PRECISION NETWORKS AND INFORMATION RETRIEVAL FOR DESIGNING AND ANALYZING CLINICAL STUDIES

By

ELLIE SMALL

A dissertation submitted to the

School of Graduate Studies

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Statistics and Biostatistics

Written under the direction of

Javier Cabrera

And approved by

_____

_____

_____

_____

New Brunswick, New Jersey

May 2019

ABSTRACT OF THE DISSERTATION

Precision Networks and Information Retrieval for Designing and Analyzing Clinical Studies

By ELLIE SMALL

Dissertation Director:

Javier Cabrera

A Bayesian network is a probabilistic graphical model that represents a set of variables and their conditional dependencies via one directed acyclic graph. For example, a Bayesian network could represent the probabilistic relationships between diseases and symptoms.

However, in some cases, the situation at hand does not lend itself to the single network model. Sometimes each observation represents a network, and so we are dealing with many networks rather than just one. We refer to these individual networks as ***precision networks***. As an example, we may have a set of patients, each of which suffered multiple symptoms, conditions, and diseases referred to as ***events***. These events may or may not be related to each other. A precision network, here called a precision disease network or PDN, may be created for each patient, and the total set of such PDNs can be stored and analyzed together.

In order to build such a PDN for each patient, we need to establish when events are related and when they are not. We developed a nonparametric algorithm that will

determine whether such a relationship likely exists for two events, based on a data set with patients who experienced both. If such a relationship appears likely, we can provide an estimate of the proportion of dependent observations based on the time period between the two events. With the help of medical professionals, we may then establish an interval of time differences between those events within which we consider the events related, and outside of which we consider the events to be independent.

We note that medical researchers are often in need of finding new and interesting ideas for research within a topic. Those researchers will access the PubMed database and extract publications for the desired topic, usually resulting in a large amount of publications. They will then spend significant amounts of time perusing the abstracts of these publications in order to find an interesting idea that may be a candidate for a new clinical study.

We have developed a new method and computer application that examines all abstracts that fulfill the general search terms from bibliographic databases such as PubMed, mines those extracts for non-trivial, frequently occurring words, and allows for clustering of the abstracts using those words. By clustering and repeatedly re-clustering interesting clusters, a researcher can find an interesting subject for a new clinical study in a fraction of the time they spent previously.

We have also developed a new method to extract principal phrases from large volumes of text. Using this method, we have created an extension to the mining of abstracts that allows the clustering of principal phrases rather than words.

Acknowledgements

I am sincerely grateful for the support, encouragement, and advice I have received from my advisor, Dr. Javier Cabrera. His contributions, creativity, patience, and guidance were essential for the completion of this dissertation, as well as for my academic and professional growth.

I would like to thank Dr. John Kostis for his invaluable insight into the medical aspects of my dissertation and his continuous support throughout my time at Rutgers, and I would also like to thank him for serving on my dissertation committee.

I would like to thank Dr. John Kolassa and Dr. David Tyler for serving on my dissertation committee as well as for their constant willingness to share their expertise; thanks to their superior teaching methods I have been able to achieve the best that I can be.

I would like to thank all the other faculty members and staff in the department of Statistics at Rutgers University for their unwavering help, support, guidance, and attention throughout the past four years.

Finally, I want to thank my family for their unwavering love and support. Especially my daughter Pearl, who during her high school years willingly shared my attention with my studies.

Dedication

This dissertation is dedicated to Rob, Nikki, Hazel, and Pearl.

Table of Contents

List of Tables

List of Illustrations

PRECISION NETWORKS AND EVENT RELATIONSHIPS

1. Introduction

A ***precision network*** is a network in the form of a directed acyclic graph (DAG) that applies to one observation in a data set and is associated with an outcome for that observation. A collection of precision networks can be clustered, and the cluster an observation belongs to used as a predictor for the outcome.

Each precision network consists of a set of nodes, which we refer to as ***events***, with directed relationships between some of them. We are particularly interested in ***precision disease networks*** (PDNs), where our goal is to model medical/clinical outcomes as a function of patients' information, comorbidities, and network features.

Part of this dissertation deals with the task of determining whether a directed relationship exists between each pair of events for one observation. We developed a nonparametric algorithm that will analyze a data set consisting of observations (usually patients) for which two events occurred at specific moments in time and determines whether a relationship exists between the two events. Should such a relationship exist, it will determine, for each moment in time, an estimated proportion of observations that is likely to be dependent. Using this information, which we provide using a Shiny application, we can let medical experts determine at what interval in time between the two events we would consider there to be a relationship. We refer to this interval as the ***relation interval***.

Alternatively, we have also established a method that will determine the relation interval automatically. A proportion would need to be specified indicating the dependent observations that need to be present at the time between the events in order to consider a relationship to be existing. This proportion will then be used for many events to determine the relation interval for each ordered pair of those events.

Once those intervals have been determined for all possible events related to the disease progression, we can complete the PDNs for each patient by selecting a relationship if the time between two events falls within the relation interval. The relationship in that case will be directed from the first occurring event to the later occurring one.

We present simulations providing proof of concept, and we demonstrate its application to real-world data obtained from the Cardiovascular Institute of New Jersey.

2. Precision Networks

Bayesian networks (Koski & Noble, 2011), generally deal with single networks that model a complete data set. However, there are times when, instead of analyzing one network modeling a data set, we need to analyze multiple networks together, one for each observation in the data set. We refer to each of these individual networks as a precision network.

A precision network consists of multiple nodes (events) with directed relationships between some of them and models just one observation in a data set. Each precision network is associated with a specific outcome. The data set is to be analyzed in its totality in order to discover similarities in the precision networks that would influence the outcome. The goal is to provide inference on the outcome for groups of precision networks with similar relationships, and ultimately to predict outcome with a reasonable degree of certainty when presented with a specific precision network.

2.1 Motivation

2.1.1 Precision Brain Networks

Precision networks naturally occur in the case of the human brain. A brain can be separated into a set of areas. Often it is divided into 68 of those areas, 34 in the left hemisphere and 34 in the right. Axons transmit information and follow very precise paths in the nervous system. Each person has axonal pathways between areas in the brain, but the locations of those pathways vary. These axonal pathways describe the relationships between the different brain areas. Use of electroencephalography (EEG) and functional Magnetic Resonance Imaging (fMRI) (among others) produces a spatial map of activity in the brain. In addition, Diffusion Tensor Imaging (DTI) technologies have made it possible to estimate the location of white matter fibers. These methods allow us to estimate axonal pathways, resulting in a set of data for each subject describing the relationships between the various brain regions.

These brain areas and the axonal pathways between them result in a ***precision brain network*** for an individual. Studies (Durante & Dunson, 2018) relating these networks to the membership of these individuals to either a high or low creative reasoning group have shown a strong association between these networks and creativity, significantly more so than previous research linking creativity to region-specific activity in isolation.

Note that precision brain networks are currently undirected.

2.1.2 Precision Disease networks

In cardiology and more generally in medicine, researchers build graphs or pathways that summarize the evolution of disease in a patient (evolutionary disease pathway). A patient may suffer from a series of symptoms, diagnoses, medical interventions, procedures, and conditions (which we call ***events***) that may or may not be related to each other.

A precision disease network or PDN, created for an individual patient, would show events that occurred in the patient's lifetime and the connections between them. See Figure 1 for examples of two such PDNs.

a) Patient 1 b) Patient 2

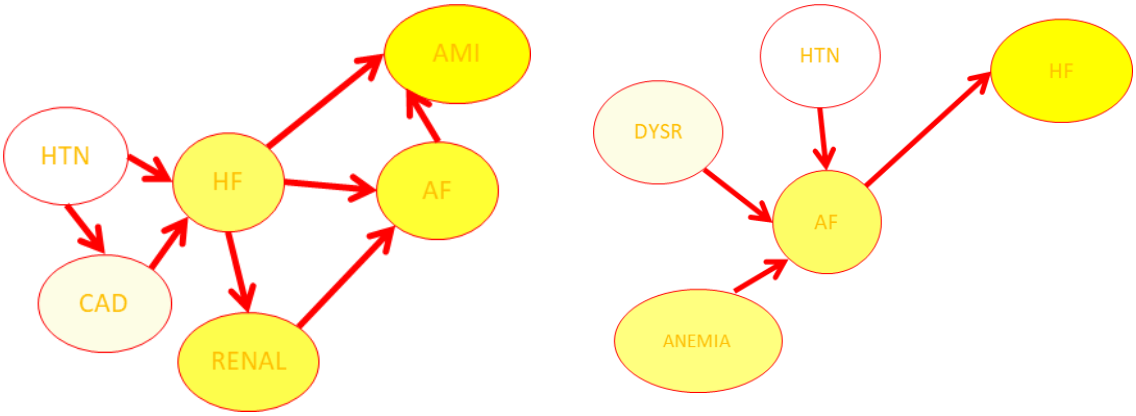

Figure 1: PDNs for Two Different Patients

The first person started with hypertension (HTN) that caused coronary artery disease (CAD). These two then caused heart failure (HF), followed by renal failure (RENAL),

atrium fibrillation (AF), and finally, acute myocardial infarction (AMI), i.e. heart attack. This path indicates an extremely serious situation.

The second person started with hypertension (HTN), dysentery (DYSR), and anemia (ANEMIA) causing atrial fibrillation (AF), which then in its turn caused heart failure (HF). This path is indicative of a much less serious situation.

Aside from the heart attack, the events are similar. But the structure of these conditions makes the situation very different in each of these patients, so the structure has information that is not present in the raw data.

For PDNs we would look for time to death, time to heart attack, or time to recurrence of disease as the outcome. We would like to see if specific paths of disease progression influence the outcome, and ideally, we would like to establish a reliable prediction of the outcome based on the relationships between the events as experienced by a patient.

2.2 Analysis

In order to create a precision network for an observation, we need the events associated with the observation and the path between those events. The events themselves are usually readily available; however, the path between them usually is not. The data often consists of the events and the moment in time at which they occurred; see Table 1 as an example of a (partial) data set we would typically receive for a group of 10 patients for which PDNs need to be created.

Table 1: Events Data for a Group of Patients

MR	CMTHY	CHD	HTN	DM	NEO	COPD	RENAL	STROKE	AMI
8/23/06	10/4/11	8/29/95	8/30/95	2/27/98	NA	3/20/03	9/9/10	NA	NA
5/22/95	NA	5/22/95	1/13/95	NA	NA	5/14/95	11/30/10	4/27/07	NA
NA	NA	9/25/07	9/18/99	2/12/07	NA	9/18/99	4/21/09	NA	NA
12/27/11	NA	10/14/97	10/19/01	8/3/07	3/8/03	1/22/02	10/14/97	10/19/01	4/30/03
NA	8/14/08	12/19/00	10/18/01	12/19/00	11/18/10	12/15/01	11/5/08	NA	NA
NA	NA	10/27/97	10/21/04	10/31/97	NA	NA	NA	NA	NA
NA	4/3/95	4/3/95	4/3/95	10/16/95	NA	3/31/02	9/3/09	NA	NA
NA	NA	11/22/96	11/22/96	NA	8/29/05	3/9/11	NA	NA	NA
NA	NA	5/2/95	2/25/98	NA	1/15/08	2/25/98	9/8/10	NA	NA
NA	11/18/07	6/2/06	3/26/02	3/26/02	NA	NA	11/18/07	NA	NA

Note that sometimes, instead of a date of occurrence, we receive number of days since

birth. This is the number we need for our analysis, and if it is not available, we need the birthdate of each patient so we can calculate it.

We see in this table that the fifth patient experienced hypertension (HTN) on or around October 18, 2001, and cardiomyopathy (CMTHY) on or around August 14, 2008. The PDN for this patient will include nodes for the events HTN and CMTHY; an important question we need to answer is whether we should include an arrow going from HTN to CMTHY.

We provide two different methods that will allow us to answer this question; the first one is a supervised method described in section 2.2.1 which depends on the outcome, the second is an unsupervised method described in the next chapter of this dissertation (chapter 3). Using the latter method, we establish the proportion of related occurrences of CMTHY that occur 7 years after HTN versus all occurrences of CMTHY that occur 7 years after HTN, related and unrelated. If this proportion is high, we will include a relationship represented by an arrow. If this proportion is low, we will not.

Once the precision networks have been completed, the information can be summarized in one row per observation (patient), and the full set of precision networks may be represented by a matrix. Each row in this matrix contains a column for every ordered combination of each pair of events, with a 1 indicating a relationship exists, and a 0 indicating such a relationship does not exist. See Table 2 for an example of this matrix for 15 patients showing a subset of the PDN for each of them.

Table 2: Matrix with PDNs for 15 Patients

AF to HF	HF to AF	AF to MR	MR to AF	AF to CMTHY	CMTHY to AF	AF to CHD	CHD to AF	AF to HTN	HTN to AF
0	0	0	0	0	0	0	0	0	0
0	1	0	0	0	1	0	1	0	1
0	0	0	0	0	0	0	0	0	0
0	1	0	0	1	1	0	1	1	0
0	1	0	0	0	1	0	1	0	1
0	1	0	0	1	0	0	1	0	1
0	0	0	0	0	0	0	0	0	0
0	1	1	0	1	0	0	1	0	1
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	1	0	1	0	1	0	1	0	1
0	1	0	0	1	0	0	1	0	1
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0

We can tell from this matrix for example that the second patient experienced atrial fibrillation (AF), which was related and possibly caused by heart failure (HF), cardiomyopathy (CMTHY), coronary heart disease (CHD), and hypertension (HTN), each of which occurred prior to the occurrence of AF.

The dimensionality of this matrix can be large, in which case we can reduce it via principal component analysis (PCA). Using the principal components, we may then cluster the precision networks, establishing an optimal number of clusters which may be visualized by representing the commonality of the relationships using different colors. See Figure 2 which represents one such a cluster; the red arrows represent relationships that are present most often in this cluster, the green arrows represent relationships that are present reasonably often, but not as often as the red ones, while the yellow arrows

represent relationships that are present in many patients in the cluster, but not as many as the red and green ones.

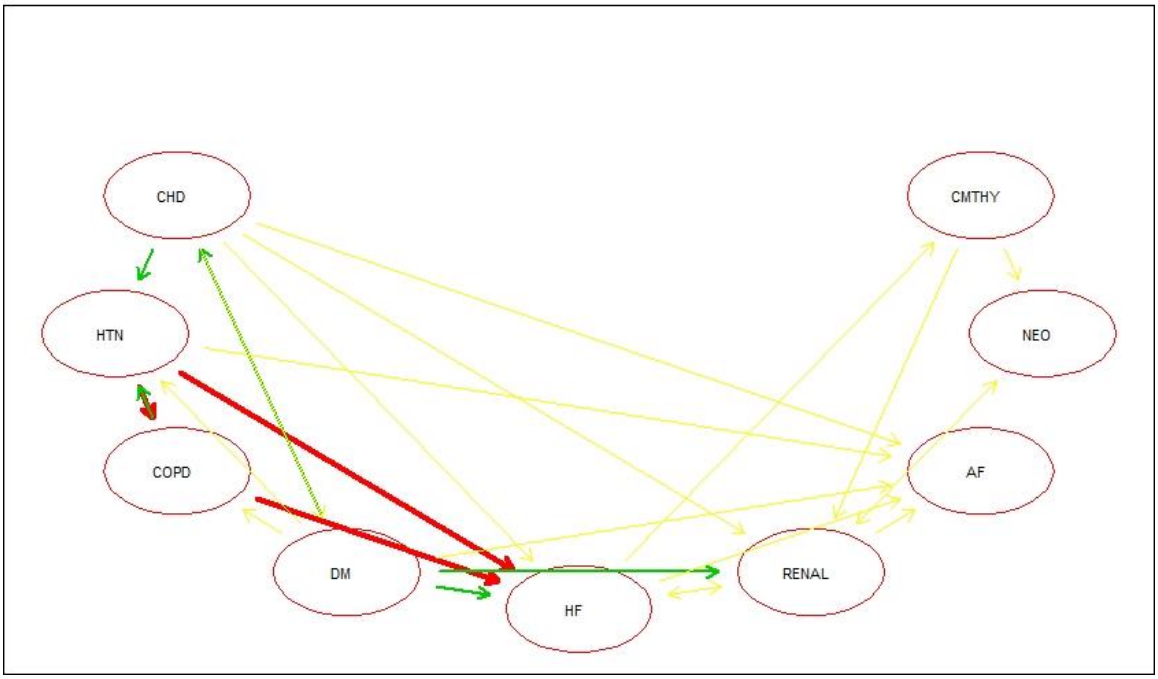

Figure 2: A Clustered set of Precision Disease Networks

For example, in this particular cluster, most people experienced hypertension which was usually followed and likely the cause of an occurrence of HF and COPD at a later time.

Finally, which cluster an observation belongs to can be used as a predictor in addition to other predictors such as age, gender, and comorbidities, to be regressed on the outcome. The outcome usually consists of time to an event (such as death, or recurrence of disease), suggesting we should use the cox proportional hazard model to perform the regression.

The performance of this method was tested on a data set with multiple events obtained from the MIDAS database (see section 3.2.2), while determining the existence of a relationship between two events using the supervised method described in section 2.2.1, and using the cox proportional hazard model to perform the regression. The addition of the PDN in the regression appeared to significantly improve its R^2. We intend to redo this and other tests using the unsupervised method for determining relationships as described in chapter 3.

2.2.1 Supervised Method for Determining Event Relationships

This is the current PDN method to determine whether a relationship between two events occurring in a patient's lifetime are related. We define a ***relation interval*** as an interval of time differences between which we consider two events related, and outside which we do not consider them related. This routine establishes relation intervals for a collection of events.

