8-2017

Data Mining in Large-Scale Clinical Visit Data for Rett Syndrome Patients

Neela Saranya Avudaiappan
Clemson University, neela.3513@gmail.com

Data Mining in Large-Scale Clinical Visit Data for Rett Syndrome Patients

---

A Dissertation
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
Master of Science
Computer Science

---

by
Neela Saranya Avudaiappan
August 2017

Accepted by:
Dr. Brian Dean, Committee Chair
Dr. Ilya Safro
Dr. Alexander Herzog

Abstract

Rett syndrome (RTT) is a rare neurological disorder that predominantly affects girls. Research on RTT has mostly centered around gene mutations and possibility of cure using gene therapy. In this thesis we perform the first large scale systematic study of RTT patient records. The thesis has two major goals. One is to identify behavioral groups and the other is to study the association of medications and behavior or conditions. To achieve the first goal we apply standard clustering techniques like non-negative matrix factorization and k-means. We identify behavioral groups which could be used by clinicians for formulating better treatments. For the second goal we start with the most popular existing technique, disproportionality analysis, and make necessary adaptations for our data set. We then generalize this method and suggest an alternate approach which efficiently answers which medication caused the most change in a behavior. We test both approaches and show that the medications shown to decrease seizures the most are indeed those prescribed for the same. Using this as a tool, clinicians can identify possible side effects of medications.

Acknowledgments

I would like to thank Dr. Brian Dean for being my advisor, whose valuable guidance and encouragement made my research experience truly enjoyable. I would also like to thank my committee members Dr. Ilya Safro and Dr. Alexander Herzog for being supportive throughout. This research was possible only because of the data we received by collaboration with Dr. Walter Kaufmann. I would like to extend my thanks to him, Dr. Carrie Buchanan, Dr. Jennifer Stallworth, Dr. Aubin Tierney and all other doctors in Greenwood genetic center who took their time out to answer all our data related queries. I would also like to thank Aditya Bettadapura for helping us with data cleaning.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Rett syndrome (RTT) is a neurobiological disorder that predominantly affects females and was first described by Andreas Rett in 1960s. No cure has been found for Rett syndrome though many medications and therapies improve the quality of the life of patients. Most of the research on RTT has been on gene mutations and possibility of cure using gene therapy. No large scale systematic study of RTT patient records to detect abnormalities, patterns, or prominent sub-populations with consistent behavioral traits has been done to the best of our knowledge. A plethora of information hidden in the medical records are underutilized which if used well could facilitate better treatment. This is particularly true for non curable diseases where the only aim is to improve the standard of life as much as possible and a better treatment would be a life long benefit for the patients. In this thesis we try various data mining techniques on a large set of patient records and attempt to address these questions.

1.1 Preliminaries

1.1.1 Rett Syndrome

RTT is a neurobiological disorder that is caused by mutations in MECP2 gene which is located on the X chromosome at Xq28 [16]. Subjects are diagnosed to have either (1) classic / typical RTT or (2) variant / atypical RTT. Atypical RRT is when the patient does not meet all diagnostic criteria but has some symptomns of RTT. RTT is considered unique among other developmental

disorders because of its "usually sporadic occurrence, extreme female gender bias, early normal development and subsequent developmental regression, autonomic dysfunction, stagnation in brain growth and distinctive neuropathology" [16]. The diagnostic criteria for RTT as per [21] is shown in Table 1.1.

1.1.2 Medical Data Mining

The concept of storing health care data and information in electronic formats, now popularly known as Electronic Medical Records (EMRs), dates back to early 1970s [6]. EMRs have facilitated better viewing, ordering, care management, analysis, reporting and patient-directed functionality [18]. One of the major uses of EMRs is in Safety Signal Detection which is defined by WHO as detection of possible casual relationship between a drug and an adverse event. The Council for International Organizations of Medical Sciences extended this definition to include beneficial events. Adverse effects, often called "side effects" among non-health professionals are also known as Adverse Drug Events (ADEs) or Adverse Drug Reactions (ADRs). Administrative health databases are maintained by hospitals for administrative purposes and contain information on hospital admissions and drug prescriptions. The major difference between these databases and EMRs if any is that EMRs capture low level information on health care details of the patient. The branch of pharmacological science that encompasses detection and assessment of adverse events is called *pharmacovigilance* [24]. Pharmacovigilance evaluates how safe a drug is by evaluating it after its release in the market, which is referred to as *postmarket surveillance*. Passive surveillance relies on reports by health professionals and manufacturers while active surveillance aims to automatically generate safety reports from medical records and databases.

1.2 Related Work

Much of Rett-based research has been on gene mutations, studying involvement of MECP2 in biological, neurochemical and neurotransmitter/receptor systems. Notably, recent pre-clinical studies on mouse models have indicated that the condition is potentially treatable [27]. Hence most of Rett-based research now revolves around possible treatments like gene therapy, MECP2 reacivation, RNA editing and protein replacement. A large-scale study of patient records to analyze patterns and clusters has never been done before.

Required for typical or classic RTT
Consider diagnosis when postnatal deceleration of head growth observed
Required for typical or classic RTT
1. A period of regression followed by recovery or stabilization
2. All main criteria and all exclusion criteria
3. Supportive criteria are not required, although often present in typical RTT
Required for atypical or variant RTT
1. A period of regression followed by recovery or stabilization
2. At least 2 of the 4 main criteria
3. 5 out of 11 supportive criteria
Main criteria
1. Partial or complete loss of acquired purposeful hand skills.
2. Partial or complete loss of acquired spoken language
3. Gait abnormalities: Impaired (dyspraxic) or absence of ability.
4. Stereotypic hand movements such as handwringing/squeezing, clapping/tapping, mouthing and washing/rubbing automatisms.
Exclusion criteria for typical RTT
1. Brain injury secondary to trauma (peri- or postnatally), neurometabolic disease, or severe infection that causes neurological problems
2. Grossly abnormal psychomotor development in first 6 months of life
Supportive criteria for atypical RTT
1. Breathing disturbances when awake
2. Bruxism when awake
3. Impaired sleep pattern
4. Abnormal muscle tone
5. Peripheral vasomotor disturbances
6. Scoliosis/kyphosis
7. Growth retardation
8. Small cold hands and feet
9. Inappropriate laughing/screaming spells
10. Diminished response to pain
11. Intense eye communication - eye pointing

Table 1.1: RTT Diagnostic criteria reproduced from [21]

Though data mining on patient records specific to Rett syndrome has never been done, work on safety signal detection started in late 1970s. Most of the initial works were based on Spontaneous Reporting Systems (SRSs). Spontaneous reports are reports with conclusions that a particular drug may be responsible for an adverse event, drawn by clinicians during diagnostic appraisal of a patient. While SRSs are reliable, they generally fail to detect most of ADEs because of duplication, underreporting and reporting bias. A study in 1991 by A.P. Fletcher did a direct comparison between event monitoring system and ADE reporting of over 44000 patients. He showed that under-reporting could be as high as 98% for many ADEs. He argued that SRS suffered from reporting bias caused by prejudices of medical staff, and other methods need to be explored to detect ADEs [3]. Though the limitations of SRSs were discussed by Naranjo et al. [20] 10 years before Fletcher, no system was introduced as an alternative until 1991. The first system that detected ADRs based on actions recorded in patient records like decrease on dosage or discontinuation of medication was implemented in 1991 by Classen [2]. The shortcomings of SRS have been studied many times hence (for example, [1], [5] and [19]), in 2006 Joel Lexhin dicussed ways to improve quality and quantity of reporting and argued that SRS would continue to play an important role in ADE detection.

The traditional method of detecting ADEs before EMRs was with chart reviews that were prepared by nurses using patient data. In 2001 Honigman confirmed that computerized systems were useful in detecting ADEs by comparing them with chart reviews and shifting to EMRs was valuable [7]. Later in 2008 Hwang showed that computer based ADE monitoring was successful in identifying most of ADRs with a study conducted in a 1300-bed tertiary care teaching hospital in Seoul, Korea. Compared against a chart review by pharmacists to identify ADEs, the computer-based system was shown to have 79% sensitivity and all severe ADRs were captured[11]. In 2009, Zhengwu Lu reviewed the benefits, challenges and future of information technology in pharmacovigilance. He identified that data mining signals were not always indicators of problems but were often good indicators of possible problems. He concluded that data mining techniques could be used to improve efficiency of pharmacovigilance and not replace it[15].

Through the years various statistical tests and data mining techniques were introduced for ADE detection, most of which were based on Disproportionality Analysis (DA). The major methods used were Reporting Odds Ratio (ROR), Proportional Reporting Ratio (PRR), Yule's Q ratio and Information Component (IC). A study in 2002 compared these methods to detect ADRs on SRSs on the Netherlands Pharmacovogilance Foundation Lareb dataset, and concluded that all methods

were broadly comparable. They also highlighted the efficiency of IC in higher dimensions and for large calculations [29]. Andrew et al. [30] recognized the exclusive use of disproportionality analysis and listed other data mining techniques that could be useful in pharmacovigilance. Predictive modeling, clustering, association mining and other visualization techniques were suggested as possible techniques.

A year later Marc Suling and Iris Pigeot studied all data mining algorithms used in SRS databases and how they can be extended for ADR detection in longitudinal databases. Other existing algorithms for ADR detection in longitudinal databases are MUTARA and an improved version of it called HUNTS [12]. Both approaches use Temporal Association Rules (TARs) to mine for patterns as an extension of association rules. Noren et al. [22] proposed a method that extended DA in SRSs to longitudinal patient records. In a time interval t, on drug of interest x, medical event of interest y they define:

- n_x^t is the number of first prescriptions of x with follow up in time period t.
- n_y^t is number of first prescriptions of any drug followed by y in time t.
- n^t is the number of first prescriptions of any drug with follow up in time t.
- n_{xy}^t, is the number of first prescriptions of x followed by y in time t.

The expected value of n_{xy}^t under a simple null model assuming no association between x and y is given by

$$E_{xy}^t = n_x^t \cdot \frac{n_y^t}{n^t}.$$

The logarithm

$$\log_2 \frac{n_{xy}^t}{E_{xy}^t}$$

gives an association score that if positive can mean the event occurs disproportionately often and if negative occurs disproportionately rarely. Noren et al. proposed the information component (IC) measure of disproportionality as

$$IC = \log_2 \frac{n_{xy}^t + \frac{1}{2}}{E_{xy}^t + \frac{1}{2}}$$

which reduces sensitivity to outliers due to rare events using shrinkage. To account for temporal variation they introduced IC_Δ. The follow up period of primary interest was u, and the control period to contrast was v. If we define $E^{u*}_{xy} = \frac{n^v_{xy}}{E^v_{xy}} \cdot E^u_{xy}$, then

$$IC_\Delta = \log_2 \frac{n^u_{xy} + \frac{1}{2}}{E^{u*}_{xy} + \frac{1}{2}}.$$

Since these methods were implemented on different databases, it was hard to tell which method was superior in ADR detection. A comparison of these algorithms was performed in [25] and [26] on The Health Improvement Network (THIN) for six drugs with known adverse effects. The conclusion of the study was that no algorithm was superior and all of them failed to detect rare ADRs. The results indicated that HUNT could be more optimal when number of patients prescribed the drug was small. The authors suggested that more than one algorithm must be used for ranking ADEs. It was recently shown that disproportionality analysis in SRSs should be used for hypothesis generation only and more robust methods were required to influence clinical decisions [17].

