# A DATABASE FOR EXPLORATORY ANALYSIS OF HUMAN SLEEP

by

Shivin Misra

A Thesis

Submitted to the Faculty

of the

WORCESTER POLYTECHNIC INSTITUTE

in partial fulfillment of the requirements for the

Degree of Master of Science

in

Computer Science

May 2008

APPROVED:

_____

Professor Carolina Ruiz, Thesis Advisor

_____

Professor Sergio A. Alvarez, Thesis Co-advisor

Computer Science Department, Boston College

_____

Professor Elke A. Rundensteiner, Thesis Reader

_____

Professor Michael A. Gennert, Head of Department

# Acknowledgements

I would like to thank my parents and little brother for all the support and encouragement they give me. I am thankful to my advisors, Prof. Ruiz and Prof. Alvarez for always being there to give me good advice that helped me overcome the obstacles in this thesis, for their patience in going over the details of this project, their valuable feedback that helped me improve the quality of the work, and their encouragement that always kept me motivated. Prof. Ruiz, thank you for carrying hundreds of patient files all the way from the sleep clinic to WPI and back so that our massive data collection could steadily progress. I'd like to thank my thesis reader, Prof. Rundensteiner for her precious time in reading my thesis, the Day Kimball Sleep clinic technicians for the patient files and information on human sleep data, J. Banning, the system administrator at WPI-CS (Computer Science) for his great help in system related arrangements during the data collection, database build and backup phases of this project, the WPI-CS and WPI-CCC (Computing and Communications Center) departments for providing us with computing facilities, and the developers of the excellent PostgreSQL database system. This section would be incomplete without the mention of my friends, Swaraj, Kaus, SK, Salma, Ketki, Venks, Nithin, Natalie, Ryan, Debu, Avraj, Amit with whom I felt like I was at home and had the best time during my Masters.

# Abstract

This thesis focuses on the design, development, and exploratory analysis of a human sleep data repository. We have successfully collected comprehensive data for 1,046 sleep disorder patients and created a Terabyte-scale database system to handle it. The data for each patient was collected from the patient's medical records, and from the patient's allnight sleep study (for a total of about 0.6 Gigabytes per patient). Data collected from the patient's medical record contain more than 70 attributes, including demographic data, smoking, drinking, and exercise habits, depression and daytime sleepiness questionnaires, and overall medical history. Data collected from the patient's all-night sleep study consist of 50-55 time-series signals recorded during a period of 6-8 hours at the hospital's sleep clinic. These signals include among others an electroencephalogram, electromyogram, electrooculogram, electrocardiogram, and signals tracking blood oxygen level, body position, limb movements, snoring and blood pressure. 350 additional attributes summarize sleep related events taking place during the night long study, including sleep stages, arousals, and respiratory disturbances.

Particular attention during the development of our database system was paid to a database design that effectively handles the data size and complexity, that describes the structure of sleep data in clinically meaningful terms, and that will facilitate the discovery of patterns in sleep data using machine learning algorithms. We have interfaced our database with Weka, a well known data mining system. To the best of our knowledge, our database is one of the world's largest and most comprehensive in the domain of human sleep disorders.

# Contents

# List of figures

# List of tables

# 1 Introduction

Every year, clinics conducting studies on patients suffering from sleep irregularities collect large amounts of clinical and survey data. Clinical instruments attached to a patient via electrodes are able to record time-series signals that originate from biochemical and physical processes taking place within the sleeping patient's body.

Polysomnography (PSG/sleep study) is a clinical procedure that records physiologic attributes during sleep. A polysomnogram reads the electrical potential of the brain (using EEG - Electroencephalogram); electrical activity of muscles (using EMG – Electromyogram); and electric potentials resulting from eye movements (using EOG – Electrooculogram) into time series signals. Also, signals tracking body processes like electrical activity of heart (using ECG – Electrocardiogram), blood oxygen level, body position, limb movements, snoring and blood pressure are registered. A clinical technician generates a patient's summary report after the end of each sleep study session.

As described in[7], before a sleep study is performed, the patients are asked to fill out survey questionnaires. These questionnaires carry demographic details like age, height, weight, body mass index, and collar size. They also have information on the patient's medical history, the medicines administered, sleep, exercise, smoke, caffeine and alcohol consumption habits. Epworth Sleepiness Scale (ESS) is a set of questions that determines level of daytime sleepiness in a patient. Beck Depression Inventory (BDI) gives insight into the existence and severity of symptoms of depression. Both, ESS and BDI are included in the survey questionnaire.

## *1.1 Scope of the thesis*

There exists a need to collect, organize, represent and store all of the above data for the purpose of computational analysis. To achieve this goal, we develop a platform that can store every patient's data over time. Moreover, the data organization should conform to

the conceptual model of the domain as understood by the medical experts. This thesis work aims at correctly representing and storing clinical sleep data for its analysis with a scope to accrue data of approximately 1000 patients. The data used in this project came from the patients undertaking sleep studies at the Sleep Disorder Center at Day Kimball Hospital, CT.

The sleep clinic registers the polysomnographic time-series signal recordings using the medical software REMbrandt [13], which is a Windows® based sleep monitoring and analysis system for clinical and research applications.

This research work is being conducted in collaboration with Dr. Majaz Moonis from the University of Massachusetts Medical School.

# 2 Background

## *2.1 Databases in Computer Science*

A database is an organized and well-structured repository of data. Database Management systems (DBMS) help in managing this information by enabling us to systematically add, delete, update, and query the data stored in the database. They also keep a check on the consistency of data.

Within a database, information is stored in tables. A table is a collection of records, with each record containing data, whose properties are identified by the attributes (names of table's columns) that are defined in the table's design. In a good design, a table models a group of related data with each record (also known as a row or tuple) identified by special attribute(s) known as key attribute(s). In a practical database, there exist many tables that may be related to each other by referencing attributes (foreign keys). The overall design of the database is called a database schema [16].

This thesis involves understanding the structure of real human sleep data, modeling it into database tables along with constraints, creating the database, populating it with sleep data and performing pre-processing to prepare the data for computational analyses. The end result is a fully functional database system built from scratch, that hosts nearly 500 GB of rich sleep data for knowledge discovery: the wealth of information that existed only in the form of patient files and compact discs is now available to us scientists for intense computational research.

To design the database for sleep data, the logical data model should reflect the conceptual understanding of the sleep domain by capturing the relationships between the entities and describing their attributes. The physical design phase involves creating a database, defining schemas, tables and primary and foreign keys for organizing and storing sleep data. Apt data types for the attributes should be chosen, and the design must be validated by normalization procedures. Also, it should facilitate exploratory analysis.

The clinical sleep data has time-series information collected overnight for each patient. Also, for every patient, we have set valued information like patient's medical history, list of medicines and patient's response to probable reasons for sleep disorder. We need a DBMS equipped with data types that can model this information.

## *2.2 Selection of the database*

We selected our DBMS keeping in mind the nature of data we want to represent:

- Time-series information
- List of values for an attribute

To handle the above data, we needed a database system that could support an array-like data structure with variable length.

**Candidate DBMS**

From our research into the most popular DBMSs in the industry, we chose PostgreSQL 8.2.5 (PGSQL) as the DBMS for this thesis.

- MySQL 5.1 was ruled out because of its inability to model arrays or user data types.
- Microsoft SQL Server 2005, Express Edition is a freely downloadable DBMS engine, but with a limited set of features.
- Oracle10g1 DBMS was not selected because even though it had facility for custom data types and arrays (`VARRAYS`), the size of the arrays was fixed and needed to be specified in the table definition. Since our data has varying lengths of sleep study sequences, this was not a feasible choice.

Due to the enormity of data we deal with, further exploration into the abilities of the PGSQL DBMS to scale up to high data sizes yielded the following interesting figures:

**PostgreSQL data handling capacity** [10]

| **Limit** | **Value** |
|---|---|
| Maximum Database Size | Unlimited |
| Maximum Table Size | 32 Terabytes |
| Maximum Row Size | 1.6 Terabytes |
| Maximum Field Size | 1 Gigabytes |
| Maximum Rows per Table | Unlimited |
| Maximum Columns per Table | 250 - 1600 depending on column types |
| Maximum Indexes per Table | Unlimited |

**Table 2.1 Postgres DBMS limits**

As from the table above, the data handling capabilities of PGSQL are well suited for the type and amount of data we deal with in this research. The features of PGSQL can support both lists and custom data types. It is an open source, lightweight DBMS that has proven to be capable of satisfying all the needs of this research project.

Note: In the context of PostgreSQL DBMS, a database contains one or more named schemas, which in turn contain tables. Schemas also contain other kinds of named objects, including data types, functions, and operators [11].

## *2.3 Sleep Medicine*

### *2.3.1 Study Types*

There are four types of studies that a encountered in the patient files. The number of signals recorded can vary for different studies.

- Full Polysomnogram (PSG)

  Full PSG is the regular night long sleep study which records limb, arm and eye movements, activity of the brain, breathing rate, oxygen level and heart rate into time series data. Patients suffering from sleep related disorders like narcolepsy, sleep walking, restless leg syndrome, nocturnal seizures are advised to take a full PSG study to analyze their sleep.

- CPAP titration

  CPAP stands for Continuous Positive Airway Pressure and is usually administered in cases of sleep apnea. A small, comfortable mask is fitted on the patient's nose. The mask is connected to a CPAP unit that can deliver air pressure through the nose to the air passage of the patient. The pressure is titrated (controlled) in order to keep the back of the airway open during the patient's sleep. This allows the patient to breathe in all body positions and helps in achieving a restful sleep.

- BiPAP titration

  BiPAP stands for Bilevel Positive Airway Pressure. It is meant for patients who are not tolerant/ responsive of CPAP because of the need to exhale against the extra pressure in CPAP. BiPAP delivers what CPAP does, but keeps two levels of pressure, one for inhaling and one for exhaling, thus decreasing the effort made in breathing against the air pressure.

- Split Night PSG with CPAP titration

  Split night PSG study is conducted when the patient is diagnosed with moderate to severe apnea during the first part of nights study. During the later half of the night, CPAP titration is administered. The technical summary report consists of baseline/ treatment sections corresponding to these two halves of the Split night study, respectively.

### *2.3.2 Signal Information*

EEG, EOG, EMG, and ECG are multidimensional PSG signals that are recorded by keeping sensor electrodes at different places on the body.

*Electroencephalogram (EEG)* records the electrical activity in the brain. EEG electrodes are placed at different locations on the scalp. EEG is comprised of the following signals named: C3-A2 (left-central), C4-A1 (right central), O1-A2 (left occipital), and O2-A1 (right occipital).

The figure below shows the placement of various electrodes on a human cranium:

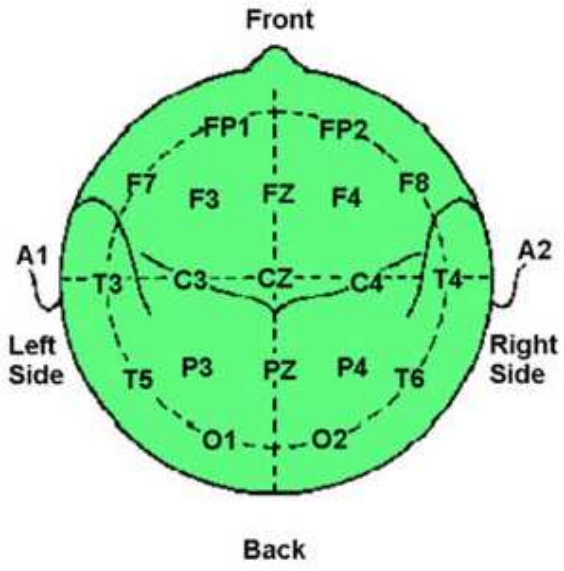



**Figure 2.1 Electrodes on a human cranium**

(Note that the letters are technical names of electrodes [14])

*Electrooculogram (EOG)* electrodes are placed slightly above the outer canthus of the right eye and slightly below the outer canthus of the left eye. These electrodes help in determining when sleep happens and when REM stage is reached. The EOG signals are: ROC-A1, LOC-A2, XFlow, XSum, RMI, Phase, and RR.

*Electromyogram (EMG)* electrodes are placed over the chin, arms and legs to track muscle tensions and leg movements. EMG is composed of signals named CHIN1, L LEG2, R LEG3, ARMS4.

*Electrocardiogram (ECG)* electrodes are placed on the chest to record electrical activity of the heart with every single heart beat (units: μV). The heart beat rate is also recorded in beats per minute (BPM). The heart activity is tracked by the signals called HEARTRATE and EKG8.

The patient's respiratory response is tracked by electrodes named PSNORE (snoring), CFLOW (CPAP flow), FLOW5 (nasal-oral airflow), CHEST (thoracic), ABDM (abdominal), CPRESS (CPAP pressure) placed near the nose and on the chest of the patient.

*SaO2, Oxygen Saturation*, is measured by a probe called Oxymeter that is placed on a finger to record the percentage of oxygen level in the blood stream. The probe is like a clip that has a small red light on one side, and a detector on the other to measure the amount of oxygen in the blood.

**Note:** Electrode names have been referred from [2].

The figure below shows some time-series signals recorded by the different electrodes monitoring patient's sleep, as seen in the REMbrandt [13] Viewer application.

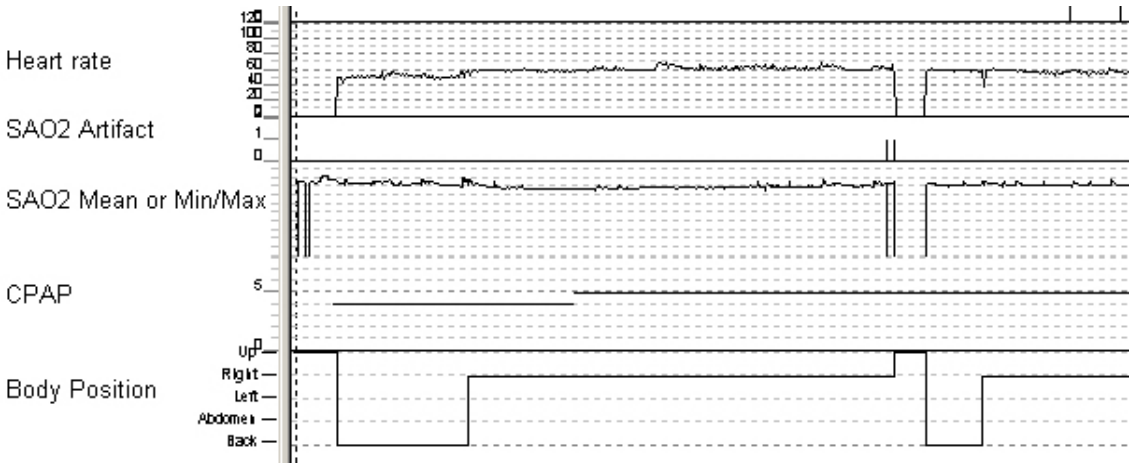



**Figure 2.2 Signal Waveforms**

### *2.3.3 Sleep domain definitions*

**Events**

Events are patterns in data that capture time-related occurrences of interest [7]. Some examples of events are: the class of sleep stage (see sleep stages below); body position - back, prone, tight side, or left side; and heart rate – normal, mildly-reduced or low.

**Epochs**

Every thirty second piece of a sleep recording constitutes an epoch. Sleep stage scoring involves classifying the polysomnogram signals on an epoch – to – epoch basis [7]. The REMbrandt system [13] displays events for every epoch.

**Sleep Stages**

Sleep is classified into six stages: Wake, Stage 1, Stage 2, Stage 3, Stage 4, and REM stage as per the Rechtschaffen and Kales (R&K) system [12]. The Non-REM stage is made up of Stages 1, 2, 3 and 4.

Some definitions of attributes present in the Chapter 4 Technical Summary Report are given below. These definitions are taken from [19].

**Time in Bed (TIB)**

Time spent in bed without attempting to sleep or unsuccessfully trying to sleep. Time spent in resting or napping during the day.

**Sleep Period Time (SPT)/Total Sleep Period**

This refers to the time from sleep onset to the final awakening from the main sleep period of the day. Total sleep period increases with age because of the increase in the number of awakenings.

**Total Sleep Time (TST)**

This refers to the total sleep period minus the time spent awake during the sleep period. Studies have found the total sleep time to be either reduced or unchanged in the older population compared with younger age groups.

**Sleep latency**

This is the time from the decision to sleep to the onset of sleep. Studies have found considerable variability in individuals. In females, sleep latency has been related to both age and hypnotic drug use.

**Wake after sleep onset (Waso)**

This is the time spent awake from sleep onset to final awakening. An increase occurs in the time spent awake after sleep onset in the older population.

**Sleep efficiency**

This is the ratio of total sleep time to nocturnal time in bed. Most studies have found sleep efficiency to be decreased in the older population.

### *2.4 Patient file organization*

A typical patient file is organized in the following way:

1. One or more CDs (backup copy) containing overnight sleep data and technical summary report. The overnight sleep data is in '.msr' format, and can be read with REMBrandt Viewer. There can be different sets of CDs corresponding to different types of sleep study the patient might have undertaken.

2. Log of any communication history with the patient.

3. Analysis report written by the interpreting physician.

4. Medical history of the patient.

5. Pre-sleep study form.

6. Survey questionnaires.

7. Post-sleep study form.

8. Print out of technical summary report.

9. Technician's notes.

10. Miscellaneous information.

### *2.5 Existing sleep database repositories*

Few sleep data repositories seem to exist. We describe below all of those that we could find. Our data repository is by far the largest among them.

Several of the databases listed below are part of PhysioBank, an archive of physiologic signals developed by PhysioNet for use by the biomedical research community. PhysioNet is formed by a group of computer scientists, physicists, mathematicians, biomedical researchers, clinicians, and educators at MIT, the Beth Israel Deaconess Medical Center/Harvard Medical School, Boston University, and McGill University.

**a. The Sleep-EDF Database - Sleep Recordings and Hypnograms in European Data Format (EDF)** [17]**.**

This is a publicly available sleep data in the form of EDF files. There are a total of 8 sleep studies recorded from the years 1989 and 1994, consisting of 7 signals and 5 signals respectively. The sleep observations were conducted on Caucasian males and females (21 - 35 years old) without any medications. The signals in the studies are sampled at frequencies 100 Hz and 1 Hz.

The technical names of the signals that appear in the studies are EEG Fpz-Cz, EEG Pz-Oz, EOG horizontal, Resp oro-nasal, EMG submental, Temp body, Event marker, EEG Fpz-Cz, EEG Pz-Oz, EOG horizontal, EMG submental.

**b. The Sleep Heart Health Study Polysomnography Database** [18]

This database consists of studies conducted on subjects 40 years or older to study the relationship of sleep disordered breathing and cardiovascular disease, with no history of treatment of sleep apnea, no tracheotomy, and no current home oxygen therapy. The studies are in EDF format along with PhysioBank annotation files. From the sample EDF study posted online, there exist 11 signals recorded during the study. The number of data records could not be known by inspecting the header content (see EDF Header) of this EDF file (`0000.edf`), as this values was "-1". The signals are sampled at the rates 250 Hz, 125 Hz, 50 Hz, 10 Hz, and 1 Hz.

At present, we could find only one such study posted online at the web resource [18], and it is said that upto 1000 PSG studies will be posted in the future.

The technical names of the signals that appear in the study are SaO2, PR, EEG (sec), ECG, EMG, EOG(L), EOG(R), EEG, AIRFLOW, THOR RES, ABDO RES.

**c. St. Vincent's University Hospital / University College Dublin Sleep Apnea Database** [20]**.**

This repository consists of 25 full night PSG studies conducted on subjects who were 18 years or older, with probable sleep-disordered breathing. From one of the sample EDF study posted online (`ucddb003.rec`), there are 14 signals recorded having sampling rates 256 Hz, 128 Hz, 64 Hz, and 8 Hz.

The technical names of the signals that appear in the study are EEG C4-A1 [C4A1], EEG C3-A2 [C3A2], EOG E1-A1 [EOGL], EOG E2-A1 [EOGR], EMG-Chin [EMYG], ECG [ECG1], SaO2 [OSAT], Sound, Flow [AFLO], Sum, Thorax [CHMV], Abdomen [ABMV], Position, Pulse.

**d. MIT-BIH Polysomnographic Database** [5]

From [5], this is a collection of recordings of multiple physiologic signals during sleep. 18 subjects were monitored in Boston's Beth Israel Hospital Sleep Laboratory for evaluation of chronic obstructive sleep apnea syndrome, and to test the effects of constant positive airway pressure (CPAP), a standard therapeutic intervention that usually prevents or substantially reduces airway obstruction in these subjects. The database contains over 80 hours' worth of four-, six-, and seven-channel PSG recordings, each with an ECG signal annotated beat-by-beat, and EEG and respiration signals annotated with respect to sleep stages and apnea.

**e. The Sleep Heart Health Study** [21]

There were 71 PSG studies to evaluate in this study. The patients underwent overnight EEG-based polysomnography at home. Data was collected for 12 signals (oximetry, heart rate, chest wall and abdomen movement, nasal/oral airflow, body position, EEG, EOG, EMG, and ECG). Along with signal data, the patients were also asked to fill out a sleep

habits questionnaire. The respiratory abnormalities like apneas and hypopneas were the main focus of analysis in this study.

Compared to all the above existing studies that we could find, our database is the largest and most comprehensive, as it has 1319 full night PSG, CPAP and SPLIT type studies of 1046 patients amounting to 7914 hours or 329.75 days (considering that patient has an average of 6 hours of sleep per night) of sleep study data, with each study comprising of 50 to 55 different signals of varying sampling rates. Moreover, we collect data on patient demographics, daytime sleepiness and depression survey, medical history, smoke, drink and exercise habits (total of 70 attributes) and more than 350 attributes in the technical sleep summary generated at the end of every sleep study.

### *2.6 Prior work at WPI*

In previous work at WPI [7][7] and [8], analysis was performed on subjective (surveys) and objective (clinical observations) sleep data to mine statistically significant rules, by implementing a time window-based association rule mining technique within the WPI-WEKA system. This work made an organization of sleep data into macro (see Chapter 3 Macro dataset, also called subjective data) and micro (see Chapter 5 Micro Dataset, also called objective data) categories with consultation from the medical domain expert, Dr. Moonis at U. Mass Medical School.

For the macro dataset, there were 242 patient records and 63 attributes. These attributes consisted of patient demographics, epworth, depression, habitual attributes as well as technical sleep summary information like time spent in various sleep stages, sleep efficiency, arousal index, periodic leg movement syndrome, etc. The micro data consisted of 6 signals (Heart rate, Epochs, Sleep stages, Oxygen potential in the blood, CPAP and BiPAP pressure, and body position) recorded for a night long study for 120 patients.

Three types of analyses, macro level, micro level and mixed (macro mixed with micro) level were performed on the datasets. As an example, an interesting result from the macro level analysis was the discovery of a statistically significant association between the body mass index and snoring of the patient. The patients in overweight and obese categories exhibited moderate to heavy snoring. The micro level analysis employed a window-based rule mining technique to discover statistically significant associations between events of interest. As an example, one of the results stated with high confidence that the heart rate remains normal when the patient is in wake stage or in stage 1 or in stage 2. The dataset for mixed level analysis had 81 records and consisted of numeric, nominal, set and sequence type data. As an example result from this type of analysis, it was discovered that obese patients with moderate levels of depression frequently experienced sleep stage 2 and REM stage towards the middle and terminal stages of the sleep, while the patients with mild or insignificant depression experienced stage 2 in the first hour of sleep. Hence, stage 2 could be used as an identifier to segregate obese patients suffering from mild or moderate levels of depression (see p113 in [7]).

# 3 Macro dataset

## *3.1 Context*

*What is macro data?*

Before a night long sleep study is performed, the patients are asked to fill out survey questionnaires. These questionnaires carry demographic details like age, height, weight, body mass index, and collar size. They also have information on the patient's medical history, the medicines administered, sleep, exercise, smoke, caffeine and alcohol consumption habits. A survey like Epworth Sleepiness Scale (ESS) is a set of questions that determines level of daytime sleepiness in a patient. Beck Depression Inventory (BDI) gives insight into the existence and severity of symptoms of depression. Both, ESS and BDI are included in the survey questionnaire.

*When is the macro data collected by the sleep clinic?*

The patient is asked to fill out the questionnaires either the day of the test, or sometime before when the study has been scheduled. Even though there may be more than one type of sleep study that the patient has undertaken over time, there is only one survey in every patient file. Exceptions may arise, for example, when two studies are separated by long periods of time (2 to 3 years), they can have 2 different surveys. In this case, we only track the survey that was conducted the latest for which we have the complete micro data. Also, there is a high probability that the old survey was conducted with different (old) structure of micro data, which we are not dealing with in this thesis.

*How is the macro data collected for this research?*

The macro data is present in paper format. We analyze every section of the survey, eliciting and naming the attributes. The naming of an attribute is done in such a way that one can closely identify it with the sentential question in surveys. Depending on the

response to the question, the data type of the attribute is chosen. As part of good database design, we choose the narrowest column that can hold the values of the attribute. To keep the data as complete as possible, no discretization of the attributes is done at this stage.A list of medical disorders is created by reading the medical history of a patient.

For some patient cases, the micro data is incomplete due to missing CDs, CDs that could not be read, corrupt data or CDs having some other patient's data. These cases are eliminated for completeness and correctness of data collection.

The rest of the chapter deals in detail with the description and collection of macro data, how the structure of the data was mapped to the database design, and how we build the macro database.

## *3.2 Data Description*

The macro data consists of following categories:

3.2.1 Patient demographics
3.2.2 Epworth questionnaire
3.2.3 Beck Depression Inventory
3.2.4 Drink, Exercise, and Smoke habits

### *3.2.1 Patient demographics*

This section contains information about physical characteristics of the patient (imperial units of measure), sleep/wake up times, residence location, the medical and family history, and the date of survey.

The following table lists the attributes and their corresponding names in the database table - `macro.demographics`.

**Table 3.1 Demographics**

| No. | Attribute Name | Description | Attribute name (in database) | PostgreSQL data type |
|---|---|---|---|---|
| 1 | Age | Calculated from the data of birth of the patient | age | smallint |
| 2 | Gender | Gender of the patient, male or female | gender | "char" |
| 3 | Height | X feet, Y inches | height | real |
| 4 | Weight | X lbs | weight | real |
| 5 | BMI | Body mass index (BMI) [1] | bmi | real |
| 6 | Collar size | X inches | collar_size | real |
| 7 | Bed Time | Time when patient sleeps (24 hour scale)[2] | bedtime | time |
| 8 | Wake up time | Time when patient wakes up (24 hour scale)[2] | wakingtime | time |
| 9 | City | The city in which patient resides | city | text |
| 10 | State | The state in which the patient resides | state | text |
| 11 | Indication of Study | The indication for undertaking of the study | indication | text[] |
| 12 | Medical history[3] | Handwritten/printed past pathological and surgical reports of the patient | medical_hx | text[] |
| 13 | Family history[3] | Handwritten/printed past pathological and surgical reports of the patient's family members | family_hx | text[] |
| 14 | Medications[4] | List of current medications administered | medications | text[] |

| | | | | |
|---|---|---|---|---|
| | | to the patient | | |
| 15 | Survey date[5] | Date when the survey data was obtained | survey_date | date |
| 16 | Need for sleep study (by patient) [6] | Patient's reason on why he/she is undertaking a sleep study | need_study_by_patient | text[] |

**Note:**

[1]Body mass index is a measure of body fat in adult men and women.

- This attribute is calculated from the height and weight of the patient.
- Units: $kg/m^2$
- BMI formula:

$$BMI\ (kg/m^2) = (weight\ in\ pounds * 703) / (height\ in\ inches)^2$$

- BMI Classification [22]

```
Underweight = <18.5
Normal weight = 18.5 to 24.9
Overweight = 25 to 29.9
Obesity (Class 1) = 30 to 34.99
Obesity (Class 2) = 35 to 39.99
Morbid obesity = 40 or greater
```

- BMI calculations, if required, were done using the utility at [23].

[2]If Bedtime and Wake up time are both unknown, then they are both entered as: 0:00. If only bed time is known, then the waking time is entered as 00:01 which is our way to signify that the latter is not known, and vice-versa.