The procedure accepts a data set as in Table 1 as well as corresponding survival data for the same patients and determines the relation interval for every pair of events. For each such pair of events, say event A and event B, we select all datapoints in the data set with patients who experienced both, and extract the matching survival data for those patients, indicating for each patient the time until death (or any other major event), or censoring.

We create a new data set with time between event B and A (time to event A minus time to event B). We call this data set the observed differences data set. We determine the smallest multiple of 100 larger than the largest value in the observed differences data set

and denote this by m. We then take steps of size 100 from 100 to m. For each step, we do the following:

- Create a predictor variable that equals 1 for patients for whom the observed difference value lies between 0 and the step size, and 0 otherwise.
- Run a cox proportional hazard model fit on the survival data for the patients using the newly created predictor variable.

The step size that has the largest predictive value (largest z-score) obtained from the estimated cox proportional hazard model is the one we use to create the negative limit of the relation interval. We repeat the process reversing the events (i.e. time to event B minus time to event A) to obtain the positive limit of the relation interval. Note that for the interval, positive values indicate that A appeared first.

This method is considered supervised since it uses patients' survival data, which is the response variable, to determine the relation intervals. This makes the relationship between events dependent on the survival of the patients in the data set which is undesirable.

The PDN cluster values obtained using this method appear to be useful for predicting patient survival rates. We intend to repeat the analysis using our new method for determining the relation intervals as described in chapter 3, in order to establish if the method is useful for prediction when the unsupervised method is used.

2.3 Future Direction

We plan to improve the PDN project by comparing three different methods for determining the existence of a relationship between two events that occurred in a patient's lifetime:

- Via the relation interval obtained by the supervised method in section 2.2.1
- Via the relation interval obtained by the unsupervised method in chapter 3
- Via the cutoff interval obtained by the unsupervised method in chapter 3

We also intend to improve on the results by incorporating cross validation to determine the predictive value of the PDN clusters.

Furthermore, we plan to improve and expand an existing R package that performs the following functionality:

- Construct the relation interval via one or more of the discussed methods
- Construct PDNs from data such as presented in Table 1, and the relation intervals
- Provide functions used to cluster and analyze the PDNs
- Provide functions that allow visualization of individual PDNs as well as clustered groups of PDNs.

3. Event Relationships

3.1 Introduction

We developed an algorithm that provides information about the relationship between two events that may be used to determine whether two events occurring in a precision network are related.

We assume we have a data set available that contains time of occurrence for two events in subjects that experienced both. We refer to these two events as event A and event B.

Using this data set we determine if it is likely the two events are related. In many cases this means that event A may have caused event B. However, even though we can determine a potential directed relationship, we are not claiming causality; among other possibilities, causality could be indirect and event A may have resulted in measures taken that caused event B to occur.

Should we find a likely relationship between the two events, we will determine an interval of time differences between the events during which we consider it likely that there is a relationship, and outside of which we are not confident of such a relationship existing. We refer to this interval as the ***cutoff interval***.

In addition, we provide a new data set like the original containing time of occurrence for the two events by subject, but with all observations (subjects) that we estimate as dependent during the process, removed. This data set may be used to visualize an estimate of the shape of the independent densities of time to each of the two events. In

addition, we use this data set to estimate the proportion of dependent observations for groups of time differences between the two events.

We also developed a Shiny application that accepts a data set with time of occurrence for two events in subjects that experienced both, performs the algorithm, and presents the outcome of the algorithm in text as well as visually. Researchers may then inspect the information presented to determine the interval of time differences between the events within which we will create a directed link between the events for an observation/subject, and outside of which we will not. We refer to this interval as the ***relation interval***.

We also provide an automated process to determine these relation intervals for groups of events, which may be used when medical experts are not available to determine them. In this case we expect as input a set of observations (subjects) with time of occurrence of all events of interest for each observation, with the value NA indicating the event did not occur for that subject. A specific proportion (set by default to one half) may be specified indicating the minimum proportion of dependent observations that need to be present for time differences to be included in the relation interval. The interval will then be calculated for each pair of events present in the data, provided enough data is available for such a pair.

We note that there has been extensive research in bivariate survival analysis investigating nonparametric tests of association between two event times (Zhu & Wang, 2014) (Schemper, Kaider, Wakounig, & Heinze, 2013). However, these types of analyses aim to describe association between gap times, i.e. the relationship between the length of time between two events (the first gap) and the length of time between two other events (the

second gap), often dealing with a sample of patients where the gap times for the two different situations are ranked and the association is determined using a ranked statistic such as Kendall's Tau. For example, they would attempt to determine if age at onset of HIV is indicative of the time between onset of HIV and development of AIDS. Our research differs from this since we aim to investigate the probability that a second event, which has already happened, occurred due to the occurrence of a first event. In the case of HIV and AIDS, our research would be unnecessary since AIDS never occurs independently from HIV.

3.2 Proposal and Examples

Starting with a data set consisting of time to event A and time to event B in subjects that experienced both, we compare the time between the events as they occurred with the time between the events in the case that the events are independent. We estimate the latter by treating the data for the two events as separate and apply a correction in case a dependency exists. We use bootstrapping on the independent time differences to perform the comparison.

This will allow us to determine if a relationship between the two is likely and allows us to obtain an estimate of the independent observations as well as an estimate of the shape of the independent densities.

We will now discuss two examples that we will make extensive use of in order to explain the process more easily; one a simulation, the other a real-world situation.

3.2.1 Example 1 (Simulation)

We let $X \sim Weibull(\alpha = 6, \lambda = 700)$ and $Y \sim Weibull(\alpha = 2, \lambda = 600)$ represent time to respectively event A and event B. We took 2000 random and independent observations from each, matching each observation of A with an observation of B, simulating subjects experiencing both events independently. This provided our *independent* data set for this example.

We found 656 observations where B occurred after A, so we took another 656 observations from X. We let $W \sim \exp(\lambda = \frac{1}{20})$. We matched each of these 656 observations of X as time to event A, with X plus an observation of $W|W \leq 30$ as time to event B, simulating subjects experiencing event B from 1 to 30 days after event A, with more occurring close to A. We mixed this data set with the independent data set obtained before, giving a data set where 25% of the data is dependent with time differences between 0 and 30 days, with more dependent data shortly after the zero time difference and less dependent data closer towards the 30 days difference. This provided our *dependent* data set for this example.

Figure 3a) shows the relative frequency distributions for the two independent data sets, which may be interpreted as an estimate of the probability densities of the independent distributions, while Figure 3b) shows the relative frequency distributions for the two data sets in the dependent example. For comparison we also show, in Figure 3c), the relative frequency distributions for the two data sets in the dependent example after our algorithm removed those observations it deemed dependent. It is clear from the picture that the latter strongly resembles the frequency distributions of the independent data sets in Figure 3a).

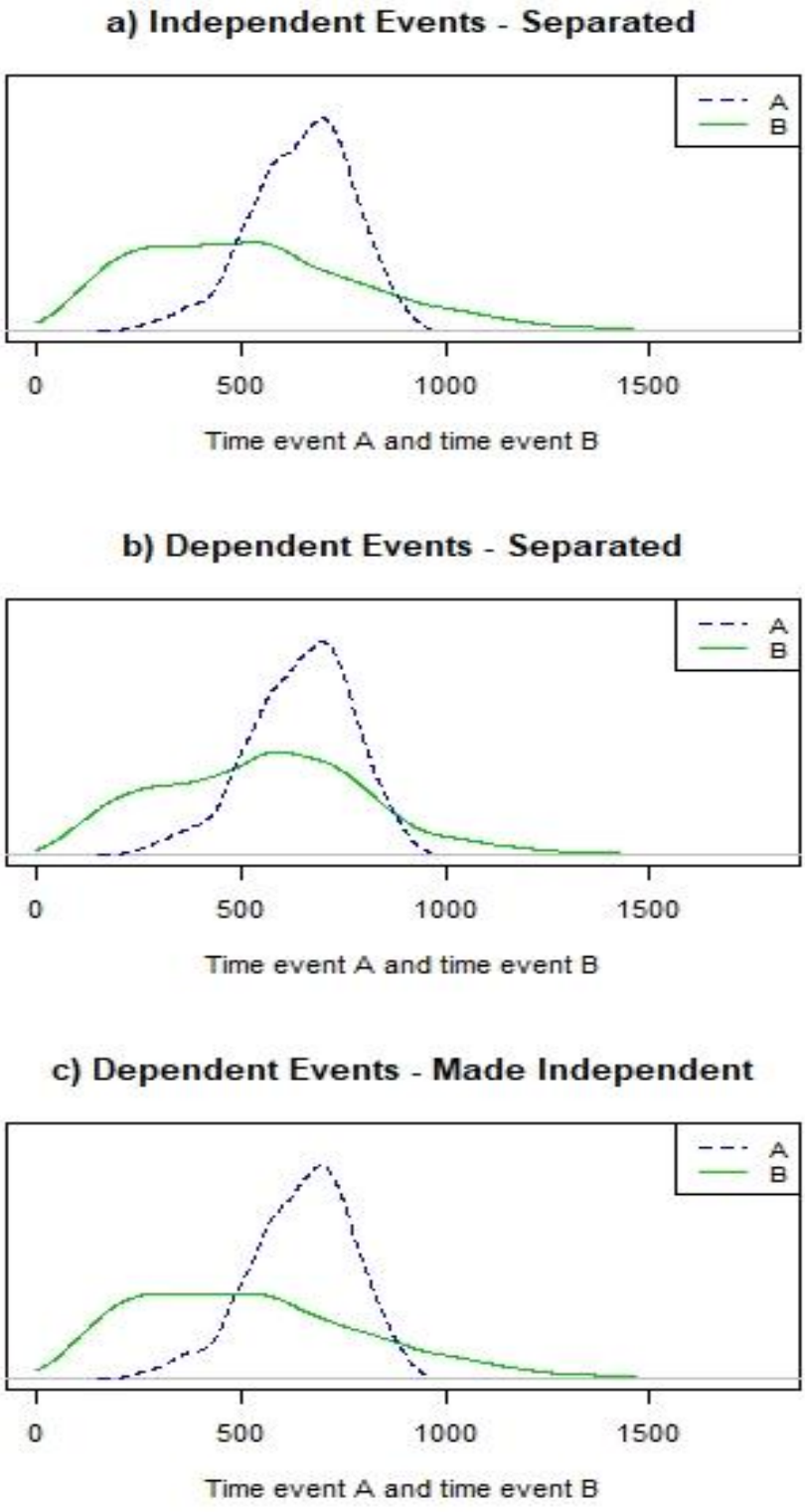

Figure 3: Frequency Distributions for Example 1

Our algorithm determined that for the independent data set we could not reject the null hypothesis of independence with a p-value of .50. For the dependent data set, it determined that the probability of the data sets being independent was 0% (p-value of .00), and thus the null hypothesis of independence was rejected. The algorithm determined that there was a significant likelihood that there were dependent observations when event B occurred between 0 and 23 days after event A. It estimated that about 26% of the data was dependent.

3.2.2 Example 2 (Real-World)

For our second example we consider patients who have experienced atrial fibrillation (AF) as well as heart failure (HF). See (Anter, Jessup, David, & Callans, 2009) who discuss the association between the two in detail. Atrial fibrillation is often caused by heart failure, but it can also be a precursor to HF. In addition, AF may occur independent from any heart failure the subject may or may not have experienced. When a patient experiences heart failure, and then experiences atrial fibrillation a few days later, the atrial fibrillation is almost certainly caused by the heart failure. But what if it occurs one year, or even 10 years after heart failure? At this point, can we still be as certain that its occurrence is related to the heart failure incident? The answer to this question is useful beyond its application to PDNs; (Kotecha & Piccini, 2015) explain how treatment for atrial fibrillation caused by heart failure should be different from treatment for independent atrial fibrillation. In order to establish appropriate treatment, we wish to

determine the likelihood of its independence based on the number of days between the occurrences of the events as experienced by the patient.

We obtained data from the Myocardial Infarction Data Acquisition System (MIDAS) database, discussed in detail by (Kostis, Deng, Pantazopoulos, Moreyra, & Kostis, 2010). This database includes the hospital discharge records of patients with myocardial infarction and invasive cardiovascular procedures who were admitted to New Jersey (NJ) non-federal acute care hospitals since 1986, plus hospital discharge records for all admissions of patients with any cardiovascular diagnosis since 1994. It contains abstracted discharge data, including the primary reason for admission and up to eight additional diagnoses, derived from the NJ statewide hospital uniform billing system, of 15 million hospitalizations for 5 million patients with cardiovascular diagnoses. We observed 93,162 patients with both heart failure and hypertension, and 55,323 patients that also experienced atrial fibrillation. Of these, 39,697 experienced AF before the onset of HF, 6,820 were recorded as having the events occurring the same day, and 8,806 were diagnosed with AF after they were diagnosed with HF.

Figure 4 shows the relative frequency distributions for $\boldsymbol{x} =$ time to heart failure (solid green) and $\boldsymbol{y} =$ time to atrial fibrillation (dashed black) for patients in our data set with both. The data is depicted as relative frequency of the event versus days since birth of the patients experiencing the event, and is an estimate of the probability densities of the distributions for time to AF and time to HF. It displays the distributions before (a) and after (b) dependent observations were removed.

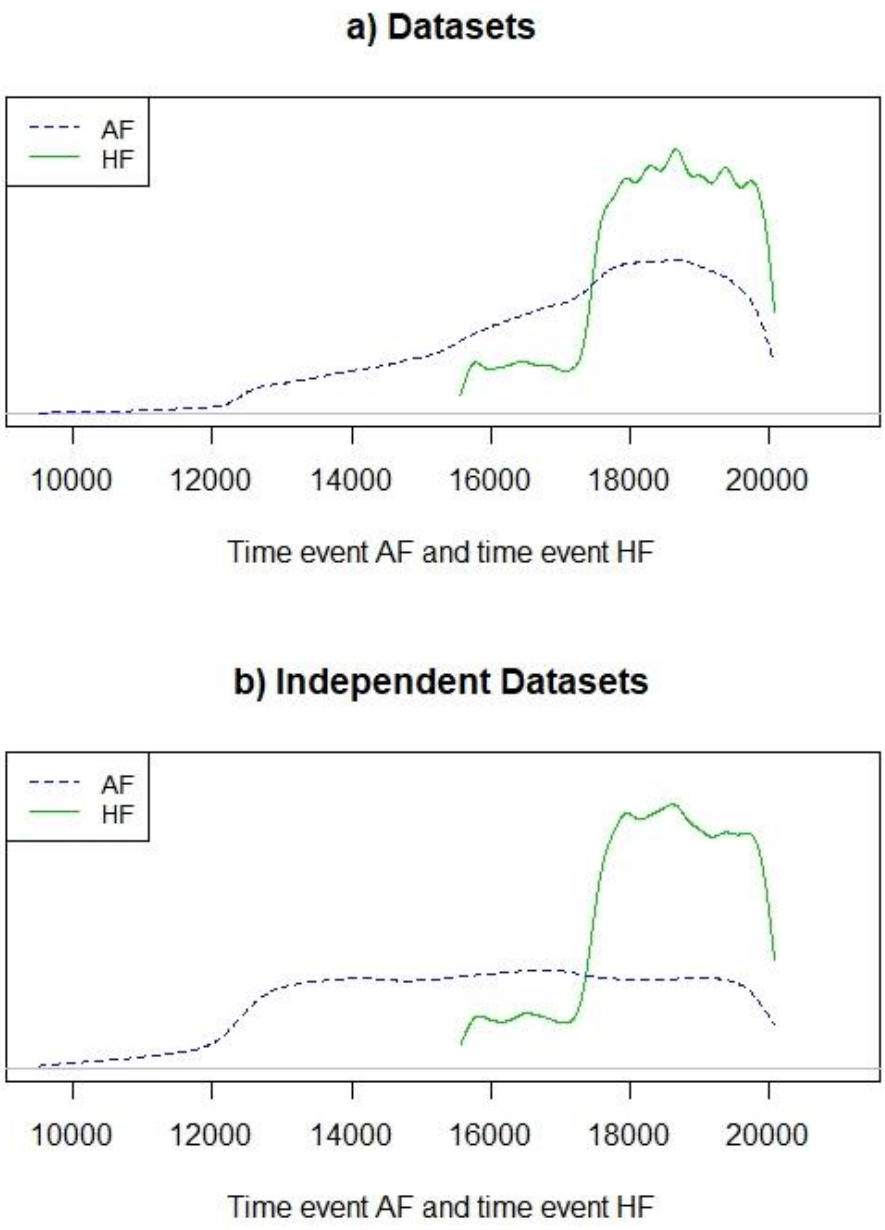

Figure 4: Frequency Distributions for Example 2

We see here that generally AF occurs before, and at times well before HF. Removing the dependent observations does not appear to significantly change the graph for HF.

However, the graph for AF changes significantly, showing that after the removal, the independent data set for AF is close to uniform at least for patients between the ages of about 35 and 53 years old (note that we have no data for patients over 55).

The algorithm determined that the probability of the data sets being independent was 0% (p-value .00). The cutoff interval, i.e. the interval beyond which we no longer have sufficient evidence to reject the null hypothesis of independence, was determined to be $(-5,16)$ years where AF is event A, and 73% of the data was deemed dependent. This indicates that some observations of HF occurring within 16 years after AF likely have been caused by the occurrence of AF, and some observations of AF occurring within 5 years after HF likely have been caused by the occurrence of HF.

Note, that this does not indicate that at any time within that interval we have a high number of dependent observations; it could be that at certain times the percentage of dependent observations is so low it would not be considered relevant. Due to this, it is likely that the relation interval that determines whether to create a link between the two events for a precision network is significantly smaller.

When we use the automated determination of the relation interval (see section 3.3.5.1) with the proportion set to the default of one half, we found the relation interval in this case to be equal to $(-2,10)$ years, which is indeed significantly smaller than the cutoff interval.