Apart from this, studies have been conducted to find correlation between clinical concepts and laboratory test results. In 2011 George Hripcsak et al. [9] used lagged linear correlation to reveal associations between clinical concepts extracted from sign-out notes and laboratory tests. They found many interesting associations including low blood potassium preceding 'hypokalemia'. As an extension of their work, they used multivariate distributed lag models in their lagged linear analysis. The addition of context-related variables was intended to facilitate better characterization of intended and unintended effects [14]. Univariate lagged linear regression (ULLR) is used to compute coefficients β_τ, where y_t represents laboratory value at time t and x represents drug value at $t - \tau$:

$$y_t = c_t + \beta_\tau x_{t-\tau} + \varepsilon.$$

Multivariate lagged regression (MLLR) for L time lags and N variables is used to find $\beta_{\mu^i,\tau}$, which is the coefficient for lag τ of the variable μ^i. Many such models are explored in [14], one of which is the following.

$$y_t = c + \sum_{i=1}^{N} \sum_{\tau=1}^{L} \beta_{\mu^i,\tau} \mu^i_{t-\tau} + \varepsilon$$

Lagged linear correlation has also been used as a metric in clustering clinical concepts [8].

Chapter 2

Data Acquisition and Preprocessing

2.1 Data Acquisition

The data sharing agreement between Greenwood Genetic Center and Clemson University gave us access to 19 data files in the form of spreadsheets with medical records of patients with Rett syndrome. As per the agreement the shared data had masked patient identification information and replaced it with a unique *maskid* per patient. There were a total of 1194 patients involved in the study. A brief overview of the documents is provided in Table 2.1. Along with the data files, *data dictionaries* that contained a list of variable names, types and labels in the data files were provided.

2.2 Data Description and Preprocessing

The details of prescribed medications along with the indications for which it was prescribed like cold, seizures etc., were maintained in two files. The details were maintained in the *medicationlog* data file until a a point of time. The rest of the visits were recorded in the *concomitantMeds* file in a different format. The medicationlog file contains 396 columns, the details of which are in Table 2.2. This format assumes that a maximum of 49 medications would be prescribed on any visit. Patients typically visited once every year or once every 6 months. On each visit the list of medications

File	*No of fields*	*length*	*Details*
adverseEvents	6	1	Has list of adverse events
childQuestionnaire	53	3971	Details of behavioral pattern filled out parents on visits
clinicalCriteria	30	1225	Details on clinical criteria filled during baseline visit
clinicalSeverityScale	16	6539	Details on clinical severity recorded every visit
conclusionOfStudyParticipation	8	798	Details on when and why patients concluded being a part of the study
concomitantMeds	1111	1016	Details on prescribed medications filled during every visit in new format since
currenthistory	177	6545	Details on some clinical and behavioral characteristics filled during every visit
deathrecord	22	61	Details on death of patients
demographicinfo	18	1225	Demographic information of patients
diagnosticmecp2status	28	1225	Details on diagnosis of patients
ekg	7	1614	Details of QT and QTc Intervals
eligibility	7	1228	consensus and Gene positive/negative
initialhistory	479	1213	Details on health during baseline visit
measurements	35	6557	Height, weight and other measurements taken on every visit
medicationlog	396	2953	Prescribed medications in old format
motorbehavioralassessment	45	6531	Categorical details on behavior
registration	5	1234	Details filled during first registration
sf36	40	4381	Details of how patients felt about their health

Table 2.1: Overview of data Files

were entered one per column. The corresponding columns for indication, dose, etc., contains further details. There are 49 columns, one for every medication.

- *MedRxNormInput* contains the name of the prescribed medication.
- *Indication* contains the condition for which the medication was prescribed.
- *Units* contains the units of the quantity of medication to be consumed if applicable.
- *Frequency* contains the frequency of medication like 'once per day', 'AM;PM' etc.
- *Dose* contains dosage of medication if applicable.
- *Start_age* contains the age at which the medication was first prescribed.
- *Stop_Age* contains the age when the patient stopped using the medication.
- *MedRxCode* contains the RxCode of the medication which is the code provided for each medication by *RxNorm*.

RxNorm is a normalized naming system for many branded drugs and is provided by National Library of Medicine (NLM) [23]. The name of the medication filled in MedRxNormInput contained many errors. It was written in different formats, with different abbreviations, special characters, and often contained spelling mistakes. This made the *RxCodes* a more reliable source to identify medications, since they were free from any bias of medical practitioner. However, the MedRxCodes field had many missing entries and other kinds of errors.

For example, the first two pairs of medications mentioned below are the same but have different values in 'code' and other fields. To clean these and consider them the same medication we only consider the concept unique identifier (CUI), i.e., C0875952 for Aciphex. We create a map where we map CUIs with RxCodes such that the values are RxCodes that appear most number of times for that CUI. The RxCodes do not have to be the error free; they only need to be the same for all instances of same medication for our application. Using this technique, 204 unique medications which were listed for a total of 841 times were corrected. Some entries have the CUI typed wrong like the third pair below. These can be identified as the same using the "code". Using this technique 32 unique drugs with 78 repetitions were cleaned. Other common errors included extra white spaces like in the fourth pair. That can be easily solved by replacing all instances of more than one space

with one space.

Aciphex [C0875952 code:261440 100.0 [RxNorm]
Aciphex [C0875952 code:RX10261440 100.0 [RxNorm]

Fiber [C0225326 code:70727 100.0 [RxNorm]
Fiber [C0225326 code:70727 95.0 [RxNorm]

Iron Supplement [217790] code:217790 100.0 [RxNorm R]
Iron Supplement [C0721124 code:217790 100.0 [RxNorm]

oxcarbazepine [C0069751 code:32624 100.0 [RxNorm]
oxcarbazepine [C0069751 code:32624 100.0 [RxNorm]

We only used codes to correct spellings and not standard techniques like *edit distance*. This is because some drugs with low edit distance scores can be completely different and wrong interpretations could cause serious discrepancies in the results. For example consider the following pairs of drugs. Acetylcarnitine is a dietary supplement generally taken by patients with developmental disorders while Acetylcysteine is a medication taken for cough.

Acetylcarnitine [C0001040 code:193 100.0 [RxNorm]
Acetylcysteine [C0001047 code:197 100.0 [RxNorm]

Sodium Chloride [C0037494 code:9863 100.0 [RxNorm]
Sodium Fluoride [C0037508 code:9873 100.0 [RxNorm]

Wafer [C0991560 code:316989 100.0 [RxNorm]
Water [C0043047 code:11295 100.0 [RxNorm]

In many cases the MedRxCode is unfilled and in some cases it is filled with name of medication or junk text. It is also to be noted that not all medications have RxCode. For example

MedRxNormInput1..49
Indication1..49
Units1..49
Frequency1..49
MedRxCode1..49
Dose1..49
Start_Age1..49
Stop_Age1..49 visit_age
visit
makid

Table 2.2: medicationlog fields

ConRmed1..79
ConRcode1..79
Conunits1..79
Confreq1..79
Conint1..79
Conroute1..79
ConSmed1..79
ConScode1..79
ConContinuing1..79
ConDose1..79
MoreThan3Months1..79
Conassess_age1..79
Constart_Age1..79
Constop_Age1..79
Visit
visit_age
Partcipant_cycle_number
maskid

Table 2.3: concomitantMeds fields

RxNorm does not provide codes for some nutritional supplements like “Children’s Multivitamin”, “Gummy bears”, etc. To fill in the missing values for medicines that have RxCodes we used the existing pairs of “MedRxNormInput” and “MedRxCode”. Such pairs can also be derived from the *concomitantMeds* data file. We used this data to fill out most of the missing values. This also fixes some misspelled medications if the error was already present in the data files. For the ones that do not have RxCodes, we filled in our own code. For example for *Ranitidine*, the code generated is *Ranitdine* [K35252627 code:35252627 100.0[RxNorm], we used ‘K’ instead of ‘C’ to distinguish the legitimate RxCodes from those custom made.

In the concomitantMeds data file,

- *conRMed* contains the name of the medication.
- *conRcode* contains the RxCode.
- *Conunits* contains the units of medication like ‘milligram’.
- *Confreq* is the frequency at which the medication has to be taken.
- *Conint* contains the frequency of medications using medical terminology.
- *Conroute* contains how the medication needs be taken like “oral”, “inhaled”, etc.

- *ConContinuing* indicates whether the medication is still used.
- *ConDose* is the dosage details.
- *MoreThan3Months* is true if the medication has been prescribed for more than three months.
- *Conassess* is the age the patients were at the assessment when the medication was recorded in the database.
- *Constart* is the age the medication started.
- *Constop* is the age medication was stopped.
- *ConSmed* contains the indication.
- *conScode* contains the snomed code for the indication.

SNOMED CT (Systematized Nomenclature of Medicine Clinical Terms) or Snomed Clinical Term provides codes, terms and definitions for medical terms that was introduced for improvement of patient care [28].

While the medicationlog file's rows correspond to visits, the concomitantMeds file's rows correspond to patients. That is, for each patient only one row / record is maintained and updated on every visit. The start and stop of medication has to be inferred from *Constart* and *Constop*. *ConContinuing* doesn't have any meaning as it is modified on every visit and we only get to see the last update. The other question is that of *Constart* and *Constop*: it is unclear how this representation would capture a medication being prescribed on first and last visit but not in the intermediate visits. To add to the challenge, most of *Constart* and *Constop* are unfilled, which makes even partial assumptions invalid. Since using this file seemed to involve too many assumptions, we only used the file in data cleaning of the medicationlog file.