[3]Appendix A contains abbreviations related to this attribute.

[4]On paper, this attribute's response is present in the drink, smoke, and exercise habits questionnaire, however, it is grouped along with the attributes of the demographics table schema in the database.

[5]This date is near before/on the date of micro study, which, in majority of cases, is a regular PSG study. If this date is missing, then we assume that the survey was conducted on the day of sleep study.

[6]Examples: Wake up frequently at night, snoring problem, physician's recommendation, etc.

### *3.2.2 Epworth questionnaire*

The Epworth sleepiness scale (ESS) is a measure of general level of daytime sleepiness in a patient [9]. The questionnaire follows the demographic section of the survey. The ESS is a collection of first 8 questions in this questionnaire. The remaining questions don't directly refer to the ESS, they were developed by Dr. Moonis and Dr. Baillargeon at the Day Kimball Hospital, CT [24]. All the questions having 4 choices pertaining to the frequency of occurrence of a particular sleep problem. The significant other (spouse, close friend, or a family member) of the patient is also asked to provide the answers to some of these questions to get an observed score. In this thesis, we will refer to the complete questionnaire as "Epworth questionnaire", while the subset of Epworth questionnaire that are the first 8 questions will be referred to as "ESS questionnaire".

**Questionnaire scale**:

0: never or very seldom
1: once or twice a month
2: one or two times a week
3: very often

The following table lists the attributes and their corresponding names in the database table - `macro.epworth`

**Table 3.2 Epworth**

| No | Attribute text | Attribute name (in database) [1] |
|---|---|---|
| 1. I fall asleep when I am | a. Sitting and Reading | fa_reading |
| | b. Watching TV | fa_watching_tv |
| | c. In a movie theater or meeting | fa_theater_meeting |
| | d. As a passenger in a car for an hour | fa_passenger |
| | e. Lying down to rest in the afternoon when circumstances permit | fa_pm_rest |
| | f. Sitting quietly after lunch without alcohol | fa_after_lunch |
| | g. Driving my car while stopped in traffic | fa_as_driver |
| | e. Sitting and talking to someone | fa_talking |
| 2 | a. Does anyone in your family have a sleep problem? | sleep_problem_family |
| | b. Are you tired during the day, even after a nights sleep? | tired |
| | c. Do you fall asleep easily during meetings or watching TV? | fa_meeting_watching_tv |
| | d. Do you fall asleep while talking with someone or performing routine tasks? | fa_routine_tasks |
| | e. Do you eat, work, or worry in bed? | eat_work_worry_bed |
| | f. Do you wake up with a morning headache? | wa_am_headache |
| 3 | a. Do you wake up gasping for breath? | wa_gasp |
| | b. Has anyone told you that you snore | snore_loud |

| | | |
|---|---|---|
| | loudly? | |
| | c. Has anyone told you that you seem to stop breathing while you sleep? | stop_breath |
| 4 | a. Do you get uncomfortable feeling in your legs that makes it difficult to fall asleep? | leg_uneasy |
| | b. Do you have uncomfortable "crawly" feelings in your legs that is relieved by walking? | leg_crawly |
| | c. Has anyone noticed your legs or arms twitching during the night? | limbs_twitch |
| 5 | a. Do you have episodes of muscular weakness or paralysis when laughing, angry, or in other emotional situations? | paralysis |
| | b. Do you "act out" your dreams? | act_out_dreams |
| | c. Have you experienced being unable to move when falling asleep or upon waking up? | unable_move |
| | d. Do you have realistic, vivid dreams and/or nightmares? | vivid_dreams_nightmares |
| 6 | a. Do you walk or talk in your sleep? | walk_talk_sleep |
| | b. Do you wake up confused? | wa_confused |
| | c. Do you grind your teeth while sleeping? | grind_teeth |
| | d. Do you remember your dream upon awakening? | recall_dream |
| 7 [2] | Total Epworth score | epworth_score |
| 8 [3] | Epworth Sleepiness Scale (ESS) - 1 | ess1 |
| 9 [4] | Epworth Sleepiness Scale (ESS) - 2 | ess2 |
| 10[5] | Remarks | remarks |

**Note:**

[1]All the attributes, except for `remarks` (text) are of type - `smallint` in Posrgresql.

[2]Calculated sum of patient response to questions in 1 - 6.

[3]ESS - 1 is the total of patient response to questions in 1.

The following are the ranges for the ESS

0-8: mild

9-16: moderate

17-24: severe

[4]ESS - 2 is the total of patient's significant other/close relative's observed response to questions in 1.

[5]Remarks indicate different Epworth survey format, pediatric cases and missing surveys.

**Note A:** The missing values are filled as -1. Sometimes, the patient response to a question is "Yes". In this case, the scale value is assumed to be 3. Similarly, for "No" as an answer, the scale value of 0 is assumed. See Note A after depression table.

**Note:** There are 3 patient cases with old format of Epworth survey (2000, 2001). The old format contains a subset of questions similar to the new format, but with response only as Y (Yes) and N (No). We have used Y as 3, N as 0.

**Note B:** Sometimes the patient may circle more than one option in a question. In that case, the option indicating higher degree of occurrence is chosen.

**Note C:** In the above table,

wa = wake up; fa = fall asleep. The complete sorted list of acronyms is in Appendix A

### *3.2.3 Beck Depression Inventory*

This questionnaire is a set of 21 questions pertaining to the depression characteristics in a patient [1]. Each question is scaled from 0 - 3 depending upon the severity of a particular depression feature. At the end, a cumulative score is calculated, which is an indicator of level of depression in a patient:

0-9 is minimal
10-16 is mild
17-29 is moderate
30-63 is severe

The Beck Depression Index was added by Dr. Moonis at the Day Kimball Hospital a few years ago. [24]

The following table lists the attributes and their corresponding names in the database table - `macro.depression`.

**Table 3.3 Depression**

| No | Attribute text | Attribute name (in database) [1] |
|---|---|---|
| 1 | 0-I don't feel sad<br>1-I feel sad<br>2-I feel sad all the time<br>3-I am unbearably sad | feel_sad |
| 2 | 1-I'm not discouraged<br>2-I feel discouraged about the future<br>3-I have nothing to look forward to<br>4-My future is hopeless | discouraged_future |
| 3 | 0-I do not feel like a failure<br>1-I feel like I failed more than an average person<br>2-I can see lot of failures in retrospect<br>3-I feel like a complete failure | feel_failure |

| | | |
|---|---|---|
| 4 | 0-I don’t get as much satisfaction as before<br>1-I don’t enjoy things as before<br>2-I don’t get real satisfaction out of anything<br>3-I’m dissatisfied and bored with everything | degree_satisfaction |
| 5 | 0-I don’t feed guilty<br>1-I feel guilty for a good part of time<br>2-I feel guilty for most of the time<br>3-I feel guilty all the time | feel_guilty |
| 6 | 0-I don’t feel like being punished<br>1-I feel that I may be punished<br>2-I expect to be punished<br>3-I feel that I’m being punished | feel_punished |
| 7 | 0-I don’t feel disappointed in myself<br>1-I am disappointed in myself<br>2-I am disgusted with myself<br>3-I hate myself | feel_disappointed |
| 8 | 0-I don’t feel I’m worse than anyone else<br>1-I am critical of myself for my weaknesses or mistakes<br>2-I blame myself all the time for my faults<br>3-I blame myself for everything bad that happens | feel_worse |
| 9 | 0-I don’t have any thoughts of killing myself<br>1-I have thoughts of killing myself, but would not carry them out<br>2-I would like to kill myself<br>3-I would kill myself if I had the chance | suicide |
| 10 | 0-I don’t cry anymore than usual<br>1-I cry more now than I used to<br>2-I cry all the time now<br>3-I used to be able to cry, but now I can't cry even though I want to | crying |
| 11 | 0-I’m no more irritated now than I ever am<br>1-I get annoyed or irritated more easily than I used to<br>2-I feel irritated all the time now<br>3-I don’t get irritated at all by the things that used to irritate me | irritable |
| 12 | 0-I have not lost interest in people<br>1-I’m less interested in other people than I used to be | interest_people |

| | | |
|---|---|---|
| | 2-I've lost most of my interest in other people | |
| | 3-I've lost all of my interest in other people | |
| 13 | 0-I make decisions about as well as I ever could | make_decisions |
| | 1-I put off making decisions more than I used to | |
| | 2-I have greater difficulty in making decisions than before | |
| | 3-I can't make decisions at all anymore | |
| 14 | 0-I don't feel I look any worse than I used to | appearance |
| | 1-I'm worried that I am looking old or unattractive | |
| | 2-I feel that there are permanent changes in my appearance that make me look unattractive | |
| | 3-I believe that I look ugly | |
| 15 | 0-I can work about as well as before | work_capacity |
| | 1-It takes an extra effort to get started at doing something | |
| | 2-I have to push myself very hard to do anything | |
| | 3-I can't do any work at all | |
| 16 | 0-I can sleep as well as usual | sleep_habit |
| | 1-I don't sleep as well as I used to | |
| | 2-I wake up 1-2 hours earlier than usual and find it hard to get back to sleep | |
| | 3-I wake up several hours earlier than I used to and cannot get back to sleep | |
| 17 | 0-I don't get more tired than usual | get_tired |
| | 1-I get tired more easily than I used to | |
| | 2-I get tired from doing almost anything | |
| | 3-I'm too tired to do anything | |
| 18 | 0-My appetite is no worse than usual | appetite |
| | 1-My appetite is not as good as it used to be | |
| | 2-My appetite is much worse now | |
| | 3-I have no appetite at all now | |
| 19 | 0-I haven't lost much weight, if any, lately | lost_weight |
| | 1-I have lost more than 5 pounds | |
| | 2-I have lost more than 10 pounds | |
| | 3-I've lost more than 15 pounds | |
| 20 | 0-I'm more worried about my health than usual | worry_health |
| | 1-I'm worried about physical problems such as aches | |

| | | |
|---|---|---|
| | and pains; or upset stomach; or constipation<br>2-I'm very worried about physical problems and its hard to think of much else<br>3-I'm so worried about physical problems that I cannot think about anything else | |
| 21 | 0-I've not noticed any recent change in my interest in sex<br>1-I'm less interested in sex than I used to be<br>2-I'm much less interested in sex now<br>3-I've lost interest in sex completely | interest_sex |
| 22 | Beck Depression Inventory, total score | bdi |
| 23[2] | I'm purposely trying to lose weight by eating less | if_eat_less |
| 24[3] | Remarks | remarks |

**Note:**

[1]All the attributes, except attribute 23, ANSD REMARKS are of type - `smallint` in PGSQL.

[2]Attribute - 23 is of type "char".

- y: yes
- n: no
- u: unknown/missing

**Note A:** The missing values are filled as '-1'. Sometimes, the patient response to a question is "Yes". In this case, the scale value of 3 is assumed. Similarly, for "No" as an answer, the scale value of 0 is assumed. We made these assumptions after checking the total depression score for patients with such response. The total score on paper was calculated by a clinic technician, and the Yes responses were interpreted as 3 and the No as 0.

**Note B:** Sometimes the patient may circle more than one option in a question. In that case, the option with higher degree of severity is chosen.

**Note C:** For some pediatric cases, the BDI attributes are not applicable. Such records have been highlighted as "pediatric" in the *remarks* column, which will be removed after pre-processing the dataset.

[3]Remarks indicate different depression survey format, pediatric cases and missing surveys.

### *3.2.4 Drink, Exercise, and Smoke habits*

The following table lists the attributes and their corresponding names in the database table - `macro.drink_exercise_smoke`

**Table 3.4 Drink, Smoke, Exercise habits**

| No. | Attribute Description | Attribute Name | Data Type |
|---|---|---|---|
| 1. Why do you feel you have a sleep problem? Check all that apply. | __ stress at work<br>__ financial problems<br>__ poor sleep habits<br>__ poor eating habits<br>__ relationship with spouse<br>__relationship with children<br>__sexual<br>__other social<br>__other | reason_sleep_problem | text[] |
| 2. Do you exercise regularly? | If yes, then how often? | exercise_rate<br>exercise_remarks | real<br>text |
| 3. Do you drink caffeinated beverages | For example, coffee, tea, cola.<br>If so, how many per day ? | caffeinated_day<br>caffeinated_remarks | real<br>text |
| 4. Do you drink | If so, what do you drink on an | glasses_wine | real |

| | | | |
|---|---|---|---|
| alcoholic beverages? (scale: per week) | average day? I usually drink - - glasses of wine, sherry - bottles of beer - shots of hard liquor | bottles_beer | real |
| | | shots_hard_liquor | real |
| | | alcohol_remarks | text |
| 5. Do you now or ever smoked? | If yes, how many packs per day? | smoke | real |
| | | quit_years | real |
| | | smoke_remarks | text |
| 6. Heart disease[1] | Does the patient suffers or has suffered from heart disease? | heart_disease | text |

[1]This attribute is present in the demographics section of the Technical Summary report (see 4 Technical Summary Report), but is grouped with the attributes in this section.

**Note A:** The missing values are filled as -1.

**Note B:** There are many cases with ambiguous responses to attributes 2, 3, 4, 5. The *remarks* column has the values of these attributes exactly like how were written by the patient. The actual values that we want to interpret from these responses was decided in the pre-processing stage and is documented below.

**Some examples of vague responses:**

1. exercise - "yes", "occasionally on treadmill, but easily tired"; "smoked socially during middle age"
2. smoke - "yes", "cigar + pipe"
3. alcohol - 4 glasses - wine + beer / week; smoke - 1-2 ppd, quit for 18 yrs

**Note C:** Values of 'heart_disease': Yes, No, N/A, MurMur, Mitral Valve Prolapse, A Fib (Atrial fibrillation), HTN, triple bypass

**Modeling Smoke response -**

In order to model the smoking habits response of patients, we created two attributes corresponding to the "Do you now or ever smoked?" field -

1. smoke (quantity of smoking, in packs-per-day(ppd))
2. quit_yrs (number of years quit smoking)

If quit_yrs = 0, we only look at 'smoke' column (and the quantity becomes effective for current time). Otherwise, the patient has a history of quit smoking at some point (and the quantity corresponds to the smoke habits when the patient quit smoking). If a past year is a response to when the patient quit smoking, then the number of years quit smoking are calculated with reference to the year the survey was conducted.

An additional attribute called "smoke_remarks" is also present that has the actual response as written by the patient.

The following table gives interpretation of special case responses to the smoking habits of patients:

| **smoke** | **quit_yrs** | **Interpretation** |
|---|---|---|
| -1 | 0 | Field is left blank, we don't know if the person used to smoke in the past, we don't know if he/she smokes today. |
| -1 | -1 | If a patient quit smoking, but we don't know when and how much he/she smoked. |
| 0 | -1 | Patient used to smoke very rarely in the past. |
| 0 | 0 | If a patient has never smoked at all. |

| | | |
|---|---|---|
| -1 | 3 | If a patient quit smoking 3 years back, but we don't know how much he/she used to smoke. |
| 3 | -1 | If a patient quit smoking unknown number of years back, and smoked 3 ppd. |
| 1 | 0 | If a patient smokes 1 ppd in the present. |
| -2 | 0 | If response = "yes" |

**Table 3.5 Modeling smoke response**

-2 is used to indicate where the response is "yes", that is, we do not know how much the patient smokes. Unknown information is indicated by -1.

One pack of cigarettes is made of 20 cigarettes. If the patient smokes rarely, then the number of ppd is assumed to be 0. For patients who responded with smoking cigars or pipes, we do not know how to quantify their habit in terms of ppd. Hence, such responses are filled with smoke = -2, quit_yrs = 0.

Sometimes, the comments in remarks column can be appended with some calculation information done by us on the patient's response. This is for reference only, in order to understand how the numeric attributes got their values.

Format:
remarks - column
<remarks>; <our evaluations> (if any)

Example: Patients who smoke 2 packs / week => 40 cigarettes / week => ~6 cig / day => 0.3 ppd

**Modeling Exercise and Caffeine response -**

Many responses to caffeine and exercise habits are answered by a "yes". Such responses are filled with "-2".

For caffeine related conversions, 1 US gallon = 16 US cups

**Modeling Alcohol response**

Patient ticked on wine

| glasses_wine | bottles_beer | shots_hard_liquor |
|---|---|---|
| -1 | 0 | 0 |

Patient responded with "yes" on beer

| glasses_wine | bottles_beer | shots_hard_liquor |
|---|---|---|
| 0 | -2 | 0 |

For responses like "2 per week", where the quantity is specified, but not the type of drink, we assume that the patient drinks 2 bottles of beer in a week.

### *3.3 Macro Database*

Each questionnaire of the survey maps to a database table, and its attributes to the fields of the table. We first modeled the tables and its attributes in Microsoft Excel sheets for data entry. This made it convenient for us to collect, view, and edit the data. Later, the content of each table in Excel is saved in comma separated value format (CSV) that was exported to the PGSQL database.

A schema called `macro` is created within the database. All the macro data related tables are created in this schema:

```
macro.demographics
macro.epworth
macro.depression
macro.drink_smoke_ex
```

By using the PostgreSQL `COPY` operator, each CSV file can be exported to the PGSQL database. Every record of every macro dataset table is identified by a unique key that is the identifier of the patient.

There are 1046 patients for whom we have complete survey data.

## *3.4 Statistics on Macro data*

### *3.4.1 Number of patients, average and standard deviations of Epworth and Depression scores categorized by patients' gender and age group*

| Age groups | Number of patients | | Epworth score Average/Std. dev | | Depression score Average/Std. dev | |
|---|---|---|---|---|---|---|
| | Male | Female | Male | Female | Male | Female |
| 20 and below | 42 | 25 | 21.28/12.79 | 23.64/9.822 | 7.02/9.11 | 8.00/8.53 |
| (20,30] | 50 | 24 | 29.68/14.14 | 29.67/9.62 | 13.74/11.74 | 15.88/9.60 |
| (30,40] | 100 | 96 | 29.59/12.04 | 30.74/11.57 | 12.43/10.55 | 15.02/9.32 |
| (40,50] | 142 | 127 | 28.45/12.78 | 30.95/12.26 | 10.99/8.09 | 15.31/9.48 |
| (50,60] | 111 | 112 | 26.96/11.12 | 27.36/11.18 | 11.63/9.02 | 13.44/8.83 |
| (60,70] | 71 | 63 | 23.34/10.20 | 24.33/11.65 | 7.92/5.59 | 10.17/6.60 |
| (70,80] | 38 | 25 | 20.02/8.47 | 24.52/10.09 | 9.53/7.43 | 12.84/10.00 |
| above 80 | 12 | 8 | 27.41/10.46 | 19.63/8.57 | 11.58/9.67 | 12.25/5.23 |

**Table 3.5 Patients, Depression, Epworth by age**

Figure 3.1 below graphically depicts the distribution of the male and female populations for the various age groups.

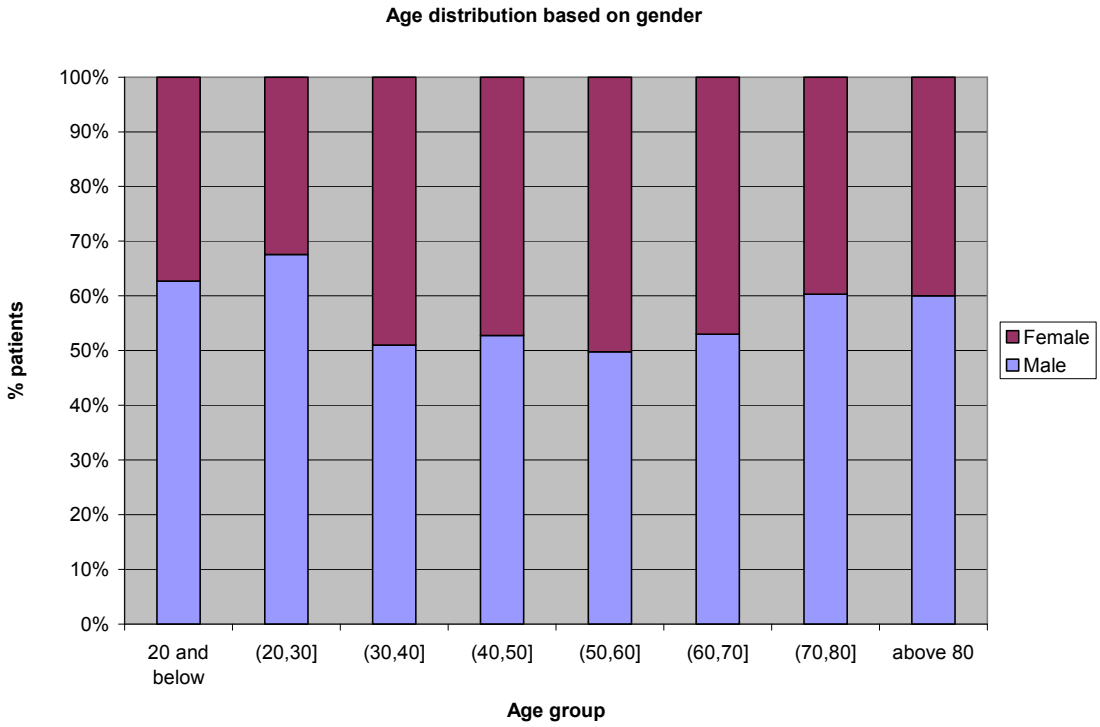



**Figure 3.1 Age distribution based on gender**

**Example queries:**

Query 1 - Number of males with age above 20 and age less than or equal to 30.

```
select count(*) from macro.demographics where age > 20 and age<=30 and
gender = 'm'
```

Query 2 - Average Epworth score for males with age more than 30 and less than or equal to 40.

```
select avg(epworth_score) from macro.epworth e, macro.demographics d
where age>30 and age <=40 and gender = 'm' and e.pid=d.pid
```

Query 3 - Standard deviation from the mean of Beck depression index for females with age above 80.

```
select stddev(bdi) from macro.depression dep, macro.demographics dev
where age>80 and gender = 'f' and dep.pid=dem.pid
```

### *3.4.2 Patient distribution with respect to BMI Classification*

BMI ranges, from Section 3.2.1 Patient demographics.

```
Underweight =<18.5
Normal weight = 18.5 to 24.9
Overweight = 25 to 29.9
Obesity (Class 1) = 30 to 34.99
Obesity (Class 2) = 35 to 39.99
Morbid obesity = 40 or greater
```

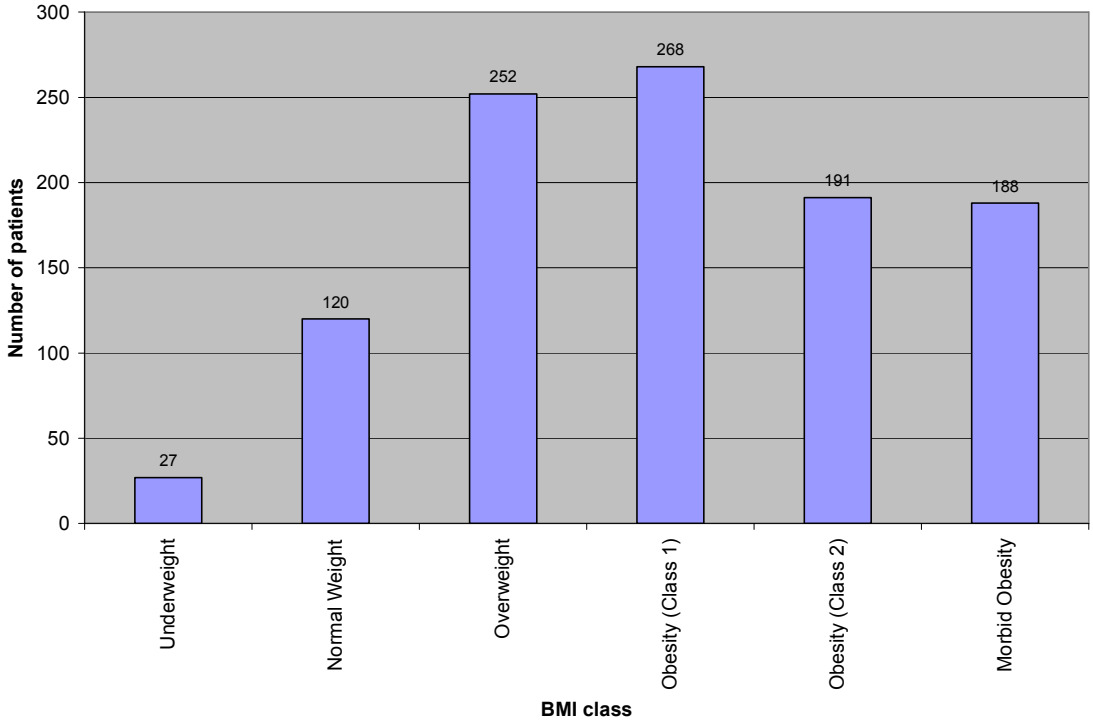



**Figure 3.2 Patient distribution by BMI class**

From the graph above, it is seen that a majority of patients with sleep disorders are overweight or obese.

**Example query:**

Query - Number of patients having BMI between 30 to 34.99

```
select count(*) from macro.demographics where bmi>=30 and bmi<=34.99
```

### *3.4.3 Epworth questionnaire score distribution*

| Epworth score sheet attributes | 0-% | 1-% | 2-% | 3-% | missing % |
|---|---|---|---|---|---|
| fa_reading | 43.69 | 15.87 | 15.20 | 23.80 | 1.43 |
| fa_watching_tv | 16.73 | 15.97 | 23.61 | 42.64 | 1.05 |
| fa_theater_meeting | 71.80 | 11.38 | 6.50 | 6.98 | 3.35 |
| fa_passenger | 52.68 | 18.16 | 8.89 | 18.36 | 1.91 |
| fa_pm_rest | 19.50 | 20.27 | 19.31 | 39.58 | 1.34 |
| fa_after_lunch | 48.37 | 17.21 | 14.63 | 17.69 | 2.10 |
| fa_driver | 83.27 | 6.31 | 3.15 | 2.87 | 4.40 |
| fa_talking | 83.94 | 7.74 | 3.92 | 2.87 | 1.53 |
| sleep_problem_family | 52.96 | 8.22 | 6.69 | 23.90 | 8.22 |
| tired | 5.26 | 10.23 | 19.98 | 63.58 | 0.96 |
| fa_meeting_watchingtv | 26.96 | 15.20 | 19.22 | 37.48 | 1.15 |
| fa_routine_tasks | 81.26 | 8.51 | 5.07 | 4.02 | 1.15 |
| eat_work_worry_bed | 41.68 | 19.31 | 16.73 | 20.94 | 1.34 |
| wa_am_headache | 40.34 | 25.14 | 17.69 | 15.58 | 1.24 |
| wa_gasp | 60.42 | 16.44 | 13.10 | 7.84 | 2.20 |
| snore_loud | 15.87 | 9.18 | 10.61 | 62.43 | 1.91 |
| stop_breath | 48.76 | 8.41 | 11.09 | 28.97 | 2.77 |
| leg_uneasy | 47.99 | 18.45 | 12.72 | 19.12 | 1.72 |
| leg_crawly | 63.29 | 12.62 | 10.80 | 11.28 | 2.01 |
| limbs_twitch | 54.97 | 11.28 | 14.15 | 16.54 | 3.06 |
| paralysis | 80.02 | 8.22 | 5.07 | 3.92 | 2.77 |
| act_out_dreams | 71.51 | 11.47 | 6.21 | 5.83 | 4.97 |
| unable_move | 78.20 | 11.85 | 4.59 | 3.54 | 1.82 |
| vivid_dreams_nightmares | 29.16 | 28.30 | 19.41 | 21.51 | 1.63 |
| walk_talk_sleep | 62.05 | 15.97 | 9.37 | 10.42 | 2.20 |
| wa_confused | 66.54 | 18.16 | 8.32 | 5.35 | 1.63 |
| grind_teeth | 58.70 | 12.05 | 9.08 | 14.72 | 5.45 |
| recall_dream | 23.71 | 34.80 | 20.84 | 19.02 | 1.63 |

**Table 3.6 Epworth questionnaire score distribution**

From the table above, we see many patients (more than 60%) that very often feel tired during the day. Similarly, the number patients who snore very often are is also very high.