3.3 Method

We let X and Y indicate the time (which may be measured continuously or in days) from some reference date to events A and B respectively. Often, in the case where the subjects are individuals and the events are medical conditions, a good reference date would be the subject's birth date. This reference date will allow us to compare data for different subjects.

To determine the existence of a dependency between X and Y we need to determine time between the two events in subjects for whom both events occurred, i.e. we are interested in the distribution of $Y - X$. We assume we have access to a data set that contains time to event A and time to event B for n subjects.

Let vector $\boldsymbol{x}$ consist of all data points for X, i.e. x_i contains the time to event A for subject i, where $i = 1, \cdots, n$. Let vector $\boldsymbol{y}$ consist of all data points for Y, i.e. y_i contains the time to event B for subject i, where $i = 1, \cdots, n$.

3.3.1 The ODD

The random variable $O = Y - X$ gives us the time differences between event A and event B for subjects who experienced both events. We create a data set for this random variable as follows.

$$o_i = y_i - x_i, i = 1, \cdots, n, \qquad \boldsymbol{o} = \begin{bmatrix} o_1 \\ \vdots \\ o_n \end{bmatrix}$$

At this point $\boldsymbol{o}$ contains a set of observed data points from the distribution of O, which we shall refer to as the observed differences distribution, or ***ODD***.

When Y and X are measured in days, it is likely that the difference will have many repeated values. We will build an empirical probability density function for O as follows:

$$\hat{f}_n(t) = \frac{frequency\ of\ t\ in\ \boldsymbol{o}}{n}, t \in o_1, \cdots, o_n$$

When n is large, we can interpolate to obtain an estimate of the density of O, since:

$$\hat{f}_O(t) = \lim_{n\to\infty} \hat{f}_n(t), \min(o_1, \cdots, o_n) \leq t \leq \max(o_1, \cdots, o_n)$$

where t indicates the time between the two events.

3.3.2 The IDD

We let the independent differences distribution, or ***IDD***, be the distribution of $T = Y - X$ for the case where $X \perp\!\!\!\perp Y$. We need to estimate the IDD.

Note that if subject i experienced event A at time t_X, i.e. $X = t_X$, then under the assumption of independence, $P(Y = t_Y | X = t_X) = P(Y = t_Y)$. Thus we define

$$T_0 = \begin{bmatrix} t_{11} & \cdots & t_{n1} \\ \vdots & \ddots & \vdots \\ t_{1n} & \cdots & t_{nn} \end{bmatrix} where\ t_{ij} = y_j - x_i, i = 1, \cdots, n, j = 1, \cdots, n$$

$$and\ \boldsymbol{t} = \begin{bmatrix} t_1 \\ \vdots \\ t_{n^2} \end{bmatrix} = vec(T_0)$$

where $vec(T_0)$ combines the columns of the matrix T_0 to form the vector $\boldsymbol{t}$. At this point vector $\boldsymbol{t}$ contains observed data points for the distribution of $T = Y - X$ under the assumption that $X \perp\!\!\!\perp Y$.

We build the empirical probability density for the IDD from the vector $\boldsymbol{t}$ as follows:

$$\hat{f}_n(t) = \frac{frequency\ of\ t\ in\ \boldsymbol{t}}{n_g}, t \in t_1, \cdots, t_{n^2}$$

When n is large, we can interpolate to obtain an estimate of the probability distribution of the IDD since:

$$\hat{f}_T(t) = \lim_{n\to\infty} \hat{f}_n(t)\,, \min(t_1, \cdots, t_{n^2}) \leq t \leq \max(t_1, \cdots, t_{n^2})$$

where t once again indicates the time between the two events.

As an example, if subject i experiences event A at time 1600 days, then the probability that this subject experiences event B independently a day later is given by the probability that any subject experiences event B at time 1601 days, i.e. $P(Y = 1601)$. This probability is estimated by the data set contained in $\boldsymbol{y}$. Furthermore, the probability of A occurring one day before B is estimated by the number of ways we can combine an observation x in $\boldsymbol{x}$ with an observation y in $\boldsymbol{y}$ so that $y = x + 1$ divided by the number of ways we can combine an observation in $\boldsymbol{x}$ with an observation in $\boldsymbol{y}$.

However, the IDD we have so constructed would be a true independent differences distribution only if $\boldsymbol{x}$ and $\boldsymbol{y}$ are, in fact, independent. In other words, if the data set in $\boldsymbol{y}$ would consist of random observations of Y. But if there is a relationship between $\boldsymbol{x}$ and $\boldsymbol{y}$, $\boldsymbol{y}$ would instead be a set of observations of $Y|X$!

For example, say that A occurs mostly uniformly between days 50 and 80, whereas *independent* B occurs uniformly between day 10 and 80. Now if B depends on A such that B occurs close to A, eg. within 1 unit of A 50% of the time, then $P(50 < B < 80) \geq 0.5$, i.e. we will have more of B between day 50 and 80, and B is no longer uniform. In this case, the probability of B occurring between 50 and 80 would be overstated when calculating the IDD.

Due to this issue, when we do find a dependency, we will use the IDD thus constructed only as a first-pass estimate of the true IDD. When a dependency exists, there will be many more observations around the peak (of zero) than expected from the IDD. We remove random observations with time differences in the section where the probability of observations in the ODD exceeds the probability of the observations in the IDD. We then have a new set of observations, the original minus the removed, and recreate the IDD from this. We compare the two again and, if necessary, remove more observations. We repeat this until we no longer find a dependency between the two events.

This process is illustrated using the dependent data set of example 1. Figure 5 shows the relative frequency distributions of $\boldsymbol{y - x}$ (the ODD) versus the original IDD constructed on our simulated dependent data set.

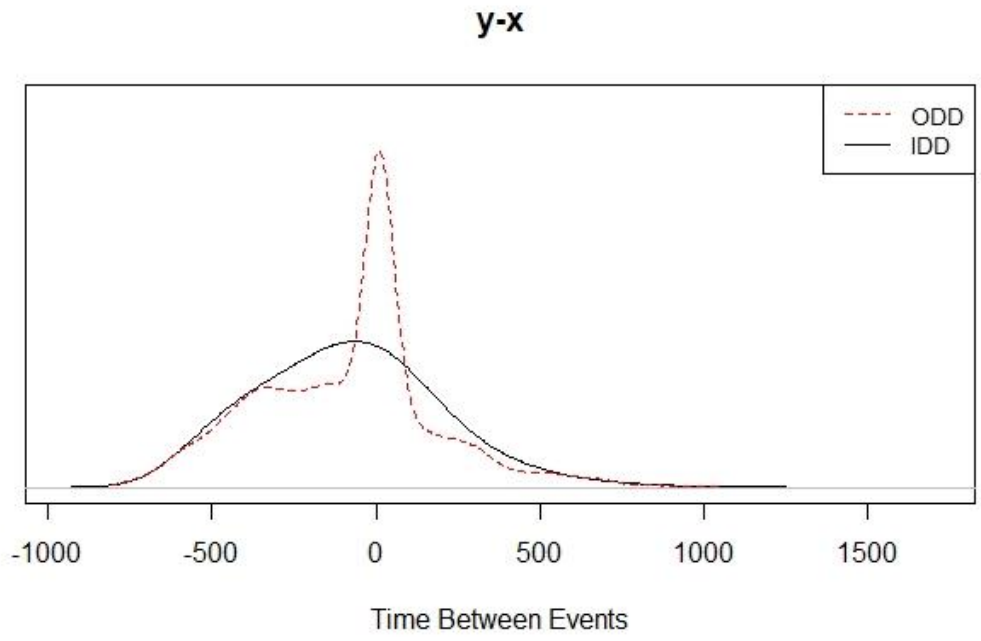

Figure 5: y-x with First Pass IDD

It is clear from this picture that the ODD does not follow the IDD, and so there is a dependency. Approximately, the section between the solid line and the dashed line where the dashed line lies above the solid one indicates observations that should not be there if the data was independent. We therefore remove random observations with time differences that fall between these lines, resulting in a new data set of xs and ys that is likely to be more independent than the original, and create a new IDD from this data set. Note, that the observations to be removed are chosen randomly, so if two observations are very close with respect to the time difference between the occurrences of the two events, one may be removed whereas the other one may not. For interpretation purposes,

what matters is the resulting distribution, not the individual observations, as there is no way to tell exactly which observations are dependent and which ones are not.

We repeat this process until the "cleaned" data set no longer contains enough evidence of dependence to reject the null hypothesis of independence. Figure 6 shows the final estimated probability density of the IDD with the (original) ODD.

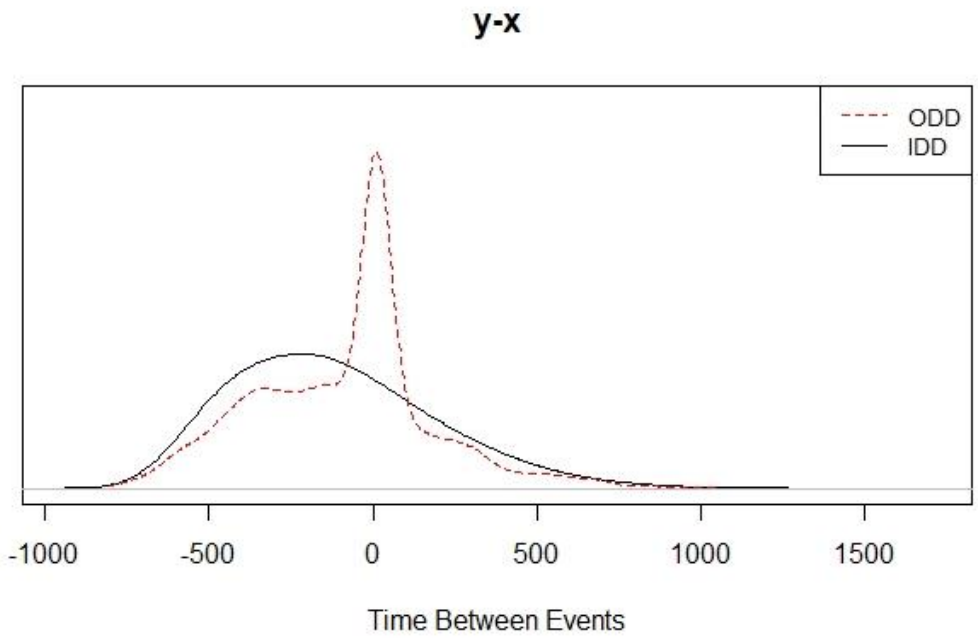

Figure 6: y-x with Final IDD

In addition, we end up with a new data set with all suspected dependent observations removed, which we may use to determine the proportion of likely dependent observations as well as create a visualization as to what the individual distributions may look like had they been independent.

3.3.3 Comparison

In order to obtain a quantitative result, we determine the probability that the ODD is a sample from the IDD. For this, we use the bootstrap method which was originally introduced by (Efron, 1979).

First, we take the absolute value of the difference between the areas under the curve of the two estimated density functions:

$$D = \int_{-\infty}^{\infty} \left| \hat{f}_T(t) - \hat{f}_O(t) \right| dt$$

We define

$$h(t) = \left| \hat{f}_T(t) - \hat{f}_O(t) \right|, \qquad a = \min(t), b = \max(t)$$

We split the distance between a and b into n equally sized steps, where n is even:

$$a = t_1 < t_2 < \cdots < t_n < t_{n+1} = b$$

We then use Simpson's Rule (Larson & Edwards, 2010) to estimate the absolute difference in area D between the two distributions

$$D = \int_a^b h(t)\, dt \approx \frac{b-a}{3n} [1\ 4\ 2\ 4\ 2 \cdots 4\ 2\ 4\ 1] \begin{bmatrix} h(t_1) \\ \vdots \\ h(t_{n+1}) \end{bmatrix}$$

this will give us the absolute difference in area between the ODD and the IDD.

We then compare this difference to the expected absolute difference in area: we obtain a number $(= bs)$ of bootstraps on the IDD, of the same size as the ODD. Each bootstrap b_i

results in an estimated density function for the bootstrap, $\hat{f}_{b_i}(t)$. We calculate D_i^*, $i = 1, \cdots, bs$ for each in the same manner (using Simpson's Rule) as before:

$$D_i^* = \int_{-\infty}^{\infty} \left| \hat{f}_T(t) - \hat{f}_{b_i}(t) \right| dt$$

The D_i^* so obtained will give us a distribution of expected differences in area for samples from the IDD.

Finally, we obtain a nonparametric p-value for the likelihood that the ODD is a sample from the IDD

$$p\text{-value} = \frac{\#\left(D_i^* \geq D\right)}{bs}$$

If it is likely (judging by the *p*-value, which will be relatively large) that the ODD is a sample from the IDD, then we do not have enough evidence that event A is related to event B, the null hypothesis of independence cannot be rejected, and the analysis is complete.

If it is unlikely (low *p*-value) that the ODD is a sample from the IDD, then there appears to be a relationship between the two events. The nature of this relationship, however, still needs to be investigated further, so further analysis will be required.

3.3.4 Further Analysis

At this point we know that there is a relationship between events A and B. In order to find the most common time difference between the events, we simply determine the peak at which the difference between the probability of the ODD and the probability of the IDD is the largest. In many cases this will be zero, i.e. when there is a relationship, event B will most often appear right before or after event A. For the remaining analysis we will concentrate our attention on a peak of zero; further analysis with respect to different peaks is referred to a later time (see Section 3.8).

With event B likely occurring close to the occurrence of event A (peak = 0), we wish to find the cutoff interval, i.e. the interval outside of which we no longer have enough evidence that two events are related. We refer to the limits of a cutoff interval as the cutoff points. Since we have a peak equal to about 0, we should have a nonnegative cutoff point as well as a nonpositive one. For the nonnegative one, event A occurs first, while for the nonpositive one, event B occurs first.

To start with, we attempt to find the nonnegative cutoff point. We repeat our initial analysis, but only use data for B that occurred more than a specific number of days (t) after A occurred, and once again compare the distribution of differences to the (equally restricted) IDD. We compare the estimated probability densities for $O|O \geq t$ and $T|T \geq t$ for some $t \geq 0$.

We run the analysis with several different time differences t, spaced out over the range of time differences between the events starting with $t = 0$, and register the p-value for each.

When the p-value exceeds 2.5% (half of 5% since we have two directions), we have reached the desired cutoff point since beyond that we no longer have sufficient evidence of a relationship between the two events.

We repeat the same process for the nonpositive cutoff point but reversing the positions of A and B.

We will illustrate this using example 2. In Figure 7a) we see that there is a relationship between AF and HF. In Figure 7b) we let $t = 0$, and graphed the estimated probability densities for $O|O \geq 0$ (ODD) and $T|T \geq 0$ (IDD). For these restricted densities it also appears highly unlikely that the ODD data set is a sample from the IDD, shown again by the fact that the dashed red line is very different from the solid black one. To find the (nonnegative) cutoff point, we need to find a t such that this is no longer the case, i.e. where the ODD is likely to be a sample from the IDD.

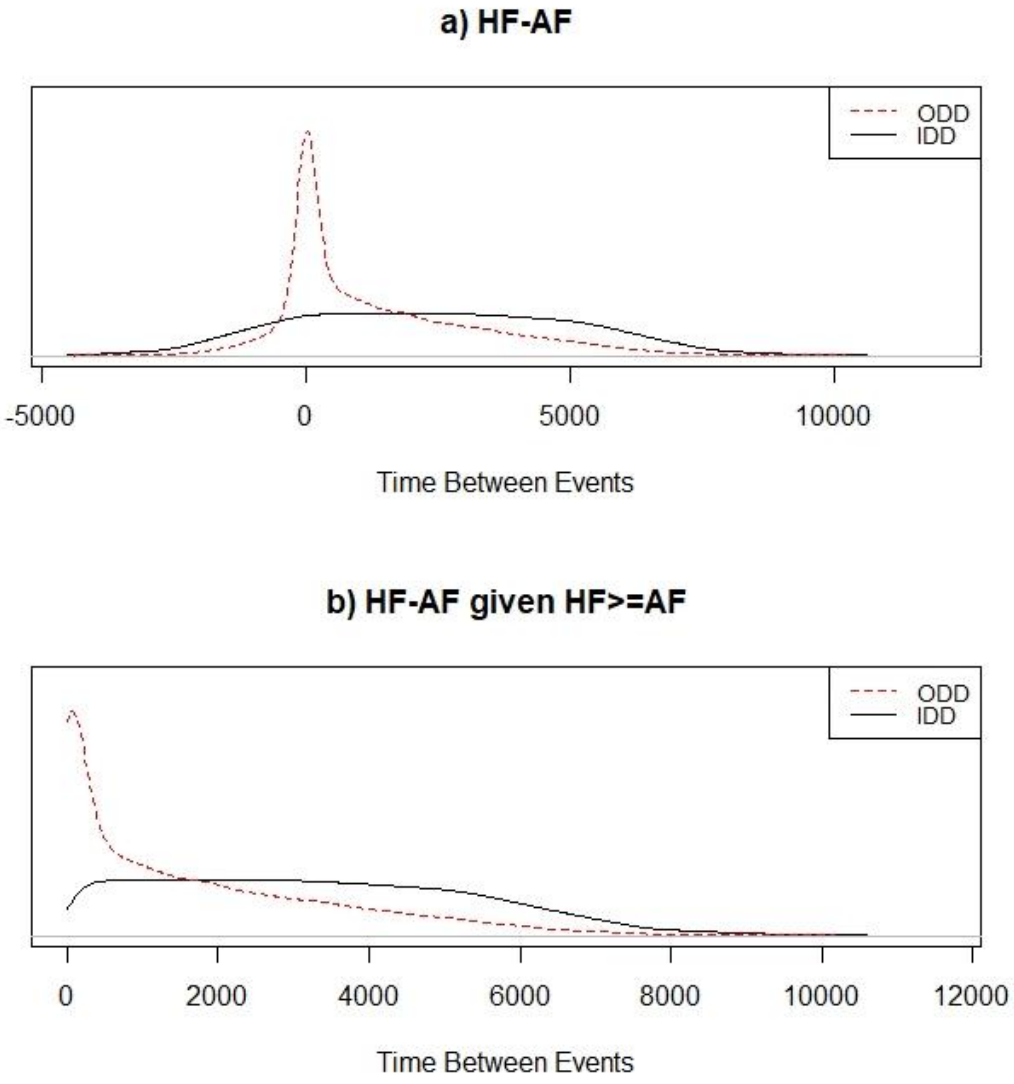

Figure 7: HF-AF in Full vs. Restricted

Figure 8 shows the graphs of the p-values we obtained when running the analysis with several different time differences t, where the dashed red line is drawn at a p-value of 0.025. Figure 8a) shows the p-values when AF occurs first, giving us the nonnegative

cutoff point, whereas Figure 8b) shows the p-values when HF occurs first giving us the nonpositive cutoff point.