One of the most interesting data files is *childQuestionnaire* since it is the only form filled out by parents. It has rich behavioral information which is generally considered reliable. Parents generally know the patients the most and convey information by daily observation over a period of one or more years. But it could suffer from a reporter bias as different parents may understand the scales differently. Unfortunately, the forms are not filled out on every visit and sometimes filled out once every 2-3 years. Since this is a long gap, it becomes difficult to analyze if any medication is causing behavioral changes using this form. The form contains mostly categorical data like (1)

Field	Description
RateChildsHealth	In general would you say your child's health is
LimitedActivityLotOfEnergy	Doing things that take a lot of energy such as playing soccer or running
LimitedActivitySomeEnergy	Doing things that take some energy such as riding a bike or skating
LimitedActivityAbilityToGetAro	Ability physically to get around the neighborhood playground or school
LimitedActivitySelfCare	Taking care of him/herself that is eating dressing bathing or going to the toilet
EmotionalDifficultyKindsOfAction	Limited in the KIND of schoolwork or activities with friends he/she could do
EmotionalDifficultyDurationOfA	Limited in the AMOUNT of time he/she could spend on schoolwork or activities with friends
BehaviorInattentive	Had difficulty concentrating or paying attention
BehaviorStealing	Stole things inside or outside the home
MoodCrying	Felt like crying
MoodLonely	Felt lonely
MoodNervous	Acted nervous
SatisfactionSchoolAbility	His/her school ability
SatisfactionFriendships	His/her Friendships
SatisfactionFamily	His/her family relationships
RateChildsFutureHealth	I expect my child will have a very healthy life
RateConcernForChildsPhysicalHe	Your child's physical health
LimitedByChildsPhysicalHealth	Your child's emotional wellbeing or behavior
FrequencyOfLimitationFamilyAct	Limited the types of activities you could do as a family
FrequencyOfLimitationEverydayA	Interrupted various everyday family activities eating meals watching tv
FrequencyOfLimitationTensionCo	Caused tension or conflict in your home
visit_age	Computed Age (in days) at visit (where DOB provided)
maskid	Masked unique participant identifier

Table 2.4: some fields in `childQuestionnaire`

Excellent, Very Good, Good, Fair, Poor, None, (2) Never, Almost Never, Sometimes, Fairly Often, Very Often, (3) No, Not limited at all, Yes limited a little, Yes limited some, Yes limited a lot ranging from scales 1-4, 1-5 or 1-6. Since the scales vary, it is important to normalize them before using them for any machine learning. For our applications we do a min-max scaling where we make all values range from 0-1 for techniques like non-negative matrix factorization where the values cannot be negative, or standardize the data by making it zero mean, unit variance for methods like k-means. This document also contains some missing values. If more than 50% of the rows are missing we do not use it or if 50% or less values are missing we fill them by taking the average of previous and next visit of the same patient if present. If it is the first visit we simply copy the value from the next visit and if it is the last value we copy the previous visit value. The idea is to not fill in values that will cause too much deviation. Another way of filling the missing values would be using mean of field values of all visits for the particular patient.

Another document that contains some behavioral attributes and is filled out on every visit is *motorbehavioralAssessment*. Though some information is clinical, it serves as a good pool of information to study the effect of medications. The values are categorical with low values meaning the patient is doing well. The scales are in the range 0-4 for every field. The missing values are filled in a way similar to how childQuestionnaire file was filled.

Some of the most important details are contained in *diagnosticmecp2status* and *demographicinfo*. The *diagnosis* field in diagnosticmecp2status tells us if the patient is "classic", "variant" or "non-rett". Since we are are interested in analyzing Rett patients and the characteristics of non-rett patients could be significantly different, we remove maskids of patients with diagnosis as 'non-rett' in all our analysis. A similar role is played by "Gender" field in diagnosticmecp2status. Since the characteristics of males with Rett is different and their count is negligible, we only analyze female data. A total of 94 patients were found to fall under "non-rett" or "male" category. This is not a huge number compared to 1193 unique patients found in the motorbehavioralAssessment file for instance.

Since most of our analysis involved using more than one data file, we created one large datafile that merged many data files to make clustering and other analysis easier. For all merges, the visit was uniquely identified using *maskid, visit_age* pairs. The tricky part is some of the files have entries on every visit, some have entries on some visits and some only on baseline visit. The files that have entries only on baseline visit, like *clinicalcriteria* for instance mostly have data that does

MotorSkillsRegression
VerbalSkillsRegression
PoorEyeSocialContact
LackOfSustainedInterest
IrritabilityCryingTantrums
OverActiveOverPassive
DoesNotReachObjectsPeople
DoesNotFollowVerbalActsDeaf
FeedingDifficulties
ChewingDifficulties
LackToiletTraining
Masturbation
SelfMutilatingScratching
AggressiveBehavior
Seizures
ApparentInsensitivityToPain
SpeechDisturbance
Bruxism
BreathHolding
Hyperventilation
AirSalivaExpulsion
MouthingHandsObjects
BitingSelfOthers
HandClumsiness
StereotypicHandActivities
AtaxiaApraxia
OculogyricMovements
Bradykinesia
Dystonia
Hypomimia
Scoliosis
Myoclonus
ChoreaAthetosis
HypertoniaRigidity
Hyperreflexia
VasomotorDisturbance
TruncalRockingShiftingWeight
visit_age
maskid

Table 2.5: Fields in `motorBehaviouralAssesment`

AgeAtDiagnosisInYears
AgeAtDiagnosisInMonths
MECP2Results
MECP2ResultsMutation1
diagnosis
SpecifyDiagnosis
DiagnosisMadeBy
MECP2ResultsMutation2
MECP2ResultsMutation3
MutationChoices_3Truncation
MutationChoices_C316TR106W
MutationChoices_C397TR133C
MutationChoices_C473TT158M
MutationChoices_C502TR168X
MutationChoices_C763TR255X
MutationChoices_C808TR270X
MutationChoices_C880TR294X
MutationChoices_C916TR306C
MutationChoices_Deletion
MutationChoices_Duplication
MutationChoices_Exon1
MutationChoices_Insertion
MutationChoices_LargeDeletion
MutationChoices_Otherpointmutati
MutationChoices_SpliceSite
Visit
visit_age
maskid

Table 2.6: Fields in `diagnosticmecp2status`

Gender
Adopted
Ethnicity
PrimaryResidenceOfParticipant
AgeAtEnrollmentInYears
AgeAtEnrollmentInMonths
Race_AmericanIndianorAlaskaNat
Race_Asian
Race_BlackorAfricanAmerican
RaceNativeHawaiianOtherPacIsland
Race_Other
Race_Refused
Race_Unknown
Race_Unknownornotreported
Race_White
Visit
visit_age
maskid

Table 2.7: Fields in `demographicinfo`

not change with time, like *NormalInitialDevelopment*, *NormalPrenatalPeriod*, etc. While merging, we repeat the same values for every visit of the patient. But for the files where the entries were made partially, like the childQuestionnaire, we leave them blank or fill in with a value like -1 to indicate they were not filled. Apart from this, for studying effect of medication on behavior we merged some fields of *medicationlog* and *motorbehaviouralassessment* using a left join. This is because the medicationlog file only contained some visits and we were not using the *concomitantMeds* file for reasons explained before.

Chapter 3

Effect of Medications

The study of effect of medication on behavior is important in post-marketing pharmacovigilance and has been done even before the advent of EMRs. In the following sections we show how to adapt and extend popular approaches for this sort of association study, such as disproportionality analysis (DA) and linear regression to our specific setting.

3.1 Disproportionality Analysis

3.1.1 Extensions to Consider Increases and Decreases

The most common method used for studying effect of medication is probably disproportionality analysis; for example, a recent work [22] discusses how to detect temporal patterns in longitudinal patient records. To recall from Chapter 1, in a time interval t, on drug of interest x, medical event of interest y we define:

- n_x^t is the number of first prescriptions of x with follow up in time period t.
- n_y^t is number of first prescriptions of any drug followed by y in time t.
- n^t is the number of first prescriptions of any drug with follow up in time t.
- n_{xy}^t, is the number of first prescriptions of x followed by y in time t.
- $E_{xy}^t = n_x^t \cdot \frac{n_y^t}{n^t}$ is the expected value of n_{xy}^t under a simple null model assuming no association between x and y.

The proposed measure of disproportionality is then

$$IC = \log_2 \frac{n^t_{xy} + \frac{1}{2}}{E^t_{xy} + \frac{1}{2}}.$$

Our goal is to study the effect of medication on behavior and clinical symptoms. To extend this method to work on our dataset we need to define *events* based on change in behavior. The data we have in *motorbehaviouralassessment* data file is on a scale of 1-4. Since we want to study if a medication helps or exacerbates a condition, we define an event as *increase in value* which means it has gotten worse and *decrease in value* which means it has gotten better. As per the algorithm, we need to run it separately for (1) medication's effect in increasing value of a behavior and (2) medication's effect in decreasing its value.

The natural way to define n^t_y is as the number of increases / decreases. One major issue with this is that we would be ignoring the scale of change by treating an increase by 1 and by 3 the same. This would be underutilization of available data and might not produce desirable effects. We will generalize our model to address this shortly, but for now we ignore the magnitude of change.

While the original algorithm considered a medical event of interest y, we now separately consider two types of events $y+$ and $y-$ that lead to the following terms.

- n^t_{xy+} is the number of first prescriptions of x where the first subsequent change in behavior y within time t is an increase.
- n^t_{xy-} is the number of first prescriptions of x where the first subsequent change in behavior y within time t is a decrease.
- n^t_{y+} is the number of first prescriptions of any drug where the first change in behavior y within time t is an increase.
- n^t_{y-} is the number of first prescriptions of any drug where the first change in behavior y within time t is a decrease.
- $E^t_{xy+} = n^t_x \cdot \frac{n^t_{y+}}{n^t}$
- $E^t_{xy-} = n^t_x \cdot \frac{n^t_{y-}}{n^t}$
- IC_+ and IC_- for events $y+$ and $y-$ respectively:

$$IC_+ = \log_2 \frac{n^t_{xy+} + \frac{1}{2}}{E^t_{xy+} + \frac{1}{2}}$$

$$IC_- = \log_2 \frac{n^t_{xy-} + \frac{1}{2}}{E^t_{xy-} + \frac{1}{2}}$$

Conveniently, the following theorem holds.

Theorem 3.1. *At most one of IC_+ and IC_- can be positive.*

Proof. For the purpose of contradiction, assume both IC_+ and IC_- are positive. Then,

$$\frac{n^t_{xy+} + \frac{1}{2}}{E^t_{xy+} + \frac{1}{2}} > 1 \tag{3.1}$$

and

$$\frac{n^t_{xy-} + \frac{1}{2}}{E^t_{xy-} + \frac{1}{2}} > 1. \tag{3.2}$$

Given non-negative values a, b, c and d, the *mediant* of two fractions $\frac{a}{b} \leq \frac{c}{d}$ is $\frac{a+b}{c+d}$ and satisfies $\frac{a}{b} \leq \frac{a+b}{c+d} \leq \frac{c}{d}$. Taking the mediant of equations in (3.1) and (3.2) therefore yields

$$\frac{n^t_{xy+} + n^t_{xy-} + 1}{E^t_{xy+} + E^t_{xy-} + 1} > 1. \tag{3.3}$$

However, since $n^t_{xy+} + n^t_{xy-} = n^t_x$ and $E^t_{xy+} + E^t_{xy-} = n^t_x$, the LHS of (3.3) is 1, a contradiction. □

As a cautionary remark, suppose we had defined a third type of event, $y0$ indicating a measurement that indicates no change in behavior y. This would introduce two new terms:

- n^t_{xy0} denotes the number of first prescriptions of x where the first subsequent measurement of behavior y within time t indicates no change.
- n^t_{y0} denotes the number of first prescriptions of any drug where the first subsequent measurement of behavior y within time t indicates no change.

Using these we can define IC_0 just like IC_- and IC_+. However, while one can easily prove a generalization of Theorem 3.1 that all three of these cannot be simultaneously positive, it is quite possible for both IC_+ and IC_- to be positive (and we have observed this in our data). This is clearly not desirable and we do not recommend extending the model to ternary events in this way.