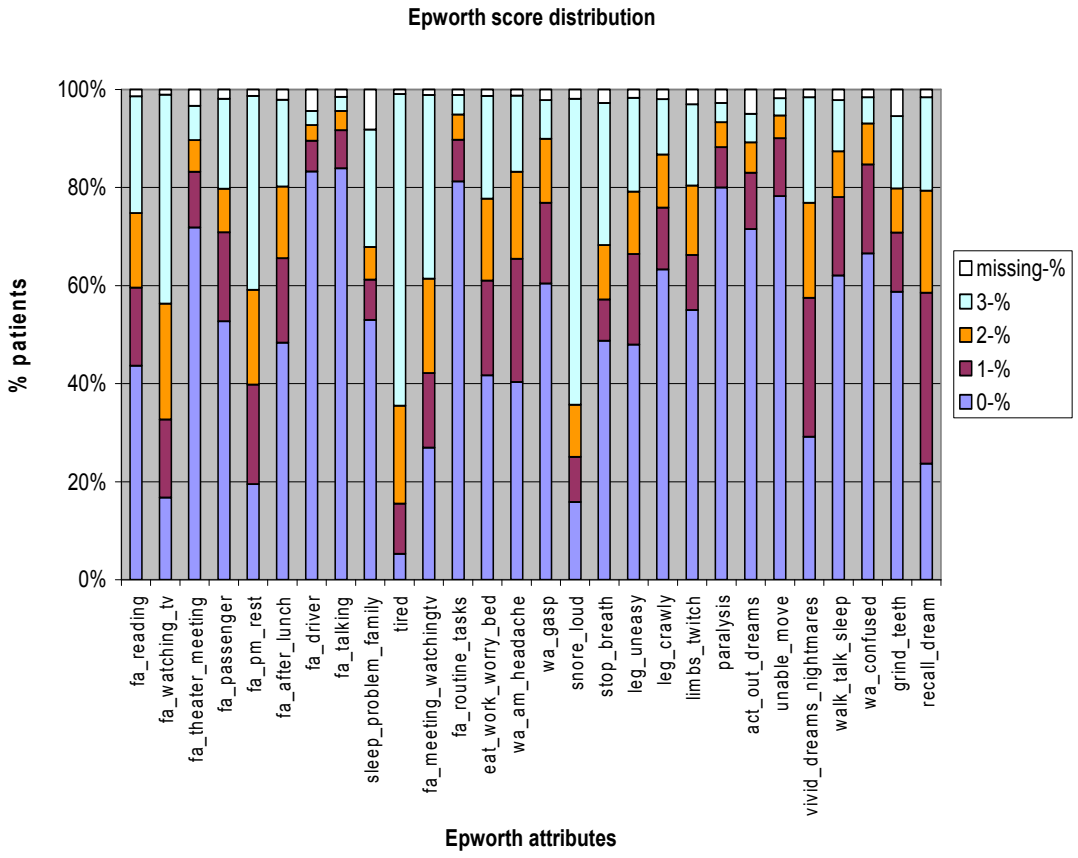



**Figure 3.3 Epworth score distribution**

**Example query:**

Query - Number of patients who fall asleep while reading very often.

```
select count(*) from macro.epworth where fa_reading=3
```

(See Section 3.2.2 Epworth questionnaire for table and attribute descriptions)

### *3.4.4 Depression score distribution*

| Depression | 0-% | 1-% | 2-% | 3-% | missing-% |
|---|---|---|---|---|---|
| feel_sad | 63.00 | 27.25 | 4.97 | 2.96 | 1.82 |
| discouraged_future | 67.69 | 22.94 | 4.97 | 2.58 | 1.82 |
| feel_failure | 71.13 | 17.21 | 7.55 | 1.72 | 2.39 |
| degree_satisfaction | 43.31 | 42.16 | 6.98 | 5.93 | 1.63 |
| feel_guilty | 75.14 | 17.40 | 3.44 | 1.91 | 2.10 |
| feel_punished | 82.31 | 7.93 | 1.15 | 5.93 | 2.68 |
| feel_disappointed | 57.93 | 32.22 | 6.02 | 1.91 | 1.91 |
| feel_worse | 53.44 | 33.94 | 7.27 | 3.35 | 2.01 |
| suicide | 86.81 | 10.42 | 0.67 | 0.00 | 2.10 |
| crying | 71.89 | 18.16 | 1.91 | 6.50 | 1.53 |
| irritable | 39.67 | 47.13 | 8.22 | 3.15 | 1.82 |
| interest_people | 62.14 | 27.63 | 6.98 | 1.24 | 2.01 |
| make_decisions | 60.80 | 22.28 | 14.44 | 0.96 | 1.53 |
| appearance | 56.50 | 22.18 | 13.67 | 5.64 | 2.01 |
| work_capacity | 33.65 | 41.59 | 19.02 | 3.92 | 1.82 |
| sleep_habit | 22.18 | 49.04 | 15.30 | 11.09 | 2.39 |
| get_tired | 13.38 | 52.77 | 25.62 | 6.50 | 1.72 |
| appetite | 74.19 | 15.77 | 6.12 | 1.72 | 2.20 |
| lost_weight | 73.90 | 9.75 | 4.97 | 5.74 | 5.64 |
| worry_health | 45.41 | 40.34 | 10.52 | 1.43 | 2.29 |
| interest_sex | 44.93 | 22.85 | 15.97 | 10.52 | 5.74 |

**Table 3.7 Depression score distribution**

(See Section 3.2.3 Beck Depression Inventory for table and attribute descriptions)

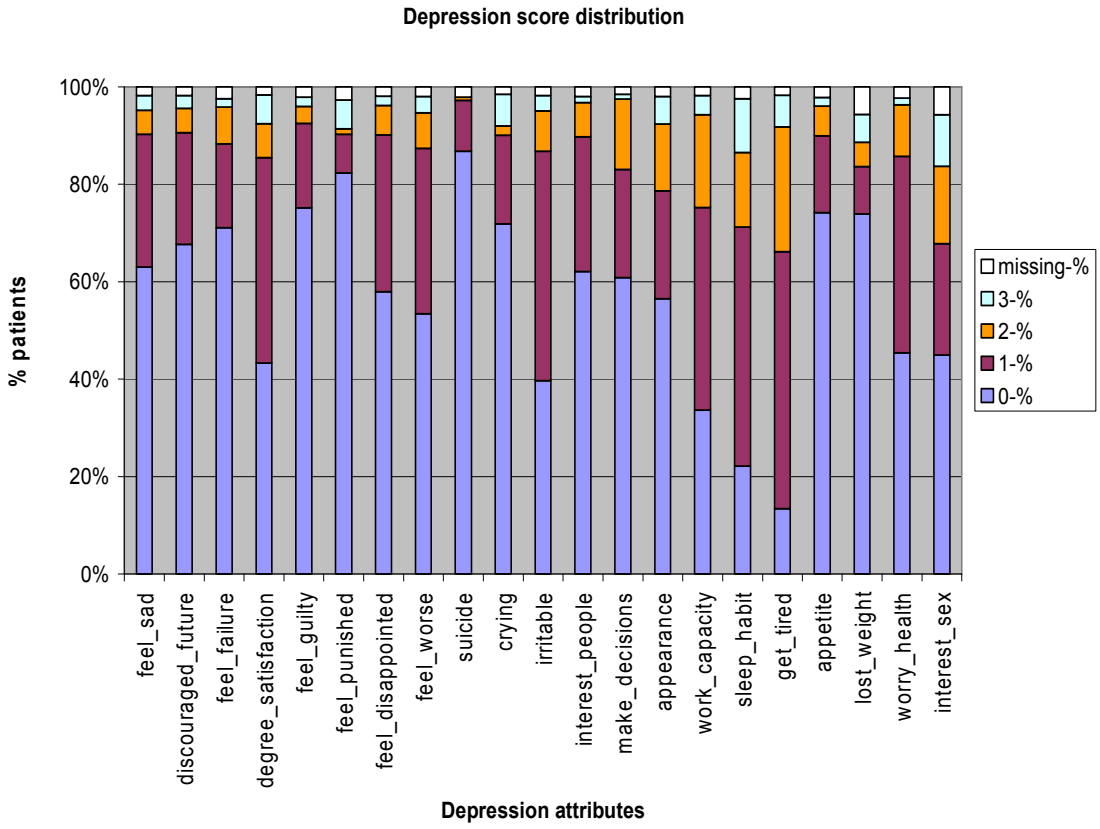



**Figure 3.4 Depression score distribution**

**Example query:**

Query - Number of patients who hate themselves

```
select count(*) from macro.epworth where fa_reading=3
```

### *3.4.5 Beck depression index with respect to patient's age*


**Figure 3.5 BDI with respect to patient's age**

From the above graph, patients with majority of severe (30 to 63) depression scores belong to 20 to 60 year age bracket. There are few instances of high depression seen in the senior group of patients of 75 to 90 years of age too. A majority of patients who are less than 20 years of age have minimal (0 to 9) to mild (10 to 16) rate of depression.

# 4 Technical Summary Report

The technical summary (of a PSG/CPAP/Split Night type study) is a summary of the recordings made during the overnight sleep study of the patient. This technical summary is stored by the REMBrandt system [13] as a Microsoft Excel file within the sleep study.

## *4.1 Summary Sections*[1,2]

**a. Patient's demographic information**

See Section 3.2.1 Patient demographics.

**b. Sleep and Sleep Stages**

This section contains information on total sleep time tracked, the efficiency of sleep, percent sleep period time spent in each of the sleep stages, number of awakenings during the sleep, etc.

**c. Respiratory Disturbance (RD)**

The number of obstructive, central and mixed - Apneas/Hypopneas, and UARS (Upper Airway Respiratory syndrome) counted during the patients sleep are entered into this section. Time and counts are tracked for REM (Rapid eye movement), Non - REM, Supine, Non - Supine and MT events. Some attributes that are aggregates of these counts (e.g., total number of Supine events) are also present.

**d. Limb Movement (LM)**

This table contains the counts and index of limb movements, periodic limb movements, and respiratory related limb movements for REM, NREM, Arousal, and Non - Arousal events.

**e. Arousals**

The counts and index of arousals during sleep events like - Apnea, Hypopnea, Snore, Limb Movements, etc. are present.

**Note:**

[1]All the attributes corresponding to each of the sections are listed later in this chapter. The complete forms of acronyms encountered in the summary reports can be read in Appendix B (a).

[2]SPLIT and CPAP type studies have another report called `SPLIT/CPAP/BILEVEL TABLE` along with the standard summary report. We are not storing this data in the database in this thesis. However, they are available for any future research.

**f. EKG**

This section has the mean heart rate of the patient during events - REM, NREM, Wake, and MT. An overall mean heart rate from the entire length of sleep is also present.

**g. Oxygen Saturation (OS)**

The oxygen desaturations during NREM, REM, Wake and MT events are tracked in this section. Also present is the information on mean oxygen saturation, mean wake oxygen saturation, lowest desaturation, and time in REM stage in which oxygen saturation was less than 90%.

**h. Body Position (BP)**

The counts of Apnea, Hypopnea, UARS events and RDI (Respiratory disturbance index) against the Back, Left, Right, Prone/Abdomen, Upright body positions appear in this section.

### *4.2 Summary Format Types*

## A

Completely old format that has PSG or CPAP studies only. There are a lot of missing attributes present in this type of format, when compared to the standard Type-C:

- No Total Sleep time (TST) for REM and NREM events in Sleep summary.
- No RD summary for Obstructive, Central and Mixed type Hypopnea events.
- No time attribute for REM, NREM, Wake and MT events in RD summary.
- No UARS for REM, NREM, Wake and MT events in RD summary.
- No attributes for LM, periodic LM (PLM), respiratory related LM (RRLM), and total LM (TLM) for REM, NREM, Arousal, and No arousal events in LM summary.
- No attributes for left and right body positions for LM, PLM and RRLM, during REM and NREM events in LM summary.
- No Mean wake Oxygen Saturation attribute in OS summary.
- No body position summary.

## B

Similar to format C, but with different cell indexing and some missing attributes:

- No time attribute for REM, NREM, Wake and MT events in RD summary.
- No UARS for Left, Right, Back, Abdomen and Up body positions in Body position summary.
- No Body position/Time/Stage summary, hence no time attributes for Abdomen, Back, Left, Right, Up body positions for REM, NREM, Wake and MT events.
- No Advanced Signal Time-table in Split type-B studies.

C

The most current and default format. The majority of the studies (psg/cpap/split) are in this format.

**Table 4.1 Number of reports in each format type**

| Type | Format identifier[1] | Count | % |
|---|---|---|---|
| PSG A | psgA | 6 | 0.45 |
| CPAP A | cpapA | 6 | 0.45 |
| PSG B | psgB | 39 | 2.95 |
| CPAP B | cpapB | 12 | 0.91 |
| SPLIT B | splitB | 5 | 0.38 |
| PSG C | psg | 830 | 62.92 |
| CPAP C | cpap | 309 | 23.43 |
| SPLIT C | split | 112 | 8.49 |
| | **TOTAL** | **1319** | **100** |

**Note:** [1]The format identifier is a part of the technical summary filename, and has information needed for distinction between different formats of the reports. See Section 4.5.1 File naming scheme.

The two figures below give an abstract layout of the structure of Type - C PSG and SPLIT summary sheets. Note that CPAP summary has the same format as PSG.

For Split Night however, the structure has to incorporate baseline and treatment attributes (hence there are two tables for every section). The split studies also have two extra sections - Sleep Parameters and Advanced saturation summary. However, note that there are some attributes that have values for the entire length of the SPLIT study (for

example, Sleep Period Time (SPT), see B Technical summary report, part (b) for complete list). The figure below shows the layout of a Type-C split study. In the figure,

ALL means total of diagnostic and treatment values
NDX O means baseline/diagnostic/DX
RX means treatment

O is Obstructive, C is Central, M is Mixed type of apnea or hypopnea. The notation <ALL, NDX O> means the entire summary section is divided into two sub-sections, one corresponding to ALL and other to NDX O for a given event (like REM). Thus, given ALL and NDX O, we can calculate RX = ALL - NDX O.

**Figure 4.1 PSG/CPAP Type C Layout**

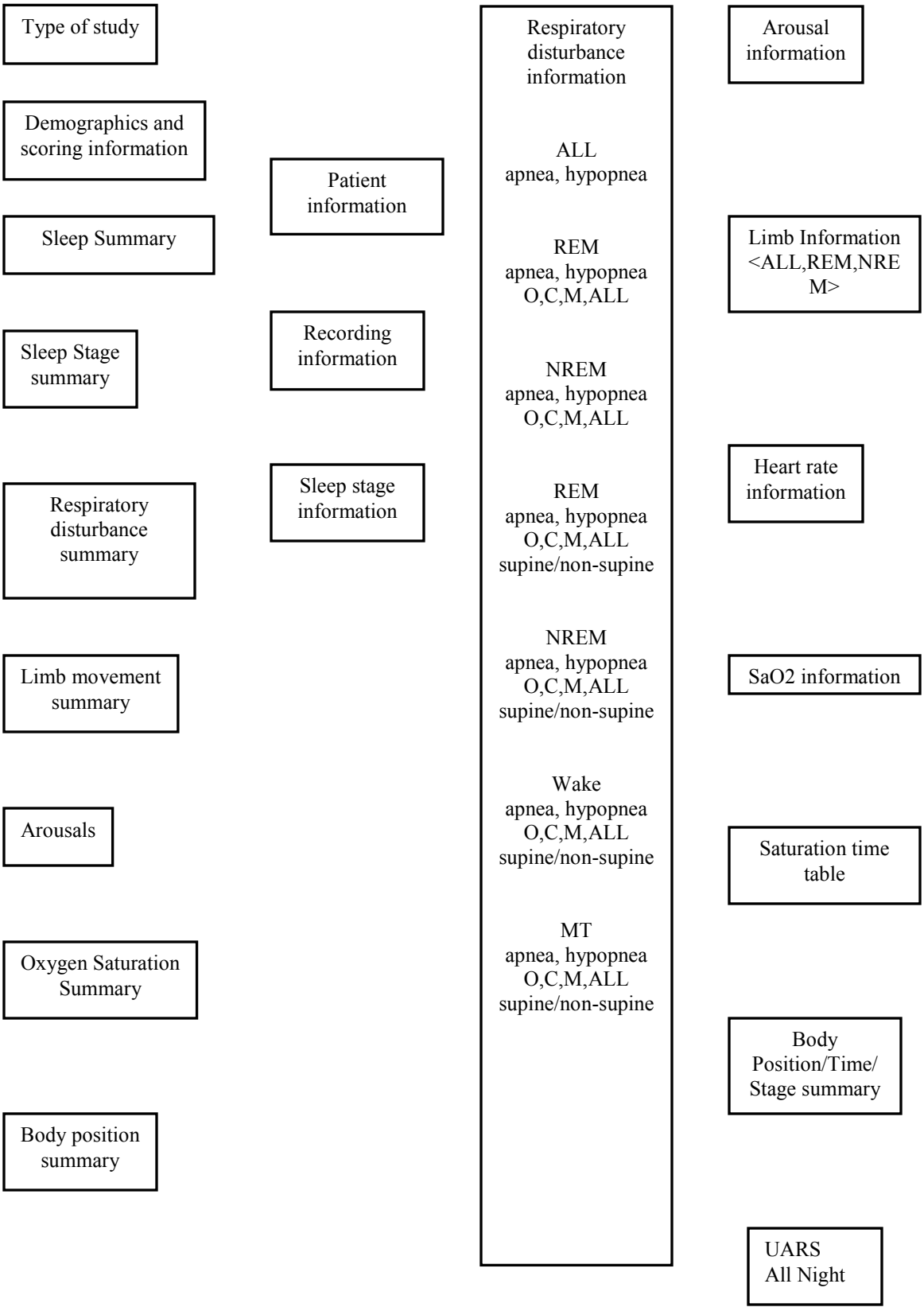
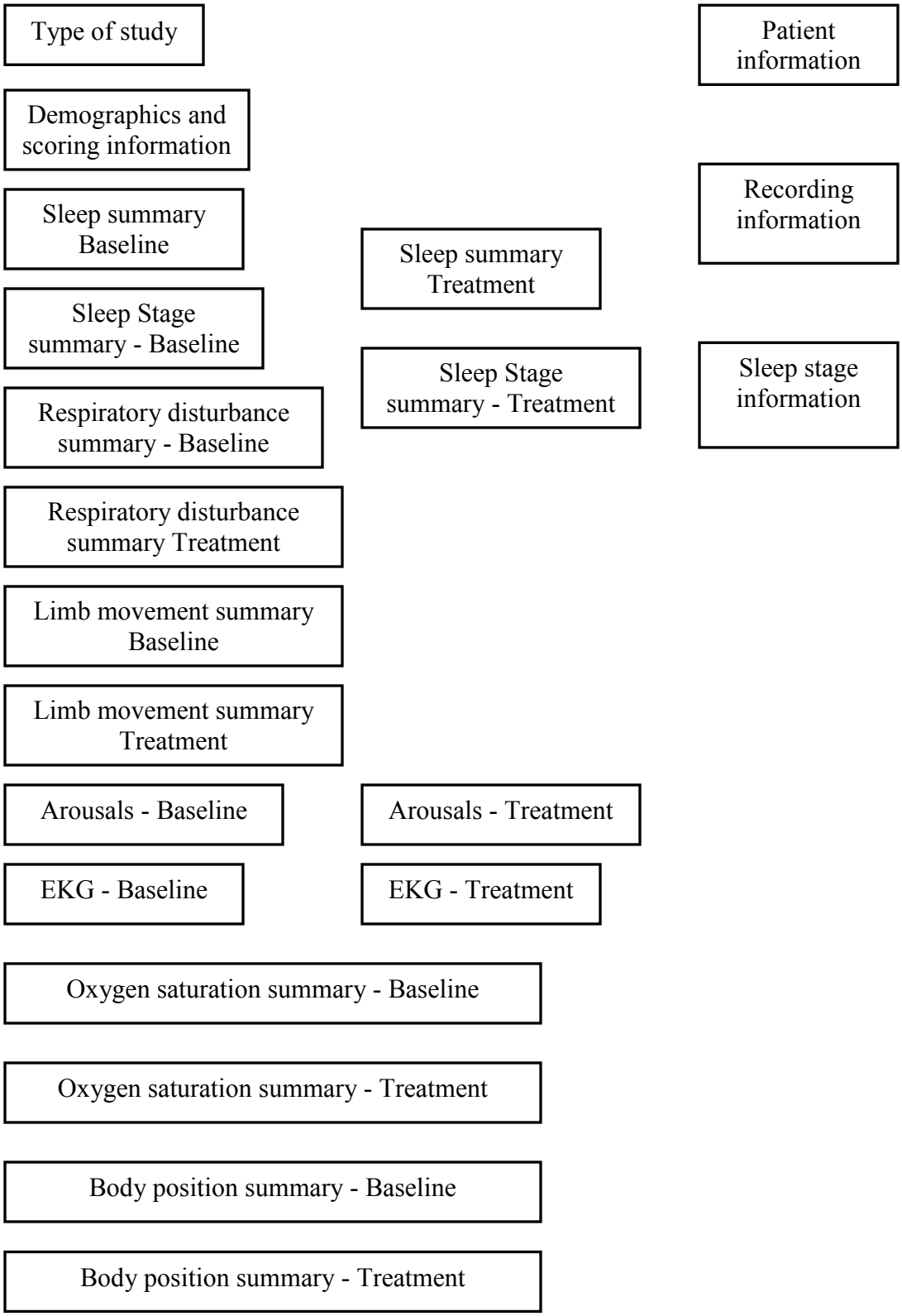
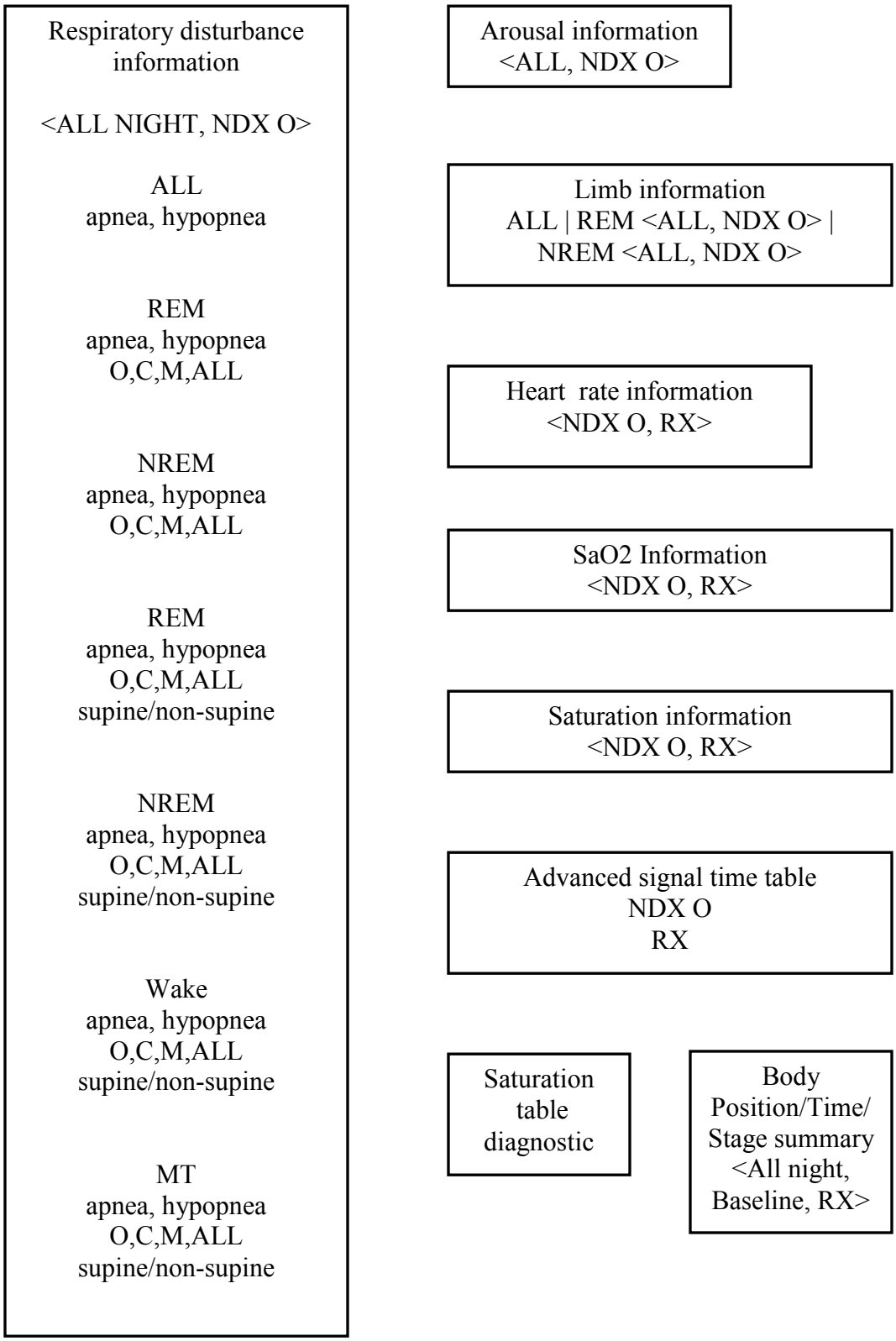


**Figure 4.2 Split Type C Layout (Left - half)**



**(Right - half)**


Respiratory disturbance information
<ALL NIGHT, NDX O>
ALL
apnea, hypopnea
REM
apnea, hypopnea
O,C,M,ALL
NREM
apnea, hypopnea
O,C,M,ALL
REM
apnea, hypopnea
O,C,M,ALL
supine/non-supine
NREM
apnea, hypopnea
O,C,M,ALL
supine/non-supine
Wake
apnea, hypopnea
O,C,M,ALL
supine/non-supine
MT
apnea, hypopnea
O,C,M,ALL
supine/non-supine
Arousal information
<ALL, NDX O>
Limb information
ALL | REM <ALL, NDX O> |
NREM <ALL, NDX O>
Heart rate information
<NDX O, RX>
SaO2 Information
<NDX O, RX>
Saturation information
<NDX O, RX>
Advanced signal time table
NDX O
RX
Saturation table diagnostic
Body Position/Time/ Stage summary <All night, Baseline, RX>

### *4.3 Attributes: PSG and CPAP studies*

The following tables show the attributes for PSG and CPAP type-C studies only. Since we deal with different formats and types of summary reports, we created a Microsoft Excel file that has the Summary sections, their attributes, cell Indexes and database attribute names for the all possible summary structures.