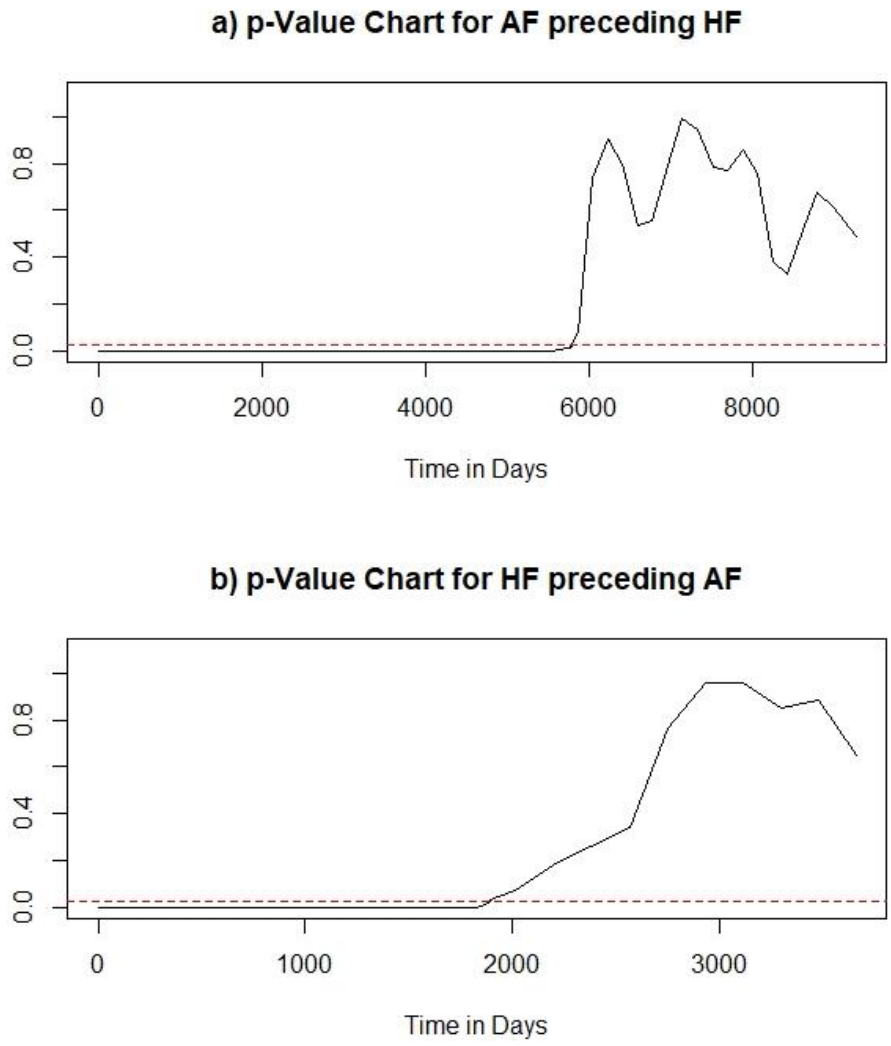

Figure 8: p-Value Charts

In a), after approximately 5,800 days it starts to become less likely that the two events are related. The cutoff point would be the point of intersection between the graph and the dashed line. In this case, we found it to be 5,784 days, or about 16 years.

In b), the cutoff point is determined to be 1,885 days, or just over 5 years. This means that some occurrences of AF that happened within 5 years after HF are likely related. Combined, we say that the cutoff interval is (-1885, 5784), containing the HF-AF differences between which we are reasonably confident (at least 95%) that there is a relationship.

It is important to understand that the cutoff interval gives us the interval outside of which we no longer have sufficient evidence of a relationship between the two events; however, this does NOT mean that at any time within that interval, a large number of observations are dependent.

In example 2 the nonnegative cutoff point is 5,784. However, based on the dependent observations that the process removed when creating the IDD, only about 15% of observations in which HF occurred more than 5000 days after AF, are dependent.

When determining whether a relationship exists based on this process, both the cutoff interval and the proportion of dependent data should be considered. This is why we will use the relation interval.

3.3.5 The Relation Interval

We wish to find a more reasonable interval for determining whether we would consider there to be a relationship between two events, which we denote as the ***relation interval***, limited by two ***relation points***.

The relation interval may be determined manually via the Shiny application, which allows the user to try out different relation points and see the accompanying estimated proportion of dependent observations. A medical expert would be able to augment their own knowledge of the various events and their effects with the information obtained from the data and should be able to make an informed decision as to where the relation points should lie for each pair of events.

Alternatively, if such experts are not available, we may set a specific proportion of dependent observations needed for a time difference to be part of the relation interval, and we use that proportion to automatically calculate those intervals for all pairs of events from a data set where each observation contains all events and the time they took place (or NA if they never took place for that observation).

3.3.5.1 Automatic Determination of the Relation Interval

For each pair of events in the data set, we select all observations for which a time is available for both events. We determine if there is a relationship using the method described previously. If there is no relationship, the relation points are both zero.

If there IS a relationship between the two events, we perform the process as described before to obtain the true IDD and remove the observations that are estimated to be dependent. This will give us two data sets; one with all the observations, and one with only the independent observations.

We then produce estimates for the proportion of dependent observations at various time differences. We do this by calculating running proportions for blocks of data that overlap significantly and calculate those at a reasonably large number of points.

Using interpolation, we can then determine at what point in time the proportion of dependent observations becomes too small as compared to a user-selected proportion.

We used the dependent data set from example 1 and set the required proportion to one half. Figure 9 shows a histogram where the full data set is shown with the independent data colored red.

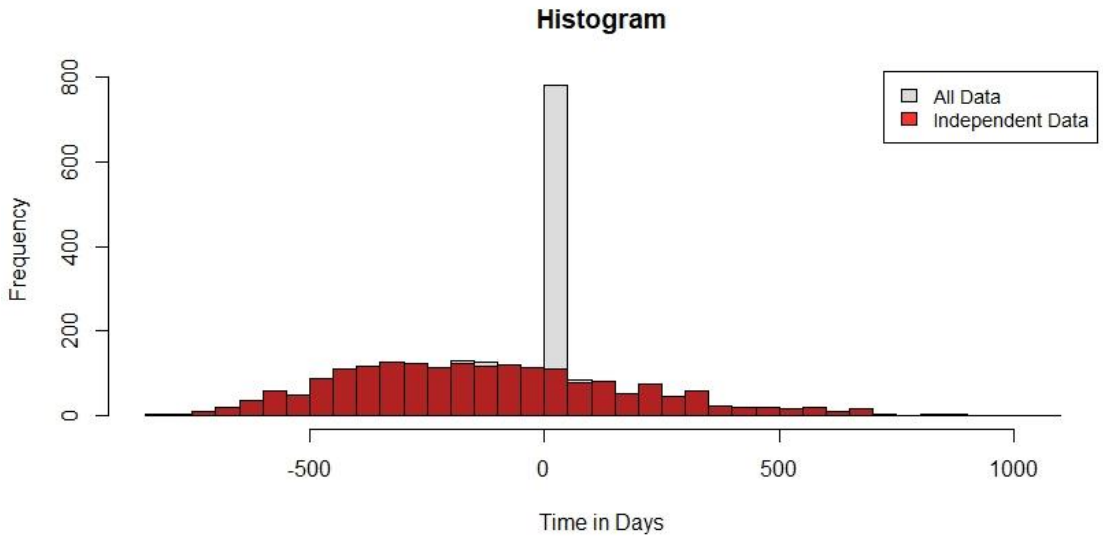

Figure 9: All Data vs. Independent Data for Example 1

We used 76 data points spread over the set of time differences and created a proportion for each of them using the data between the data point and the next 32 independent observations. We found the relation interval in this case to be (0, 27).

We also performed the same analysis on example 2. Figure 10 shows the histogram for that data set.

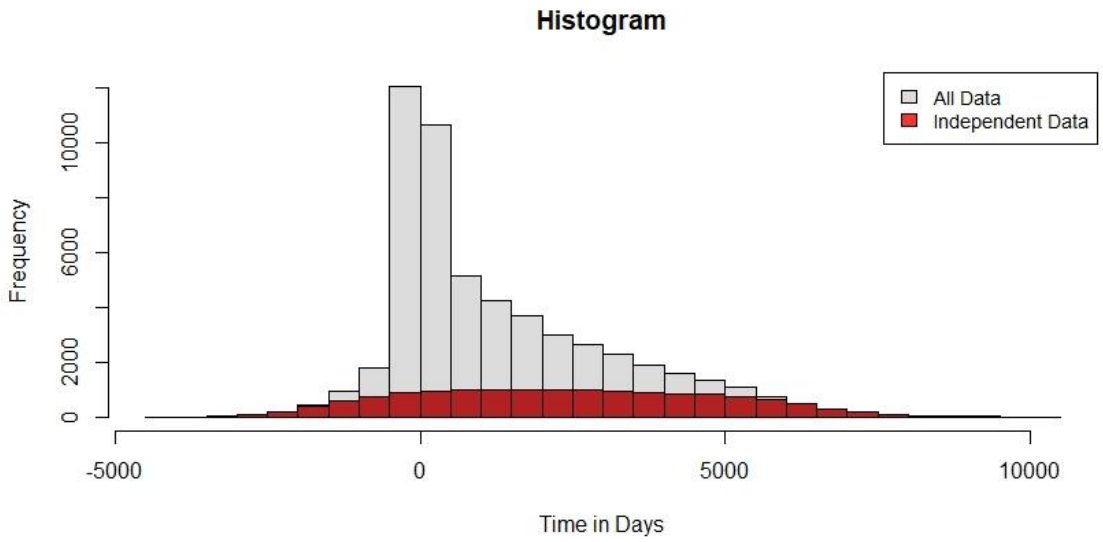

Figure 10: All Data vs. Independent Data for Example 2

We set the proportion once again to one half, used 76 datapoints spread over the range of time differences, and used all data from each data point to the next 607 independent observations to calculate the proportion of dependent observations. We found the relation interval for this case to be approximately (-800, 3800).

In other words, if a patient experienced AF within about 2 years after experiencing HF, we would add a connection from AF to HF in the patient's PDN. If the patient

experienced HF within about 10 years after experiencing AF, we would add a connection from HF to AF in the patient's PDN. If the time difference was longer than that, we would not include a connection between the two.

3.3.6 Method Summary

We start with two data sets $\boldsymbol{x}$ and $\boldsymbol{y}$ that indicate time to two separate events in subjects who experienced both. We perform the following tasks:

- Create the ODD data set $\boldsymbol{y} - \boldsymbol{x}$
- Create the first pass IDD by permuting $\boldsymbol{x}$ and $\boldsymbol{y}$
- Compare the ODD with bootstraps of the IDD; if it appears the ODD could be a sample of the IDD, the process is complete and the null hypothesis of independence is not rejected.
- If it appears unlikely that the ODD is a sample of the IDD, the null hypothesis of independence is rejected. In addition, we note that the IDD was built with dependent data and as such does not consist of truly independent differences. A correction is necessary.
- The IDD is corrected and new, independent, data sets are created for $\boldsymbol{x}$ and $\boldsymbol{y}$ based on the originals with estimated dependent observations removed.
- Using the corrected IDD and the original ODD, the cutoff interval is established.
- The relation interval is determined by either of the following two methods:
    - Experts using the Shiny application

- Providing a proportion of dependent data that should be present at any point inside the relation interval which is then used by our automated system to provide the relation interval

Any subject who experienced both events, for whom the difference between the two events lies within the relation interval, is estimated to have experienced two related events rather than two independent events.

Note that the user-supplied proportion of dependent observations needed within the relation interval is by default set to .5, suggesting that at any point within the interval at least half of the observations are dependent. Another good proportion would be .8, which would ensure that at any point within the interval most observations are dependent. The appropriate choice for this proportion should be made by experts in the field.

3.4 Simulations (Proof of Concept)

Using example 1 as a starting point we created several simulations. Recall that the independent data is drawn from $X \sim Weibull(\alpha = 6, \lambda = 700)$ and $Y \sim Weibull(\alpha = 2, \lambda = 600)$. Figure 11 shows the ODD and the IDD. Since the dashed red line is very close to the solid black line, it is clearly likely that the ODD is indeed a sample from the IDD. The *p*-value equaled .50, indicating it is likely that the ODD is a sample of the IDD.

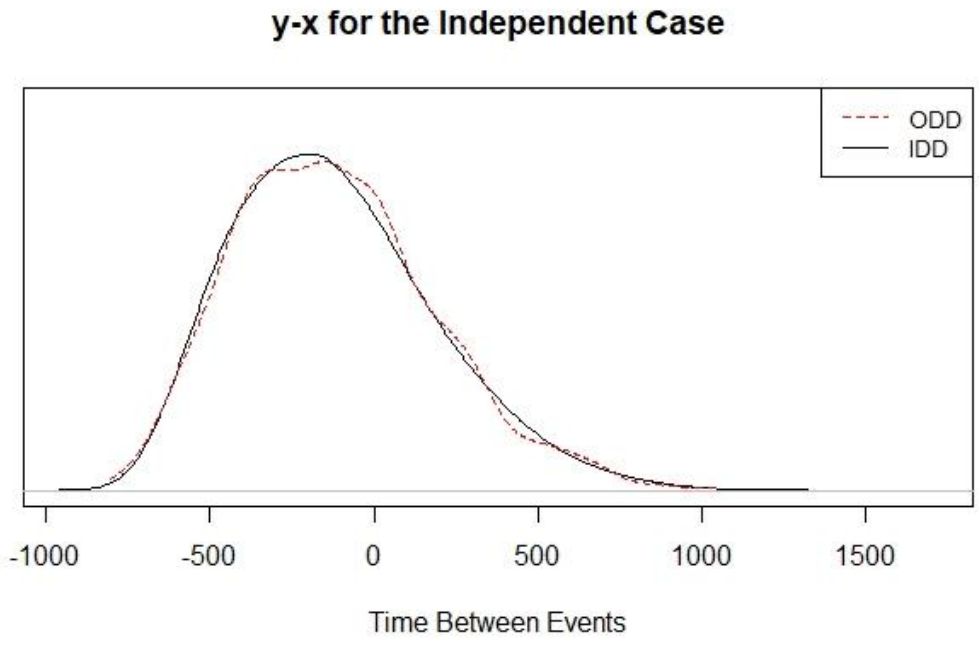

Figure 11: Time Differences for the Independent Case of Example 1

When analyzing the dependent data set the application removed 694 observations, close to the number of dependent observations that we added (we added 656). We saw that the "cleaned" data looked very similar to the independent data set we started with (see Figure

3). In Figure 6 we saw a comparison of the IDD with the ODD for this case. Figure 12 shows the estimated densities of the IDD and the ODD 23 days out ($t = 23$). It is clear that beyond 23 days the ODD could indeed be a sample of the IDD.

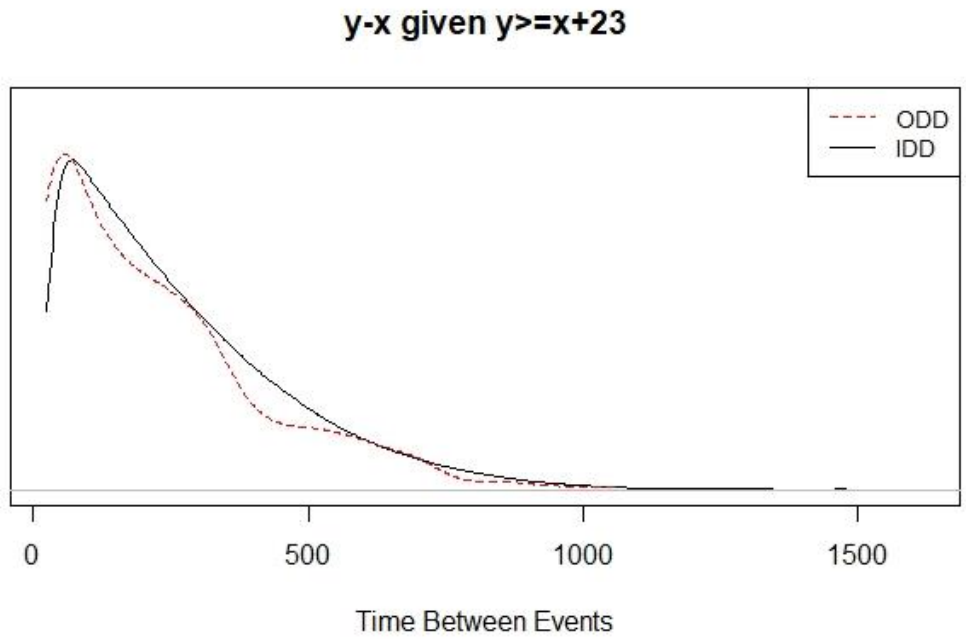

Figure 12: Restricted Time Differences for Example 1

We then added another 656 dependent observations where B preceded A between 0 and 30 days (more towards 0) to the dependent data set, for a total of 1312 dependent observations and 2000 independent ones. We again rejected the null hypothesis and obtained a cutoff interval of (-22, 23) days.