3.1.2 Extensions to Consider Magnitude of Increases and Decreases

Recall that the previous approach ignored the magnitude of change. To address this, we redefine n_y^t as the sum of all increases/decreases. Our new definition of y affects both n_y^t and n_{xy}^t . They are now defined as the following:

- n_{xy+}^t denotes the sum of magnitudes of all increases in behavior y occurring within time t following any first prescription of x.
- n_{xy-}^t denotes the sum of magnitudes of all decreases in behavior y occurring within time t following any first prescription of x.
- n_{y+}^t denotes the sum of magnitudes of all increases in behavior y occurring within time t following any first prescription of any drug.
- n_{y-}^t denotes the sum of magnitudes of all decreases in behavior y occurring within time t following any first prescription of any drug.

Though this captures the change well, there is still an issue: both IC_+ and IC_- can be positive if we are not careful in how we define E_{xy+}^t and E_{xy-}^t via an appropriate null model. We show this with a simple counter example.

Suppose there is an increase of 10 in behavior following one prescription of x and a decrease of 10 following another. In addition, there is an increase of 1 and decrease of 1 in two cases where x was not prescribed. Here,

$$\begin{aligned} n_{xy+}^t &= 10 \\ n_{y+}^t \cdot \frac{n_x^t}{n^t} &= \frac{11}{2} \\ IC_+ &= \log_2 1.909 \\ n_{xy-}^t &= 10 \\ n_{y-}^t \cdot \frac{n_x^t}{n^t} &= \frac{11}{2} \\ IC_- &= \log_2 1.909 \end{aligned}$$

Hence, both IC_+ and IC_- are positive. This, if not handled well could lead to seemingly contradictory results. This can be fixed by checking if the value of n_{xy} we get is significant by testing it with the expected value of a slightly different, more nuanced null model. A null model is a model

that is generated by randomizing the samples while preserving some of its statistical properties. The null model we used so far was simplistic in assuming that x and y were unrelated.

Let us now calculate the expected value for n^t_{xy+}. There are n^t_x visits with first prescriptions of x overall. Let us call these the x *associated* visits. The goal of our null model is to take each individual increase in behavior (n^t_{y+} in total) and redistribute them. One way is to randomly shuffle them such that the mass gets randomly distributed among all visits. In this case, the probability of any value now being x *associated* would be $\frac{n_x}{n_t}$ as before. But in our new model, we would like to preserve how much *activity* is present in x *associated* visits. An activity could mean an increase or a decrease. Hence, we now redistribute the positive weights by the ratio of $\frac{n^t_{xy+}+n^t_{xy-}}{n^t_{y+}+n^t_{y-}}$. The expected value is

$$E^t_{xy+} = n^t_{y+} \cdot \frac{n^t_{xy+} + n^t_{xy-}}{n^t_{y+} + n^t_{y-}}.$$

The overall activity after the first prescription of a medication is preserved in expectation in this model. If a medication both increased and decreased a behavior a lot, the shuffling preserves the activity and lands more positive mass there, thereby increasing the expected value. We use this E^t_{xy+} to calculate IC and we no longer get contradicting results. We get a proof similar to the one seen in the previous section. Taking the mediant yields us the following.

$$\frac{n^t_{xy+} + n^t_{xy-} + 1}{E^t_{xy+} + E^t_{xy-} + 1} > 1. \tag{3.4}$$

However, here $E^t_{xy+} + E^t_{xy-} = n^t_{xy+} + n^t_{xy-}$, and the LHS of (3.4) is 1, a contradiction.

Another potential issue is due to very low values of n_x: if n_{xy} is close to n_x and there is high increase/ decrease in those instances, the medication could have a high IC value thereby giving false positives. This could result from rarely prescribed medications.

The adaptation of disproportionality analysis for our problem has some caveats that when handled well could give us better results. To highlight, the main issues are

- DA considers effect of a medication on every behavior separately and hence does not give any insight on possible interaction between drugs. For example the model does not capture the possibility of a drug being usually prescribed with another drug, to combat the associated 'side effects'.

Condition	IC_-
ChoreaAthetosis	1.48093926656
Myoclonus	1.33866160932
HandClumsiness	1.31739452258
LackOfSustainedInterest	1.27875507442
AggressiveBehavior	1.25196850394
Seizures	1.23184665352
HypertoniaRigidity	1.19536661094

Condition	IC_+
SelfMutilatingScratching	1.43816254417
Masturbation	1.34066624764
Scoliosis	1.27561374795
OculogyricMovements	1.25
StereotypicHandActivities	1.15051320882
BreathHolding	1.12733129291
Hypomimia	1.11691536406

Table 3.1: Conditions affected the most by Keppra, a well known commonly prescribed seizure medication

- It handles only binary events well.

3.1.3 Results from generalized DA

We have no standard way of knowing whether the method outputs reliable results other than by using spot checks against results that would be anticipated based on strong clinical belief. We first run our analysis and test if seizure medications have significant IC_- values for decreasing seizures.

The scatterplot in Figure 3.1 shows plot of the IC_- values and n_x values associated with decrease in seizures. To get this list, a list of all unique medication names was generated, and for each medication, IC_- values for seizures was calculated. The plot in Figure 3.1 contains all medications that have a positive IC_-. A list of all seizure medications was retrieved using the indications mentioned in the medicationlog data file and marked in the figure. There were a total of 96 seizure medications and only 52 of which were prescribed for a total of 5 times or more. Out of these, 42 seizure medications were found to have positive association with decrease in seizures. Some of the false positives that seem to have high IC_- values are the ones with very low n_x^t values as seen in the plot. The reliable region is probably the one with higher n_x^t values where most of the commonly used seizure medications like 'Keppra' lie. Another interesting observation is IC_- values for *Myoclonus* being high. Myoclonus is involuntary twitching of muscles and is related to seizures.

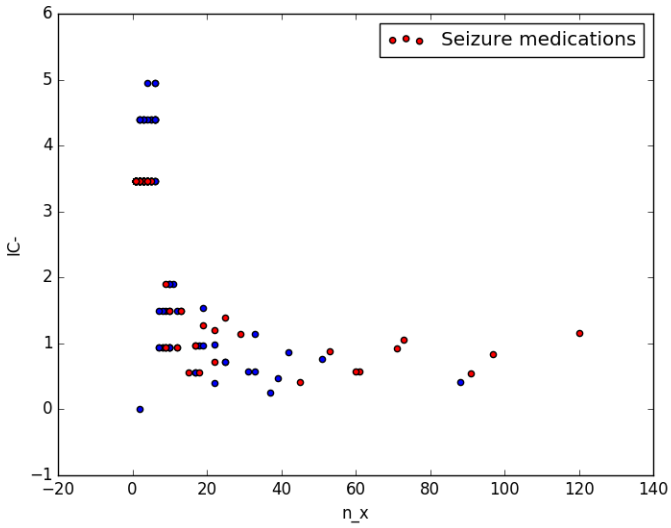

Figure 3.1: Scatterplot indicating the n_x and $IC-$ values for medications that decrease seizures

Condition	IC_-
HypertoniaRigidity	1.48662763095
AggressiveBehavior	1.4175060785
Seizures	1.30380601044
VerbalSkillsRegression	1.27870701295
LackOfSustainedInterest	1.26194360289
Bradykinesia	1.24692226521
DoesNotReachObjectsPeople	1.23475223414

Condition	IC_+
Masturbation	1.62998624484
OculogyricMovements	1.25
LackToiletTraining	1.24337185588
ChoreaAthetosis	1.23564763562
SpeechDisturbance	1.23018203171
MouthingHandsObjects	1.21170831008
ChewingDifficulties	1.21012409969

Table 3.2: Conditions affected the most by Lamictal, another common seizure medication

Condition	IC_-
LackOfSustainedInterest	1.56383367882
Myoclonus	1.54473832302
LackToiletTraining	1.42738085606
HypertoniaRigidity	1.37614297589
Seizures	1.344268665
Bradykinesia	1.30325558596
Masturbation	1.30076838639
ChoreaAthetosis	1.2922938198

Condition	IC_+
MouthingHandsObjects	1.29983255081
VerbalSkillsRegression	1.25025804818
AggressiveBehavior	1.20569680749
ChewingDifficulties	1.15825595478
VasomotorDisturbance	1.14379844961
FeedingDifficulties	1.08958918456
Dystonia	1.06413225376

Table 3.3: Conditions affected the most by Trileptal, another common seizure medication

3.2 Linear Regression Model

3.2.1 Methodlogy

Disproportionality analysis considers every behavior and medication separately, and cannot simultaneously test all medications affecting a behavior elegantly. It is also difficult to interpret what the meaning of IC is. To address these and other issues discussed in the previous section, we propose an unconstrained linear regression model. We define the following terms:

- b is a vector that represents change in behavior for all visits j.
- M is a matrix that contains values representing change in medication i in visit j.
- α and β are regularization parameters.

Our model assumes prescription of medication i results in change a_i in the behavior under consideration. Letting a be the vector of these values (the output we wish to compute), we would like to ideally satisfy $Ma = b$; that is, we would like to be able to express the change in behavior on each visit as precisely the linear combination of influences resulting from changes in medication at that visit. The statement of the objective we need to optimize in order to best express a change in behavior as a linear combination of medications is the following:

$$\underset{a}{\text{minimize}} \quad f(a) = \|b - Ma\|_2^2 + \alpha\|a\|_1 + \beta\|a\|_2^2. \tag{3.5}$$

In the above equation, for all non-initial visits, b_j is the change in behavior on visit j relative to visit $j-1$. If lower values indicate being healthier, a negative difference would indicate improvement in condition and vice versa. M_{ij} can take one of three values 1, -1 or 0 depending on whether a medication was introduced, removed or unchanged. "Introduced" means the medication was not prescribed in visit $j-2$ and prescribed in visit $j-1$, "removed" means the medication was prescribed in visit $j-2$ and not prescribed in visit $j-1$, and "unchanged" could mean two things (1) medication was not prescribed in visit $j-2$ and $j-1$, or (2) medication was prescribed in visit $j-2$ and $j-1$. Because of the way we have defined M, we ignore change of behavior in the first visit of every patient. The regularization parameters α and β encourage sparsity and uniformity of the resulting vector. To fully characterize the effect of changing values of these parameters on the resulting vector is a challenging and deep question on its own, and left as a topic of future study.

To calculate the gradient needed to minimize the function f, we simplify it:

$$\begin{aligned} f(a) &= \|b - Ma\|_2^2 + \alpha\|a\|_1 + \beta\|a\|_2^2 \\ &= (b - Ma)^T(b - Ma) + \alpha\|a\|_1 + \beta\|a\|_2^2 \\ &= b^Tb - 2b^TMa + a^TM^TMa + \alpha\|a\|_1 + \beta\|a\|_2^2 \\ &= b^Tb - 2b^TMa + a^T(M^TM + \beta I)a + \alpha\|a\|_1. \end{aligned} \tag{3.6}$$

We find the gradient as follows:

$$\nabla f(a) = -2b^TM + 2(M^TM + \beta I)a + \alpha \cdot sign(a). \tag{3.7}$$

The sign of a_i tells us whether the medication i improved a behavior or made it worse. As per our definition, a negative value of a_i would indicate that the medication i helped in making the condition get better and vice versa. A high absolute value would indicate that a particular medication affected the behavior a lot. Because of this, there can be a better distinction of whether a medication helps improve a condition or makes it worse.