**Table 4.2 Sleep Stage Summary**

| Attribute Name | Cell - Index | Database attribute | Type |
|---|---|---|---|
| Total Recording Time - min | E24 | total_rec_time_min | real |
| Sleep Period Time - min | E25 | sleep_period_time_min | real |
| TST NREM - min | E26 | tst_nrem_min | real |
| TST REM - min | E27 | tst_rem_min | real |
| Sleep Latency - min | E28 | sleep_latency_min | real |
| Latency to REM (from Sleep onset) - min | E29 | latency_to_rem_min | real |
| Number of Awakenings(NW) | E30 | awakenings_count | smallint |
| Sleep Efficiency - % | E31 | sleep_efficiency_perc | real |
| | | | |
| Waso - min | C37 | waso_min | real |
| Stage 1 - min | C38 | stage1_min | real |
| Stage 2 - min | C39 | stage2_min | real |
| Stage 3 - min | C40 | stage3_min | real |
| Stage 4 - min | C41 | stage4_min | real |
| REM - min | C42 | rem_min | real |
| Total NREM - min | C43 | nrem_min | real |
| MT - min | C44 | mt_min | real |
| Supine and REM - min | C45 | supine_rem_min | real |
| | | | |
| SPT[1] | | | |
| Waso - % spt | D37 | waso_spt | real |
| Stage 1 - % spt | D38 | stage1_spt | real |
| Stage 2 - % spt | D39 | stage2_spt | real |
| Stage 3 - % spt | D40 | stage3_spt | real |
| Stage 4 - % spt | D41 | stage4_spt | real |
| REM - % spt | D42 | rem_spt | real |
| Total NREM -% spt | D43 | nrem_spt | real |

| | | | |
|---|---|---|---|
| MT - % spt | D44 | mt_spt | real |
| Supine and REM - % spt | D45 | supine_rem_spt | real |
| | | | |
| TIB[2] | | | |
| Waso - % tib | C37/AF41 * 100 | waso_tib | real |
| Stage 1 - % tib | C38/AF41 * 100 | stage1_ tib | real |
| Stage 2 - % tib | C39/AF41 * 100 | stage2_ tib | real |
| Stage 3 - % tib | C40/AF41 * 100 | stage3_ tib | real |
| Stage 4 - % tib | C41/AF41 * 100 | stage4_ tib | real |
| REM - % tib | C42/AF41 * 100 | rem_ tib | real |
| Total NREM -% tib | C43/AF41 * 100 | nrem_ tib | real |
| MT - % tib | C44/AF41 * 100 | mt_ tib | real |
| Supine and REM - % tib | C45/AF41 * 100 | supine_rem_ tib | real |
| | | | |
| Lights Off[3] | AE29 | lights_off_time | time |
| Lights On[3] | AE30 | lights_on_time | time |
| Wake[3] | AE38 | | |
| Time in Bed | AE41 | time_in_bed | real |
| Total Sleep Time[3] | AE42 | total_sleep_time | real |
| Wake /SPT[3] | AE53 | wake_spt | real |
| TST Supine[3] | AF55 | tst_supine | real |
| TST Non Supine[3] | AF56 | tst_non_supine | real |
| | | | |
| Referring Physician[3] | C21 | refer_phy | text |
| Interpreting Physician[3] | I21 | interpret_phy | text |

**Note:**

The **[1]%SPT** (Sleep Period Time) and **[2]%TIB** (Time in bed) calculated for sleep stages in PSG/CPAP studies are for the entire length of the study. However, for SPLIT type studies, the % TIB for sleep stages corresponds to baseline or treatment study duration, while the % SPT corresponds to the entire length of the study (because the SPLIT summary does not give the SPT values exclusively for baseline and treatment). For

example, `waso_spt` for a SPLIT studycorresponds to waso durtation for baseline on SPT.

[3] These attributes correspond to the entire length of the sleep study for SPLIT studies. See Appendix B (b) for a complete list of such attributes.

**Table 4.3 Respiratory Disturbance Summary**

| Attribute Name | Cell - Index | Database attribute | Type |
|---|---|---|---|
| Apneas Obstructive REM | D68 | apnea_obs_rem | integer |
| Apneas Obstructive REM Supine | D69 | apnea_obs_rem_supine | integer |
| Apneas Obstructive REM Non Supine | D70 | apnea_obs_rem_non_supine | integer |
| Apneas Obstructive NREM | D71 | apnea_obs_nrem | integer |
| Apneas Obstructive NREM Supine | D72 | apnea_obs_nrem_supine | integer |
| Apneas Obstructive NREM Non Supine | D73 | apnea_obs_nrem_non_supine | integer |
| Apneas Obstructive Wake | D74 | apnea_obs_wake | integer |
| Apneas Obstructive Wake Supine | D75 | apnea_obs_wake_supine | integer |
| Apneas Obstructive Wake Non Supine | D76 | apnea_obs_wake_non_supine | integer |
| Apneas Obstructive MT | D77 | apnea_obs_mt | integer |
| Apneas Obstructive MT Supine | D78 | apnea_obs_mt_supine | integer |
| Apneas Obstructive MT Non Supine | D79 | apnea_obs_mt_non_supine | integer |
| Apneas Central REM | E68 | apnea_central_rem | integer |
| Apneas Central REM Supine | E69 | apnea_central_rem_supine | integer |
| Apneas Central REM Non Supine | E70 | apnea_central_rem_non_supine | integer |
| Apneas Central NREM | E71 | apnea_central_nrem | integer |
| Apneas Central NREM Supine | E72 | apnea_central_nrem_su | integer |

| | | | |
|---|---|---|---|
| | | pine | |
| Apneas Central NREM Non Supine | E73 | apnea_central_nrem_non_supine | integer |
| Apneas Central Wake | E74 | apnea_central_wake | integer |
| Apneas Central Wake Supine | E75 | apnea_central_wake_supine | integer |
| Apneas Central Wake Non Supine | E76 | apnea_central_wake_non_supine | integer |
| Apneas Central MT | E77 | apnea_central_mt | integer |
| Apneas Central MT Supine | E78 | apnea_central_mt_supine | integer |
| Apneas Central MT Non Supine | E79 | apnea_central_mt_non_supine | integer |
| Apneas Mixed REM | F68 | apnea_mixed_rem | integer |
| Apneas Mixed REM Supine | F69 | apnea_mixed_rem_supine | integer |
| Apneas Mixed REM Non Supine | F70 | apnea_mixed_rem_non_supine | integer |
| Apneas Mixed NREM | F71 | apnea_mixed_nrem | integer |
| Apneas Mixed NREM Supine | F72 | apnea_mixed_nrem_supine | integer |
| Apneas Mixed NREM Non Supine | F73 | apnea_mixed_nrem_non_supine | integer |
| Apneas Mixed Wake | F74 | apnea_mixed_wake | integer |
| Apneas Mixed Wake Supine | F75 | apnea_mixed_wake_supine | integer |
| Apneas Mixed Wake Non Supine | F76 | apnea_mixed_wake_non_supine | integer |
| Apneas Mixed MT | F77 | apnea_mixed_mt | integer |
| Apneas Mixed MT Supine | F78 | apnea_mixed_mt_supine | integer |
| Apneas Mixed MT Non Supine | F79 | apnea_mixed_mt_non_supine | integer |
| Hypopneas Obstructive REM | G68 | hypopnea_obs_rem | integer |
| Hypopneas Obstructive REM Supine | G69 | hypopnea_obs_rem_supine | integer |
| Hypopneas Obstructive REM Non Supine | G70 | hypopnea_obs_rem_non_supine | integer |

| | | | |
|---|---|---|---|
| Hypopneas Obstructive NREM | G71 | hypopnea_obs_nrem | integer |
| Hypopneas Obstructive NREM Supine | G72 | hypopnea_obs_nrem_supine | integer |
| Hypopneas Obstructive NREM Non Supine | G73 | hypopnea_obs_nrem_non_supine | integer |
| Hypopneas Obstructive Wake | G74 | hypopnea_obs_wake | integer |
| Hypopneas Obstructive Wake Supine | G75 | hypopnea_obs_wake_supine | integer |
| Hypopneas Obstructive Wake Non Supine | G76 | hypopnea_obs_wake_non_supine | integer |
| Hypopneas Obstructive MT | G77 | hypopnea_obs_mt | integer |
| Hypopneas Obstructive MT Supine | G78 | hypopnea_obs_mt_supine | integer |
| Hypopneas Obstructive MT Non Supine | G79 | hypopnea_obs_mt_non_supine | integer |
| Hypopneas Central REM | H68 | hypopnea_central_rem | integer |
| Hypopneas Central REM Supine | H69 | hypopnea_central_rem_supine | integer |
| Hypopneas Central REM Non Supine | H70 | hypopnea_central_rem_non_supine | integer |
| Hypopneas Central NREM | H71 | hypopnea_central_nrem | integer |
| Hypopneas Central NREM Supine | H72 | hypopnea_central_nrem_supine | integer |
| Hypopneas Central NREM Non Supine | H73 | hypopnea_central_nrem_non_supine | integer |
| Hypopneas Central Wake | H74 | hypopnea_central_wake | integer |
| Hypopneas Central Wake Supine | H75 | hypopnea_central_wake_supine | integer |
| Hypopneas Central Wake Non Supine | H76 | hypopnea_central_wake_non_supine | integer |
| Hypopneas Central MT | H77 | hypopnea_central_mt | integer |
| Hypopneas Central MT Supine | H78 | hypopnea_central_mt_supine | integer |
| Hypopneas Central MT Non Supine | H79 | hypopnea_central_mt_non_supine | integer |
| Hypopneas Mixed REM | I68 | hypopnea_mixed_rem | integer |
| Hypopneas Mixed REM Supine | I69 | hypopnea_mixed_rem_supine | integer |

| | | | |
|---|---|---|---|
| Hypopneas Mixed REM Non Supine | I70 | hypopnea_mixed_rem_non_supine | integer |
| Hypopneas Mixed NREM | I71 | hypopnea_mixed_nrem | integer |
| Hypopneas Mixed NREM Supine | I72 | hypopnea_mixed_nrem_supine | integer |
| Hypopneas Mixed NREM Non Supine | I73 | hypopnea_mixed_nrem_non_supine | integer |
| Hypopneas Mixed Wake | I74 | hypopnea_mixed_wake | integer |
| Hypopneas Mixed Wake Supine | I75 | hypopnea_mixed_wake_supine | integer |
| Hypopneas Mixed Wake Non Supine | I76 | hypopnea_mixed_wake_non_supine | integer |
| Hypopneas Mixed MT | I77 | hypopnea_mixed_mt | integer |
| Hypopneas Mixed MT Supine | I78 | hypopnea_mixed_mt_supine | integer |
| Hypopneas Mixed MT Non Supine | I79 | hypopnea_mixed_mt_non_supine | integer |
| REM Supine Time - minutes | C69 | rem_supine_min | real |
| REM Non Supine Time - minutes | C70 | rem_non_supine_min | real |
| NREM Supine Time - minutes | C72 | nrem_supine_min | real |
| NREM Non Supine Time - minutes | C73 | nrem_non_supine_min | real |
| Wake Supine Time - minutes | C75 | wake_supine_min | real |
| Wake Non Supine Time - minutes | C76 | wake_non_supine_min | real |
| MT Supine Time - minutes | C78 | mt_supine_min | real |
| MT Non Supine Time - minutes | C79 | mt_non_supine_min | real |
| Apneas Obstructive Total | D80 | apnea_obs_total | integer |
| Apneas Central Total | E80 | apnea_central_total | integer |
| Apneas Mixed Total | F80 | apnea_mixed_total | integer |
| Hypopneas Obstructive Total | G80 | hypopnea_obs_total | integer |
| Hypopneas Central Total | H80 | hypopnea_central_total | integer |
| Hypopneas Mixed Total | I80 | hypopnea_mixed_total | integer |
| REM UARS | J68 | uars_rem | integer |
| REM Supine UARS count | J69 | uars_rem_supine | integer |
| REM Non Supine UARS count | J70 | uars_rem_non_supine | integer |
| NREM UARS | J71 | uars_nrem | integer |
| NREM Supine UARS count | J72 | uars_nrem_supine | integer |
| NREM Non Supine UARS count | J73 | uars_nrem_non_supine | integer |

| | | | |
|---|---|---|---|
| Wake UARS | J74 | uars_wake | integer |
| Wake Supine UARS count | J75 | uars_wake_supine | integer |
| Wake Non Supine UARS count | J76 | uars_wake_non_supine | integer |
| | | | |
| MT UARS | J77 | uars_mt | integer |
| MT Supine UARS count | J78 | uars_mt_supine | integer |
| MT Non Supine UARS count | J79 | uars_mt_non_supine | integer |
| | | | |
| UARS Event totals | J80 | uars_events_total | integer |
| | | | |
| Total Events Supine - Count | D83 | resp_events_supine_total | integer |
| Total Events Non- Supine - Count | D84 | resp_events_non_supine_total | integer |
| Total Respiratory Events - Count | D85 | resp_events_total | integer |
| | | | |
| Total Events Supine - Index | E83 | resp_events_supine_index | real |
| Total Events Non- Supine - Index | E84 | resp_events_non_supine_index | real |
| Total Respiratory Events - Index | E85 | resp_events_index | real |

**Respiratory Info**

**ALL**

| | | | |
|---|---|---|---|
| All Apneas (Event Totals: obs+central+mixed) | AL4 | apneas_total | integer |
| All Hypopneas (Event Totals: obs+central+mixed) | AL5 | hypopneas_total | integer |
| All Apneas B | AL6 | apneas_supine_total | integer |
| All Hypopneas B | AL7 | hypopneas_supine_total | integer |
| All Apneas NS | AL8 | apneas_non_supine_total | integer |
| All Hypopneas NS | AL9 | hypopneas_non_supine_total | integer |

**REM Respiratory Info**

| | | | |
|---|---|---|---|
| All Hypopneas | AL16 | hypopneas_rem_total | integer |
| All Apneas * | D68 + E68 + F68 | apneas_rem_total | integer |

**NREM Respiratory Info**

| | | | |
|---|---|---|---|
| All Hypopneas | AL26 | hypopneas_nrem_total | integer |
| All Apneas * | D71 + E71 + F71 | apneas_nrem_total | integer |

**REM Supine / Non Supine**

| | | | |
|---|---|---|---|
| All Hypopneas - Supine | AL37 | hypopneas_rem_supine_total | integer |
| All Hypopneas - Non Supine | AL44 | hypopneas_rem_non_supine_total | integer |
| All Apneas * - Supine | D69 + E69 + F69 | apneas_rem_supine_total | integer |
| All Apneas * - Non Supine | D70 + E70 + F70 | apneas_rem_non_supine_total | integer |

**NON-REM Supine / Non Supine**

| | | | |
|---|---|---|---|
| All Hypopneas - Supine | AL53 | hypopneas_nrem_supine_total | integer |
| All Hypopneas - Non Supine | AL60 | hypopneas_nrem_non_supine_total | integer |
| All Apneas* - Supine | D72 + E72 + F72 | apneas_nrem_supine_total | integer |
| All Apneas* - Non Supine | D73 + E73 + F73 | apneas_nrem_non_supine_total | integer |

**WAKE**

| | | | |
|---|---|---|---|
| All Hypopneas | AL70 | hypopneas_wake_total | integer |
| All Apneas * | D74 + E74 + F74 | apneas_wake_total | integer |
| All Hypopneas - Supine | AL77 | hypopneas_wake_supine_total | integer |
| All Hypopneas - Non Supine | AL84 | hypopneas_wake_non_supine_total | integer |
| All Apneas * - Supine | D75 + E75 + F75 | apneas_wake_supine_total | integer |

| | | | |
|---|---|---|---|
| All Apneas * - Non Supine | D76 + E76 + F76 | apneas_wake_non_supine_total | integer |
| **MT** | | | |
| All Hypopneas | AL93 | hypopneas_mt_total | integer |
| All Apneas * | D77 + E77 + F77 | apneas_mt_total | integer |
| All Hypopneas - Supine | AL100 | hypopneas_mt_supine_total | integer |
| All Hypopneas - Non Supine | AL170 | hypopneas_mt_non_supine_total | integer |
| All Apneas * - Supine | D78 + E78 + F78 | apneas_mt_supine_total | integer |
| All Apneas * - Non Supine | D79 + E79 + F79 | apneas_mt_non_supine_total | integer |

* Apnea aggregate attributes were absent in Microsoft Excel sheet. Hence, they were made up by us to complement the "hypopnea" aggregate attributes.

**Table 4.4 Limb Movement Summary**

| Attribute Name | Cell - Index | Database attribute | Type |
|---|---|---|---|
| Limb movements - REM | E91 | lm_rem | integer |
| Limb movements - NREM | F91 | lm_nrem | integer |
| Limb movements - Arousal | G91 | lm_arousal | integer |
| Limb movements - No Arousal | H91 | lm_no_arousal | integer |
| Limb movements - Total | I91 | lm_total | integer |
| Limb movements - Index | J91 | lm_index | real |
| Periodic Limb movements - REM | E92 | plm_rem | integer |
| Periodic Limb movements - NREM | F92 | plm_nrem | integer |
| Periodic Limb movements - Arousal | G92 | plm_arousal | integer |
| Periodic Limb movements - No Arousal | H92 | plm_no_arousal | integer |

| | | | |
|---|---|---|---|
| Periodic Limb movements - Total | I92 | plm_total | integer |
| Periodic Limb movements - Index | J92 | plm_index | real |
| | | | |
| Respiratory related Limb movements - REM | E93 | rrlm_rem | integer |
| Respiratory related Limb movements - NREM | F93 | rrlm_nrem | integer |
| Respiratory related Limb movements - Arousal | G93 | rrlm_arousal | integer |
| Respiratory related Limb movements - No Arousal | H93 | rrlm_no_arousal | integer |
| Respiratory related Limb movements - Total | I93 | rrlm_total | integer |
| Respiratory related Limb movements - Index | J93 | rrlm_index | real |
| | | | |
| Total Limb movements - REM | E94 | total_lm_rem | integer |
| Total Limb movements - NREM | F94 | total_lm_nrem | integer |
| Total Limb movements - Arousal | G94 | total_lm_arousal | integer |
| Total Limb movements - No Arousal | H94 | total_lm_no_arousal | integer |
| Total Limb movements - Total | I94 | total_lm_total | integer |
| Total Limb movements - Index | J94 | total_lm_index | real |

**Limb Information**

| | | | |
|---|---|---|---|
| **ALL** | | | |
| LM Both | AU16 | lm_both | integer |
| LM Left | AU17 | lm_left | integer |
| LM Right | AU18 | lm_right | integer |
| PLM Both | AU19 | plm_both | integer |
| PLM Left | AU20 | plm_left | integer |
| PLM Right | AU21 | plm_right | integer |
| RRLM Both | AU22 | rrlm_both | integer |
| RRLM Left | AU23 | rrlm_left | integer |
| RRLM Right | AU24 | rrlm_right | integer |
| | | | |
| **REM** | | | |
| LM Both | BB16 | lm_rem_both | integer |
| LM Left | BB17 | lm_rem_left | integer |
| LM Right | BB18 | lm_rem_right | integer |
| PLM Both | BB19 | plm_rem_both | integer |
| PLM Left | BB20 | plm_rem_left | integer |
| PLM Right | BB21 | plm_rem_right | integer |
| RRLM Both | BB22 | rrlm_rem_both | integer |

| | | | |
|---|---|---|---|
| RRLM Left | BB23 | rrlm_rem_left | integer |
| RRLM Right | BB24 | rrlm_rem_right | integer |
| | | | |
| **NREM** | | | |
| LM Both | BI16 | lm_nrem_both | integer |
| LM Left | BI17 | lm_nrem_left | integer |
| LM Right | BI18 | lm_nrem_right | integer |
| PLM Both | BI19 | plm_nrem_both | integer |
| PLM Left | BI20 | plm_nrem_left | integer |
| PLM Right | BI21 | plm_nrem_right | integer |
| RRLM Both | BI22 | rrlm_nrem_both | integer |
| RRLM Left | BI23 | rrlm_nrem_left | integer |
| RRLM Right | BI24 | rrlm_nrem_right | integer |

**Table 4.5 Arousals**

| Attribute Name | Cell - Index | Database attribute | Type |
|---|---|---|---|
| Apnea - count | C100 | apnea_count | integer |
| Hypopnea - count | C101 | hypopnea_count | integer |
| Snore - count | C102 | snore_count | integer |
| Desaturation - count | C103 | desat_count | integer |
| Spontaneous - count | C104 | spontaneous_count | integer |
| Limb Movement - count | C105 | lm_count | integer |
| Periodic LM - count | C106 | plm_count | integer |
| Respiratory RLM - count | C107 | rrlm_count | integer |
| Total - count | C108 | arousal_total | integer |
| | | | |
| Apnea - index | D100 | apnea_index | real |
| Hypopnea - index | D101 | hypopnea_index | real |
| Snore - index | D102 | snore_index | real |
| Desaturation - index | D103 | desat_index | real |
| Spontaneous - index | D104 | spontaneous_index | real |
| Limb Movement - index | D105 | lm_index | real |
| Periodic LM - index | D106 | plm_index | real |
| Respiratory RLM - index | D107 | rrlm_index | real |
| Total - index | D108 | arousal_index | real |

**Table 4.6 EKG Summary**

| Attribute Name | Cell - Index | Database attribute | Type |
|---|---|---|---|
| Mean Heartrate | J100 | heartrate_mean | real |
| Mean REM Heartrate | J101 | heartrate_rem_mean | real |
| Mean NREM Heartrate | J102 | heartrate_nrem_mean | real |
| Mean Wake Heartrate | J103 | heartrate_wake_mean | real |
| Mean MT Heartrate | J104 | heartrate_mt_mean | real |

**Table 4.7 Oxygen Saturation Summary**

| Attribute Name | Cell - Index | Database attribute | Type |
|---|---|---|---|
| Mean SaO2 - % | D121 | mean_sao2_perc | real |
| Lowest Desat - % | D123 | lowest_desat_perc | real |
| Minutes TRT SaO2 < 90% | D125 | trt_sao2_LT90 | real |
| Mean Wake SaO2 | D126 | mean_wake_perc | real |
| **Desaturations 4% or > [1]**- Count | D129 | desat_MT4 | integer |
| NREM Desaturations - Count | D130 | nrem_desat | integer |
| REM Desaturations - Count | D131 | rem_desat | integer |
| Wake Desaturations - Count | D132 | wake_desat | integer |
| MT Desaturations - Count | D133 | mt_desat | integer |
| Desaturations 4% or > - index | E129 | desat_MT4_index | real |
| NREM Desaturations - index | E130 | nrem_desat_index | real |
| REM Desaturations - index | E131 | rem_desat_index | real |
| Wake Desaturations - index | E132 | wake_desat_index | real |
| MT Desaturations - index | E133 | mt_desat_index | real |
| **Saturation time table** | | | |
| 110- 95 minutes | AR49 | sttmin_110_95 | real |
| 95- 90 minutes | AS49 | sttmin_95_90 | real |
| 90- 85 minutes | AT49 | sttmin_90_85 | real |
| 85- 80 minutes | AU49 | sttmin_85_80 | real |
| 80- 75 minutes | AV49 | sttmin_80_75 | real |
| 75- 70 minutes | AW49 | sttmin_75_70 | real |
| 70- 65 minutes | AX49 | sttmin_70_65 | real |

| | | | |
|---|---|---|---|
| 65- 60 minutes | AY49 | sttmin_65_60 | real |
| 60- 55 minutes | AZ49 | sttmin_60_55 | real |
| 55- 50 minutes | BA49 | sttmin_55_50 | real |
| 50- 45 minutes | BB49 | sttmin_50_45 | real |
| 45- 40 minutes | BC49 | sttmin_45_40 | real |
| 40- 35 minutes | BD49 | sttmin_40_35 | real |
| 35- 30 minutes | BE49 | sttmin_35_30 | real |
| 30- 0 minutes | BF49 | sttmin_30_0 | real |
| | | | |
| 110- 95 % time | AR50 | perc_time_110_95 | real |
| 95- 90 % time | AS50 | perc_time_95_90 | real |
| 90- 85 % time | AT50 | perc_time_90_85 | real |
| 85- 80 % time | AU50 | perc_time_85_80 | real |
| 80- 75 % time | AV50 | perc_time_80_75 | real |
| 75- 70 % time | AW50 | perc_time_75_70 | real |
| 70- 65 % time | AX50 | perc_time_70_65 | real |
| 65- 60 % time | AY50 | perc_time_65_60 | real |
| 60- 55 % time | AZ50 | perc_time_60_55 | real |
| 55- 50 % time | BA50 | perc_time_55_50 | real |
| 50- 45 % time | BB50 | perc_time_50_45 | real |
| 45- 40 % time | BC50 | perc_time_45_40 | real |
| 40- 35 % time | BD50 | perc_time_40_35 | real |
| 35- 30 % time | BE50 | perc_time_35_30 | real |
| 30- 0 % time | BF50 | perc_time_30_0 | real |

**Note:** [1]Type-A studies have "Desaturations 3% or >" instead of "Desaturations 4% or >". "4% or >" can be read as "4% or more".

**Table 4.8 Body Position Summary**

| Attribute Name | Cell - Index | Database attribute | Type |
|---|---|---|---|
| TST - Back | C144 | tst_back | real |
| TST - Left | D144 | tst_left | real |
| TST - Right | E144 | tst_right | real |
| TST - Prone | F144 | tst_prone | real |
| TST - Upright | G144 | tst_upright | real |
| | | | |
| Total events - Back | C145 | back_events_total | integer |
| Total events - Left | D145 | left_events_total | integer |
| Total events - Right | E145 | right_events_total | integer |
| Total events - Prone | F145 | prone_events_total | integer |
| Total events - Upright | G145 | upright_events_total | integer |

| | | | |
|---|---|---|---|
| Apneas - Back | C146 | back_apnea_total | integer |
| Apneas - Left | D146 | left_apnea_total | integer |
| Apneas - Right | E146 | right_apnea_total | integer |
| Apneas - Prone | F146 | prone_apnea_total | integer |
| Apneas - Upright | G146 | upright_apnea_total | integer |
| | | | |
| Hypopneas - Back | C147 | back_hypopnea_total | integer |
| Hypopneas - Left | D147 | left_hypopnea_total | integer |
| Hypopneas - Right | E147 | right_hypopnea_total | integer |
| Hypopneas - Prone | F147 | prone_hypopnea_total | integer |
| Hypopneas - Upright | G147 | upright_hypopnea_total | integer |
| | | | |
| UARS - Back | C148 | back_hypopnea_uars | integer |
| UARS - Left | D148 | left_hypopnea_uars | integer |
| UARS - Right | E148 | right_hypopnea_uars | integer |
| UARS - Prone | F148 | prone_hypopnea_uars | integer |
| UARS - Upright | G148 | upright_hypopnea_uars | integer |
| | | | |
| **RDI**[1]- Back | C149 | back_rdi | real |
| RDI - Left | D149 | left_rdi | real |
| RDI - Right | E149 | right_rdi | real |
| RDI - Prone | F149 | prone_rdi | real |
| RDI - Upright | G149 | upright_rdi | real |
| | | | |
| REM Abdomen | AZ58 | abdomen_rem_min | real |
| REM Back | AZ59 | back_rem_min | real |
| REM Left | AZ60 | left_rem_min | real |
| REM Right | AZ61 | right_rem_min | real |
| REM Up | AZ62 | up_rem_min | real |
| | | | |
| NREM Abdomen | AZ63 | abdomen_nrem_min | real |
| NREM Back | AZ64 | back_nrem_min | real |
| NREM Left | AZ65 | left_nrem_min | real |
| NREM Right | AZ66 | right_nrem_min | real |
| NREM Up | AZ67 | up_nrem_min | real |
| | | | |
| Wake Abdomen | AZ68 | abdomen_wake_min | real |
| Wake Back | AZ69 | back_wake_min | real |
| Wake Left | AZ70 | left_wake_min | real |
| Wake Right | AZ71 | right_wake_min | real |
| Wake Up | AZ72 | up_wake_min | real |
| | | | |
| MT Abdomen | AZ73 | abdomen_mt_min | real |
| MT Back | AZ74 | back_mt_min | real |
| MT Left | AZ75 | left_mt_min | real |
| MT Right | AZ76 | right_mt_min | real |

| | | | |
|---|---|---|---|
| MT Up | AZ77 | up_mt_min | real |

**Note:** [1]**RDI** appears as **A+H** (apnea plus hypopnea) in type-B format (psg/cpap/split) type summaries.

### *4.4 Attributes: Split Night studies*

The attributes for split night studies are the same as CPAP and PSG, but there exist two tables for every section (baseline and treatment). Also, there are 2 extra sections - Sleep parameters and Advanced signal saturation time table that are exclusive to Split Night studies only. These are documented below.