Next, we started from the independent data set and added 656 dependent observations, but this time we drew from $W|W \leq 150$ where $W \sim \exp(\lambda = \frac{1}{100})$, thus creating a data

set with a dependency that goes further out. We once again found a dependency, this time with a cutoff interval of (0,115) days.

To simulate a situation with a great many dependent observations even further out, we once again started with the independent data set and added 5248 dependent observations, eight times the number of independent observations with differences greater than zero. We drew from $W|W \leq 400$ where $W \sim \exp\left(\lambda = \frac{1}{200}\right)$. We found a dependency and a cutoff interval equal to (0, 384) days. 5193 observations were removed, which is 72%, and slightly less than the 5248 dependent observations we added.

In order to determine how well the process works, we created the independent data set 1000 times, analyzed each of them and recorded their p-values; Figure 13 shows those p-values in two separate plots.

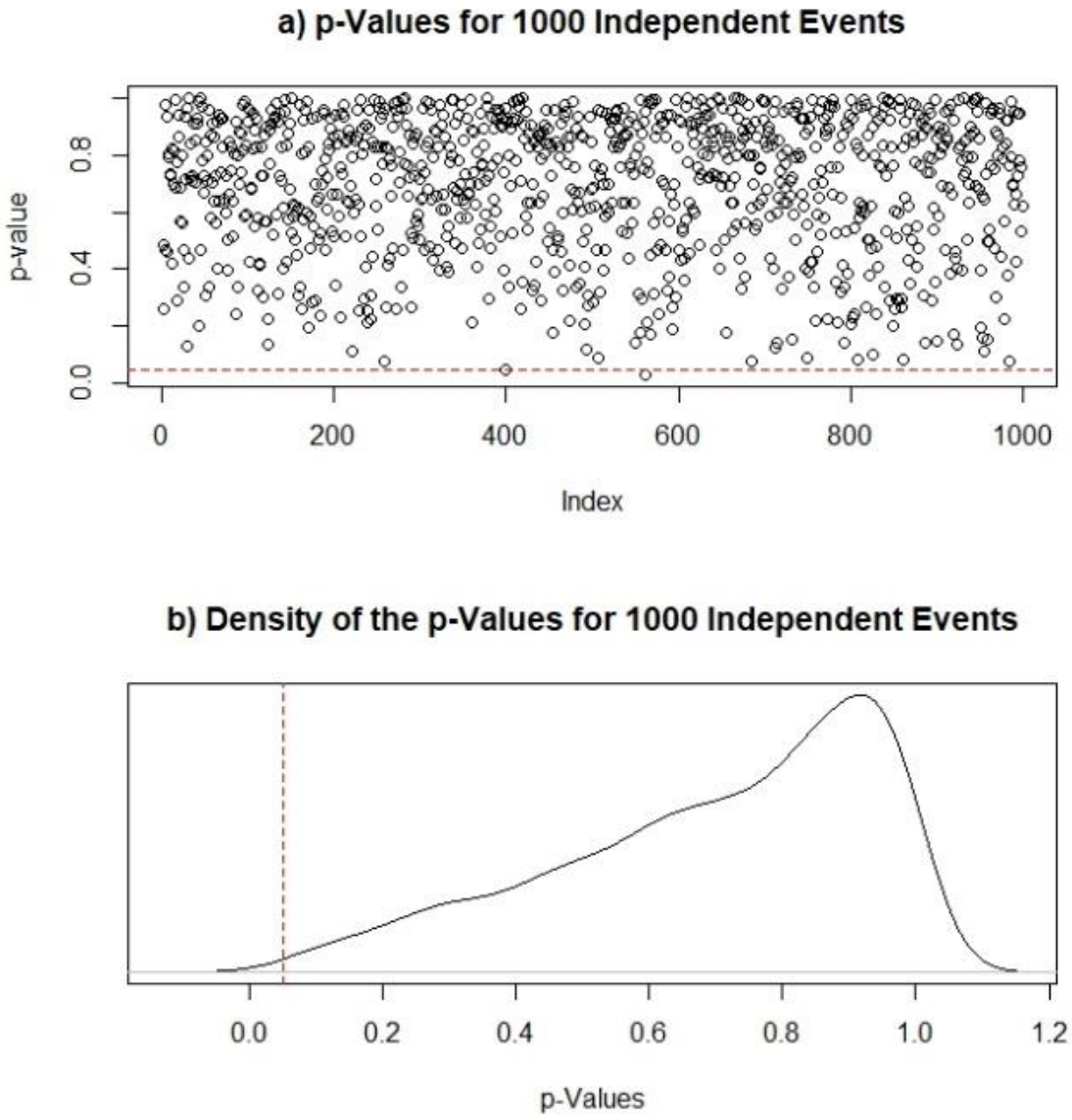

Figure 13: p-values for 1000 Independent Events

The dashed red line is drawn at a p-value of 5%. Figure 13a) plots every p-value, while Figure 13b) shows the relative frequency distribution for the p-values. We found that 2 p-values, or .2%, were less than 5%. In other words, there is about a .2% chance of a type I

error, i.e. there is a .2% chance that our method will falsely reject the null hypothesis of independence.

Note that the p-value indicates the probability that a sample taken from the independent time differences between the two events is as different or more different than the observed time differences between the events.

We also analyzed 100,000 independent observations from the two Weibull distributions to get an excellent estimate of the true IDD for independent differences between them. We then analyzed the independent data set versus this IDD 1000 times. We again recorded the p-value for each; Figure 14 shows those p-values in the same two separate plots.

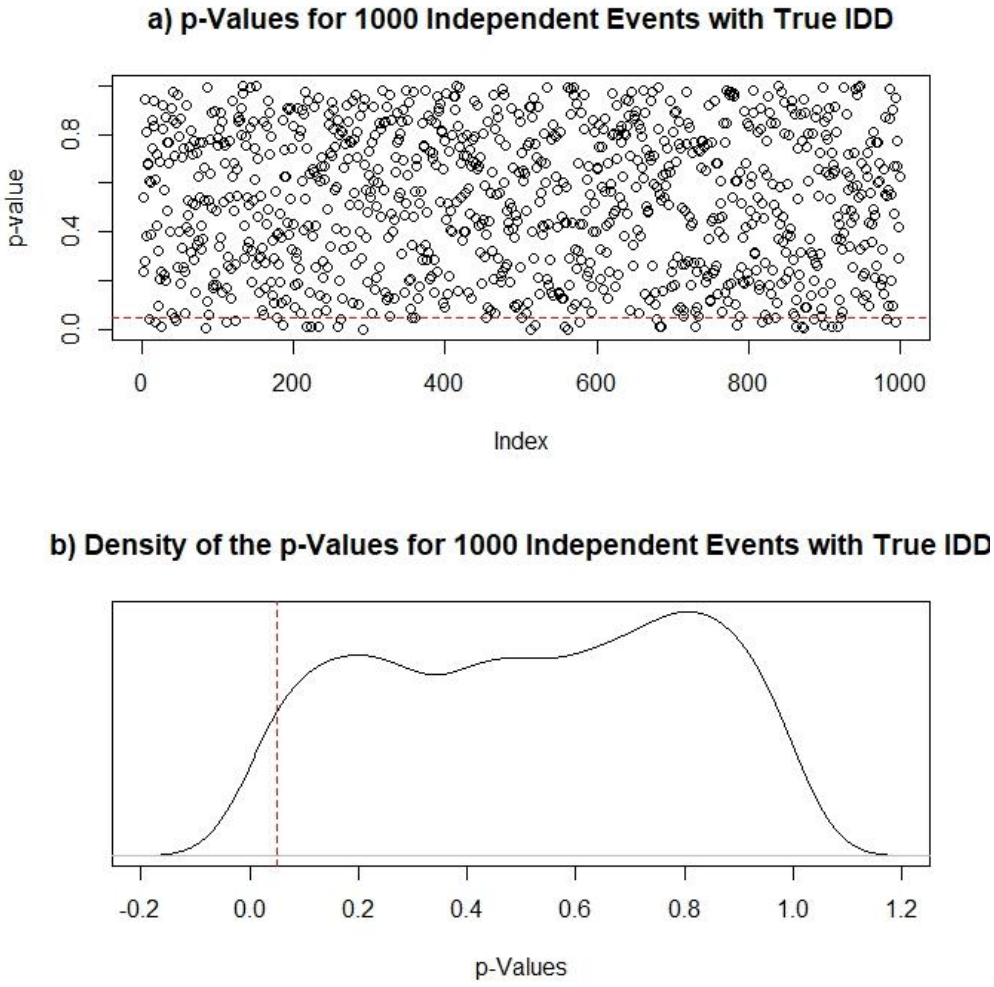

Figure 14: p-values Calculated from the True IDD

In this case we found that 43 *p*-values, or 4.3%, were less than 5%. This is when we have access to the true IDD, instead of an estimate obtained from the observed differences. There is a bias towards independence due to the IDD being an estimate constructed from the observed data rather than the true IDD. Note that generally the true IDD is unknown.

We also created a situation where we started with the independent data set and added 200 dependent observations where x was simulated from X and $y = x \pm w$, where w was simulated from $W|W \leq 150$ and $W \sim \exp(\lambda = \frac{1}{100})$, resulting in about 9% of the data being dependent with time differences between the events being between -150 and 150 days. We did this 100 times and found that in 92% of the tests the dependency was discovered. This situation is shown in Figure 15.

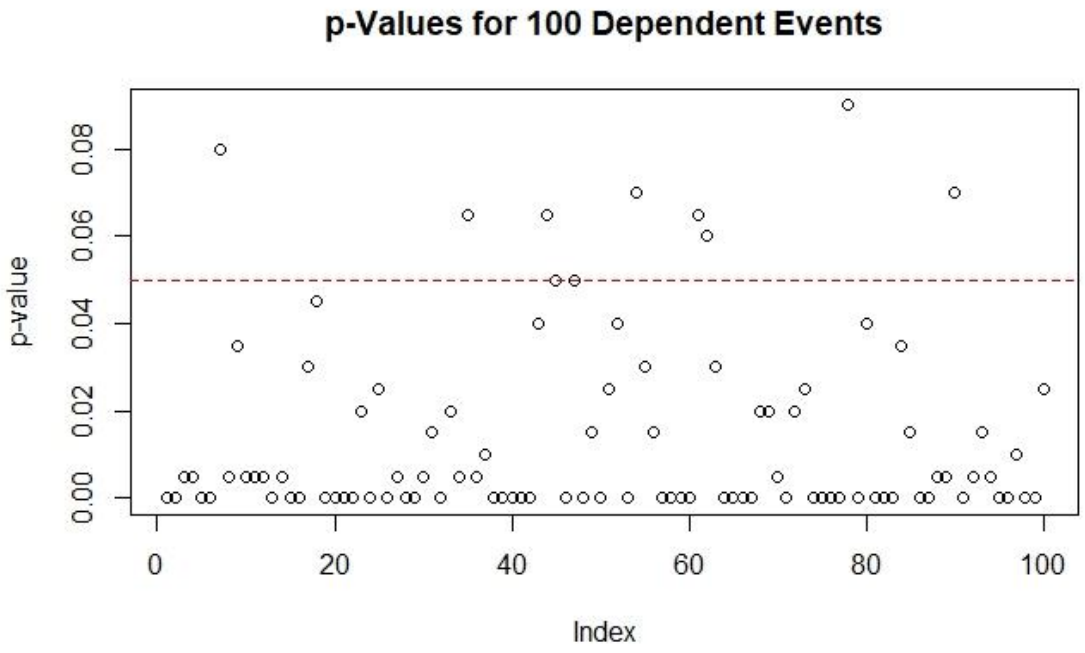

Figure 15: 100 Data Sets with a 9% Dependency

Figure 16 shows the density for the p-values in the same situation.

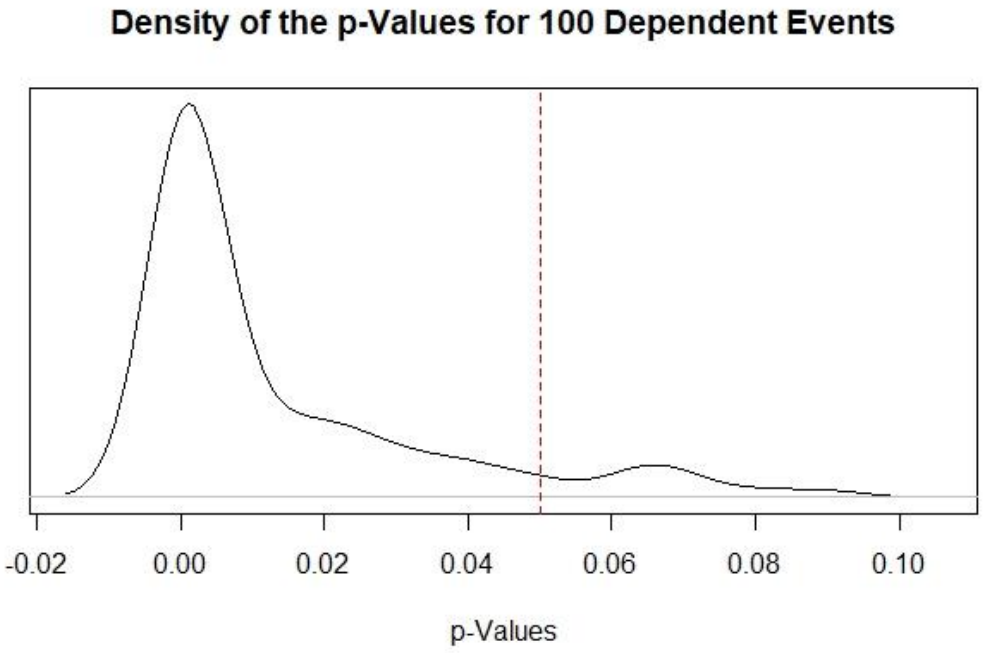

Figure 16: p-value Density for 100 Data Sets with a 9% Dependency

When we increased the number of dependent observations to 300 so that 13% of the data was dependent, 100% of the tests found the dependency.

3.5 Shiny Implementation

The Shiny application takes a text file with two columns, the first with time to the first event, the second with time to the second event. The header of the file (first line) should have (short) names for the two events. Processing will then proceed on the data according to the previously described method. It will be determined whether there is enough evidence of a relationship between the two events, and if so, the cutoff points will be calculated and a data set with estimated independent observations created. Note that this is a somewhat time-consuming process; the bottom right corner will display a progress bar and will also show the task being performed. The tasks are:

- Getting the IDD
- Getting Steps (determine the p-value for different values of t)
- Determine Cutoff Points

When this process is complete, the file ae.RData is created. This file may be entered into the application at a later time in order to access the results immediately. It is recommended, however, to rename this file in order to prevent it from being overwritten during a subsequent analysis.

3.5.1 The Main Tab

The application will then display the results of the analysis on the Main tab of the program: the number of observations in the file, the p-value indicating the probability

that the two events are independent, and the cutoff interval. It will also show a graph displaying the two estimated densities of time to the events and a graph displaying the ODD and the IDD. Finally, if a relationship is found, a graph of the two estimated densities of time to the events after dependent observations have been removed. The number and percentage of observations determined to be dependent and thus removed is also displayed. Figure 17 shows this information for example 2.

Figure 17: Shiny Main Tab for Example 2

3.5.2 The Details Tab

Additional information is available on the “Details” tab. The "Details" tab will split the information into two parts, one where *B* occurs first, and one where *A* occurs first. In Figure 18 we see this information displayed for the data in example 2.