One noticeable trait of a is its dependence on the scale of b in Equation 3.6. Normalization of values might be better for some applications but for our purpose where the scales are even, we prefer not to normalize. In our model, we consider M_{ij} to be 0 when the medication i was prescribed in both visit $j-2$ and $j-1$. This can be changed by assigning appropriate levels of values in case of increase in dosage. We did not do this since the data on dosage we had seemed to be mostly unfilled. The definition can also be changed based on whether the medication would have continued effect throughout the prescription period.

The major drawback of our model can be said to be the following:

- We need a minimum of three visits to calculate effect of medication.
- We consider a change of behavior from 4 to 2 and from 2 to 0 as the same, which may not be good.
- Assuming there is no effect when a medication continues to be prescribed.

3.2.2 Results

To check if our model produced reasonable results, we check if seizure medications helped decrease seizures. Two seizure medications 'Keppra' and 'Klonopin' show up in the top five medications decreasing seizures in Table 3.4. In this setting we do not consider that a drug can continue to have an effect if it is prescribed over a period of time. If we change that assumption and assign a value of 0.4 to M_{ij} when it was prescribed in both visit $j-2$ and $j-1$, we get slightly better results. It could mean the medications continue to show effect over prescriptions or the dosage is a valuable data that is missing. This time, three of the top five, and four of the top six medications are seizure medications. The four seizure medications that show up in Table 3.6 are Keppra, Klonopin, Lamictal and Topomax. The other medications that show up could possibly be the ones often prescribed together with seizure medication, they can also be seen as the medications taken to suppress the side effects of seizure medications. However, we have no data to back up this hypothesis and would require a clinician to analyze it further. Another interesting observation is 'Zantac' and 'Singulair' showing up as medications that increase seizures in Tables 3.5 and 3.7 . 'Zantac' and 'Singulair' can produce dizziness and vertigo which could appear like fainting in a motor-impaired, non-verbal subjects and interpreted as seizures. These are called non-epileptic spells and is indeed an issue with Rett syndrome patients. This could be an indication of non-epileptic spells that are interpreted as seizures.

Medication	*a*
Keppra [C0876060 code:261547 100.0 [RxNorm]	-0.0325
ATROPINE @ 1% @ DROPS [C1353673 code:424405 100.0 [RxNorm]	-0.0222
Klonopin [C0699315 code:202585 100.0 [RxNorm]	-0.0194
Albuterol [C0001927 code:435 100.0 [RxNorm]	-0.0164
Childrens Formula [C0719284 code:216009 91.4 [RxNorm]	-0.0138

Table 3.4: Medications that helped decrease seizures the most

Medication	*a*
Tegretol [C0700087 code:203029 100.0 [RxNorm]	0.0335
Singulair [C0595724 code:153889 100.0 [RxNorm]	0.0305
Zantac [C0592278 code:152523 100.0 [RxNorm]	0.0301
Benefiber, 100% oral powder for reconstitution [C1828775 code:686929 100.0 [RxNorm]	0.0277
Simethicone [C0037138 code:9796 100.0 [RxNorm]	0.0249

Table 3.5: Medications that increased seizures the most

Medication	*a*
Keppra [C0876060 code:261547 100.0 [RxNorm]	-0.1027
Klonopin [C0699315 code:202585 100.0 [RxNorm]	-0.0668
Lamictal [C0678180 code:196502 100.0 [RxNorm]	-0.0466
ATROPINE @ 1% @ DROPS [C1353673 code:424405 100.0 [RxNorm]	-0.0434
Albuterol [C0001927 code:435 100.0 [RxNorm]	-0.0432
Topamax [C0723778 code:220343 100.0 [RxNorm]	-0.0357

Table 3.6: Medications that helped decrease seizures the most, assuming they have continued effect

Medication	*a*
Prevacid [C0286036 code:83156 100.0 [RxNorm]	0.0674
Zantac [C0592278 code:152523 100.0 [RxNorm]	0.0638
Simethicone [C0037138 code:9796 100.0 [RxNorm]	0.0586
Singulair [C0595724 code:153889 100.0 [RxNorm]	0.0533
AQUAPHOR OINT,TOP C0715870 code:212929 100.0 [RxNorm]	0.0521

Table 3.7: Medications that increased seizures the most, assuming they have continued effect

Chapter 4

Clustering

Patients with Rett syndrome are currently classified only at a very coarse-grained level as *classic* or *variant*. While these are based on clinical criteria, patients show a huge variation in behavior within each group. It is hence believed by many clinicians that identification of behavioral groups could lead to better understanding of the nature of RTT as well as improved treatments. For example, the techniques of the previous chapter could be applied to determine how effects of medication vary across different subgroups, if at all. Our goal is to automatically discover natural subpopulation of patients using the behavioral attributes in the *childQuesstionnaire* data file.

4.1 Non-Negative Matrix Factorization

Non-negative matrix factorization (NMF or NNMF) is a commonly used method in analysis of high-dimensional non-negative data. Given a matrix X with non-negative entries, NMF factors it into two non-negative matrices, W and H:

$$X = WH.$$

The problem is NP-hard in general but many heuristics exist to obtain reasonable solutions in practice [4]. We use alternating non-negative least squares using projected gradients. NMF tends to extract sparse and interpretable factors. In our data, X is a matrix with values of behavioral attributes for each patient, where $X_{ij} \geq 0$ indicates the measured level in patient i of behavior j.

Accordingly, the rows of output matrix W correspond to patients and H's columns correspond to behavior. The number of columns in W and rows in H, say k, is the number of clusters. The entry W_{ir} indicates the extent of membership of patient i in cluster r, and H_{rj} indicates the extent to which cluster r is associated with behavior j. Note that, NMF provides a "non-crisp" partition with data points being assigned multiple memberships. We assign a cluster to a patient based on values in W and find the behavioral characteristics of the cluster by analyzing values in H. In our application, we assign a patient the cluster of maximum value in W.

We cluster patients based on the first time point in the *childQuestionnaire* file. We perform min-max scaling to normalize the data, so that all values are affinely mapped in the range $[0, 1]$, with 0 being minimum and 1 being maximum. Since we do not have any specific number of groups to consider, we try different values of k and test using an *alluvial diagram* (Figure 4.1), if the memberships change haphazardly. An alluvial diagram is a flow diagram that is typically used to represent any change in grouping over time. In the alluvial diagram, we do not see anything suspicious as the clusters are relatively consistent with different values of k. The group sizes are uneven, and that could be because of how we assign membership. For example consider the following row in W {0.582720202, 0.435967205, 0.357546551, 0.527948532, 0.1458655805}. Though group 4 has a value close to group 1, we assign the patient to group 1 because it has higher weight.

The values in H indicate the general behavior of the cluster. For example, the following are the highest valued behaviors for $k = 4$. Recall that higher value means the patient more strongly exhibits a behavior.

1. FrequencyOfLimitationTensionCo(0.828182182864), LimitedByChildsBehavior(0.82627590182), FrequencyOfLimitationEverydayA(0.781461241073), FrequencyOfLimitationCausingAr(0.77294554474)

2. EmotionalDifficultyKindsOfActi(1.1317638759), EmotionalDifficultyDurationOfA(1.12263561587), EmotionalDifficultyLimitedPerf(1.11631556281), PhysicalDifficultyDurationOfAc(0.598676459154)

3. RateChildsHealth2(1.29715255416), RateWorryingOverChildsHealth(1.28070790312), RateConcernForChildsPhysicalHe(1.25580501932), RateConcernForChildsBehavior(1.15555134634)

4. LimitedActivityAbilityToGetAro(1.23856340219), LimitedActivityWalkOneBlock(1.20752390029), LimitedActivityBendingLifting(1.20730119618), LimitedActivityLotOfEnergy(1.15556835598)

The behaviors grouped together are quite related. The first group seem to have those who

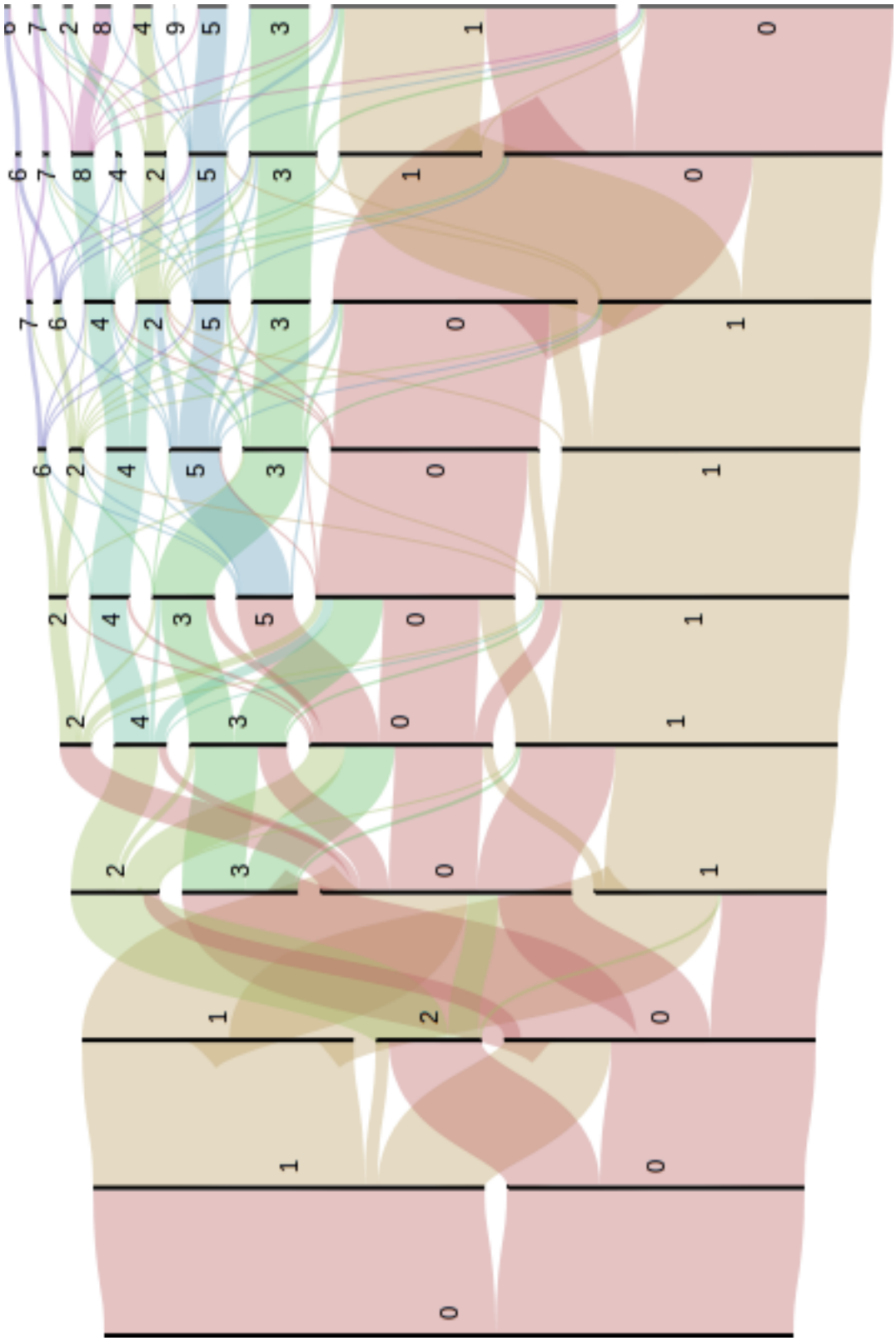

Figure 4.1: Alluvial diagram indicating how membership changes as the number of groups increases

are not doing well with family activities, the second group can be seen as group with emotional difficulty, the third group seems to be a group where parents (who fill out the questionnaire) are very worried about the patient's health, the fourth group seems to be physically less active. For different values of k, the cluster characteristics remains well grouped as seen below.
Prominent behaviors for $k = 3$:

1. LimitedActivityAbilityToGetAro(0.998506665748), LimitedActivityBendingLifting(0.983315645867), LimitedActivityWalkOneBlock(0.97624963901), LimitedActivityLotOfEnergy(0.952576240262), LimitedActivitySomeEnergy(0.951527061286)

2. LimitedByChildsBehavior(0.807689063063), LimitedByChildsAttention(0.766592208097), FrequencyOfLimitationTensionCo(0.701325638214), FrequencyOfLimitationEverydayA(0.683725185299), FrequencyOfLimitationCausingAr(0.680785371673)

3. RateWorryingOverChildsHealth(1.15449174153), RateChildsHealth2(1.09213101982), RateConcernForChildsPhysicalHe(1.08609374692), RateConcernForChildsAttention(1.04416907966), RateConcernForChildsBehavior(1.02175899223)

Prominent behaviors for $k = 5$:

1. LimitedByChildsBehavior(0.757245096774), FrequencyOfLimitationTensionCo(0.747230398082), FrequencyOfLimitationEverydayA(0.704523031906), LimitedByChildsAttention(0.700402019604), FrequencyOfLimitationCausingAr(0.698069822565)

2. EmotionalDifficultyKindsOfActi(1.00274067291), EmotionalDifficultyLimitedPerf(1.00250459265), EmotionalDifficultyDurationOfA(0.997982579861), PhysicalDifficultyDurationOfAc(0.525015336672), PhysicalDifficultyTypesOfActiv(0.507251190695)

3. RateChildsHealth2(1.49896420531), RateWorryingOverChildsHealth(1.25809531975), RateChildsImmunity(1.17926506191), SatisfactionFamily(1.08216746941), SatisfactionAppearance(0.930997087904)

4. LimitedActivityAbilityToGetAro(1.26322113242), LimitedActivityWalkOneBlock(1.23083118088), LimitedActivityBendingLifting(1.22480018183), LimitedActivitySomeEnergy(1.16406773482), LimitedActivityLotOfEnergy(1.15974750996)

5. RateConcernForChildsBehavior(1.59496215911), RateConcernForChildsAttention(1.58982783581), RateConcernForChildsPhysicalHe(1.40218241777), RateChildsSeriousIllness(0.895057100373), RateChildsHealth(0.591996626239)

Prominent behaviors for $k = 6$:

1. LimitedByChildsBehavior(0.783326855918), LimitedByChildsAttention(0.770036909512), FrequencyOfLimitationTensionCo(0.731928092866), FrequencyOfLimitationCausingAr(0.685830819289), FrequencyOfLimitationEverydayA(0.679485398607)

2. EmotionalDifficultyLimitedPerf(0.93128804871), EmotionalDifficultyKindsOfActi(0.926261597888), EmotionalDifficultyDurationOfA(0.923682973306), PhysicalDifficultyDurationOfAc(0.456761099707), PhysicalDifficultyTypesOfActiv(0.438924509671)

3. RateChildsHealth2(1.43687814834), RateWorryingOverChildsHealth(1.24290350046), RateChildsImmunity(1.08242358721), RateConcernForChildsPhysicalHe(1.06012087007), RateConcernForChildsBehavior(0.876373329574)

4. LimitedActivityAbilityToGetAro(1.20810066284), LimitedActivityWalkOneBlock(1.19851387921), LimitedActivityBendingLifting(1.18887347413), LimitedActivityLotOfEnergy(1.13531506862), LimitedActivitySomeEnergy(1.13228072249)

5. SatisfactionFamily(1.51027010532), RateChildsSeriousIllness(1.41440534207), SatisfactionLife(1.40673013643), SatisfactionAppearance(1.374140241), RateChildsFutureHealth(1.05418720439)

6. RateConcernForChildsAttention(1.32833482305), RateConcernForChildsBehavior(1.3088834011), RateConcernForChildsPhysicalHe(1.09948704489), RateChildsSeriousIllness(0.808017188025), FrequencyOfLimitationChangingP(0.713178797721)

Prominent behaviors for $k = 7$:

1. SatisfactionFamily(0.965115213896), SatisfactionAppearance(0.870626888226), SatisfactionLife(0.853732952555), RatePainOrDiscomfort(0.275121452545), RateChildsHealth(0.274404164043)

2. EmotionalDifficultyLimitedPerf(0.879151893793), EmotionalDifficultyKindsOfActi(0.878492199458), EmotionalDifficultyDurationOfA(0.872821984619), PhysicalDifficultyDurationOfAc(0.47586864916), PhysicalDifficultyTypesOfActiv(0.456587511262)

3. RateChildsHealth2(2.08557094761), RateChildsImmunity(1.68769084361), RateWorryingOverChildsHealth(1.58897757466), FamilyAbilityToGetAlong(0.757768889384), FrequencyOfLimitationCausingAr(0.733720374956)

4. LimitedActivityAbilityToGetAro(1.09187417833), LimitedActivityWalkOneBlock(1.08429289377), LimitedActivityBendingLifting(1.07044537243), LimitedActivityLotOfEnergy(1.02521504438), LimitedActivitySomeEnergy(1.02202877444)

5. RateConcernForChildsBehavior(1.79487180259), RateConcernForChildsPhysicalHe(1.74619360865), RateConcernForChildsAttention(1.68573192327), RateWorryingOverChildsHealth(0.518757717521), FamilyAbilityToGetAlong(0.416431623893)

6. RateChildsSeriousIllness(2.60416212182), RateChildsFutureHealth(1.16457190631), RateFrequencyOfPain(1.1368678442), RatePainOrDiscomfort(1.09371646958), RateChildsHealth(1.01064680601)

7. LimitedByChildsBehavior(1.06068369416), LimitedByChildsAttention(1.02275961415), FrequencyOfLimitationEverydayA(1.00861980031), FrequencyOfLimitationTensionCo(0.986834948409), FrequencyOfLimitationFamilyAct(0.936842896854)

While we have only looked at the high values, the low values can also be a valuable source of information. This way of clustering patients can be useful to the clinicians who might use this information to devise treatments and therapies.

4.2 K-Means Clustering

K-Means in one of the most popular clustering objective functions. When given a set of points in d-dimensional space for a given k, the goal is to find k centers by minimizing squared distance of every point to its closest center [13]. This problem is also NP-hard but a commonly used heuristic approach usually finds reasonable solutions. The algorithm has two major parts, executed in alternation until convergence:

- Assigning centers by finding mean of points in each cluster
- Updating membership of each point based on its closest center

The most popular distance metric used is the Euclidean distance, which is well suited for our application. This is because when two patients differ in some behavioral attribute by a larger value we want their distance to be substantially higher. We standardize the data by making it zero mean, unit variance. It is to be noted that this makes all coordinates contribute the same weight to our distance calculation. This could be an issue if we use multiple data files for clustering where one coordinate is split into several, thereby gaining extra weight just due to the extra level of detail. A common way to overcome this issue is to use a weight vector in distance calculation using apriori knowledge. Many recent methods suggest ways to automatically calculate weights [10].

An alluvial diagram shows that the patients remain well clustered even when clustered into a different number of groups (Figure 4.2). An analysis of prominent behaviors in clusters reveal that like NMF these clusters also show consistency across different values of k.

Prominent behaviors for $k = 3$:

- RateChildsHealth2(0.595537312314), RateWorryingOverChildsHealth(0.32573977003), RateConcernForChildsPhysicalHe(0.253485586357), RateChildsImmunity(0.177329760222), RateConcernForChildsBehavior(0.151506695955)
- LimitedByChildsBehavior(0.67152503964), LimitedByChildsPhysicalHealth(0.660253019372), FrequencyOfLimitationFamilyAct(0.62062734622), FrequencyOfLimitationEverydayA(0.619662831285), LimitedByChildsAttention(0.586539401273)
- LimitedActivityLotOfEnergy(0.476972515015), LimitedActivityBendingLifting(0.468941996382), LimitedActivitySomeEnergy(0.44163456975), LimitedActivityWalkOneBlock(0.432748873763), LimitedActivityAbilityToGetAro(0.431747178913)

Prominent behaviors for $k = 4$:

- RateChildsHealth2(0.589883018244), RateWorryingOverChildsHealth(0.347403599608), RateConcernForChildsPhysicalHe(0.233953774768), RateChildsImmunity(0.161016752873), RateConcernForChildsBehavior(0.148464367011)
- EmotionalDifficultyLimitedPerf(0.914456927468), EmotionalDifficultyKindsOfActi(0.912534561357), EmotionalDifficultyDurationOfA(0.885193953464), PhysicalDifficultyTypesOfActiv(0.612371037255), PhysicalDifficultyDurationOfAc(0.609986633633)