**Table 4.9 Sleep parameters**

| Attribute Name | Cell - Index | Database attribute | Type |
|---|---|---|---|
| Back and REM | BP4 | back_rem | real |
| Wake after Sleep per CPAP Index | BP5 | wake_after_sleep | real |
| End of CPAP Index | BP7 | end_time | time |
| Duration of CPAP Index | BP8 | duration | real |
| Total Sleep Time per CPAP Index | BP9 | tst | real |
| Sleep Efficiency per CPAP Index | BP10 | sleep_efficiency | real |
| Number of Awakenings per CPAP Index | BP11 | awakenings | smallint |
| MT per CPAP Index | BP12 | mt | real |
| REM per CPAP Index | BP13 | rem | real |
| Stage 1 per CPAP Index | BP14 | stage1 | real |
| Stage 2 per CPAP Index | BP15 | stage2 | real |
| Stage 3 per CPAP Index | BP16 | stage3 | real |
| Stage 4 per CPAP Index | BP17 | stage4 | real |
| Wake per CPAP Index | BP18 | wake | real |
| TRT | BP19 | trt | real |
| TST Supine | BO21 | tst_supine | real |
| TST Non Supine | BQ21 | tst_non_supine | real |
| SPT | BP6 | spt | real |

| | | | |
|---|---|---|---|
| **CPAP ON** | | | |
| TST | BO61 | tst_cpap_on | real |
| TIB | BO62 | tib_cpap_on | real |
| SE | BO63 | se_cpap_on | real |
| so_cpap_on | BO65 | so_cpap_on | real |
| REM Onset | BO66 | rem_onset | real |

**Table 4.10 Advanced signal table for Oxygen Saturation**

| Attribute Name | Cell - Index baseline | Cell - Index treatment | Database attribute | Type |
|---|---|---|---|---|
| 20-30 minutes | AR68 | AR86 | min_20_30 | real |
| 30-40 minutes | AS68 | AS86 | min_30_40 | real |
| 40-50 minutes | AT68 | AT86 | min_40_50 | real |
| 50-60 minutes | AU68 | AU86 | min_50_60 | real |
| 60-70 minutes | AV68 | AV86 | min_60_70 | real |
| 70-80 minutes | AW68 | AW86 | min_70_80 | real |
| 80-90 minutes | AX68 | AX86 | min_80_90 | real |
| 90-100 minutes | AY68 | AY86 | min_90_100 | real |
| | | | | |
| 20-30 % time | AR69 | AR87 | perc_time_20_30 | real |
| 30-40 % time | AS69 | AS87 | perc_time_30_40 | real |
| 40-50 % time | AT69 | AT87 | perc_time_40_50 | real |
| 50-60 % time | AU69 | AU87 | perc_time_50_60 | real |
| 60-70 % time | AV69 | AV87 | perc_time_60_70 | real |
| 70-80 % time | AW69 | AW87 | perc_time_70_80 | real |
| 80-90 % time | AX69 | AX87 | perc_time_80_90 | real |
| 90-100 % time | AY69 | AY87 | perc_time_90_100 | real |
| | | | | |
| MIN<90% 20-90 min | AR56 | AR61 | min_lt90_20_90 | real |
| MIN<90% 20-90 % time | AR57 | AR62 | perc_time_lt90_20_90 | real |

### *4.5 Data Extraction*

The technical summary files are Microsoft Excel format sheets embedded in the REMBrandt format data that can be viewed using the REMBrandt Viewer utility. We manually copied and saved them in a Microsoft Excel Workbook. A workbook is a collection of Excel sheets. For our research, each workbook has Excel sheets for 100

patients. There can be more than one sleep study conducted for each patient, and thus, more than one summary sheet per patient (Format Type A, Type B, or Type C).

### *4.5.1 File naming scheme*

**(a) Sheet naming**

The Excel sheets within a workbook are named using the following naming convention. Note that this name is the same as the name of the folder that contains the raw sleep data. Each sheet name is pre-appended by a study identifer that is an incrementally generated integer.

**For type C:**

`1316.56.cpap.2006_07_02`

study identifier = `1316`

folder name = `56.cpap.2006_07_02`

patient identifier = `56`

study type + format type = `cpap, C`

study date = `2006_07_02 (YYYY_MM_DD)`

**Note:** Study type can be: cpap, psg, split. There is no need to specify the format type as Type C is the default format.

**For type B:**

`60.416.psgB.2004_01_02`

study identifier = `60`

folder name = `416.psgB.2004_01_02`

patient identifier = `416`

study type, format type = `psg, B`

study date = `2004_01_02 (YYYY_MM_DD)`

Similarly, for type B cpap and split studies, the study type is "cpapB" and "splitB" respectively.

**For type A:**

`345.123.cpapA.2003_11_17`

study identifier = `345`
folder name = `123.cpapA.2003_11_17`
patient identifier = `123`
study type, format type = `cpap, A`
study date = `2003_11_17 (YYYY_MM_DD)`

Similarly, for type A cpap study, the study type is "cpapA". Note that there are no split type A studies in the data we have.

**(b) Workbook naming**

The workbooks are named with regards to the patients they contain.

`301.400.summaryWorkbook.xls`

The above workbook contains sheets from studies of patients whose identifiers go from 301 - 400.

### *4.5.2 Code, Input and Output*

The code to extract data was written using the Java programming language in Eclipse IDE using the Apache POI 3.0.1 library to access the Excel files.

A run of the code reads the target Excel workbook, and iterates through all the summary sheets contained within it. Within each sheet, the values are extracted from each of the sections and appended to a CSV file corresponding to that section:

```
arousalSummary.csv
bodyPositionSummary.csv
ekgSummary.csv
limbMovementSummary.csv
oxygenSaturationSummary.csv
respiratoryDisturbanceSummary.csv
sleepSummary.csv
```

and for split only -

```
splitAdvancedSaturationSummary.csv
splitSleepParametersSummary.csv
```

Each of the CSV files above contains data for all the patients. The header of the CSV files are the database attributes listed in Section 4.3 Attributes: PSG and CPAP studies and Section 4.4 Attributes: Split Night studies. Each record is identified by the attributes study ID and type of study. The study ID does not suffice as a key attribute, because for split studies are divided into baseline and treatment observations. The study ID for these is the same, but study type differs (`split.dx` and `split.rx`, corresponding to baseline and treatment studies). Hence, the study type too is required along with Study ID and they form a composite key pair.

### *4.5.3 Handling different format types - A, B, C*

The code identifies the type of format by extracting this detail from the name of the sheet and then handles it appropriately.

### *4.5.4 Missing values*

The missing values in the data sheets are replaced by -1 for integers, -1.0 for doubles.

**(a) Blank cells**

If the cell is blank, a runtime Null Pointer exception is thrown. The code catches this exception and handles it by returning a -1 value, indicating the value was not known.

**(b) Numeric cells with '-' as values**

Sometimes a cell that is expected to contain an Integer or a Double can contain a character '-'. Thus, the code to read Integer or Double values first checks if the content of the cell is of string type. If it is, then it recognizes it as a missing value. If not, it goes ahead and reads the numeric value in the cell.

**(c) #VALUE ? or #DIV/0! as cell values**

Some split studies encountered had #VALUE ? or, #DIV/0! as cell values. These were manually replaced by -1.

## *4.6 Technical Summary Database*

Each section of the technical summary Excel sheet maps to a database table, and its attributes to the fields of the table.

### *4.6.1 Schema and tables*

A schema called `summary` is created within the database. All the summary-data related tables are created in this schema:

`summary.arousal`
`summary.bodyposition`
`summary.ekg`
`summary.limbmovement`
`summary.oxysat`
`summary.rd` (respiratory disturbance)
`summary.sleep`

split only -

`summary.advoxyst`
`summary.sleepparam`

By using the PostgreSQL `COPY` operator, each CSV file corresponding to a summary section can be exported to the PGSQL database.

There are 1319 studies for which we have complete survey data.

### *4.6.2 Patient-Micro study map table*

Using the Study ID and Patient ID information present in the name of every technical summary report, we have created a patient - study map table, since a patient can undergo more than one type of micro study over time. Study ID is the primary key `<PK>` constraint of this table.

**Table 4.11 Patient-Study Map table**

| No. | Attribute name | Description | Type |
|---|---|---|---|
| 1 | sid <PK> | Identifier of the study | int |
| 2 | pid | Identifier of the patient | int |
| 3 | study_type | Type of study | text |
| 4 | pname | Patient initials | text |
| | | | |

### *4.7 Normative technical summary data for males and females*

This is a reference table that gives durations of different sleep period types, sleep stages and sleep efficiency, number of awakenings and REM periods of various normal male and female age groups.

Note: The normative data could differ in the values of few attributes from one study to the other based on subpopulations between the patients[24]. The following tables give the normative data for a study.

**Table 4.12 Normative data for males**

**Part I**

| | 20 - 29 | | 30 - 39 | | 40 - 49 | |
|---|---|---|---|---|---|---|
| | Mean | Std.Dev. | Mean | Std.Dev. | Mean | Std.Dev. |
| TIB | 442.23 | 12.22 | 434.55 | 20.49 | 429.1 | 39.17 |
| SPT | 424.64 | 14.42 | 427.85 | 23.16 | 414.95 | 36.86 |
| TST | 419.27 | 14.51 | 421.45 | 21.93 | 389.1 | 46.5 |
| Sleep Efficiency | 0.95 | 0.04 | 0.97 | 0.02 | 0.91 | 0.06 |
| # Awakenings | 3.05 | 2.57 | 2.5 | 1.43 | 4.65 | 2.27 |
| # REM Periods | 4.05 | 0.65 | 4.5 | 0.71 | 4.35 | 0.94 |
| Sleep Latency | 14.55 | 11.38 | 5.8 | 3.85 | 10 | 7.87 |
| REM Latency | 88.27 | 21.32 | 85.35 | 30.08 | 71.65 | 32.77 |
| Wake (%SPT) | 1.26 | 1.08 | 1.47 | 1.94 | 6.29 | 5.56 |
| Stage 1 | 4.44 | 1.62 | 5.71 | 3.43 | 7.56 | 3.03 |
| Stage 2 | 45.54 | 5.17 | 56.89 | 7.36 | 54.75 | 11.14 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Stage 3 | 6.21 | 1.35 | 5.67 | 1.46 | 5.37 | 3.27 |
| Stage 4 | 14.55 | 4.38 | 6.79 | 5.2 | 3.18 | 6.25 |
| REM | 28 | 5.66 | 23.47 | 3.86 | 22.85 | 4 |
| | | | | | | |
| DELTA | 20.76 | | 12.46 | | 8.55 | |

**Part II**

| | 50 - 59 | | 60 - 69 | | 70 - 79 | |
|---|---|---|---|---|---|---|
| | Mean | Std.Dev. | Mean | Std.Dev. | Mean | Std.Dev. |
| TIB | 422.58 | 44.93 | 451.55 | 37.7 | 493.09 | 49.48 |
| SPT | 406.96 | 45.88 | 441.2 | 37.47 | 448.86 | 35.31 |
| TST | 389.79 | 49.54 | 407.3 | 44.65 | 372.86 | 87.22 |
| Sleep Efficiency | 0.95 | 0.92 | 0.04 | 0.9 | 0.07 | 0.77 |
| # Awakenings | 3.05 | 5.67 | 1.78 | 7.55 | 3.69 | 7.09 |
| # REM Periods | 4.05 | 4.17 | 0.84 | 4.55 | 0.72 | 4.14 |
| Sleep Latency | 14.55 | 11.92 | 10.52 | 8.25 | 7.2 | 32 |
| REM Latency | 88.27 | 84.79 | 19.37 | 83.9 | 38.12 | 111.14 |
| Wake (%SPT) | 1.26 | 4.33 | 2.33 | 7.73 | 6.02 | 16 |
| Stage 1 | 7.56 | 3.94 | 9.73 | 3.97 | 9.47 | 3.46 |
| Stage 2 | 61.71 | 10.3 | 56.79 | 8.76 | 55.49 | 16.62 |
| Stage 3 | 1.69 | 3.2 | 2.06 | 3.34 | 1.36 | 2.34 |
| Stage 4 | 4.92 | 7.7 | 0.6 | 1.9 | 0 | 0 |
| REM | 21.48 | 4.01 | 23.09 | 3.59 | 17.68 | 6.63 |
| | | | | | | |
| DELTA | 6.61 | | 2.66 | | 1.36 | |

## Sleep stage pie-chart distribution:

| | MEAN | STD DEV |
|---|---|---|
| Stage 1 | 7.56 | 3.03 |
| Stage 2 | 54.75 | 11.14 |
| Stage 3 | 5.37 | 3.27 |
| Stage 4 | 3.18 | 6.25 |
| REM | 22.85 | 4 |
| DELTA | 8.55 | |

## Table 4.13 Normative data for females

**Part I**

| | 20 - 29 | | 30 - 39 | | 40 - 49 | |
|---|---|---|---|---|---|---|
| | Mean | Std.Dev. | Mean | Std.Dev. | Mean | Std.Dev. |
| TIB | 445.65 | 24 | 443.65 | 18.64 | 441.45 | 22.9 |
| SPT | 432.3 | 22.4 | 433.25 | 19.48 | 432.32 | 23.82 |
| TST | 429.95 | 21.71 | 425.7 | 31.53 | 425.18 | 23.32 |
| Sleep Efficiency | 0.96 | 0.02 | 0.96 | 0.06 | 0.96 | 0.02 |
| # Awakenings | 1.1 | 0.78 | 1.4 | 1.15 | 3.09 | 2.1 |
| # REM Periods | 4.2 | 0.89 | 4.3 | 0.63 | 3.95 | 0.61 |
| Sleep Latency | 12.9 | 9.62 | 9.8 | 9.51 | 7.82 | 6.01 |
| REM Latency | 100.2 | 44.18 | 78.65 | 19.47 | 82.32 | 34.04 |
| Wake (%SPT) | 0.53 | 0.49 | 1.84 | 3.97 | 1.63 | 1.3 |
| Stage 1 | 4.18 | 2.39 | 4.17 | 1.65 | 5.64 | 2 |
| Stage 2 | 52.37 | 5.89 | 53.77 | 7.73 | 54.01 | 8.55 |
| Stage 3 | 5.27 | 1.97 | 6.42 | 3.51 | 7.51 | 3.48 |
| Stage 4 | 12.42 | 6.24 | 7.58 | 6.67 | 4.54 | 6.91 |
| REM | 25.23 | 3.63 | 26.22 | 5.27 | 26.67 | 4.1 |
| | | | | | | |
| DELTA | 17.69 | | 14 | | 12.05 | |

**Part II**

| | 50 - 59 | | 60 - 69 | | 70 - 79 | |
|---|---|---|---|---|---|---|
| | Mean | Std.Dev. | Mean | Std.Dev. | Mean | Std.Dev. |
| TIB | 466.73 | 47.2 | 465.68 | 45.21 | 507.05 | 56.29 |
| SPT | 454.77 | 44.26 | 444.14 | 47.51 | 470.7 | 46.88 |
| TST | 430.77 | 34.94 | 404.95 | 55.8 | 413.55 | 36.64 |
| Sleep Efficiency | 0.96 | 0.93 | 0.07 | 0.87 | 0.09 | 0.82 |
| # Awakenings | 1.1 | 4.64 | 2.12 | 4.36 | 2.17 | 8.35 |
| # REM Periods | 4.2 | 4.32 | 0.93 | 4.04 | 0.96 | 4.5 |
| Sleep Latency | 12.9 | 10.59 | 4.99 | 16.5 | 13.83 | 15.4 |
| REM Latency | 100.2 | 85.73 | 26.52 | 90.14 | 41.82 | 88.45 |
| Wake (%SPT) | 0.53 | 4.95 | 6.48 | 8.93 | 8.47 | 11.69 |
| Stage 1 | 4.85 | 2.2 | 7.69 | 4.12 | 6.59 | 2.28 |
| Stage 2 | 57.8 | 6.5 | 54.78 | 8.59 | 52.22 | 8.34 |
| Stage 3 | 6.49 | 2.38 | 4.5 | 3.99 | 6.3 | 4.03 |
| Stage 4 | 4.14 | 5.28 | 2.67 | 3.04 | 3.74 | 5.68 |
| REM | 21.77 | 3.26 | 21.43 | 4.04 | 19.46 | 4.23 |
| | | | | | | |
| DELTA | 10.63 | | 7.17 | | 10.04 | |

**Sleep stage pie-chart distribution:**

| | MEAN | STD DEV |
|---|---|---|
| Stage 1 | 5.64 | 2 |
| Stage 2 | 54.01 | 8.55 |
| Stage 3 | 7.51 | 3.48 |
| Stage 4 | 4.54 | 6.91 |
| REM | 26.67 | 4.1 |
| DELTA | 12.05 | |

This normative data for males and females is stored in `summary` schema, in 2 tables table called `normative.male` and `normative.female` tables corresponding to male and female data for every study (See Table 4.15 Normative data table). The attribute values of this table are the means and standard deviations of all the attributes that are mapped from the normative table. The following tables give an idea of the naming of the attributes in the database:

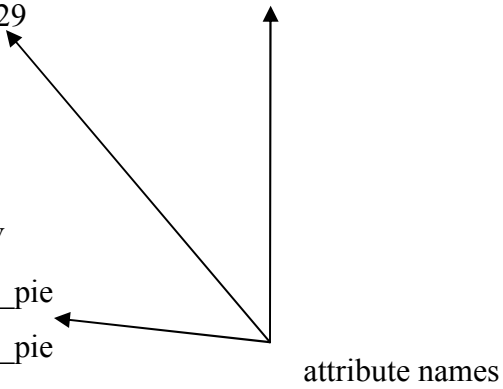
**Table 4.14 Attribute naming for normative data**

| | **20 - 29** | |
|---|---|---|
| | **Mean** | **Std.Dev.** |
| **TIB** | tib_mean_20_29 | tib_sdev_20_29 |
| **SPT** | spt_mean_20_29 | spt_sdev_20_29 |
| **TST** | tst_mean_20_29 | tst_sdev_20_29 |
| **Sleep Efficiency** | sleep_eff_mean_20_29 | sleep_eff_sdev_20_29 |
| **# Awakenings** | awakenings_mean_20_29 | awakenings_sdev_20_29 |
| **# REM Periods** | rem_periods_mean_20_29 | rem_periods_sdev_20_29 |
| **Sleep Latency** | sleep_latency_mean_20_29 | sleep_latency_sdev_20_29 |
| **REM Latency** | rem_latency_mean_20_29 | rem_latency_sdev_20_29 |
| **Wake (%SPT)** | wake_spt_mean_20_29 | wake_spt_sdev_20_29 |
| **Stage 1** | stage1_spt_mean_20_29 | stage1_spt_sdev_20_29 |
| **Stage 2** | stage2_spt_sdev_20_29 | stage2_spt_sdev_20_29 |

| | | |
|---|---|---|
| **Stage 3** | stage3_spt_sdev_20_29 | stage3_spt_sdev_20_29 |
| **Stage 4** | stage4_spt_sdev_20_29 | stage4_spt_sdev_20_29 |
| **REM** | rem_spt_mean_20_29 | rem_spt_sdev_20_29 |
| **DELTA** | delta_spt_mean_20_29 | |

**Sleep stage pie-chart distribution:**

| | **MEAN** | **STD DEV** |
|---|---|---|
| **Stage 1** | stage1_mean_pie | stage1_sdev_pie |
| **Stage 2** | stage2_mean_pie | stage2_sdev_pie |
| **Stage 3** | stage3_mean_pie | stage3_sdev_pie |
| **Stage 4** | stage4_mean_pie | stage4_sdev_pie |
| **REM** | rem_mean_pie | rem_sdev_pie |
| **DELTA** | delta_mean_pie | |




The above attributes and corresponding values exist in the table as:

**summary.normative_male**

| **sid** | **study** | **tib_mean_20_29** | **tib_sdev_20_29** | **spt_mean_20_29** | **spt_sdev_20_29** | **..** |
|---|---|---|---|---|---|---|
| 1 | psg | 442.23 | 12.22 | 424.64 | 14.42 | ... |
| 2 | cpap | 442.23 | 12.22 | 424.64 | 14.42 | ... |
| ... | ... | ... | ... | ... | ... | ... |

**Table 4.15 Normative data table**

The attributes study id (`sid`), and the study (`study`) act as a composite key to identify every row of the table.

### *4.8 Effect on sleep efficiency*

We present a couple of visualizations produced when we compared the attributes of sleep summary data and the macro data.

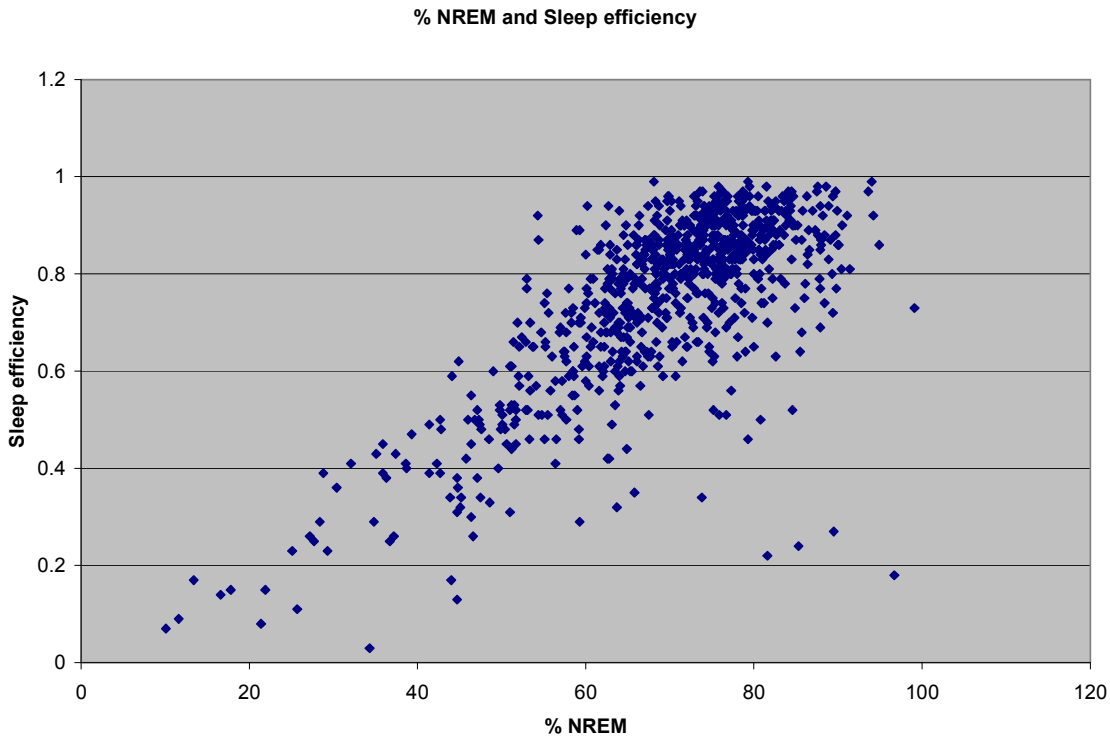



**Figure 4.3 Sleep efficiency and % NREM**

The above graph shows a high correlation between the sleep efficiency and the time patient spends in the Non-REM stage (Stages 1, 2, 3, and 4) in our data.

The correlation between the sleep efficiency and the age of the patient shows a falling scatter over time, as seen in the following scatter plot.

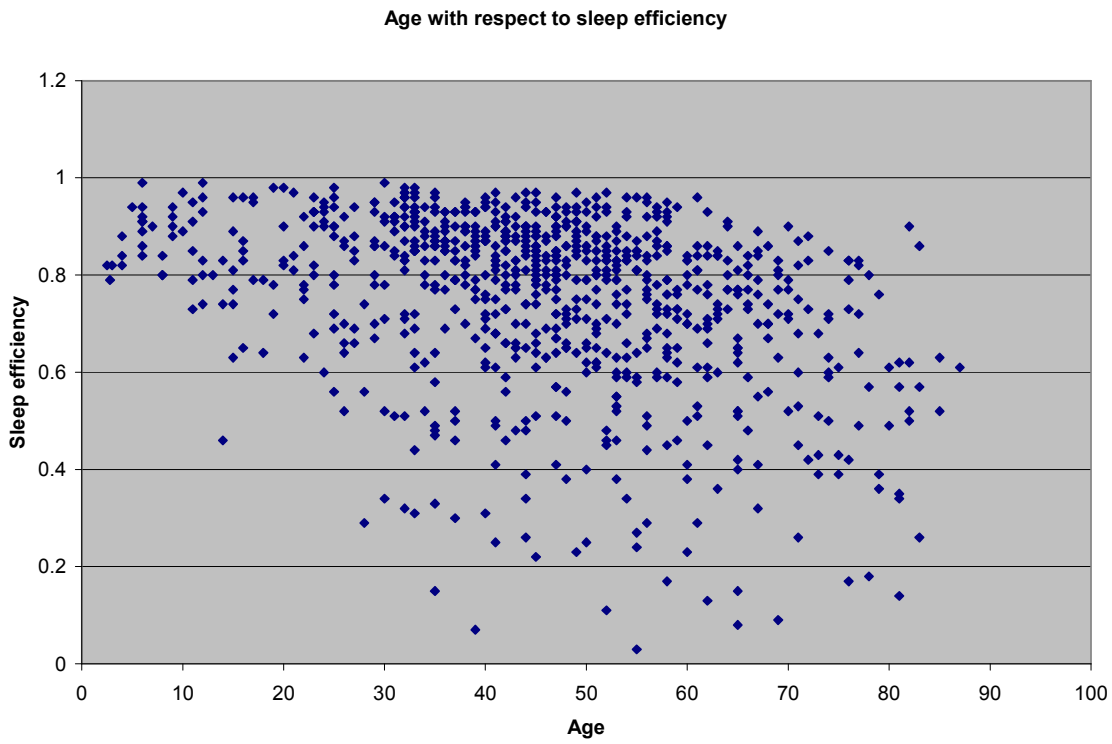



**Figure 4.4 Age and Sleep efficiency**

# 5 Micro Dataset

## *5.1 Context*

Micro level data, also known as objective data, is the time-series signal data obtained during conduct of overnight sleep observation on the patient. Clinical instruments attached to a patient via electrodes are able to record time-series signals that originate from biochemical and physical processes taking place within the sleeping patient's body. Polysomnography (PSG/sleep study) is a clinical procedure that records physiologic attributes during sleep. A polysomnogram reads the electrical potential of the brain (using EEG - Electroencephalogram); electrical activity of muscles (using EMG – Electromyogram); and electric potentials resulting from eye movements (using EOG – Electrooculogram) into time series signals. Also, signals tracking body processes like electrical activity of heart (using ECG – Electrocardiogram), blood oxygen level, body position, limb movements, snoring and blood pressure are registered.

In a typical PSG/CPAP/Split Night study, data is collected from approximately 50 - 55 different signals for an average of around 7.2 hours of sleep.

All the required signal information is stored in a REMbrandt proprietary format file [13]. Data of each patient occupies 375 MB on an average of storage on Compact Disc media.

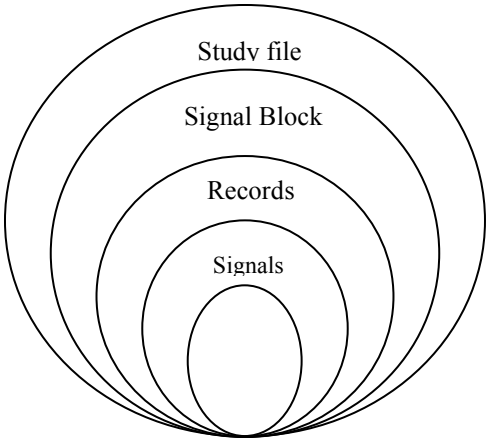



**Figure 5.1 High level representation of composition of micro data**

From the above figure, at the lowest level, the micro data is composed of signal values generated at a certain fixed cycle rate per second, which constitute a signal. Each record is composed of a series of signals (and thus, the signal data points). One record is generated every second. All the records constitute a signal block in a micro study file.

The study file also has some meta-data about the signal properties (units, dimensions, physical and digital minimums and maximums, sampling rates etc), called as signal header, and a global header (patient name, date of study, number of records, number of signals, and duration of each record) for the sleep study).

The large file size of one sleep study is due to the high sampling rates of the majority of the signals. Almost 30 signals are sampled at 200 Hz, with each sample being a short integer type. For a 7.2 hour long study (26000 seconds), storage occupied by one such signal amounts to: 200 Hz * 26000 seconds * 2 Bytes ~ 10 MB (each data point is a 2 byte short integer). For 30 such signals, the storage needed increases to 300 MB.