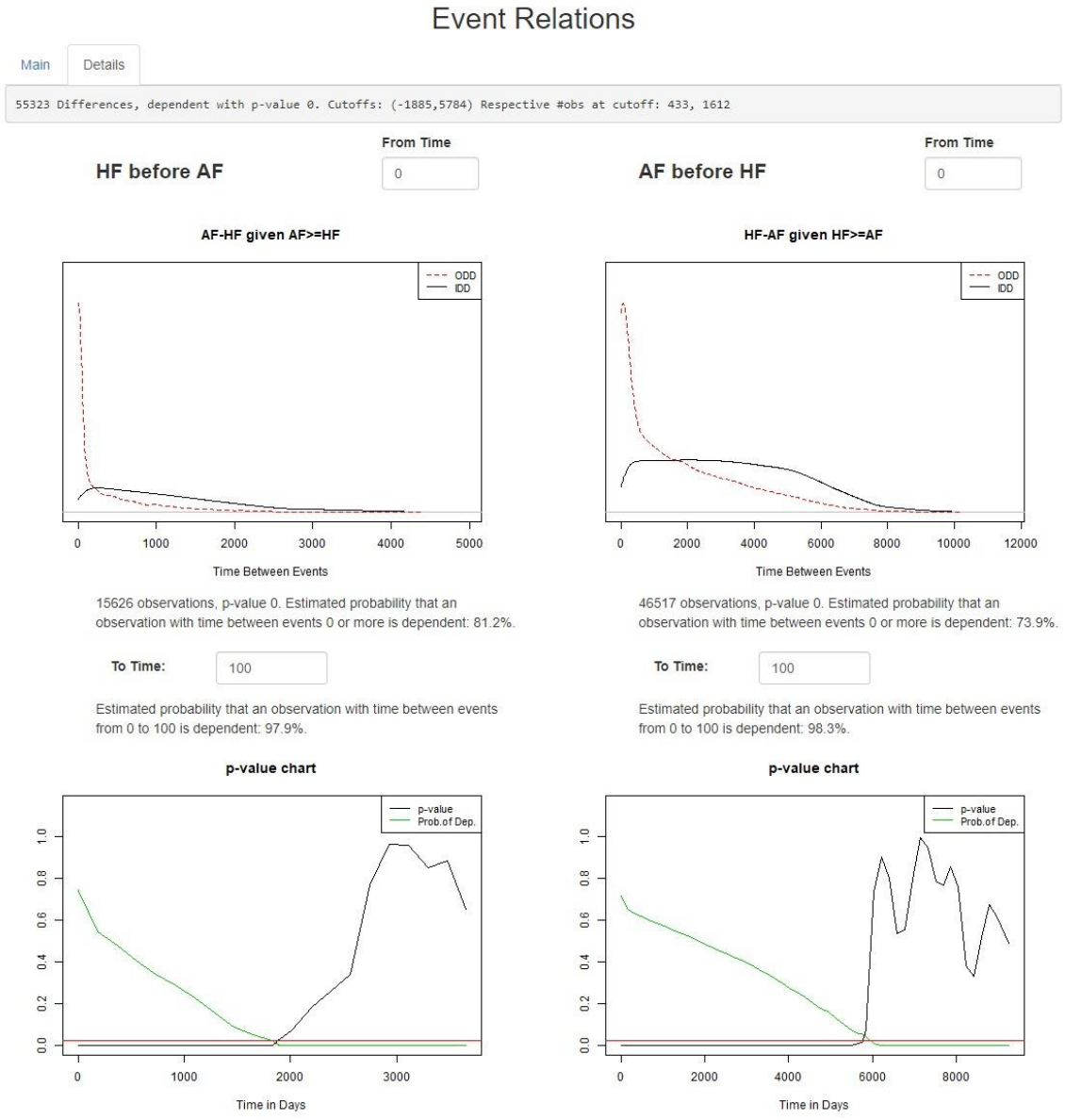

Figure 18: Shiny Details Tab for Example 2

The user may enter a time in either or both of the "From Time" fields and in either or both of the "To Time" fields. The first graphs as well as the information underneath them is dependent on these fields which we explain using example 2 in Figure 19.

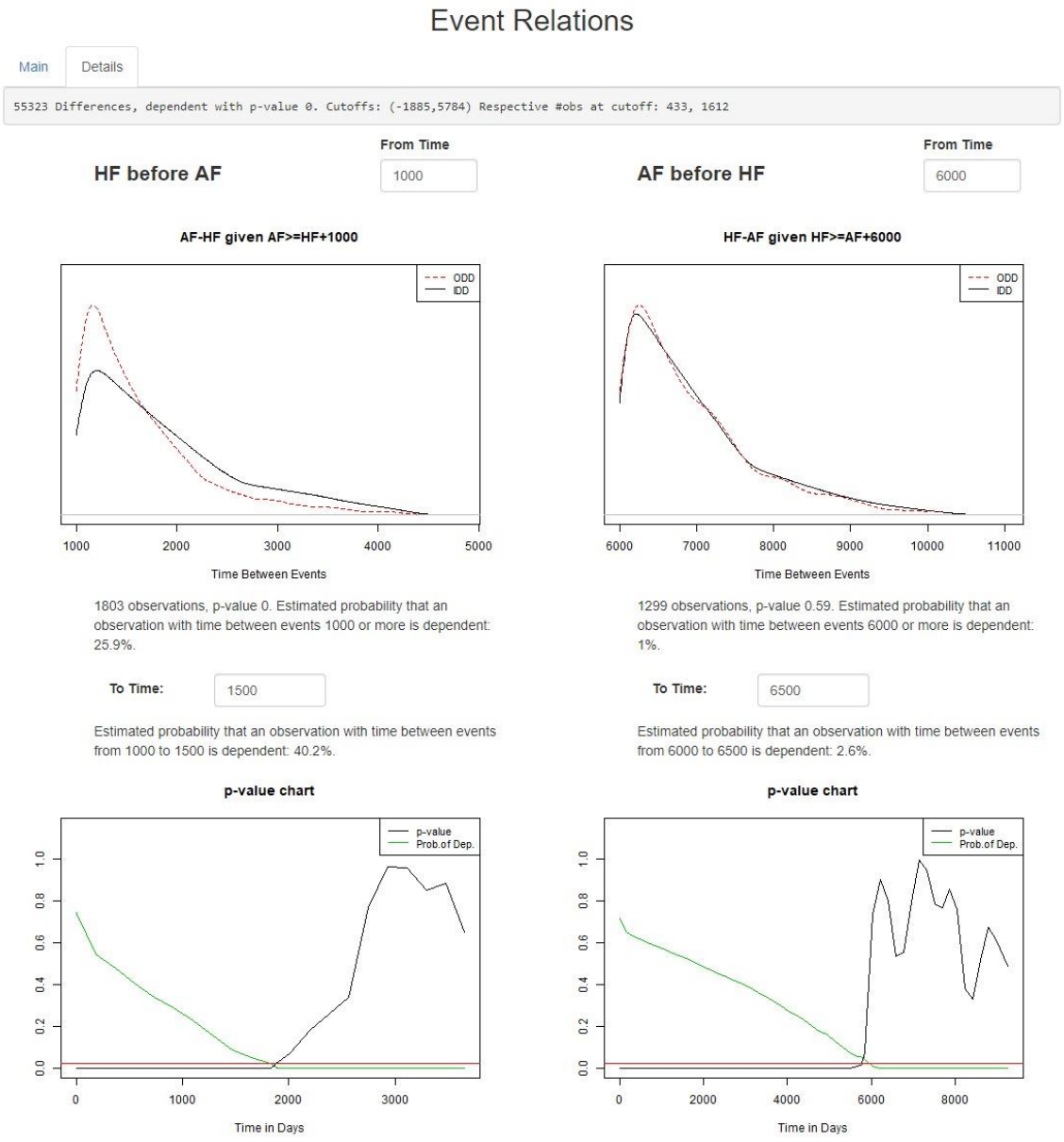

Figure 19: Shiny Details Tab with Different Times

The first graph on the left displays the time differences when AF occurs at least 1000 (the number in "From Time") days after HF. As before, the dashed red line gives the time differences as observed in the file, while the solid black line indicates the time differences when the two events are independent (the IDD).

The information underneath the graph tells us there were 1803 observations where AF occurred at least 1000 days after HF, and the p-value, i.e. the probability that the ODD (red line) is a sample of the IDD (black line), is 0% (rounded). We see that the red line is not close to the black line. It also tells us that the probability that an observation with AF occurring 1000 days or more after HF has a probability of approximately 25.9% of being dependent.

Under the "To Time" field we note that an observation with time between events from 1000 to 1500 (the number in the "To Time" field) days has a 40.2% probability of being dependent. Note that this number is an estimate obtained from the data; it will be most reliable for large sets of time differences, and unreliable for small ones.

Finally, the p-value chart is displayed; the black line gives the p-value indicating the probability that the data is dependent when considering only data equal or larger than the Time in Days on the x-axis. We see here that, for example, this p-value equals about 0% for the data when AF occurs 1000 or more days after HF, and it equals about 35% when AF occurs 2500 or more days after HF; at this latter point there is not enough evidence of a dependency. The green line gives the probability that an observation with AF occurring the number of days in "Time in Days" or more after HF is dependent. We see here, for example, that an observation with AF occurring 1000 days or more after HF has a probability of about 30% of being dependent, while at 2500 days that probability equals

about 0%. The red line is drawn at the significance level of 2.5% (5% split into two for the two directions).

The first graph on the right-hand side displays the observed differences (red line) for those observations where HF occurs 6000 ("From Time") days or more after AF. We see that there are 1299 such observations, and the p-value, i.e. the probability that the red line is a sample of the black line, is 59%. This is clear from the picture, since the red line is very close to the black line. An observation with HF occurring 6000 days or more after AF has a probability of approximately 1% of being dependent. However, since the p-value is as high as it is (59%), we assume that there are no dependent observations with HF occurring 6000 days or more after AF.

We also see that the probability that an observation with HF occurring between 6000 and 6500 days after HF is dependent can be estimated to be about 2.6%. Again, in this case, since the p-value is high, this probability can be ignored; we assume they are all independent.

From the p-value chart on the right we see that for data where HF occurs, for example, 7000 or more days after AF, the p-value, i.e. the probability that the ODD is a sample of the IDD, is about 90% (black line), while the probability that an observation with HF occurring 5000 or more days after AF has a probability of being dependent of about 20% (green line).

3.6 Additional Examples

Another interesting example compares hypertension (HTN) with heart failure (HF). Figure 20 shows the frequency distributions of the two data sets, with a) showing the originals, and b) showing the data sets after estimated dependent observations were removed. Figure 21 shows the ODD vs. the IDD for this data set.

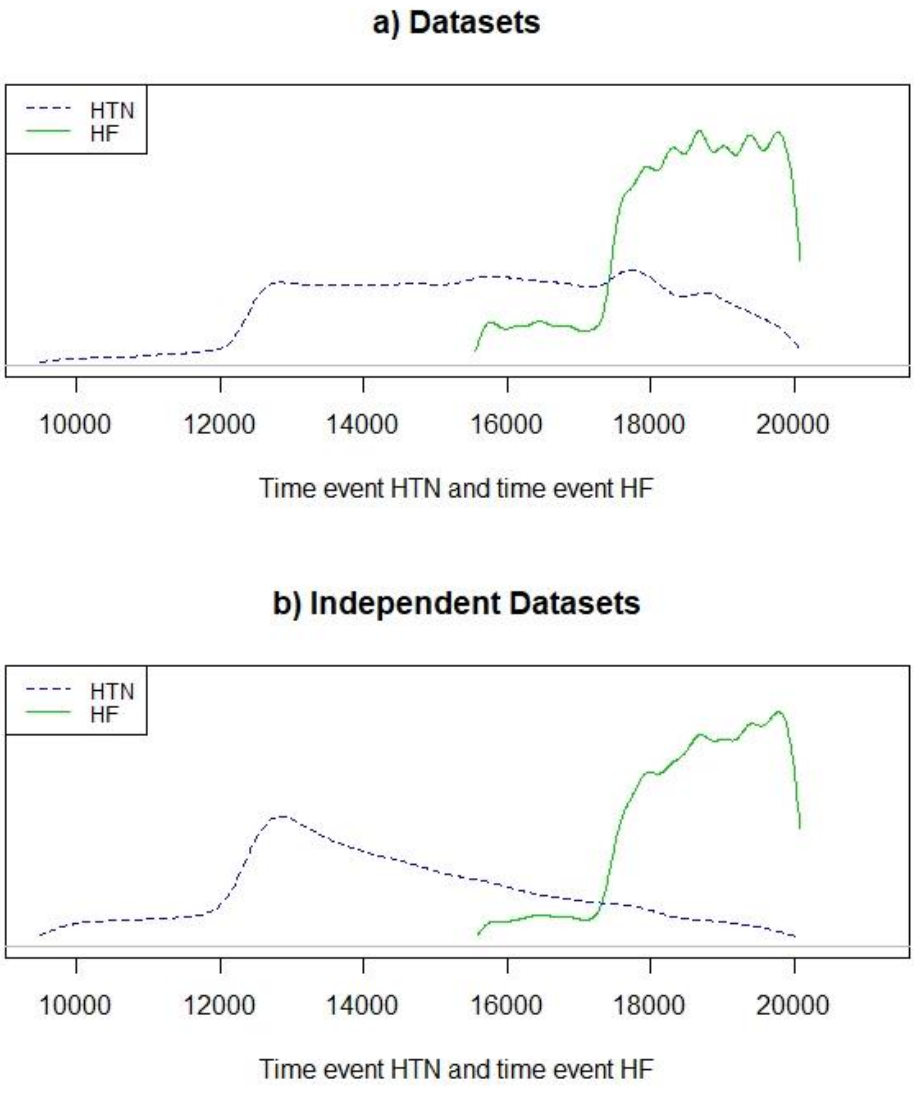

Figure 20: Frequency Distributions for HTN vs. HF

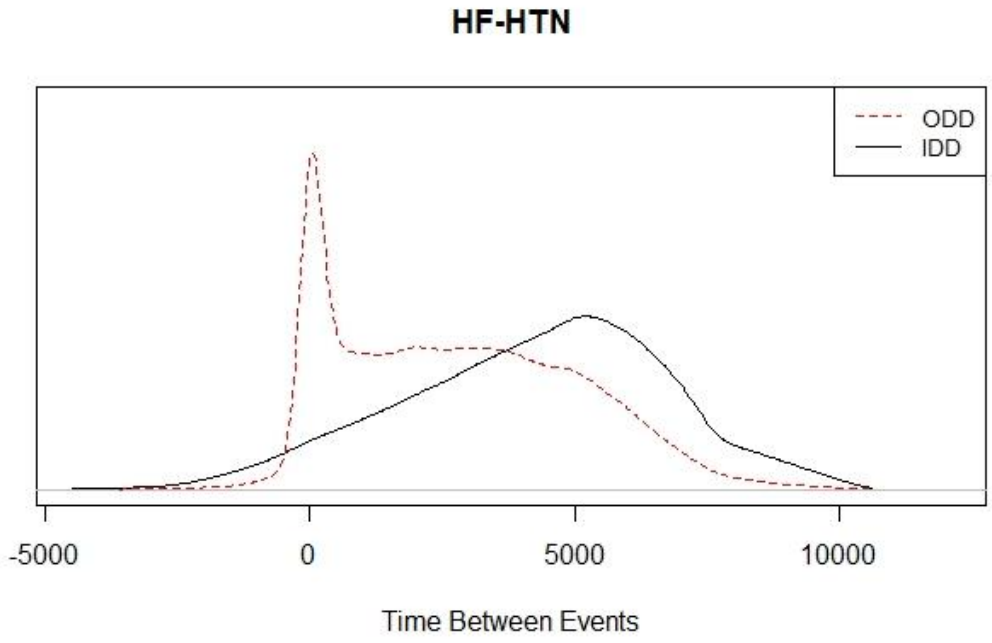

Figure 21: ODD vs. IDD for HTN before HF

We found the cutoff interval to be (-2326, 8128) days, or (-6, 22) years, indicating that for some observations such that HF-HTN lies in that interval, the events are likely related. See also (Drazner, 2011) for more information about the connection between HTN and HF.

Of the 93,162 observations, 72,707 were deemed dependent and removed from the data sets. This amounts to 78% of the data.

Automatic calculation of the relation interval with a dependent observations proportion of one half estimated the relation interval to be approximately $(-400, 6100)$ days, or $(-1, 17)$ years.

We also analyzed the combination of cancer and HF. Figure 22 shows the frequency distributions of the two data sets. While Figure 23 shows the ODD vs. the IDD.

