-
ing longitudinal behavior of multiple sclerosis lesions, reproducibility of func-
tional magnetic resonance imaging, and causes of sleep fragmentation, under
advisor Ciprian Crainiceanu. She served as a teaching assistant for Statistical
Methods in Public Health I, II and IV, and Non-Inferiority and Equivalence
Clinical Trials. She also was co-organizer of the Department of Biostatistics
Journal Club during the 2015-2016 academic year.