After completing the surveys, the patient is asked to undertake the overnight sleep study. The sleep clinics have study labs that appear like hotel suites designed to keep the patient as comfortable as possible for a good night's sleep. Electrodes are then attached on different parts of patient's body to collect the signal data. (For information on electrode placements, see Section 2.3.2 Signal Information).

Each patient file comes with a Compact Disc media that has a sleep study recorded in REMBrandt format. The average size of this data is 375 MB.

## *5.2 Data Conversion*

### *5.2.1 EDF Extraction*

The signal data from the sleep study is in a file of REMBrandt format. It can be viewed using the REMBrandt viewer utility along with other aspects of the sleep study like the

technical summary report and patient's profile. However, REMbrandt viewer does not have the feature to extract the signal data into a file so that we can store it in the database (nor the structure of data in REMBrandt format is known). Instead, the REMBrandt package comes with a utility called "REMbrandtEDFexport.exe" to extract the data into European Data Format (EDF). EDF is a single file that contains information about the study, properties of the signals and the signal data points for an overnight of observation. The European Data Format (EDF) [3]. first published in 1992, was the outcome of an effort to devise a common format to compare sleep-wake analysis results coming from different sources.

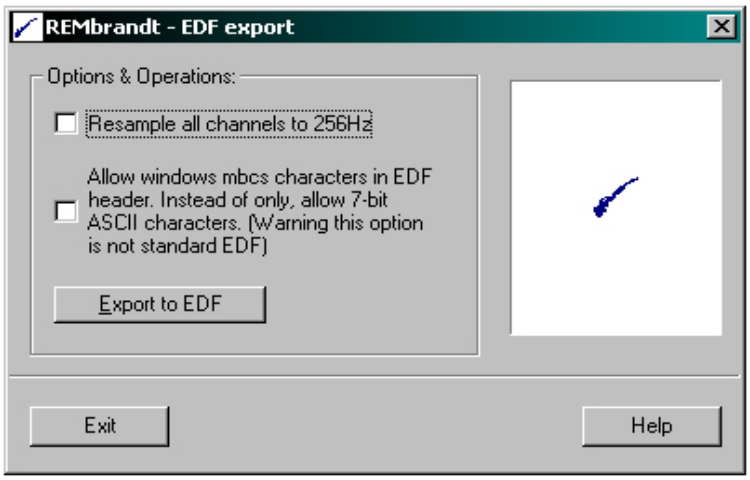



**Figure 5.2 REMbrandt EDF exporter utility interface**

### *5.2.2 Preserve the sampling rates*

During the REMbrandt -> EDF conversion, the sampling rate of all the signals is preserved. Thus, the signals are available at their natural frequency which significantly reduces the data storage costs. For instance, the signal that tracks patient's heart beat is sampled at 2 Hz per second (or 2 data points per second). For a sleep study lasting 7.2 hours (26,000 seconds), the size of signal data amounts to: 26000 sec * 2 Hz * 2 bytes = 104,000 bytes or 101.56 KB (each data point is a 2 byte short integer). If this signal is re-

sampled to 256 Hz during the conversion, the storage would go up to: 26000 sec * 256 Hz * 2 bytes = 13,31,2000 bytes or 12.7 MB.

### *5.2.3 Multiple EDF files*

As observed during EDF extraction, the conversion of a REMbrandt CPAP and Split study file to EDF may result in the creation of more than one EDF file (2 to 3 files may be generated). One of these files (the largest in size, usually 300 - 500 MB) corresponds to the overnight sleep study. The remaining files are much smaller in size (usually less than 5 MB) and may have to do with the minor test runs of the sleep monitoring system before the major night study is conducted. The minor CPAP EDF files may deal with experimenting whether the patient can be administered a CPAP procedure or not.

## *5.3 EDF Naming Scheme*

The micro data for every study belongs to a source folder that was named as per the following semantics:

`56.cpap.2006_07_02`
patient id = `56`
study type = `cpap`
study date = `2006_07_02 (YYYY_MM_DD)`

When an EDF file is extracted by using the "REMbrandtEDFexport.exe" utility, the file generated is put in a separate directory that contains EDF files and is renamed to bear the same name as that of the source folder (For instance, the name of EDF file is: `56.cpap.2006_07_02.edf`). Additionally, a study identifier is pre-appended to the EDF file name by referring to the name of the corresponding technical summary data sheet (see Section 4.5.1 File naming scheme, part (a)). Thus the study id, patient id and study type information available in the EDFs name. This will help us identify the study when it is uploaded into the database.

### *5.4 EDF Description*

An EDF file is capable of storing multi-channel biological signals. The specification of the structure of an EDF file is provided in[3]. The EDF format has undergone extensions over time, the latest format being called EDF+. However, the data that we deal with is in old EDF format (Version 0).

The body of an EDF file consists of 3 major sections, in the following order:

#### *5.4.1 Header*

The topmost section of the EDF is the header file which is 256 bytes in size, and contains global information on the EDF file.

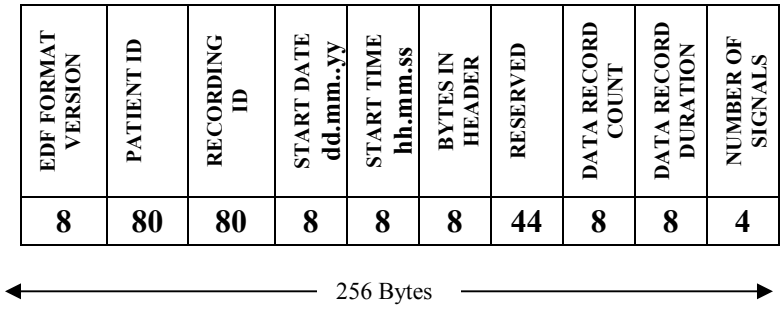



**Figure 5.3 EDF Header format (Note: the numbers denote ASCII bytes)**

**Description:**

1. EDF Format Version - 8 ASCII bytes
    - Version of the EDF format we are dealing with
    - Value: 0
2. Patient ID - 80 ASCII bytes
    - Local Patient Identification
    - Value: Patient's name

- Format: LASTNAME, FIRSTNAME

3. Recording ID - 80 ASCII bytes
    - Local Recording Identification
    - Value: This field appears blank in all the studies

4. Start Date - 8 ASCII bytes
    - The date of the study
    - Format: DD.MM.YY
    - Sample value: 23.02.06

5. Start Time - 8 ASCII bytes
    - The start time of the study
    - Format: HH.MM.SS
    - Sample value: 06.30.12

6. Bytes in Header - ASCII bytes
    - Number of bytes in the header and signal header
    - Value: (256 + 256 * Number of signals) bytes
7. Reserved - 44 ASCII bytes
    - No data

8. Data record count - 8 ASCII bytes
    - Number of data records in the EDF file

9. Data Record Duration - 8 ASCII bytes
    - Number of seconds
    - Value: 1 (for all the sleep studies)

10. Number of signals - ASCII bytes
    - 50 to 55 signals in every study

- 65 different signals that we know

### *5.4.2 Signal Properties Header*

Following EDF header is specific signal information, whose size depends upon the number of signals present in the recordings (256 bytes * number of signals).

| NAME OF SIGNAL | TRANSDUC--ER TYPE | PHYSICAL DIMENSION | PHYSICAL MINIMUM | PHYSICAL MAXIMUM | DIGITAL MINIMUM | DIGITAL MAXIMUM | PREFILTER-ING | SAMPLING RATE | RESERVED |
|---|---|---|---|---|---|---|---|---|---|
| 16 | 80 | 8 | 8 | 8 | 8 | 8 | 80 | 8 | 32 |

256 Bytes * number of signals

**Figure 5.4 EDF Signal Header format (Note: the numbers denote ASCII bytes)**

**Description:**

The information in the signal header is structured so that information in one type of field is grouped together for all the signals. For instance, for a study consisting of 55 signals, the first 16*55 = 888 Bytes are occupied by the 'Name of Signal' field for all the signals. This also defines the ordering of the signal properties for the rest of the header fields, and also the sequence of signals that make up a record of the signal block.

1. Name of Signal
    - Example: 'heartrate', 'bodyposition', 'snore', 'cflow', 'oxysat', etc.
    - Size: (16 * number of signals) Bytes

2. Transducer Type
    - Example: 'AgAgCl cup electrodes'
    - For sleep study EDF, this field is left blank
    - Size: (80 * number of signals) Bytes

3. Physical Dimension
    - Units of the signals recorded
    - Example: 'mbar' for snore pressure, '%' for oxygen saturation, 'bpm' for heart rate
    - Size: (8 * number of signals) Bytes

4. Physical Minimum
    - Minimum extreme of physical value of a signal data point
    - Unique for every signal
    - Example: '-3276.8' for Heart rate
    - Size: (8 * number of signals) Bytes
5. Physical Maximum
    - Maximum extreme of physical value of a signal data point
    - Unique for every signal
    - Example: '3276.7' for Heart rate
    - Size: (8 * number of signals) Bytes

6. Digital Minimum (-32768)
    - Minimum extreme of digital value of a signal data point
    - Same for all the signals
    - Size: (8 * number of signals) Bytes

7. Digital Maximum (+32767)
    - Minimum extreme of digital value of a signal data point
    - Same for all the signals
    - Size: (8 * number of signals) Bytes

Each signal data point is a 2 byte type short integer, and is a digital value. The conversion between digital and physical values can be done as follows:

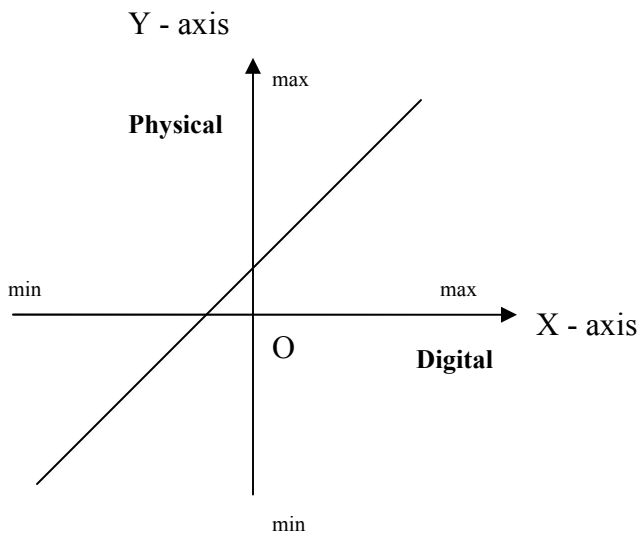



**Figure 5.5 Physical and Digital Scaling**

Using the 2-point from equation of a line:

$$y = \text{PHYmax} + \frac{x - \text{DGTLmax}}{\text{DGTLmax} - \text{DGTLmin}} * (\text{PHYmax} - \text{PHYmin})$$

**Equation 5.1 Two point form of a line**

Where,

$y$ = physical value can be calculated by plugging in $x$ (digital value from the EDF file) and the specified physical and digital extremes of the signal.

From [3], attributes 4, 5, 6, 7 often are the extreme output values of the A/D converter. They specify the offset and amplification of the signal.

8. Prefiltering

- High Pass Filtering information in Hz (HP), for example, HP: 0.1Hz
- Low Pass Filtering information in Hz (LP), for example, LP: 75Hz
- For sleep study EDF, this field is left blank
- Size: (80 * number of signals) Bytes

9. Sampling Rate

- Number of samples generated per second
- Example: 200 Hz for signal 'chest6'
- Size: (8 * number of signals) Bytes

10. Reserved

- Blank
- Size: (32 * number of signals) Bytes

The size of the signal header can be calculated as: 256 Bytes * Number of Signals

### *5.4.3 Signal Properties Table*

Following signal properties is the third and the largest part of an EDF file, the signal block - a collection of data records made of signals of data points. The table below lists all the signals (65) and their different sampling rates and physical extremes encountered while going through all the sleep studies. Any given study contains a subset of these signals (50 to 55 signals).

**Table 5.1 Signal properties**

| No. | Signal electro--des | Signal name, placements | Sampling Rate | Physical Minimum | Physical Maximum | Physical Dimens--ion (Units) |
|---|---|---|---|---|---|---|
| 1 | a1 | EEG, EOG. cranium | 200 | -3276.8 | 3276.7 | uV |
| 2 | a2 | EEG, EOG, cranium | 200 | -3276.8 | 3276.7 | uV |
| 3 | c3 | EEG, cranium | 200 | -3276.8 | 3276.7 | uV |
| 4 | c4 | EEG, cranium | 200 | -3276.8 | 3276.7 | uV |
| 5 | c3_a2 | EEG (potential difference: C3-A2) | 200 | -3276.8 | 3276.7 | uV |
| 6 | o1 | EEG, cranium | 200 | -3276.8 | 3276.7 | uV |
| 7 | o2 | EEG, cranium | 200 | -3276.8 | 3276.7 | uV |
| | | | | | | |
| 8 | fp1 | cranium | 200 | -3276.8 | 3276.7 | uV |
| 9 | fp2 | cranium | 200 | -3276.8 | 3276.7 | uV |
| 10 | f7 | cranium | 200 | -3276.8 | 3276.7 | uV |
| 11 | f3 | cranium | 200 | -3276.8 | 3276.7 | uV |
| 12 | fz | cranium | 200 | -3276.8 | 3276.7 | uV |
| 13 | f4 | cranium | 200 | -3276.8 | 3276.7 | uV |
| 14 | f8 | cranium | 200 | -3276.8 | 3276.7 | uV |
| 15 | t3 | cranium | 200 | -3276.8 | 3276.7 | uV |
| 16 | cz | cranium | 200 | -3276.8 | 3276.7 | uV |
| 17 | t4 | cranium | 200 | -3276.8 | 3276.7 | uV |
| 18 | t5 | cranium | 200 | -3276.8 | 3276.7 | uV |
| 19 | p3 | cranium | 200 | -3276.8 | 3276.7 | uV |
| 20 | pz | cranium | 200 | -3276.8 | 3276.7 | uV |
| 21 | p4 | cranium | 200 | -3276.8 | 3276.7 | uV |
| 22 | t6* | cranium | 200 | -2621440.0 | 2.62136E7 | uV |
| 23 | loc | EOG | 200 | -3276.8 | 3276.7 | uV |
| 24 | roc | EOG | 200 | -3276.8 | 3276.7 | uV |
| 25 | xflow [24] | EOG, Combined output from chest and abdomen belts (semi-quantitative signal). [24][24] | 10 | -0.26214<br>-32768.0<br>-262144.0<br>-0.03276 | 0.262136<br>32767.0<br>262136.0<br>0.032767 | V/s |
| 26 | xsum [24] | EOG, Calculated sum | 10 | -262144.0<br>-32768.0 | 262136.0<br>32767.0 | V |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | of the chest and abdomen belts. [24] | | -0.03276<br>-0.26214 | 0.032767<br>0.262136 | |
| 27 | rmi | EOG, (Respiratory Mechanics Instability) Measurement of paradoxical movements from the chest and abdomen belts. [24] | 10 | -32768.0 | 32767.0 | o |
| 28 | phase [24] | EOG, Difference between the chest and abdomen belts. Numerical value on the charts copied, now we can look at a respiratory loop. [24] | 10 | -32768.0 | 32767.0 | o |
| 29 | rr [24] | EOG, Respiratory rate, taken from PFLOW (cannula) or CFLOW (CPAP mask). [24] | 1 | -32768.0 | 32767.0 | BPM |
| 30 | chin1 | EMG | 200 | -3276.8 | 3276.7 | uV |
| 31 | lleg2 | EMG | 200 | -3276.8 | 3276.7 | uV |
| 32 | rleg3 | EMG | 200 | -3276.8 | 3276.7 | uV |
| 33 | arms4 | EMG | 200 | -2621440.0<br>-3276.8 | 2621360.0<br>3276.7 | uV |
| 34 | larm4 | EMG | 200 | -2621440.0 | 2621360.0 | uV |
| 35 | rarm5 | EMG | 200 | -2621440.0 | 2621360.0 | uV |
| | | | | | | |
| 36 | ekg8 | ECG | 200 | -2.62144<br>-6553.6 | 2.62136<br>6553.4 | V |
| 37 | heartrat-e | ECG | 2<br>3 | -3276.8<br>-3276.8 | 3276.7<br>3276.7 | bpm |
| 38 | psnore | Respiratory response | 200 | -8192.0 | 8191.75 | mbar |

| | | | | | | |
|---|---|---|---|---|---|---|
| 39 | cflow | Respiratory response, CPAP mask | 100 | -162985.0<br>-3276.8<br>-6553600.0<br>-162985.0<br>-6553600.0 | 162980.4<br>3276.7<br>6553400.0<br>162980.3<br>6553400.0 | L/m |
| 40 | flow5 | Respiratory response, Temperature gradient from a thermistor or thermocouple is placed under the nose. It has both nasal prongs and an oral prong to measure nasal/oral flow. [24] | 10<br>200 | -2621440.0<br>-32768.0 | 2621360.0<br>32767.0 | uV |
| 41 | chest | Respiratory response | 10 | -32768.0 | 32767.0 | mV |
| 42 | chest6 | Respiratory response | 200 | -32768.0<br>-2.62144<br>-2621440.0 | 32767.0<br>2.62136<br>2621360.0 | uV |
| 43 | abdm | Respiratory response | 10 | -32768.0 | 32767.0 | mV |
| 44 | abdm7 | Respiratory response | 200 | -32768.0 | 32767.0 | uV |
| 45 | snore | Respiratory response | 200 | -3276.8 | 3276.7 | uV |
| 46 | snore4 | Respiratory response | 200 | -32768.0 | 32767.0 | uV |
| 47 | snore5 | Respiratory response | 200 | -32768.0 | 32767.0 | uV |
| 48 | snoring sensor | Respiratory response | 10 | -32768.0 | 32767.0 | uV |
| 49 | cpress | Respiratory response | 10<br><br><br><br><br><br><br>100 | -529.83<br>-442.817<br>-16.384<br>-806.694<br>-496.351<br>-445.585<br>-445.195<br>-196.787<br>-177.443 | 529.8145<br>442.8041<br>16.3835<br>806.6696<br>496.3365<br>445.572<br>445.1814<br>196.7817<br>177.4383 | cmH2O |
| 50 | sao2 | Oxygen | 2 | -3276.8 | 3276.7 | % |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | saturation | 3 | -3276.8 | 3276.7 | |
| 51 | bodyp-osition | Patient unit strapped around the mid section of the patient, between the chest and abdomen belts. [24] | 10 | -3276.8 | 3276.7 | o |
| 52 | co2 [24] | One cannula under the nose. co2 is a waveform. [24] | 100 | -6.5536 | 6.5534 | V |
| 53 | etco2 [24] | Same cannula as of co2. etco2 is numerical value. [24] | 10<br>100 | -327680.0<br>-675.353<br>-6.5536 | 327670.0<br>675.333<br>6.5534 | mmHg<br>V<br>V |
| 54 | gravity x | | 10 | -32.768 | 32.767 | g |
| 55 | gravity y | | 10 | -32.768 | 32.767 | g |
| 56 | leak [24] | Calculated volume leak from CPAP machine/mask interface. [24] | 10 | -792.901 | 792.8774 | L/m |
| 57 | newch--annel | | 200 | -0.00327 | 0.003277 | V |
| 58 | pflow | | 200 | -8192.0 | 8191.75 | mbar |
| 59 | plesmo | | 75<br>100 | -3276.8<br>-3276.8 | 3276.7<br>3276.7 | V |
| 60 | pulse | Pulse rate | 10<br>100 | -6.5536<br>-6.5536 | 6.5534<br>6.5534 | BPM |
| 61 | refx1 | | 200 | -3276.8 | 3276.7 | uV |
| 62 | ribs | | 200 | -3276.8 | 3276.7 | uV |
| 63 | spo2 | Back-up oximeter in the ETCO2 machine. Probe goes on the finger when used as a back-up monitor. [24] | 2<br>10 | -3276.8<br>-1665.35 | 3276.7<br>1665.308 | % |
| 64 | thermi--stor | | 10 | -32768.0 | 32767.0 | uV |

| | | | | | | |
|---|---|---|---|---|---|---|
| 65 | vt | Inspiratory volume from the CPAP machine/mask interface. Actual volume. [24] | 100 | -33037.5 | 33036.54 | ml |

**Note:** Digital Minimum = -32768 and Digital Maximum = 32767 for all the signals

* Except for signal 't6', Digital min = -36, Digital max = 0.

Some signals in the above table have more than one row because their physical extremes and sampling rates can vary from one study to the other. For one particular study however, a signal has one physical extreme and sampling rate only.

### *5.4.4 Signal Values*

The signal header information is followed by data records. A record contains digital values of all the signals recorded with their respective sampling rates and order. Each value occupies 2- bytes and is represented in two's complement, little-endian order (least significant byte first). The duration of a data record can be determined from the global header. In our research the duration of data record is always 1 second. The figure below gives an idea about the structure of this data:

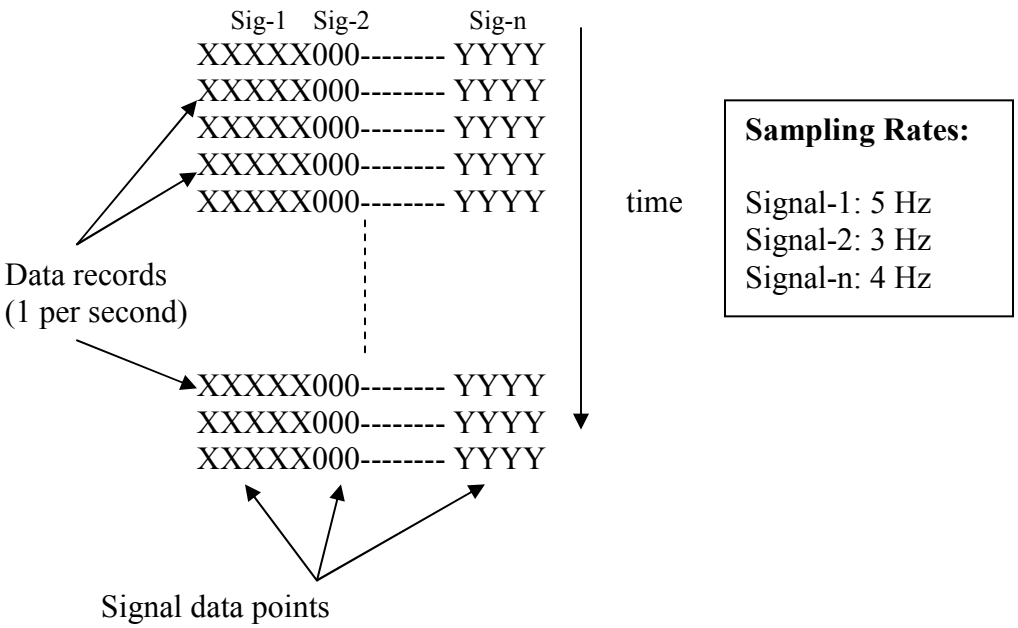



**Figure 5.6 EDF structure**

Figure below shows a screenshot of an EDF file when opened in a text editor:

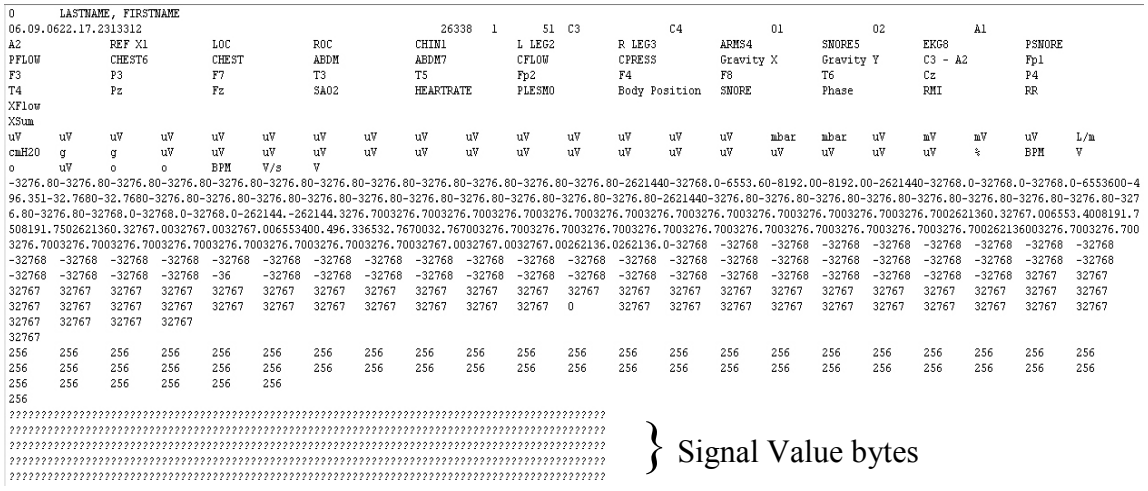



**Figure 5.7 Screenshot of an EDF file**

## *5.5 Micro database design*

The micro data consists of sleep studies, each recorded in a file in EDF format. The EDF file is further composed of a global header, signal headers and the actual signal data as described in Section 5.4 EDF Description). We need to model this information into relational tables in order to store the data in a database.

We start with creating a schema called "micro" that will store the header and signal data. Next, we create a table to model the global header information, with each tuple identified by the study number:

**Table 5.2 micro.header**

| No. | Attribute description | Database attribute name | PostgreSQL Type |
|---|---|---|---|
| 1 | Study Identifier | `sid <PK, FK>` | integer |
| 2 | Start Date of study | `start_date` | date |
| 3 | Start Time of study | `start_time` | time |
| 4 | Number of records | `number_records` | integer |
| 5 | Duration of record | `duration_record` | smallint |
| 6 | Number of signals | `number_signals` | smallint |
| 7 | List of signal names recorded in the study | `signals` | text[] |

The primary key (`PK`) Study ID attribute, as a foreign key (`FK`), refers to the `sid` attribute in the Patient-Study Map table (see Section 4.6.2 Patient-Micro study map table).

To store the signal data, one table is modeled after each type of signal. Thus, there are 65 tables to store data for all the signals (see Section 5.4.3 Signal Properties Table). The table name is same as that of the name of the signal. Since PostgreSQL supports a

variable length array data type, we can store time-series signal information in the database in a single tuple for every patient, indexed by a study identifier. Also, as observed from the Study Properties table, the Physical Minimum, Physical Maximum, and the sampling rates of a given signal may change for all the studies. Hence, information corresponding to the signal's properties is also added to each table.

**Table 5.3 A typical signal table**

| No. | Attribute description | Database attribute name | PostgreSQL Type |
|---|---|---|---|
| 1 | Study identifier | `sid <PK, FK>` | integer |
| 2 | Time-series information (signal values) | `digital_values` | smallint[] |
| 3 | Sampling rate | `sampling_rate` | smallint |
| 4 | Physical minimum | `phy_min` | real |
| 5 | Physical maximum | `phy_max` | real |
| 6 | Physical dimension (unit) | `unit` | text |

The Primary Key (`PK`) Study ID attribute, as a foreign key (`FK`), refers to the `sid` attribute in the Patient-Study Map table. (see Section 4.6.2 Patient-Micro study map table).

The figure below gives the schema as it exists in the database for micro data.