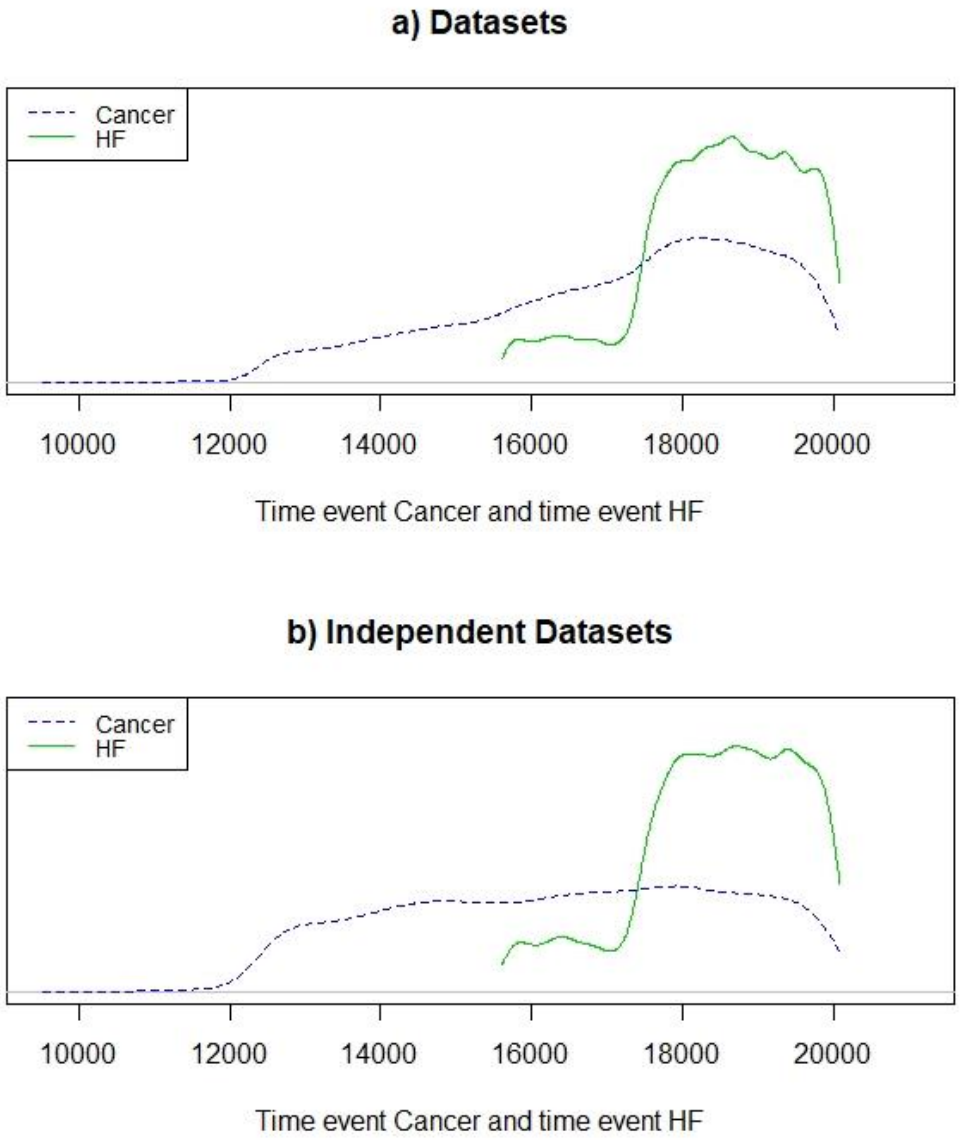

Figure 22: Frequency Distributions for Cancer vs. HF

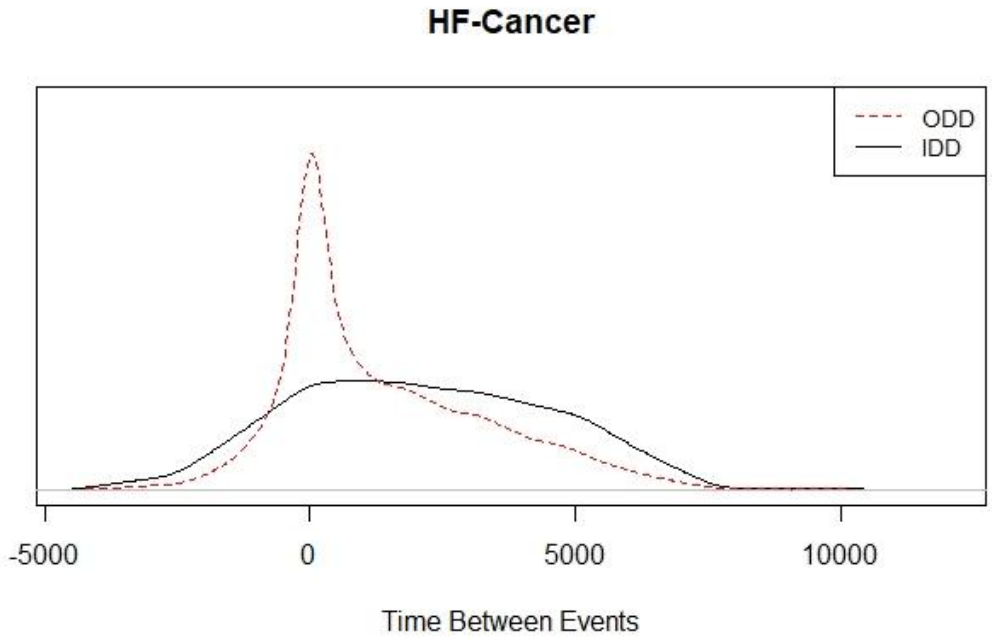

Figure 23: ODD vs. IDD for Cancer before HF

We found the cutoff interval for HF−cancer equal to (-1593, 4774) days, or (-4, 13) years. It should be noted that this indicates that, after a HF diagnosis, the probability of obtaining a cancer diagnosis within the next 4 years is increased. This perhaps somewhat unexpected result confirms studies by (Swerdel, et al., 2014) that discovered increased risk of cancer-related death following rapid decreases in blood pressure such as seen when elderly cardiovascular patients are treated with medication to reduce hypertension. Note, that in this case HF did not cause those occurrences of cancer; there is a relationship, but not a causal one. The likely explanation is that HF caused patients to be treated with medication to reduce blood pressure and the reduced blood pressure increased blood flow that in turn activated dormant cancer cells.

With a dependent observations proportion of one half, the automated calculation of the relation interval estimated it to be approximately $(-600, 1200)$ days, or $(-2, 3)$ years.

3.7 Additional Considerations

For the medical examples it is important to bear in mind that this data comes from the MIDAS database, which currently contains hospital discharge records for cardiovascular events only. As such, non-cardiovascular events may be recorded on the same day as a cardiovascular event, while they could have existed well before that. This may also be the case for many cardiovascular events, since they do not always result in hospitalization. A good example is hypertension, since this is likely to first happen not only well before hospitalization, but even before it's discovered. As such, many of these events are left-censored and should be treated as such. We aim to allow for that as an option per event in future iterations of the process.

Out of 21,135 patients with cancer in this data set, 1,321 are recorded on the same day as their HF diagnosis, which is more than 6%. This supports the above statement; a cancer diagnosis that existed before hospitalization due to heart failure will be entered into MIDAS on the same date as the date of heart failure, since hospitalizations for cancer alone are not recorded in the MIDAS database.

The algorithm described in this paper will be useful in medicine as well as many other fields. For example, one could envisage a system of purchase orders where the purchase of one item may regularly be followed by the purchase of another. One could build a network for each person who authorizes purchases in a company where the events are the purchases, connecting those that are related, determining the relationships using the methods described in this paper. The findings could then help the company predict future purchases.

Note that we now have three methods to establish a relation interval, i.e. the interval of time differences within which we consider two events to be related, and outside of which we consider them to be independent. The three methods are:

- The supervised method as described within the precision network section of this dissertation in section 2.2.1.
- The unsupervised method described in the current section of the dissertation where an expert determines the interval using the Shiny application.
- The unsupervised method described in the current section of the dissertation where several relation intervals are determined at the same time by supplying a proportion of dependent observations that need to be present at any point within the relation interval.

3.8 Future Direction

In order to implement these results into building PDNs (Precision Disease Networks, see Section 2.1.2) cardiology experts may use the Shiny application to determine the relation interval, based on the obtained cutoff interval as well as the proportion of dependent data. We found it impossible to try to get these timings from the experts previously because there are too many variables for them to decide on, and because different experts may not agree. However, if we provide them with our estimates it will be much easier for them to reach a consensus. Then when creating PDNs for patients, we will determine that a link exists between two events if the second event occurred within the relation interval set by the experts. If the experts are not available to do this for a large number of pairs of events, we may use our automated process to determine these relation intervals as described in section 3.3.5.1.

Ultimately, the goal of building these relationships is also to predict the occurrence or absence of clinical outcomes such as heart attack, stroke, and/or death. We may also wish to expand to other diseases at some point in time.

We plan to devise an R package that will allow for the calculation of event relationships as described in this section of the dissertation.

At this point, in calculating the IDD, we have only included subjects who experienced both events. Calculation of the IDD could be improved if we were to include all subjects who experienced either of the events. We wish to include this in future iterations of the process, in which case we will need to utilize survival analysis methods. This is necessary

since subjects who experienced one event may not have experienced a second event yet due to age (i.e. they are right-censored).

As described previously, an option to left-censor a particular event would improve the interpretation of the results obtained from this process.

As mentioned before, it could be of interest to determine an additional procedure dealing with situations where the "peak" of the ODD compared to the IDD is not equal to zero. For example, it may be the case that approximately one year after starting a certain medication, many people experience a certain side effect. In that case, if the secondary condition occurs shortly after the first, it may be unrelated, whereas if it occurs close to that one-year mark, it likely ***is*** related.

Finally, the process may benefit from the introduction of covariates in order to make it more precise.

INFORMATION RETRIEVAL FOR DESIGNING AND ANALYZING CLINICAL STUDIES

4. Abstract Mining

Medical researchers use PubMed and other bibliographic databases extensively to search for publications that could help them in their research. In addition, they also search these vast databases for publications that would suggest new research ideas. In the latter case they do not have specific search terms; instead, they use general search terms and look for publications that seem to indicate an unusual link between the subjects of the general search terms and other subjects that could potentially be of interest. However, those general searches usually result in very large numbers of publications, often in the thousands or more. To select the unusual or unexpected ones by reading the abstracts in that case is an extremely time-consuming, at times close to impossible task.

We have developed a solution where we will allow a researcher to obtain the abstracts for the publications that fulfill their desired search terms from those bibliographic databases. We then take those abstracts and perform text mining on them to obtain the most common, non-trivial words in those abstracts. Clustering the abstracts by those words will allow a researcher to immediately identify groups of publications whose abstracts contain words that are of interest to them. Repeated clustering will narrow down the number of interesting publications further so the researcher can identify new and interesting ideas for research in a fraction of the time taken previously for the same purpose.

4.1 Motivation

> "We are not students of some subject matter, but students of problems. And problems may cut right across the borders of any subject matter or discipline."
>
> - Karl Popper (Popper, 1963)

Complex scientific problems and socially relevant issues are challenging scientists to find new ways to integrate knowledge from multiple and disparate fields (Adams, 2007). The current explosion of information in the bibliographic databases has resulted in a total of more than 90 million records for the Web of Science and more than 160 million for Google Scholar. The citations of the medical field have become subspecialized as if they belong to different disciplines and new developments will occur at the interaction of the subspecialized fields.

Medical researchers use PubMed (PubMed, 1996) (29 million citations) and the other databases to search for publications that could help them in their research. In addition, they also search these databases for publications that would suggest new research ideas, often at those interactions of different subject matter. In the latter case they do not search for specific terms; instead, they use general search terms to identify publications that could potentially lead to new ideas for research. However, these general searches usually result in very large numbers of publications, often in the thousands or more. To select the unusual or unexpected articles that may lead to new ideas for research is extremely time-consuming, and at times nearly impossible. A researcher may spend days, weeks, and sometimes even months searching and reading through abstracts in order to find an interesting subject that is worthwhile investigating to see if it warrants a clinical trial.

4.2 Proposal

We present a method to aid the researcher in getting to those publications of interest without the need to search each individual abstract that fulfills the general search terms. The method is implemented by a computer application that will ameliorate the problem by examining all the abstracts for those publications that fulfill the general search terms and uses text mining algorithms on those abstracts to extract all non-trivial words.

The researcher may then repeatedly cluster the publications by commonality of the words in the abstracts to find unusual or unexpected combinations of words. Once a particularly interesting word or combination of words has been identified, the researcher can choose to read all published abstracts in the cluster of interest containing the selected unexpected combination of words, and if worthwhile, can devise a new clinical trial with the information so obtained.

The method is implemented via a Shiny application (Shiny, 1996) and uses the MEDLINE output of the PubMed database for medical publications. Extensions to other bibliographical databases can be made at a later point in time.

4.3 Method

We read in a file with publications obtained from a bibliographical database in response to some general search terms, and extract for each publication the date, title, abstract, and PMID. For those publications without an abstract, we use the title in place of the abstract.

We perform text mining on the abstracts using the R text mining package “tm” (Feinerer, Introduction to the tm Package. Text Mining in R, 2017) (Feinerer, An Introduction to Text Mining in R, 2008) (Feinerer, Hornik, & Meyer, Text Mining Infrastructure in R, 2008). We perform the following tasks:

- Build a Corpus
- Clean the Corpus
- Build a TermDocumentMatrix object
- Create the Term/Document Matrix
- Cluster the Documents (columns) via the Term/Document Matrix
- Re-Cluster one of the Clusters (if desired)

4.3.1 Build a Corpus

We use the text in the abstracts to build a corpus with a document for each publication. A corpus is the main structure for managing documents in the tm package and represents a collection of text documents. See (Feinerer, An Introduction to Text Mining in R, 2008) for a more detailed description.

4.3.2 Clean the Corpus

We transform all text in the corpus to lower case, remove URLs, remove anything other than English letters and spaces, remove punctuation, and remove so-called “stop words”. Stop words are trivial words that are not useful in a clustering of words. It is possible to add words to these stop words that a user might not find useful and wishes to exclude.

Finally, we remove all superfluous white space.

4.3.3 Build a TermDocumentMatrix Object

The TermDocumentMatrix is a construct central to the tm package. The function with the same name takes as input a corpus with documents, inspects the contents and finally outputs a TermDocumentMatrix object. For each document, the function records every term (word) it encounters and records it. It keeps track of which term occurs in which document and how many times it occurs in that document. The object contains several fields, see Table 3.

Table 3: Fields of the TermDocumentMatrix Object

i	An integer vector with indices of the terms as they appear in dimnames$Terms	
j	An integer vector containing document numbers for the corresponding terms	
v	An integer vector containing the number of times the corresponding term in i appears in the corresponding document in j	
nrow	Unique number of terms found	
ncol	Number of documents in the corpus	
dimnames	Terms	A character vector with all the unique terms found in the corpus
	Docs	A character vector with the document numbers

Note that the fields i, j, and v all have the same length; they indicate that term i[k] appears in document j[k], v[k] times.

Note also that the TermDocumentMatrix is secondarily a simple_triplet_matrix.

4.3.4 Create the Term/Document Matrix

The TermDocumentMatrix object can be transformed into a matrix using the as.matrix function for a simple_triplet_matrix. This matrix has a column for each document and a row for each term. Each element of the matrix has the number of instances of the term in the row for the document in the column. Many elements in this matrix will be zero, making the matrix sparse.

4.3.5 Cluster the Documents via the Term/Document Matrix

The prior tasks are performed using functions in the tm package, and we end up with a sparse matrix with non-trivial words (terms) vs. the documents as they appear in the corpus.

We then cluster the columns of this matrix using k-means clustering; columns (i.e. documents) that are most similar will be put in the same cluster. The number of clusters can be varied by the user. Each cluster will contain a subset of the documents in the corpus, and a set of terms associated with those documents. The most frequent terms in each cluster will provide useful information about the documents in that cluster.

4.3.5.1 k-Means Clustering of a Matrix

To perform k-means clustering on our sparse matrix, we first randomly assign one document (i.e. column) to each cluster. Then for each other document not assigned to a cluster, we determine its distance to each cluster. For that, we consider the columns representing the documents as vectors and determine the (Euclidean) distance between the vectors. Documents are assigned to the cluster that is nearest in distance to them.

Once all document column vectors have been assigned to a cluster, we perform the following iterations:

a) Calculate the ***centroid*** of each cluster. The cluster centroid is obtained by taking the average of the document vectors in the cluster.

b) For each document, determine its distance to each of the cluster centroids, and assign it to the cluster whose centroid is closest, using Euclidean distance.

We iterate until the cluster assignments stop changing, at which point we display the clusters together with their most common terms.

Example:

Let's assume we have a matrix as in Table 4 and we wish to create 2 clusters.

Table 4: Cluster Example

	1	***2***	***3***	***4***
A	0	0	0	1
B	1	1	1	2
C	0	0	1	0
D	0	1	3	2

We see here that the word "A" occurs only once in document 4, "B" occurs in all 4 documents, once in 1, 2, and 3, and twice in document 4. "C" only occurs once in cluster 3, while "D" appears once in document 2, 3 times in document 3, and twice in document 4.

We first assign two documents randomly to the clusters; here we will assign document 1 to cluster 1 and document 2 to cluster 2. The squared distance from document 3 to cluster 1 is 10, and to cluster 2 it is 5, so we assign document 3 to cluster 2. Document 4 also has a greater distance to cluster 1, so it is also assigned to cluster 2.

Now the centroids are calculated. For cluster 1, it is $(0,1,0,0)'$, while for cluster 2 it is $\left(\frac{1}{3},\frac{4}{3},\frac{1}{3},2\right)'$. Calculating the squared distance of document 2 to cluster 1 we get 1, while we get $\frac{4}{3}$ for cluster 2, so document 2 gets moved to cluster 1. Both documents 3 and 4 are closer to cluster 2, so they stay as they are.

At the next iteration the centroids for clusters 1 and 2 are respectively $\left(0,1,0,\frac{1}{2}\right)'$ and $\left(\frac{1}{2},\frac{3}{2},\frac{1}{2},\frac{5}{2}\right)'$, and all documents stay where they are.

In the end documents 1 and 2 are clustered together in cluster 1, while documents 3 and 4 are clustered together in cluster 2.

The most common word in cluster 1 is “B”, followed by “D”, while the most common word in cluster 2 is “D” followed by “B”.

4.3.6 Re-Cluster one of the Clusters (if desired)

When clustering is completed, the clusters may be inspected. If desired, one of the clusters may be selected and re-clustered. Abstracts for the publications in the selected cluster will be mined and processed as described in section 4.3.1 through section 4.3.5, i.e. we will build another corpus, but this time only with publications present in the selected cluster. We will clean the corpus, build the TermDocumentMatrix object, create the Term/Document matrix, and cluster its columns as before.

Once the re-clustering has completed, it may be inspected again, and if desired, one of the clusters in the re-clustering may be selected and re-clustered. This process may be repeated until one or more publications of interest are found in one of the clustered or re-clustered collections of publications.

4.4 Shiny Application

Upon entering the Shiny application, the user will arrive at a screen that will show three tabs: “Main”, “Abstracts”, and “Titles”, with the Main tab being current. See Figure 24.

4.4.1 The Main Tab

Figure 24: The Abstract Mining Application

The application at this point will require input from PubMed; the user may enter PubMed on their own accord, or they could use the link provided in the application (“Go to PubMed”). Once there, they can enter their search terms after which they must create a MEDLINE file. The field “Send to” should be clicked, the radio button “File” should be selected under “Choose Destination”, and “Format” should be set to “MEDLINE”. A file can then be created by pressing the “Create File” button. See Figure 25, where the search terms “takotsubo catecholamines” are selected.