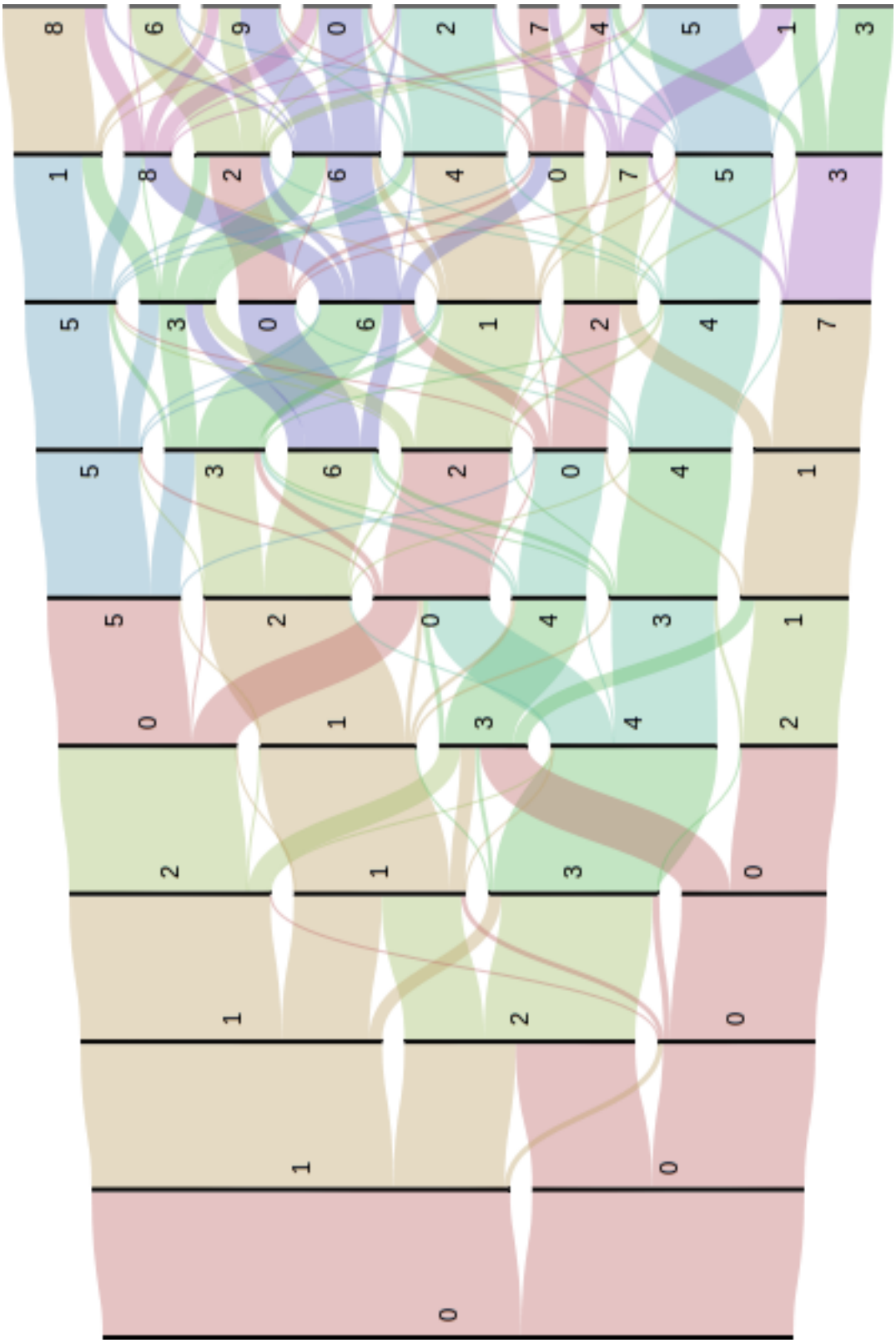

Figure 4.2: Alluvial diagram indicating how membership changes as the number of groups increases

- LimitedByChildsBehavior(0.824702121138), LimitedByChildsPhysicalHealth(0.816353279455), FrequencyOfLimitationEverydayA(0.804909692146), FrequencyOfLimitationFamilyAct(0.803526376955), FrequencyOfLimitationChangingP(0.802171106004)
- LimitedActivityLotOfEnergy(0.457799173825), LimitedActivityBendingLifting(0.453927928189), LimitedActivitySelfCare(0.414210461221), LimitedActivityWalkOneBlock(0.404291903606), LimitedActivitySomeEnergy(0.392561369369)

Prominent behaviors for $k = 5$:

- LimitedByChildsPhysicalHealth(0.829411483904), LimitedByChildsBehavior(0.793083131596), FrequencyOfLimitationChangingP(0.789153037369), FrequencyOfLimitationFamilyAct(0.777402352432),
- EmotionalDifficultyKindsOfActi(0.915002790504) EmotionalDifficultyLimitedPerf(0.910782540311), EmotionalDifficultyDurationOfA(0.900974481334), PhysicalDifficultyTypesOfActiv(0.613299322238), PhysicalDifficultyDurationOfAc(0.612868970997),
- RateChildsHealth2(0.701188981552), RateWorryingOverChildsHealth(0.448048761248), RateChildsImmunity(0.325518469865), RateConcernForChildsPhysicalHe(0.246926352465), RateConcernForChildsBehavior(0.229176309829)
- FrequencyOfLimitationCausingAr(0.735910535662), FrequencyOfLimitationTensionCo(0.688015589784), FrequencyOfLimitationEverydayA(0.575474312586), LimitedByChildsBehavior(0.555735525907), LimitedByChildsAttention(0.528983203527)
- LimitedActivityLotOfEnergy(0.483386764394), LimitedActivityBendingLifting(0.476695101983), LimitedActivityWalkOneBlock(0.447993365584), LimitedActivitySomeEnergy(0.421827041875), LimitedActivitySelfCare(0.4197648734)

Prominent behaviors for $k = 6$:

- LimitedActivityAbilityToGetAro(0.528678371971), LimitedActivityLotOfEnergy(0.524427666299), LimitedByChildsPhysicalHealth(0.520170304758), LimitedActivitySomeEnergy(0.498554866254), LimitedActivityWalkOneBlock(0.484795331807)
- RateChildsHealth2(0.61308540125), RateWorryingOverChildsHealth(0.349973611554), RateConcernForChildsPhysicalHe(0.227086204158), RateConcernForChildsBehavior(0.223114516829), RateChildsImmunity(0.215425371788)

- EmotionalDifficultyLimitedPerf(0.925215597358), EmotionalDifficultyDurationOfA(0.92017540469), EmotionalDifficultyKindsOfActi(0.904214128314), PhysicalDifficultyDurationOfAc(0.650813304666), PhysicalDifficultyTypesOfActiv(0.636407072197)
- RateChildsHealth2(0.582740647674), LimitedActivityBendingLifting(0.435214519966), RateConcernForChildsPhysicalHe(0.428321994129), LimitedActivitySelfCare(0.402711460461), RateWorryingOverChildsHealth(0.396904264482)
- FrequencyOfLimitationCausingAr(0.815699604062), FrequencyOfLimitationTensionCo(0.771345091241), FrequencyOfLimitationEverydayA(0.704961419976), LimitedByChildsBehavior(0.638381918564), LimitedByChildsAttention(0.581983060448)
- LimitedByChildsBehavior(0.99712329745), EmotionalDifficultyDurationOfA(0.943638339504), EmotionalDifficultyKindsOfActi(0.918858010143), FrequencyOfLimitationEverydayA(0.917040248891), LimitedByChildsPhysicalHealth(0.888577265914)

Prominent behaviors for $k = 7$:

- FrequencyOfLimitationCausingAr(0.914758485207), FrequencyOfLimitationTensionCo(0.868104941234), FrequencyOfLimitationEverydayA(0.689724351456), LimitedByChildsBehavior(0.623498359832), FrequencyOfLimitationChangingP(0.576868926109)
- RateChildsHealth2(0.627835541207), RateWorryingOverChildsHealth(0.353021056549), RateChildsImmunity(0.240803975565), RateConcernForChildsBehavior(0.232317341228), RateConcernForChildsPhysicalHe(0.205851908138)
- LimitedActivityAbilityToGetAro(0.544297898431), LimitedActivityLotOfEnergy(0.538344042919), LimitedActivitySomeEnergy(0.52408966323), LimitedActivityBendingLifting(0.509773437289), LimitedActivityWalkOneBlock(0.49027233031)
- EmotionalDifficultyLimitedPerf(0.872560774758), EmotionalDifficultyDurationOfA(0.838467461658), EmotionalDifficultyKindsOfActi(0.828184346392), PhysicalDifficultyDurationOfAc(0.60911683955), LimitedActivityLotOfEnergy(0.577604880491)
- RateChildsHealth2(0.59592128388), RateConcernForChildsPhysicalHe(0.453816393196), LimitedActivityBendingLifting(0.430517786482), RateWorryingOverChildsHealth(0.42052814045), LimitedActivitySelfCare(0.396445250423)

Figure 4.3: Alluvial diagram showing relationship between NMF and k-means for k=2

- LimitedByChildsBehavior(1.02228766329), FrequencyOfLimitationEverydayA(0.97102167454), RateChildsFutureHealth(0.906145462595), FrequencyOfLimitationFamilyAct(0.905266019017), FrequencyOfLimitationChangingP(0.898828765391)
- EmotionalDifficultyLimitedPerf(0.942341759205), EmotionalDifficultyKindsOfActi(0.930670525019), EmotionalDifficultyDurationOfA(0.919384500924), PhysicalDifficultyTypesOfActiv(0.655006936348), PhysicalDifficultyDurationOfAc(0.628313590314)

Analyzing the clusters, we notice the cluster centers are very similar to that of NMF. Considering $k = 4$ for example, the clusters have similar clinical interpretation as before. The first group seems to be one where parents are very worried about patient‘s health, the second group can be seen as group with emotional difficulty, third group has those who are not doing well with family activities, and the the fourth group seems to be physically less active. The alluvial diagrams shows that the clustering between NMF and k-means are similar. As k increases, NMF clusters become more uneven causing more difference between the two methods. The left horizontal line indicates NMF and the right one is for k-means in the following figures.

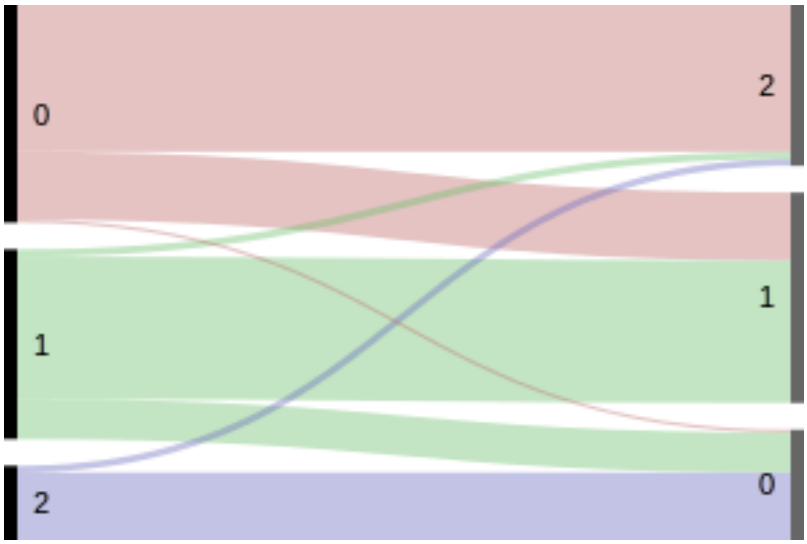

Figure 4.4: Alluvial diagram showing relationship between NMF and k-means for k=3

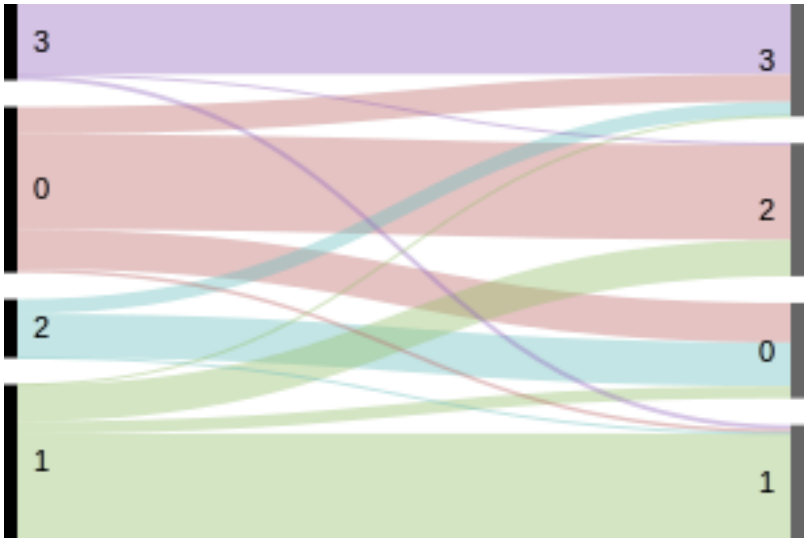

Figure 4.5: Alluvial diagram showing relationship between NMF and k-means for k=4