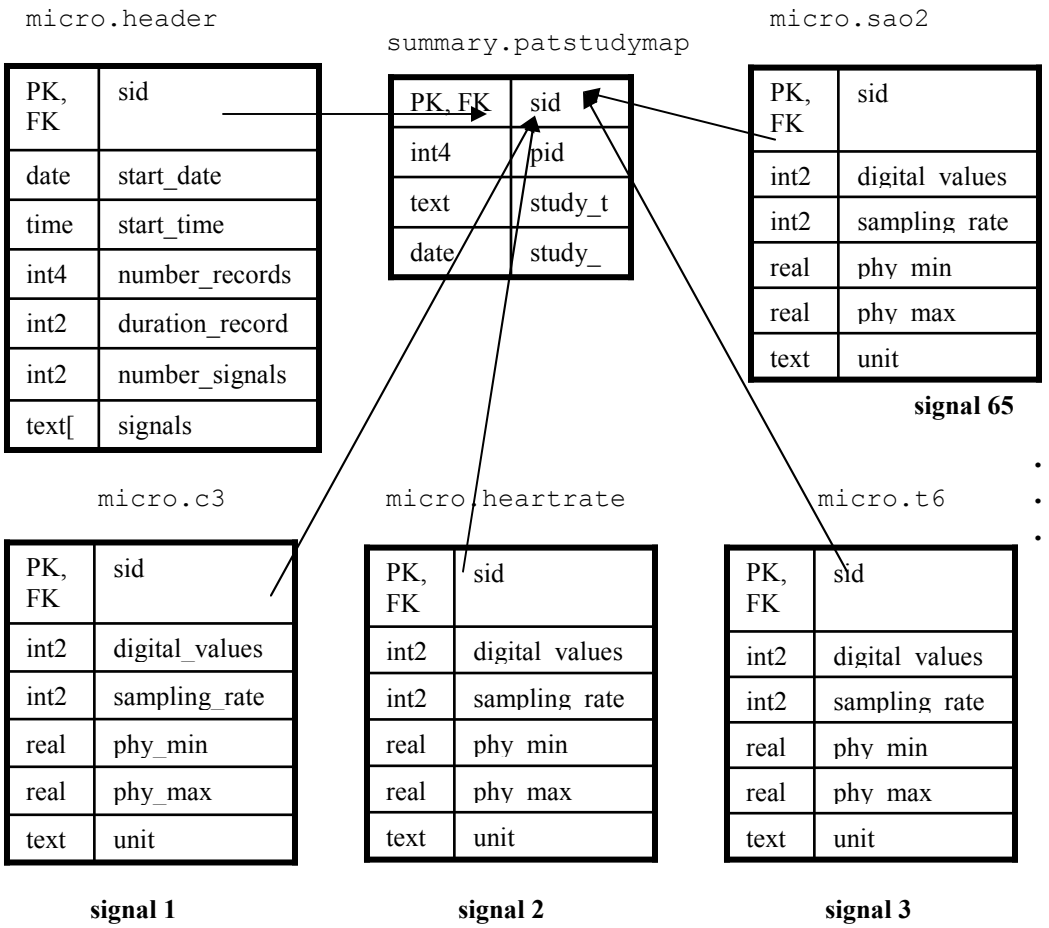



**Figure 5.8 Micro data schema**

## *5.6 Building the micro database*

### *5.6.1 Schema Management*

To create the micro database schema and the required relational tables, we wrote a Java application that establishes a database connection, and executes data definition language (DDL) SQL queries. The tables (their attributes and types, and the constraints) are defined within the *DbObjectManager* class.

The *main()* routine within this class then creates an instance of type *DbObjectManager* that sets up a connection to the `sleepdb` (see Chapter 6 System) database. The user then

can execute *DbObjectManager*'s operation of creating the `micro` schema and the objects within it.

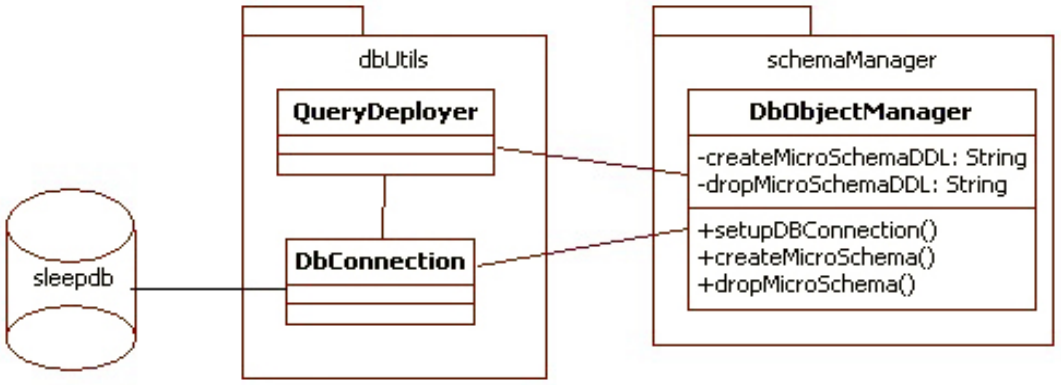



**Figure 5.9 Class diagram for Schema management**

### *5.6.2 Building the database*

**(a) Implementation**

To extract signals from the EDF and put them in the database, we wrote a standalone application "EDF2DB" that reads the EDF files, one at a time and, for each EDF, with the help of properties specified in the global and signal header (signal name, sampling rate and duration of a data record), extracts the signal values and uploads them to the database using the Java Database Connectivity (JDBC) API.

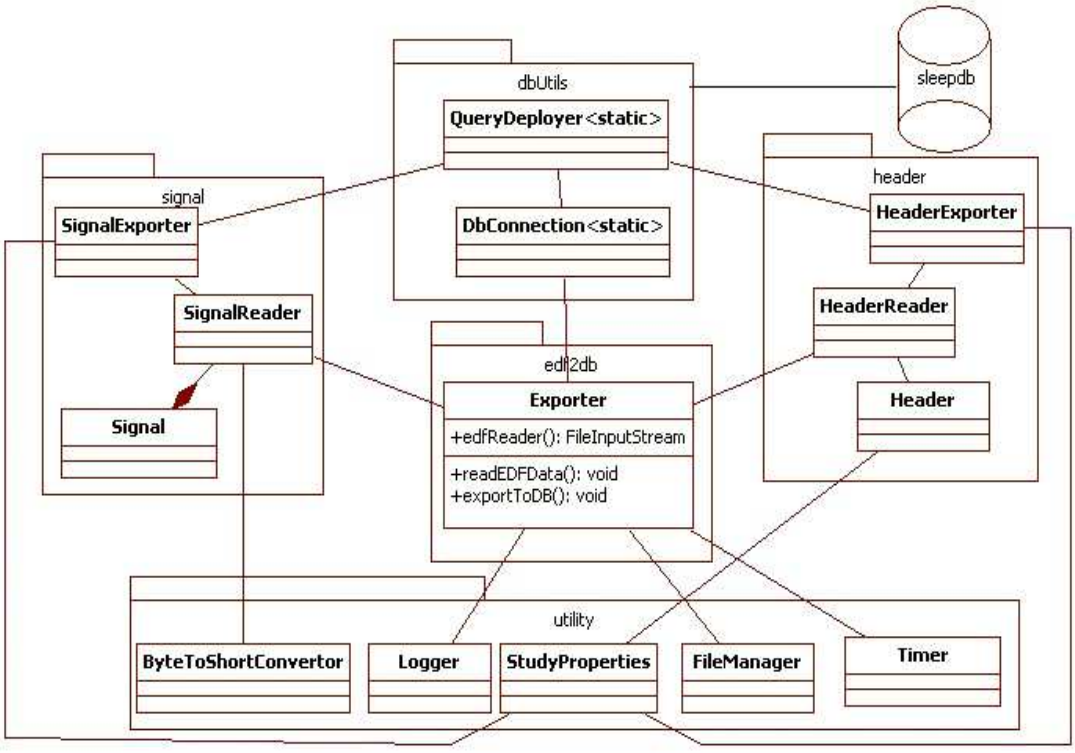



**Figure 5.10 Class Diagram of *edf2db***

This application creates a connection to the sleep database (*sleepdb,* see above Figure 5.10 Class Diagram of *edf2db)* and runs on a collection of EDF files in a directory. As per our file organization scheme, there are 100 patient cases in each folder on the storage device, which amounts to approximately 130 studies. The *main()* routine within *Exporter* fetches this collection of filenames with the help of *utility.FileManager*. For each EDF file, *main()* creates an instance of *Exporter* class. The instance sets up an input read stream to the EDF study and sets the study and patient identifiers and the study type using *utility.StudyProperties* (see Section 5.3 EDF Naming Scheme). The *Exporter* type object then uses the instance of the class *HeaderReader* to read and translate header bytes and set the *Header* object. Subsequently, it uses the class *SignalReader,* which creates a collection of *Signal* objects and sets their values. Each signal object is composed of the signal's properties and its sequence for the entire length of the study. All the contents of the EDF file are read into objects in the system's memory. The *Header* is then exported to

the database using *HeaderExporter* that calls on d*bUtil.QueryDeployer* helper class to execute an insert query on the *micro.header* table. Similarly, the signal properties and sequences are also exported to the *micro.* <*signalname*> tables.

**(b) The little, big Endian distinction**

The signal value bytes read from the EDF are present in little-endian order while Java's virtual machine uses the big-endian format (most significant byte first). Thus, we use the *utility.ByteToShortConverter* class to convert between the byte orderings and also translate bytes to a short integer value (see Appendix C Little endian and big endian distinction) [25]

**(c) Memory issues**

Since EDF2DB deals with large input file sizes, in order to prevent the Java's Virtual Machine running out of available memory, the following steps were taken:

1. Java Virtual Machine is started with 900 MB memory heap size.

2. As the exports of the signals to the database progresses, we remove signal objects containing sequences that have been uploaded. This way, as more data gets uploaded, more memory becomes available.

3. Java's garbage collector is forced to execute at strategic points during the runtime of the program to free up memory (as per the memory profiling statistics - forcing garbage collector to sweep away de-referenced objects increased available memory immediately, a phenomenon that was not observed when only the Java's virtual machine was in control of the garbage collection).

4. We split up the data upload size of every signal into half (hence first half of a signal was uploaded by an insert query and the latter half by an update query). This reduced the

object size representing the data (the object here being comma separated string of signal values), hence occupied less main memory, with little or no effect on the data upload time. The object was immediately de-referenced when its use was over.

### *5.6.3 Testing the signal data*

Since the magnitude of time and storage needed by the micro database is very high, it was necessary to test the correctness of signal values exported to the database before the actual data transfer began.

In order to test the signal values in the database, we wrote an application that compares the signal values generated from a third-party software like EDF2ASCII [4][4] (test file) with the corresponding signal in the database.

The EDF2ASCII application is capable of extracting one signal at a time from an EDF study to an ASCII format, with user choosing the delimiters types (commas, CRLF, spaces or tabs) between signal values. Given a study, we extract a signal, for example, `heartrate` at its natural frequecy of 2 Hz to a file called `heartrate.ascii`. This is our test file.

For the same study, we determine the study id in the database. The user then specifies the path of test file, name of the signal ("heartrate"), sampling rate (2 Hz) and the study identifier to the testing application. The application then fetches the target signal sequence from the database into a `database_seq` array, reads the values in the `heartrate.ascii` file into a `test_seq` array compares the two arrays.

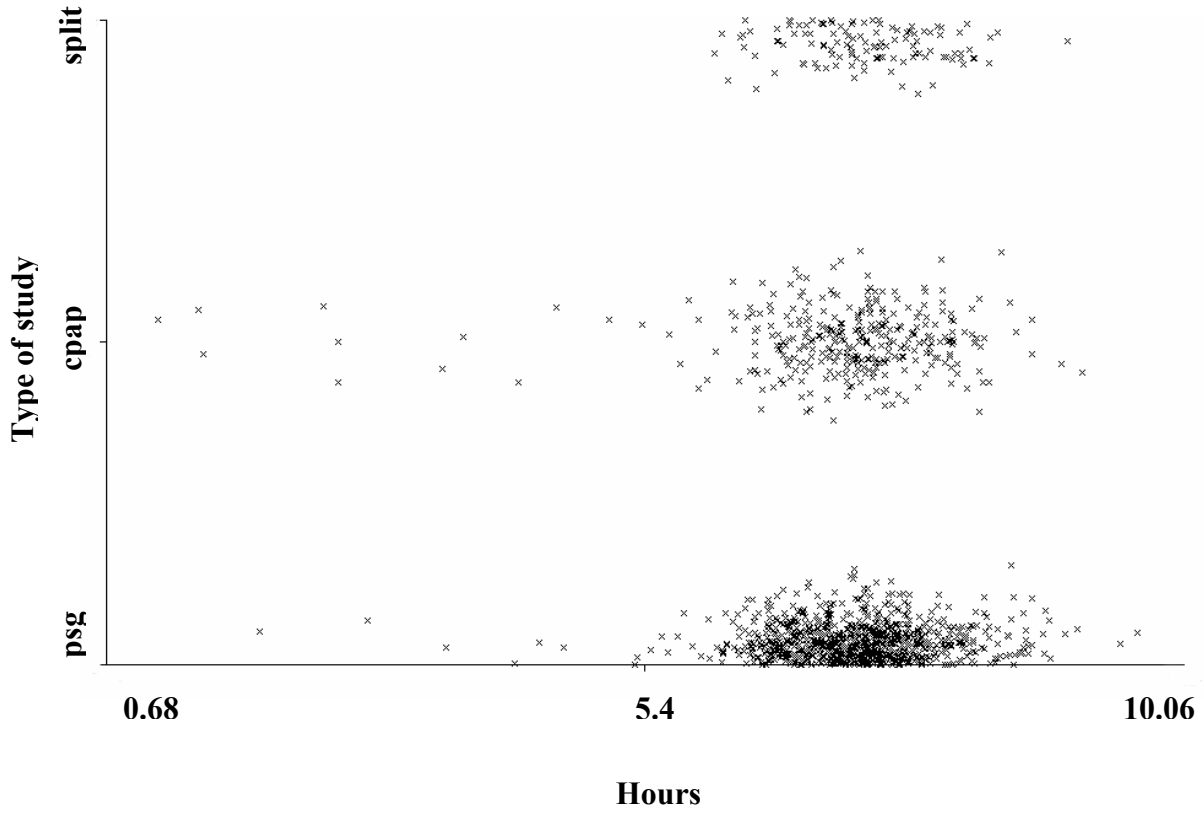
## *5.7 Statistics on micro data*

We have a total of 1319 studies for 1046 patients. There are approximately 130 studies in a folder of 100 patient cases. The upload time for a study into the database on an

average is about 10 minutes, thus for 130 studies, ~ 22 hours. The average size of a study is 360 MB.

### *5.7.1 Study distribution with respect to the length of the study*

The following graph shows the distribution of summary, cpap and split studies with respect to the duration of the study. Maximum number of instances belong to the PSG type of study, with durations from 6 - 8 hours.




(Above graph was visualized in Weka 3.5.4 system with a high degree of jitter, See Appendix E Interfacing Weka 3.5.4 to PostgreSQL 8.2.5 database)

**Figure 5.11 Study type distribution with respect to study duration**

### *5.8 Database system performance*

The database system showed impressive scalability to a massive amount of micro data that was uploaded in a period of approximately 12 days. In order to benchmark how the database performance scaled with retrieval of time-series data (the largest type of data that we have), we wrote a Java routine that pulls time-series data from random studies from the table for signal `c3` (See Table 5.1 Signal properties) that is sampled at 200Hz (Note: signal `c3` is recorded in all the studies).

The following graph shows how the database responded to incremental requests to fetch the sequence data from 25 studies:

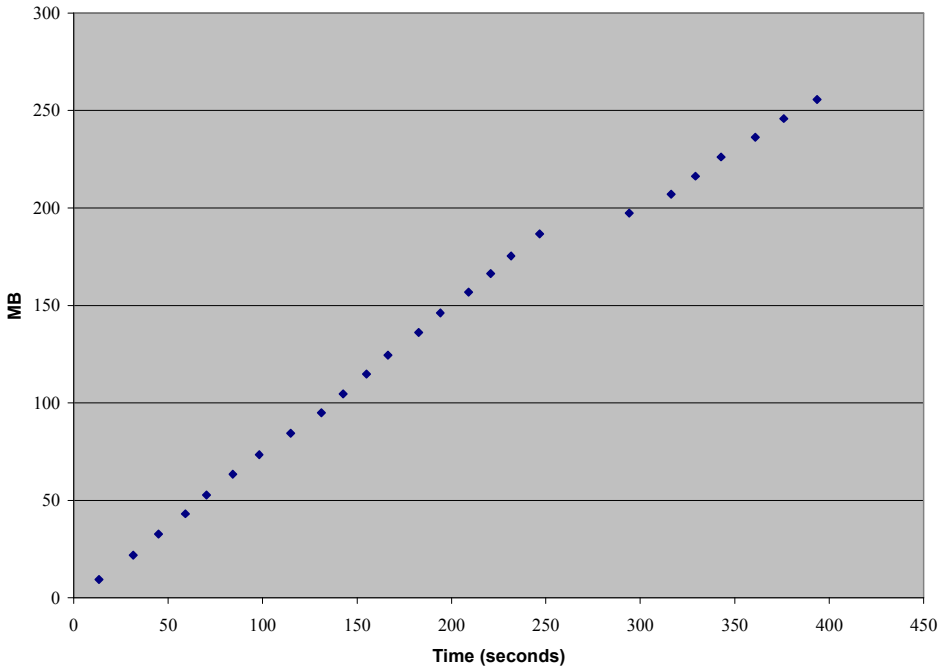



**Figure 5.12 Retrieving time-series data from the database**

As it can be seen from the above graph, the response is linear with respect to time. The slight gap in the dot plot that is seen between 250 to 300 seconds is due to Java's garbage collector clearing up the heap to make way for more data.

The PostgreSQL 8.2 DBMS supports most of the major features in SQL: 2003 standard [11]. Thus, the most common types of queries with selects, sub-selects, joins, set operations, aggregates, and database, schema, table, row and column level manipulations can be executed against the database.

The database performance could be affected by the type of query issued by the user. SQL, being a declarative language, only specifies the data that is needed as a result, not how the query itself will be executed by the database kernel. Finding an optimal way (plan) from different possibilities (plan search space) to execute the query is the responsibility of the DBMS's planner component. The search space widens as the queries become more complex. An example of a good decision made by the planner could be that it chose a hash-based join over nested-loop join when joining a large table with a relatively smaller one to give better query performance in terms of time. PostgreSQL has tunable system parameters that can affect query performance to a good extent when specific queries run slow. The plan tree selected by PostgreSQL can be visualized in the standard output and studied in detail by writing the `EXPLAIN VERBOSE` command just before the query expression. The user of the database should always keep the database statistics updated so that the planner can choose the best amongst the better possible plans for completing the query. Furthermore, query performance can benefit from the use of index objects.

# 6 System

## *6.1 Context*

Since a rich part of this thesis deals with collection, organization and movement of enormous amount of data into the database, it became inevitable for us to create a resourceful system that can act a backbone to help us attain our objectives.

## *6.2 Micro Data Collection*

During the data collection phase, we used to get batch of around 100 patient files from Day Kimball Hospital every week. The CD data corresponding to the sleep studies were downloaded to a central data repository `rous.wpi.edu`, arranged by the WPI-CS system administrator. To expedite the data download capacity, we used to keep several CCC machines (10- 15 machines, when not in use by students) busy with the job of downloading the raw micro data.

The data on `rous-server` was later transferred to our own 750 GB external storage device `kddrg-hdd-raw` in the laboratory. For the sake of file organization, the folders for each study were named as per the folder naming method described in Section 5.3 EDF Naming Scheme. Each parent folder corresponding to the data download session contained sleep studies for 100 patients.

## *6.3 System Architecture*

Before starting to build the database, we had a couple of issues to deal with - one was to have a storage space for at least 500 GB of data, and the other was the actual medium to transfer the data to the database. If the data was to be transferred from `kddrg-hdd-raw` to a remote machine hosting the database over the network, then this would imply a

large usage of network bandwidth. The most feasible solution was to create our own server, install `kddrg-hdd-raw` and the new storage device `kddrg-hdd-database` that will host the database `sleepdb` in the server, setup the PostgreSQL DBMS and the Java environment required running the micro data export application. This new server was named `kddrg`.

**About the `kddrg` server:**

The `kddrg` server runs on OpenSuse 10.3 Linux Operating system. There are two 750 GB drives installed in this system (`kddrg-hdd-raw` and `kddrg-hdd-database`), along with the machine's own 60 GB storage. The server runs on a 3 GHz Pentium 4 CPU with 1 GB of available memory. The machine is network accessible and the storage devices can be network mapped on a Windows XP operating system, however, the access has been restricted to WPI campus network due to data security reasons. Since `kddrg` is essentially a Unix machine, to run a Windows based utility like "REMbrandt to EDF data conversion" (see Section 5.2.1 EDF Extraction), an application called VMWare server was installed on `kddrg`. The VMWare server simulates machine hardware so that an operating system can be installed on it. Windows XP was installed as an operating system running on this simulated machine.

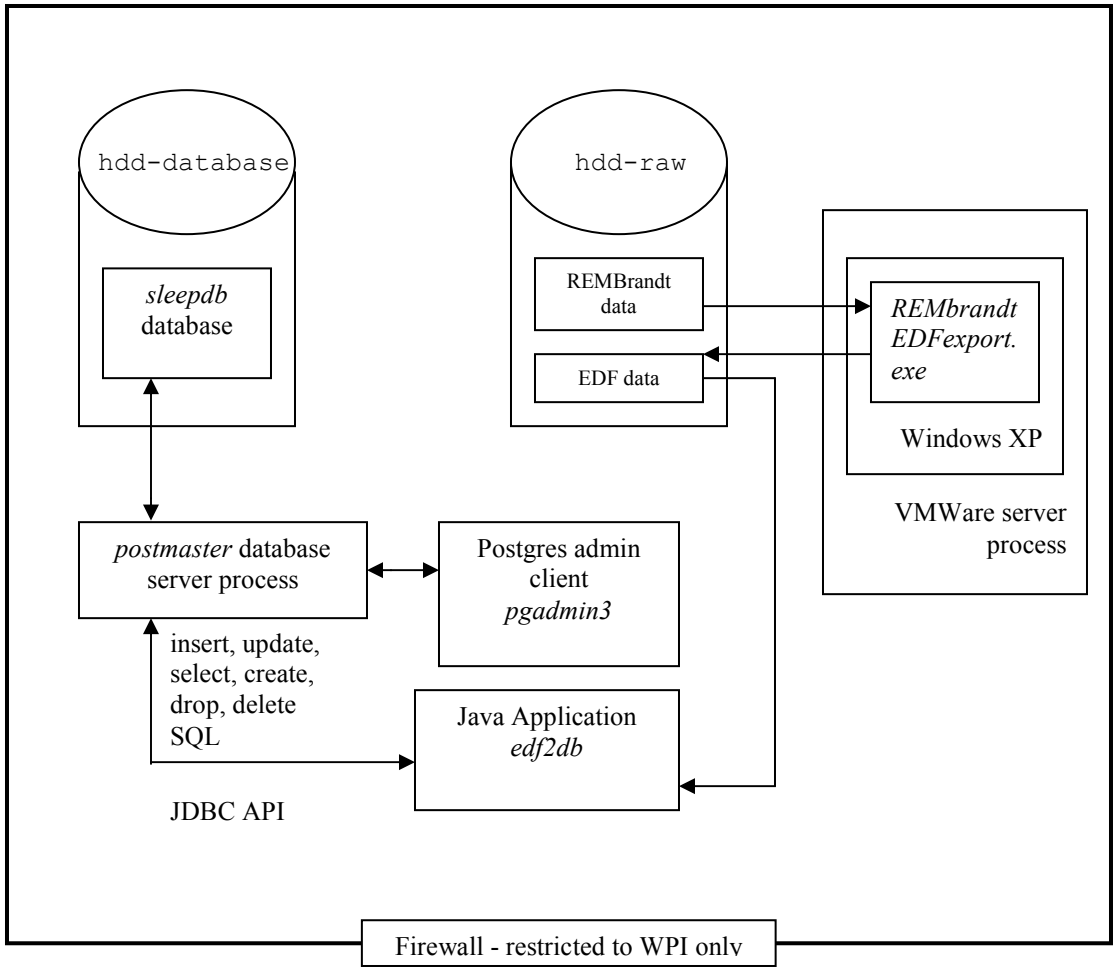



**Figure 6.1 The kddrg system architecture**

# 7 Conclusions and future work

The project aimed at designing, building, and analyzing a terabyte-scale database comprising complex data including patient questionnaire responses, their medical history, time-series data and clinical summary for 1046 patient cases. We successfully created a data repository of 1319 sleep studies of 563 GB in size that will provide a high quality dataset and one of the world's largest and most comprehensive for performing in depth computational analyses on human sleep in the future. We gathered a few exploratory facts that were presented in graphs.

## *7.1 How does the database design facilitate data analysis?*

For the analyst to get a clear understanding of the sleep database in order to perform the analysis on it, the database design features an intuitive organization of data at the schema, table and attribute levels. It grasps the relation between the different types of data we deal with, ensures that native or custom data types have been chosen to best represent the clinical information, and that they have been named for readability to benefit the analyst. A table is made to represent a single entity only.

At the schema level, there exist macro, summary and micro schemas that contain tables corresponding to the 3 distinct types of datasets we described in Chapters 3, 4 and 5 respectively. A link table called `patientstudymap` that relates patients to their studies using identifiers can help the analyst in fetching survey, technical or time-series recording data (or meta-data, like study or signal properties) for any number of patients (restricted only by the available memory in the system) on the fly. The analyst can filter out patients based on an interesting demographic feature of interest, for example, a particular disease, and then retrieve their over night studies for in-depth study.

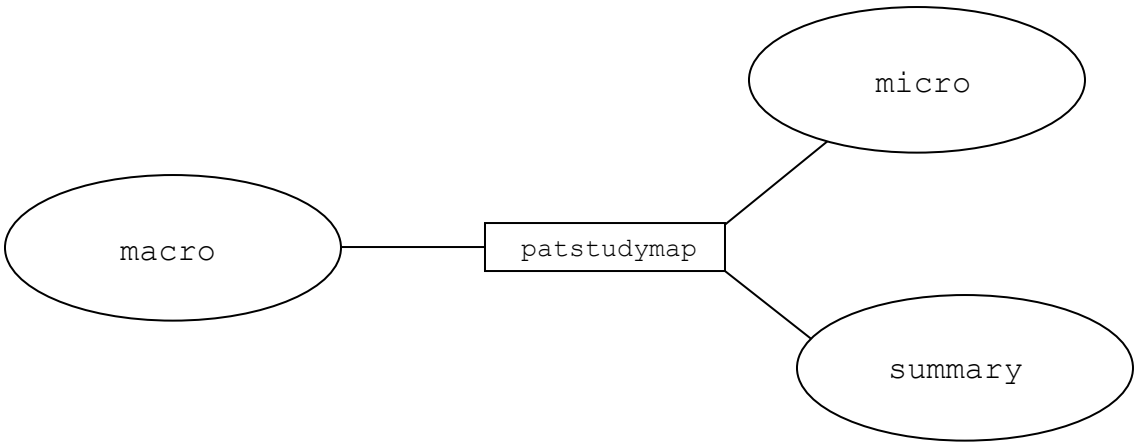



**Figure 7.1 Schema organization in the Sleep database**

## *7.2 Alternate design scenarios*

The information that was encountered in this research could be represented using different data types within PostgreSQL DBMS. Most of the design decisions in this project were influenced by the features offered by the PostgreSQL database system that help representing the information at hand, and greatly reduce the complexity involved in retrieving data of interest for its analysis. Below we discuss alternate designs that were not adopted due to analysis considerations.

### *7.2.1 Using Binary large object (`BLOB`) data type to store time-series sequence data*

If `BLOB`s are used to store the signal data, then data needs to be converted from binary to the short-integer type before it can be analyzed. There will be an extra time and space overhead during this conversion, and it will need to be done every time the analyst wants to look at the time-series recording a patient. Furthermore, based on the system's hardware architecture, there may be little endian, big endian issues involved (See Section 5.6.2 (b)) in the conversion process.

Choosing short integer array type (supported by PostgreSQL DBMS) to represent time-series data as a part of design greatly simplifies the task of storing and retrieving time-series data without any time, or storage overhead involved. The analyst also has the facility to index the array and look at any time-window of interest, thanks to the advanced features of the PostgreSQL database.

### *7.2.2 Using Character large object (`CLOB`) data type to store set based data*

Information like patient medical history is a list of diseases gathered from the patient's medical records. A `CLOB` data type can be used to store this long string type data, with commas separating the diseases' names. Usage of `CLOBs` can get inconvenient from the analysis point of view. For instance, in a scenario where the analyst wants to find the patients with a particular disease, like "type 2 diabetes", he/she has to pull the `CLOB` data for a patient, parse strings to make a list of diseases and then search this list for "type 2 diabetes" and repeat this process for every patient in the database, while keeping a track of patients who met the search criteria. Instead of `CLOBs`, we chose string type arrays to represent set-based data during the database design phase. In our example, each element of the array stores a string that is the name of a disease. PostgreSQL provides a feature to search for any element within an array structure, and we can make our search span across the entire length of the table through a simple query like:

```
    select pid from macro.demographics where 'type 2
diabetes'=ANY(medical_hx)
```

(See Section 3.2.1 Patient demographics, for descriptions of attributes in the query above).

### *7.2.3 Creating tables for every patient to store sequence data*

In this design approach, there is one table for every patient in the micro schema. The first attribute is the signal's name that is of type `text`, and the second attribute is the time series information for a given signal, of type `smallint[]`. This is an inconvenient design choice, because this would result in too many tables (equivalent to the number of studies) to manage in the schema. Moreover, the queries to retrieve signal data for an interesting group of patients would involve large number of joins (which is an expensive database operation by itself) that will be cumbersome to write or interpret. Thus, this design choice was dropped in favor of having one table dedicated to every signal type. This helps the analyst to retrieve sequence data of any signal for any number of patients (bounded by the systems memory limit).

To simplify things further, we added an attribute called `signals` (See Section 5.5 Micro database design) of type `text[]` to the header table in the micro schema. This attribute contains a list of signals that were recorded for every study. Since there are 65 different signals known to us in this research, and a typical study has around 50 -55 signals, it would become inconvenient to examine every signal table to locate whether if a signal was recorded during a particular study, or not. This problem is now solved as the analyst can quickly look into the signals list associated with every study, to find out which signals were recorded during the study.

### *7.2.4 Time series data in flat files*

The time-series data can also exist in the flat files. The drawback of this approach is that it would really miss out on of the facilities of the database to store data in an organized manner and keep its integrity. Lots of flat files can become very difficult to keep. If we decide to store all the signals in a single file (See Appendix D, ASCII - Intermediate file format), there is considerable time spent in performing I/O routines to parse and read the signal data from the files. On the other hand, if we decide to store

every signal in a separate file for a given study, then there would be 50-55 different files for every study. For 1319 studies, the number of files will become almost impossible to manage.

There could be two approaches as to how we want to store signal data in the flat file - in binary format or in ASCII text. If the signal data is in binary format, there is an overhead of converting it to short integers, as described (See Section 7.2.1) above. If they are stored in ASCII format, then the storage costs increase manifold, as each character occupies 1 byte of physical memory. This approach is not recommended if the data needs to be used for analysis purposes.

### *7.3 Summary of data*

**Table 7.1 Summary of the data statistics**

| | |
|---|---|
| Number of patients | 1046 |
| Number of clinical studies | 1319 |
| Number of PSG studies | 875 |
| Number of CPAP studies | 327 |
| Number of SPLIT studies | 117 |
| Number of schemas in the database | 3 |
| Number of tables in macro schema | 4 |
| Number of tables in summary schema | 11 |
| Number of tables in micro schema | 67 |
| Size of macro data | 1064 KB |
| Size of summary schema | 2648 KB |
| Total size of micro schema | 563 GB |
| Total size of raw (REMBrandt [13] format) data | 489 GB |

## *7.4 Future work*

To perform analysis on the sleep data in the future, we have interfaced a well known data mining tool - Weka with the Sleep database hosted by PostgreSQL DBMS. See Appendix E, that documents the Interfacing Weka 3.5.4 to PostgreSQL 8.2.5 database to fetch the dataset of interest using the power of Structured Query Language (SQL).

There are existing modifications of the Weka toolkit that enable it to mine data in relational tables. In WekaDB [28], Weka's functionality was extended to support data mining on relational database systems. There is another extension of Weka that can work with relational databases - Relational WEKA [26]. WPI-WEKA [27] [6] [15] has support for set and sequence types for data mining. All these features can help us perform data analysis on the *sleepdb*.

# References

[1] A. T. Beck, R. A. and G.K. Brown, "Manual for the Beck Depression Inventory-II". San Antonio, TX: Psychological Corporation, 1996.

[2] J.J. Harrington and Teofilo Lee-Chiong Jr. "Sleep and Older Patients". Clin Chest 28, pp 673-684, 2007.

[3] B. Kemp, A. Varri, A.C. Rosa, K.D. Nielson, J. Gade, "A simple format for exchange of digitized polygraphic recordings", *Electroencephalography and Clinical Neurophysiology*, Vol.82, p.391S, 1992.

[4] EDF2ASCII by B. Kemp and M. Roessen. Exports one of the signals in an ASCII file and all EDF header information, including calibration, about this signal in an additional textfile. 2004.
Web resource: http://www.edfplus.info/downloads/downloads.html

[5] Y. Ichimaru, G.B. Moody. Development of the polysomnographic database on CD-ROM. Psychiatry and Clinical Neurosciences 53:175-177 (April 1999). MIT-BIH Polysomnographic Database. PhysioNet. MIT. Cambridge, MA.

[6] Keith A. Pray. "Mining Association Rules from Time Sequence Attributes". MS Thesis. Department of Computer Science, Worcester Polytechnic Institute. May 2004.

[7] P. Laxminarayan. "Exploratory Analysis of Human Sleep Data". MS Thesis. Department of Computer Science, Worcester Polytechnic Institute. Jan. 2003.

[8] P. Laxminarayan, S.A. Alvarez, C. Ruiz, M. Moonis. "Mining Statistically Significant Associations for Exploratory Analysis of Human Sleep Data". *IEEE Transactions on Information Technology in Biomedicine (TITB)*. Vol. 10, No. 3, pp. 440-450, July 2006.

[9] J.W. Murray, "A new method for measuring daytime sleepiness: the Epworth Sleepiness Scale". Sleep, 1991

[10] PostgreSQL DBMS. "About PostgreSQL"

Web resource: http://www.PostgreSQL.org/about/

[11] PostgreSQL 8.3 Documentation.

Schema.
Web resource: http://www.postgresql.org/docs/current/static/ddl-schemas.html
Appendix D. SQL Conformance.
Web resource: http://www.postgresql.org/docs/8.2/static/features-sql-standard.html

[12] A. Rechtschaffen & A. Kales, "A Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages of Human Subject", US Government Printing Office, National Institute of Health Publication, Washington DC, 1968.

[13] REMbrandt Windows® based sleep monitoring and analysis system for both clinical and research applications.

Web resource: http://www.embla.com/products/analysis/rembrandt/index.asp

[14] R. Rozensky. "Doing it right. Proper electrode measurement and placement will greatly improve the quality of the polysomnogram". ADVANCE for Sleep.

Web resource: http://sleep-medicine.advanceweb.com/Editorial/Content/Editorial.aspx?CC=33781

[15] Christopher Shoemaker. "Mining Association Rules over Set-Valued Data". MS Thesis. Department of Computer Science, Worcester Polytechnic Institute. May 2001.

[16] A. Silberschatz, H.F. Korth, S. Sudarshan, "Database System Concepts", 4th Ed, McGraw-Hill, 2001.

[17] B. Kemp, The Sleep-EDF Database. Sleep Recordings and Hypnograms in European Data Format (EDF). Sleep Centre, MCH-Westeinde Hospital, Den Haag, The Netherlands. PhysioNet. MIT. Cambridge, MA.
Web resource: http://www.physionet.org/physiobank/database/sleep-edf/

[18] The Sleep Heart Health Study Polysomnography Database. Data contribution by cohort members of Atherosclerosis Risk in Communities Study (ARIC), the Cardiovascular Health Study (CHS), the Framingham Heart Study (FHS), the Cornell/Mt. Sinai Worksite and Hypertension Studies, the Strong Heart Study (SHS), the Tucson Epidemiologic Study of Airways Obstructive Diseases (TES) and the Tucson Health and Environment Study (H&E). PhysioNet. MIT. Cambridge, MA.
Web resource: http://www.physionet.org/pn3/shhpsgdb/

[19] S. Vij and A. Gentili. "Sleep Disorder, Geriatric". emedicine from WebMD.
Web resource: http://www.emedicine.com/med/topic3179.htm

[20] W. McNicholas, L. Doherty, S. Ryan, J. Garvey, P. Boyle, E. Chua, St. Vincent's University Hospital / University College Dublin Sleep Apnea Database. St. Vincent's University Hospital Sleep Disorders Clinic. PhysioNet. MIT. Cambridge, MA.
Web resource: http://www.physionet.org/pn3/ucddb/

[21] The Sleep Heart Health Study. Coordinating Center - University of Washington, the Sleep Reading Center - Case-Western Reserve University, the Project Officer - National Heart, Lung, and Blood Institute, and six Investigative Centers (University of Arizona, Boston University, University of California-Davis/University of Pittsburgh, Johns Hopkins University, University of Minnesota, and New York University).
Web resource: http://www.jhucct.com/shhs/details/manual/protocol/default.htm

[22] Classification of Overweight and Obesity by BMI, Waist Circumference, and Associated Disease Risks. National Heart Lung and Blood Institute.
Web resource: http://www.nhlbi.nih.gov/health/public/heart/obesity/lose_wt/bmi_dis.htm

[23] Calculate your body mass index. National Heart Lung and Blood Institute.
Web resource: http://www.nhlbisupport.com/bmi/

[24] Information from Sleep Disorder Center at Day Kimball Hospital, CT.

[25] A.S. Tanenbaum, "Structured Computer Organization", 4th Ed, Prentice Hall, 1999.

[26] A. Woznica. "Relational WEKA" software tool. Web resource:
http://cui.unige.ch/~woznica/rel_weka/

[27] Zachary Stoecker-Sylvia. "Mining for Frequent Events in Time Series". ". MS Thesis. Department of Computer Science, Worcester Polytechnic Institute. Aug 2004.

[28] B. Zoul, X. Mal, B. Kemmel, G. Newton, D. Precup. "Data mining using Relational Database Management Systems". McGill University, Montreal, Canada and National Research Council, Canada. 2006. Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), Singapore, April 2006.

# Appendix

## A. Macro data

### a. List of abbreviations for medical history of patient

| Acronym | Complete form |
|---|---|
| acl | Anterior cruciate ligament |
| add | Attention Deficit Disorder |
| adhd | Attention- Deficit Hyperactivity Disorder |
| adhd nos | Attention- Deficit Hyperactivity Disorder Not Otherwise Specified |
| als | Amyotrophic lateral sclerosis |
| alt | Alanine aminotransferase |
| am | Morning |
| ana | Antinuclear antibody panel |
| aodm | Adult-onset diabetes mellitus |
| ascad | Atherosclerotic coronary artery disease |
| asd | Atrial septal defect |
| ashd | Atherosclerotic heart disease |
| avm | Arteriovenous malformation |
| avn | Avascular necrosis |
| bdi | Beck's depression inventory |
| bipap | Bi-level Positive Airway Pressure |
| bph | Benign prostatic hyperplasia |
| brbpr | Bright Red Blood Per Rectum |
| bso | Bilateral salpingo-oophorectomy |
| C | Centre/central |
| cabg | Coronary Artery Bypass Graft |
| Calc | Calcium |
| cad | Coronary artery disease |
| cdd | Cervical Disc Disease |
| cfs | Chronic fatigue syndrome |
| chd | Coronary heart disease |
| chf | Congestive heart failure |
| cin | Cervical intraepithelial neoplasia |
| cho intolerance | Gestational diabetes mellitus |
| cns | Central nervous system |
| copd | Chronic obstructive pulmonary disease |
| cp | Cerebral palsy |
| cpap | Continuous positive air pressure |
| cpk | Creatine phosphokinase |
| cpr | Cardiopulmonary resuscitation |
| CREST syndrome | Limited scleroderma |
| cta-b | Clear to Auscultation Bilaterally |
| cts | Carpal tunnel syndrome |

| | |
|---|---|
| cva | Cerebrovascular accident |
| dbp | Vitamin D–binding protein |
| dcis | Ductal carcinoma in situ |
| ddd | Degenerative disc disease |
| dexa | Dual energy X-ray absorptiometry |
| djd | Degenerative joint disease |
| dm | Diabetes Mellitus |
| doe | Dyspnea on exertion |
| dvt | Deep vein thrombosis |
| ed | Erectile dysfunction |
| edf | Excessive daytime fatigue |
| eds | Excessive daytime sleepiness |
| egd | Esophagogastroduodenoscopy |
| eomi | Extra Ocular Muscles Intact |
| ess | Epworth Sleepiness scale |
| eswl | Extracorporeal shock wave lithotripsy |
| etd | Eustachian tube dysfunction |
| ett | Endotracheal tube |
| etoh | Alcohol |
| fa | Fall asleep |
| fess | Functional Endoscopic Sinus Surgery |
| ft 1 | Frederickson type 1 |
| fu | Follow up |
| gad | Generalized anxiety disorder |
| gerd | Gastroesophageal reflux disease |
| ha | Headache |
| hcle | Hypercholestrolemia |
| hctz | Hydrochlorothiazide |
| hcvd | Hypertensive cardiovascular disease |
| heent | Head, Eyes, Ears, Nose, and Throat (medical) |
| high blood P | High blood pressure |
| hpv2 | Human Papilloma Virus 2 (wart) |
| htgd | Hypertriglyceridemia |
| htn | Hypertension |
| htz | Hydrochlorothiazide |
| ibs | Irritable bowel syndrome |
| icd | Implantable cardiac defibrillator |
| iddm | Insulin-dependent diabetes mellitus |
| igt | Impaired glucose tolerance |
| ihss | Idiopathic hypertrophic subaortic stenosis |
| irbbb | Incomplete Right Bundle Branch Block |
| jra | Juvenile rheumatoid arthritis |
| L | Left (hand/leg) |
| lbp | low back pain |
| ldl | Low-density lipoprotein test |
| leep | Loop electrosurgical excision procedure |
| lft | Liver Function Test |
| lll | Left lower lobe (of the lung) |
| llq | Left lower quadrant (quarter) |
| lpr | Laryngopharyngeal reflux |

| | |
|---|---|
| lvh | Left Ventricular Hypertrophy |
| mch | Mean cell haemoglobin |
| mcv | Mean cell volume |
| meds | Medications |
| mi | Myocardial infarction |
| ms | Multiple sclerosis |
| mslt | Multiple sleep latency test |
| mva | Motor vehicle accident |
| mvp | Mitral Valve Prolapse |
| mvr | Mitral Valve Regurgitation |
| na | Not available/missing data |
| nad | No abnormalities detected |
| nash | Non-alcoholic steatohepatitis |
| niddm | Non-insulin-dependent diabetes mellitus |
| O2 | Oxygen |
| oa | Osteoarthritis |
| occ | Occasional |
| ocd | Obsessive-Compulsive Disorder |
| odd | Oculodentodigital dysplasia |
| osa | Obstructive Sleep Apnea |
| osa - h | Obstructive Sleep Apnea - Hypopnea Syndrome |
| osas | Obstructive sleep apnea syndrome |
| pap | Papanicolaou test/ Pap smear |
| pbc | Primary billiary cirrhosis |
| pdd | Pervasive developmental disorder |
| pe | Pulmonary embolism |
| perrl | Pupils Equal, Round, Reactive to Light (used by emergency room personnel) |
| pft | Pulmonary function test |
| plma | Periodic leg movement activity (doesn't imply PLM disorder) |
| plmd | Periodic leg movement disorder |
| pls | Primary Lateral Sclerosis |
| pm | Afternoon/Evening |
| pmds | Persistent Mullerian duct syndrome |
| pmr | Polymyalgia rheumatica |
| pms | Premenstrual syndrome |
| pna | Pneumonia |
| pnd | Paroxysmal nocturnal dyspnea |
| pots | Postural orthostatic tachycardia syndrome |
| ppd | Postpartum depression; Purified protein derivative (the PPD skin test for tuberculosis) |
| psa | Prostate-specific antigen; Prostate cancer screening test |
| psp | Progressive supranuclear palsy |
| ptca | Percutanerous Transluminal Coronary Angioplasty |
| ptsd | Post-traumatic stress disorder |
| pud | Peptic ulcer disease |
| pvc | Premature ventricular contraction |
| pvd | Peripheral Vascular Disease |
| pvd2 | Posterior vitreous detachment |
| R | Right (hand/leg) |
| r/o | Rule out |

| | |
|---|---|
| rad | Reactive Airways Disease |
| rap | Recurrent aspiration pneumonia |
| rbbb | Right bundle branch block |
| rbd | REM behavior disorder |
| rca | Right coronary artery |
| rdi | Respiratory disturbance index |
| rhm | Routine Health Maintenance |
| rls | Restless legs syndrome |
| rrr | Regular Rate and Rhythm |
| rsd | Reflex Sympathetic Dystrophy |
| rtc | Return to clinic |
| ruq | Right upper quadrant (quarter) |
| s/p | Status post |
| sa | Stay awake |
| sbe | Subacute bacterial endocarditis; Sporadic bovine encephalomyelitis |
| sdb | Sleep-Disordered Breathing |
| sob | Shortness of breath |
| sss | Sick sinus syndrome |
| svt | Supraventricular tachycardia |
| sx | Symptoms |
| sz | Seizure |
| tah | Total abdominal hysterectomy |
| tah-bso | Total abdominal hysterectomy with bilateral salpingo-oopherectomy |
| tb | Tuberculosis |
| tsh | Thyroid stimulating hormone |
| tia | Transient ischemic attack |
| tkr | Total knee replacement |
| tle | Temporal lobe epilepsy |
| tmj | Temporo-mandibular joint |
| tnt | Toss n turn |
| uad | Upper Airway Disease |
| uao | Upper airway obstruction |
| uaod | Upper airway Obstructive disease |
| uars | Upper Airway Resistance Syndrome |
| up3/uppp | Uvulopalatopharyngoplasty |
| uri | Upper Respiratory Infection |
| wa | Wake up |
| wbc | White blood cells |

## b. List of abbreviations for tracking family's medical history

| Abbreviation | Meaning |
|---|---|
| 2c | second cousin |
| a | aunt |
| b | brother |

| | |
|---|---|
| ch | child |
| d | daughter |
| f | father |
| fam | family |
| ft | fraternal |
| h | husband |
| m | mother |
| ma | maternal aunt |
| mc | maternal cousin |
| mgf | maternal grand father |
| mgm | maternal grand mother |
| n | nephew |
| par | parents |
| pgf | paternal grand father |
| pgm | paternal grandmother |
| pt | paternal |
| pu | paternal uncle |
| rel | relatives |
| s | sister |
| sib | sibling |
| so | son |
| sor | sororal |
| u | uncle |

**Example usage:**

1. Mother has a history of Coronary artery disease and second son suffers from migraine headaches =>

{m-CAD,s2-migraine HA}

2. Family history of high depression => {high depression} (note: no prefix for mention of general family history)

## B. Technical summary report

### a. List of acronyms seen in a summary

TST - Total sleep time

SPT - Sleep period time

TIB - Time in bed

MT - Melatonin

NDX0 - Baseline

RX - Treatment

SO - Sleep onset

Waso - Wake after sleep onset

B - Back/Supine

NS - Non - Supine

NR - Non REM

R - REM

UARS - Upper airway resistance syndrome

TRT - Total REM time

DX - Diagnosis/Baseline

RD - Respiratory disturbance

LM - Limb Movements

PLM - Periodic Limb Movements

RRLM - Respiratory related Limb Movements

TLM - Total Limb Movements

BP - Body Position

**b. Split Type-C attributes for entire duration of study**

Sleep Summary

- Total Recording Time
- Sleep Period Time

Sleep stage Information

- Lights Off
- Lights On
- Wake
- Total Sleep Time
- Wake/SPT
- TST Supine
- TST Non Supine

Saturation Time Table (All attributes)

## C. Little endian and big endian distinction

Data points in EDF are bytes in little-endian order:

Short integer

```
Byte 1    Byte 0
```

Order in memory

```
base address+0   Byte0
base address+1   Byte1
```

Java's internal representation has big-endian order

Order in memory

```
base address+0   Byte1
base address+1   Byte0
```

## D. ASCII - Intermediate file format

During the early stages while looking into the micro data, in order to overcome memory limitations in the system (512 MB) when reading the EDF, we had devised a structure to store the signal data in the ASCII format for as an intermediate format for future reading into the database. The motivation behind the structure was the need to keep all the information in a single file (and not have a file correspond to each signal, as that would increase the number of files to manage) and export the more costly data (in terms of storage) to the database as the file was read, instead of storing the signal information all at once in memory. The largest signal matrix (see below) was made of 200 Hz signals, and was read into the database in 3 block transfers. The signals grouped in lower frequencies (100, 10, 2, 1) could be read all at once in the memory and exported to the database. With the new system (`kddrg` server, see chapter 6) that had 1 GB of memory, the problem of memory limitation was solved, and the EDF was directly read into the database.

**Description of structure:**

For signals having same sampling frequencies; the code builds a rectangular matrix. The 200 Hz signals get written first, followed by 100Hz, 10Hz, and 2 Hz and 1Hz signals. Each matrix is separated by a delimiter specifying the sampling rate for following matrix. The first row of every matrix is the name of signals separated by commas, followed by their digital values. The header content is also present in the data file. ASCII file structure:

```
<<SOF>>
@header
patient_id, start_date, start_time, number_records,
duration_each_record, number_signals

@signals
name, sampling_rate, phy_min, phy_max, digi_min, digi_max,
phy_dimension

@values

@200
names of signals
      matrix [m X n]
%
@100
names of signals
      matrix      [m X n]
%
@10
names of signals
      matrix      [m X n]
%
@2
names of signals
      matrix      [m X n]
%
@1
names of signals
      matrix      [m X n]
%
<<EOF>>
```

**Note:**

Dimension of rectangular matrix = [(number of records * sampling rate) X (number of signals)] = [m X n]

where, m = rows, n = columns

SOF = start of file

EOF = end of file

## E. Interfacing Weka 3.5.4 to PostgreSQL 8.2.5 database

In Eclipse IDE 3.2.2 configured to run with Java 1.5,

1. Create a new Java project and name it: `weka_3.5.4`.
2. Create a `src` folder within this project. This folder will contain the Weka source code.
3. Import `weka-src.jar` into the `src` folder.
4. Add `PostgreSQL JDBC API` jar file to the project build path (we used `PostgreSQL-8.2-504.jdbc3.jar`)
5. After importing Weka's source code within Eclipse's Java environment, we have to set the configuration files responsible for Weka to talk with the sleep database and execute queries on it.
6. In `weka.experiment` package, locate `DatabaseUtils.prop`

In section

```
# The comma-separated list of jdbc drivers to use
```

Put:

```
jdbcDriver=org.PostgreSQL.Driver
```

And, in section:

```
# The url to the experiment database
```

Put the database connection URL:

```
jdbcURL=jdbc:PostgreSQL://hostname:port_number/database_name
```

or, you can hardcode the database authentication information as:

```
jdbcURL=jdbc:PostgreSQL://hostname:port_number/database_name?user=user_
id&password=user_pass
```

In this way, the user doesn't have to specify the username and password every time Weka has to connect to the database.

Replace fields *hostname, port_number, database_name, user_id and user_pass* with their respective values.

7. In section:

```
#the method that is used to retrieve values from the db (java datatype
+ RecordSet.<method>)
```

Add the following types:

```
int2=5
int4=5
float4=7
text=0
date=8
time=8
bpchar=0
```

Save all the modifications.

8. The Weka GUI can be started by executing the class `GUIChooser.java` within package `weka.gui`

**Running SQL `SELECT` queries from Weka:**

To execute a simple select query on the database from Weka's database interface, follow the steps below:

1. Start Weka GUI and select Explorer mode. In the Preprocess tab, locate and click on button "Open DB..."

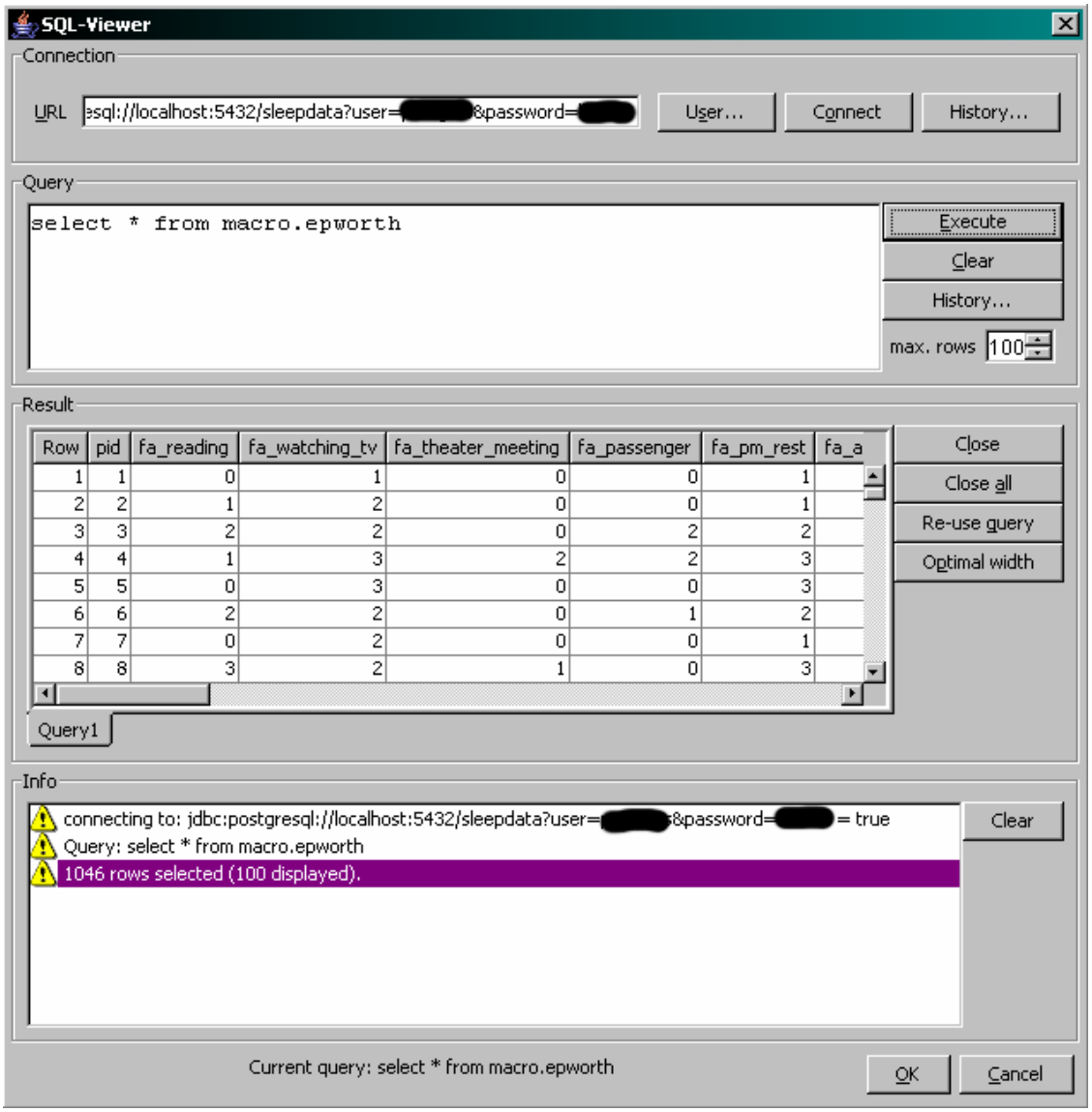
This will open Weka's SQL-Viewer interface. The database connection URL should appear in the URL text field.

2. Next, click on "User..." button next to this test field to provide the authentication information required to connect to the database (if this information was hard-coded in the `DatabaseUtils.prop` file, then skip this step).
3. Click on "Connect" button to connect with the database. If connection was successful, the following message appears in the "Info" field:

```
connecting to: jdbc:PostgreSQL://hostname:port_number/database_name =
true
```

Now we are ready to execute queries on the database.

The figure below is a screenshot of Weka executing a select query. The results of the query can be seen in the Weka Result pane:




After executing the select query, the user can come out of the SQL-Viewer by pressing the OK button.

The rows gathered after the execution of the query become instances within the Weka environment ready for further pre-processing and analysis. They can also be saved to a convenient ARFF (Attribute Relational File format) type file format if desired.