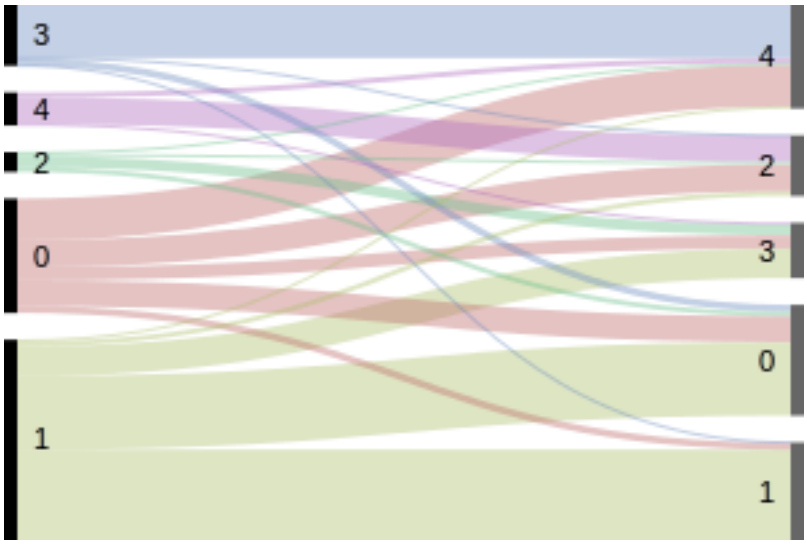

Figure 4.6: Alluvial diagram showing relationship between NMF and k-means for k=5

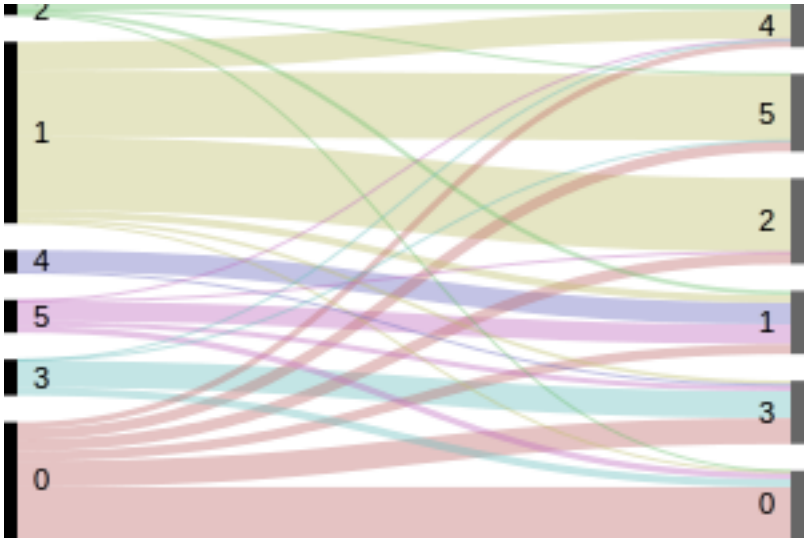

Figure 4.7: Alluvial diagram showing relationship between NMF and k-means for k=6

Chapter 5

Conclusion

In this thesis we had two major goals: (1) to study the effect of medications and (2) to identify behavioral clusters. We accomplished the first goal by using techniques like disproportionality analysis and linear regression. Both the methods were reasonable in showing that seizure medications caused decrease in seizures. In both the methods, we considered any change in behavior was an effect of medication only. A future work could be to include the effect of other therapies and diets, also considering age as a factor that affects the effectiveness of medications. Another interesting extension could be considering long term effects of using a medication, as our model does not account for it. For our second goal we used techniques like NMF and k-means to cluster patients using their behavioral characteristics. We showed that both methods have similar clusters and the clusters were fairly stable when the number of clusters was changed. A future work could use information on temporal change in behavior in clustering. Another interesting idea is to find the effect of medications on behavioral groups.

Bibliography

[1] A Alvarez-Requejo, A Carvajal, B Begaud, Y Moride, T Vega, and LH Martin Arias. Under-reporting of adverse drug reactions estimate based on a spontaneous reporting scheme and a sentinel system. *European Journal of Clinical Pharmacology*, 54(6):483–488, 1998.

[2] David C Classen, Stanley L Pestotnik, R Scott Evans, and John P Burke. Computerized surveillance of adverse drug events in hospital patients. *Jama*, 266(20):2847–2851, 1991.

[3] AP Fletcher. Spontaneous adverse drug reaction reporting vs event monitoring: a comparison. *Journal of the Royal Society of Medicine*, 84(6):341–344, 1991.

[4] Nicolas Gillis. The why and how of nonnegative matrix factorization. *Regularization, Optimization, Kernels, and Support Vector Machines*, 12(257), 2014.

[5] Stephen A Goldman. Limitations and strengths of spontaneous reports data. *Clinical Therapeutics*, 20:C40–C44, 1998.

[6] TERRY J Hannan. Electronic medical records. *Health informatics: An overview*, 133, 1996.

[7] Benjamin Honigman, Joshua Lee, Jeffrey Rothschild, Patrice Light, Russell M Pulling, Tony Yu, and David W Bates. Using computerized data to identify adverse drug events in outpatients. *Journal of the American Medical Informatics Association*, 8(3):254–266, 2001.

[8] George Hripcsak and David J Albers. Correlating electronic health record concepts with healthcare process events. *Journal of the American Medical Informatics Association*, 20(e2):e311–e318, 2013.

[9] George Hripcsak, David J Albers, and Adler Perotte. Exploiting time in electronic health record correlations. *Journal of the American Medical Informatics Association*, 18(Supplement_1):i109–i115, 2011.

[10] Joshua Zhexue Huang, Michael K Ng, Hongqiang Rong, and Zichen Li. Automated variable weighting in k-means type clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(5):657–668, 2005.

[11] Soo-Hee Hwang, Sukhyan Lee, Hyun-Kyung Koo, and Yoon Kim. Evaluation of a computer-based adverse-drug-event monitor. *American Journal of Health-System Pharmacy*, 65(23), 2008.

[12] Huidong Jin, Jie Chen, Hongxing He, Chris Kelman, Damien McAullay, and Christine M O'Keefe. Signaling potential adverse drug reactions from administrative health databases. *IEEE Transactions on knowledge and data engineering*, 22(6):839–853, 2010.

[13] Tapas Kanungo, David M Mount, Nathan S Netanyahu, Christine D Piatko, Ruth Silverman, and Angela Y Wu. An efficient k-means clustering algorithm: Analysis and implementation. *IEEE transactions on pattern analysis and machine intelligence*, 24(7):881–892, 2002.

[14] Matthew E Levine, David J Albers, and George Hripcsak. Comparing lagged linear correlation, lagged regression, granger causality, and vector autoregression for uncovering associations in EHR data. In *AMIA Annual Symposium Proceedings*, volume 2016, page 779. American Medical Informatics Association, 2016.

[15] Zhengwu Lu. Information technology in pharmacovigilance: Benefits, challenges, and future directions from industry perspectives. *Drug, Healthcare and Patient Safety*, 1:35, 2009.

[16] Toyojiro Matsuishi, Yushiro Yamashita, Tomoyuki Takahashi, and Shinichiro Nagamitsu. Rett syndrome: the state of clinical and basic research, and future perspectives. *Brain and Development*, 33(8):627–631, 2011.

[17] Christiane Michel, Emil Scosyrev, Michael Petrin, and Robert Schmouder. Can disproportionality analysis of post-marketing case reports be used for comparison of drug safety profiles? *Clinical Drug Investigation*, 5(37):415–422, 2017.

[18] Robert H Miller and Ida Sim. Physicians use of electronic medical records: barriers and solutions. *Health Affairs*, 23(2):116–126, 2004.

[19] Y Moride, F Haramburu, A Alvarez Requejo, and B Begaud. Under-reporting of adverse drug reactions in general practice. *British Journal of Clinical Pharmacology*, 43(2):177–181, 1997.

[20] Cláudio A Naranjo, Usoa Busto, Edward M Sellers, P Sandor, I Ruiz, EA Roberts, E Janecek, C Domecq, and DJ Greenblatt. A method for estimating the probability of adverse drug reactions. *Clinical Pharmacology & Therapeutics*, 30(2):239–245, 1981.

[21] Jeffrey L Neul, Walter E Kaufmann, Daniel G Glaze, John Christodoulou, Angus J Clarke, Nadia Bahi-Buisson, Helen Leonard, Mark ES Bailey, N Carolyn Schanen, Michele Zappella, et al. Rett syndrome: revised diagnostic criteria and nomenclature. *Annals of Neurology*, 68(6):944–950, 2010.

[22] G Niklas Norén, Johan Hopstadius, Andrew Bate, Kristina Star, and I Ralph Edwards. Temporal pattern discovery in longitudinal electronic patient records. *Data Mining and Knowledge Discovery*, 20(3):361–387, 2010.

[23] U.S. National Library of Medicine. Rxnorm overview. https://www.nlm.nih.gov/research/umls/rxnorm/overview.html, 2005.

[24] World Health Organization et al. The importance of pharmacovigilance. 2002.

[25] Jenna Reps, Jonathan M Garibaldi, Uwe Aickelin, Daniele Soria, Jack E Gibson, and Richard B Hubbard. Comparing data-mining algorithms developed for longitudinal observational databases. In *Computational Intelligence (UKCI), 2012 12th UK Workshop on*, pages 1–8. IEEE, 2012.

[26] Jenna M Reps, Jonathan M Garibaldi, Uwe Aickelin, Daniel Soria, Jack E Gibson, and Richard B Hubbard. Comparison of algorithms that detect drug side effects using electronic healthcare databases. 2013.

[27] Ruth R Shah and Adrian P Bird. MeCP2 mutations: progress towards understanding and treating rett syndrome. *Genome Medicine*, 9(1):17, 2017.

[28] Michael Q Stearns, Colin Price, Kent A Spackman, and Amy Y Wang. SNOMED clinical terms: overview of the development process and project status. In *Proceedings of the AMIA Symposium*, page 662. American Medical Informatics Association, 2001.

[29] Eugene P van Puijenbroek, Andrew Bate, Hubert GM Leufkens, Marie Lindquist, Roland Orre, and Antoine CG Egberts. A comparison of measures of disproportionality for signal detection in spontaneous reporting systems for adverse drug reactions. *Pharmacoepidemiology and Drug Safety*, 11(1):3–10, 2002.

[30] Andrew M Wilson, Lehana Thabane, and Anne Holbrook. Application of data mining techniques in pharmacovigilance. *British Journal of Clinical Pharmacology*, 57(2):127–134, 2004.