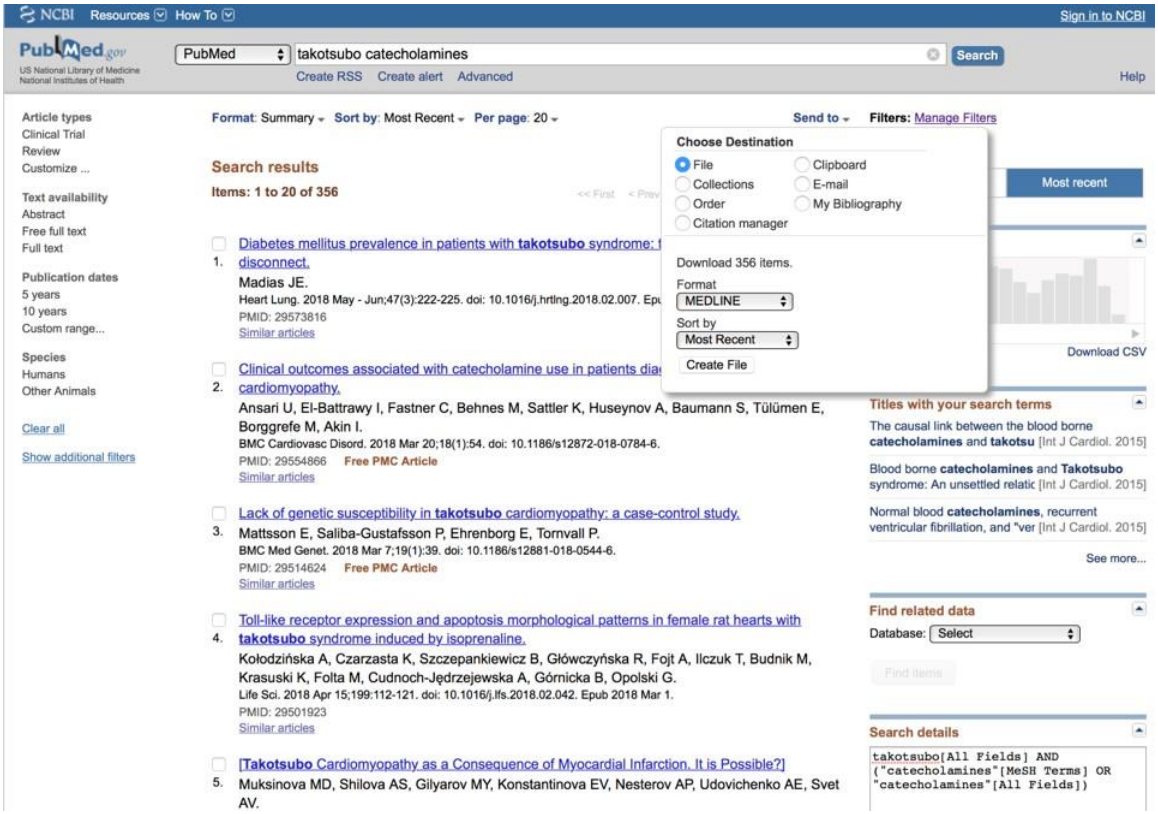

Figure 25: PubMed MEDLINE File Creation

Back in the Abstract Mining application, the MEDLINE file created by the PubMed site may be uploaded via the “Browse” button by selecting the location of the file in the user’s file system (usually the downloads directory). Before uploading the file, a number of clusters may be selected, or the user may leave the default of 6 clusters in place.

Once the MEDLINE file has been uploaded, the application uses our method to select all non-trivial words from all abstracts in the file, and clusters the publications using those words and the desired number of clusters as selected. See Figure 26 where a MEDLINE file created using the “takotsubo” search term in PubMed, has been uploaded.

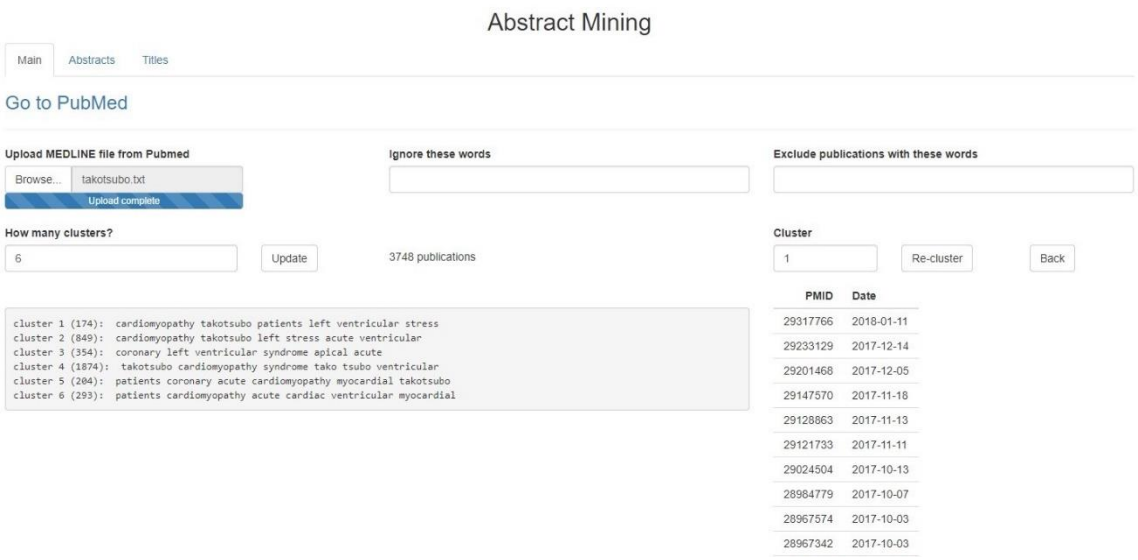

Figure 26: Clustering of the Takotsubo PubMed File

We see that the total number of publications in the file is displayed as well as the clusters. For each cluster, the number of publications in the cluster is shown in parentheses after the cluster number, followed by the most frequent words in the cluster. The field "Cluster" is by default set to 1 but can be changed. Underneath this field, the PMID and the date of all publications in this cluster are displayed (here cluster 1).

The "Re-cluster" button may be pressed to re-cluster all publications for the cluster in the "Cluster" field (here cluster 1). The "Back" button can be used to return to a previous clustering.

At this point, the user has several options for investigating the 3748 publications returned by PubMed in response to the "takotsubo" search term. We sat down with a team of cardiovascular surgeons and let them do the investigation which resulted in new and

interesting ideas for research within about 30 minutes. We will describe the options and their implementation in the takotsubo case.

4.4.1.1 Change the Number of Clusters

The number of clusters may be changed using the applicable input field. To change the number of clusters it is also necessary to press the "Update" button afterwards. See Figure 27 where the number of clusters for the Takotsubo file from PubMed is changed from 6 to 15.

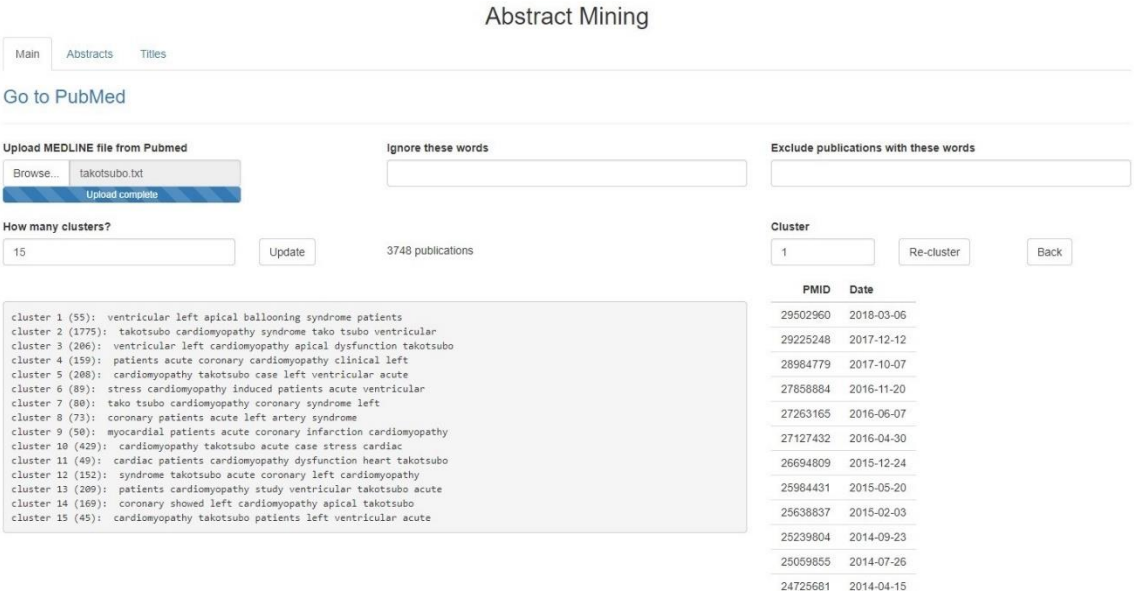

Figure 27: Changing the Number of Clusters

4.4.1.2 Select a Cluster

The user may change the "Cluster" field in order to see the PMIDs and dates for that cluster. The tab "Abstracts" may then be selected to see the abstracts for the publications in the selected cluster, and the tab "Titles" to see its titles. See Figure 28 where cluster 11 is selected.

4.4.1.3 Exclude Publications with Certain Words

The user may enter words (separated by spaces) in the "Exclude publications with these words" field. Like for the Number of Clusters, the "Update" button should then be selected, which will update the presentation of the clusters underneath it. If words are selected in the "Exclude Documents containing these words" field, all publications with abstracts that contain any of those words will be removed.

4.4.1.4 Ignore Words

The user may enter words (separated by spaces) in the "Ignore these words" field. Like for the Number of Clusters, the "Update" button should then be selected, which will update the presentation of the clusters underneath it. If words are selected in the "Ignore these words" field, those words will be removed from the clustering. Unlike when publications are excluded based on words, this will not affect the number of publications under selection.

4.4.1.5 Re-Cluster

Once the initial clustering is completed, a user may select one of the clusters (which will automatically update the list of PMIDs and dates as described previously) and select "Re-cluster". This will take all the publications in the selected cluster and cluster them according to the number of clusters requested, taking into account the excluded documents and ignored words. A new cluster presentation will be created that will replace the existing cluster presentation. See Figure 28, where cluster 10 of the Takotsubo file is re-clustered, after which cluster 11 is selected.

This process may be repeated indefinitely. If at any time the user wishes to return to a previous state, the "Back" button may be selected. Repeated use of the "Back" button will eventually return to the original set of publications from the MEDLINE file, minus any excluded publications with words in the "Exclude Documents containing these words" field.

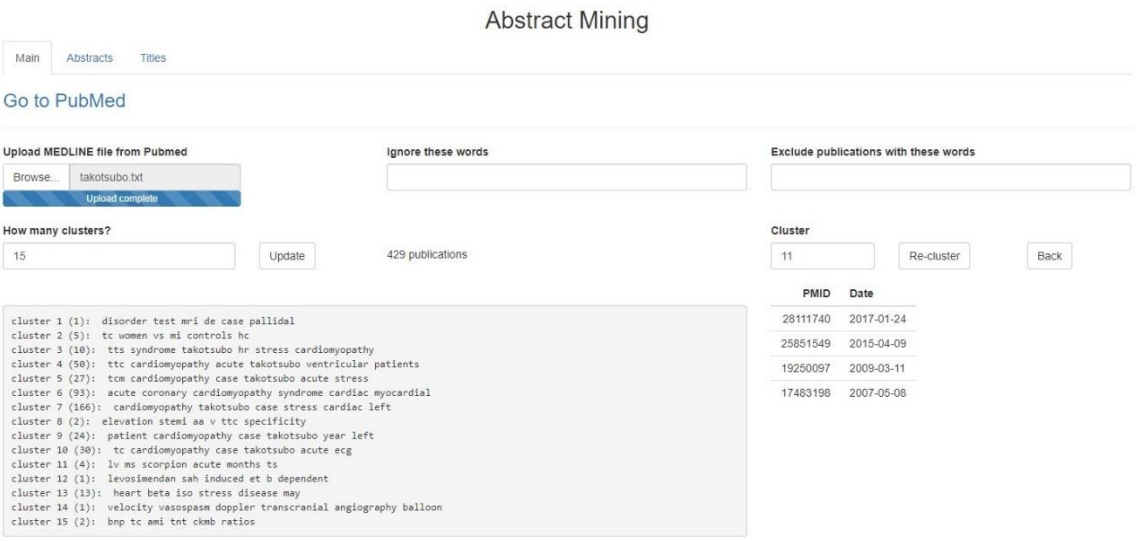

Figure 28: Re-Clustering the Takotsubo File

4.4.2 The Abstracts Tab

The "Abstracts" tab will display the PMID, date, title and Abstract of the first publication for the cluster selected on the "Main" tab. The user can scroll through all the publications in the cluster using the "Previous" and "Next" buttons underneath the abstract. In addition, the "Download Cluster" button may be selected which will create a html file with all PMIDs, dates, titles, and abstracts in the current cluster. See Figure 29 where PMID 25851549 is selected from cluster 11 of Figure 28.

Figure 29: The Abstracts Tab

4.4.3 The Titles Tab

The "Titles" tab will display the PMID, date, and title for all publications in the selected cluster. See Figure 30 where we display the titles for cluster 11 of Figure 28.

Figure 30: The Titles Tab

4.4.4 Running the Shiny Application

This Shiny application is available at https://ellie.shinyapps.io/shiny/. Note, that the application needs a MEDLINE file from PubMed; PubMed maybe accessed either via the application or independently to create a MEDLINE file to be loaded into the application. After that, clustering can commence.

4.5 Results

We include four examples showing how to use the application to perform a search for interesting publications. We started each example searching the PubMed database for publications containing the search criteria stated; the cursive words in the clusters were the ones that drew attention. The searches were performed by a team of cardiovascular surgeons.

(1) ***Search Criteria***: "embolic stroke", excluding "atrial fibrillation", 10,443 publications.

Number of Clusters 1: 23 clusters

Cluster Selected 1: 8 (576 pubs), "cerebral stroke brain *blood* artery ischemic".

Number of Clusters 2: 23 clusters

Cluster Selected 2: 12 (1 pub), "*progranulin* ischaemia ischaemic expression cerebral demonstrated".

Publications Selected: PMID **25838514**, "Multiple Therapeutic Effects of Progranulin on Experimental Acute Ischaemic Stroke" (Kanazawa, et al., 2015).

Proposed Research: Upon checking the abstract for the selected publication, the researcher determined that the link between embolic stroke and progranulin would provide a worthwhile subject for further research.

(2) ***Search Criteria***: "impedance", "mismatch", 368 publications.

Number of Clusters: 20 clusters

Cluster Selected: 9 (20 pubs), "pressure pulmonary arterial wave impedance

ventricular"

Publications Selected: PMID **21996190**, "Systemic Vascular Hemodynamics and Transplanted Kidney Survival" (Wystrychowski, et al., 2011), PMID **11281995**, "Low Compliance Rather than High Reflection of Arterial System Decreases Stroke Volume in Arteriosclerosis: A Simulation" (Sugimachi, Shishido, & Sunagawa, 2001), PMID **9709398**, "Pulmonary Impedance and Right Ventricular-Vascular Coupling in Endotoxin Shock" (D'Orio, et al., 1998), PMID **2273555**, "Aortic and Pulmonary Input Impedance in Patients with Cor Pulmonale". (Chen, et al., 1990)

Proposed Research: **21996190** relates to kidneys, **11281995** to sepsis, and the remaining two relate to the lungs.

(3) ***Search Criteria***: "aortic", "stenosis", 157 publications.

Number of Clusters: 15 clusters

Cluster Selected: 4 (2 pubs), "*pacing* leads ventricular atrial left chamber"

Publications Selected: PMID **23078085**, "Long-Term Follow-Up Impact of Dual-Chamber Pacing on Patients with Hypertrophic Obstructive Cardiomyopathy" (Yue-Cheng, et al., 2013), PMID **9121966**, "Chronic Steroid-Eluting Lead Performance: A Comparison of Atrial and Ventricular Pacing" (Hua, Mond, & Strathmore, 1997).

(4) ***Search Criteria***: "takotsubo", 3,748 publications.

Number of Clusters 1: 15 clusters

Cluster Selected 1: 10 (429 pubs), "cardiomyopathy takotsubo acute case stress cardiac" (see Figure 27).

Number of Clusters 2: 15 clusters

Cluster Selected 2: 11 (4 pubs), "lv ms *scorpion* acute months ts" (see Figure 28).

Publications Selected: PMID **25851549**, "Scorpion-Related Cardiomyopathy: Clinical Characteristics, Pathophysiology, and Treatment" (Abroug, et al., 2015) (see Figure 29 and Figure 30).

4.6 Conclusion

We describe a new method to identify new research ideas based on text mining. This may assist investigators in the health sciences. This method allows researchers to start with general search terms and find publications with unusual, unexpected findings of interest for further investigation and potential clinical studies. For example, using this method we found a link between takotsubo and scorpion and envenomation, a relationship of stroke to progranulin, and effects of impedance mismatch in kidneys, sepsis and lungs, as well as a relationship between pacing effectiveness between steroid eluting stents and non-steroid eluting stents.

This abstract mining application is helpful for creating a clustering structure of words in abstracts in order to identify interesting publications in a very short time, especially when the search criteria result in large numbers of publications. The method can be expanded to other research-oriented websites like Google Scholar, ResearchGate, etc.

The application is qualitatively different from other text mining applications since instead of describing and analyzing existing information or structures it leads to the development of new research ideas.

4.7 Limitations and Strengths

The abstract mining algorithm is operator dependent in the choice of the search criteria, the number of clusters, re-clustering and other details. Also, PubMed is dynamic with publications added daily. As a result, the findings may not always be reproducible.

However, this method has significant strengths since it provides fast access to information across many abstracts and may lead to identification of new ideas for research.

A further, important limitation is that words on their own are not always descriptive enough of the content of an abstract; replacing words with phrases would make the process more useful. We address that concern in the next chapter.

5. Abstract Phrase Mining

5.1 Introduction

A drawback to the abstract mining application is that single words at times are not descriptive enough to identify truly unique and unexpected ideas for potential new research. In order to resolve that issue, we need a method that extracts common phrases rather than words from those abstracts, after which we can perform the clustering on those phrases instead.

Phrase mining, however, comes with challenges that are not present with word mining. It is not difficult to expand the search from words to so-called "n-grams". An n-gram is any combination of n words that exists in the abstracts. We could cluster these n-grams based on frequency just like we clustered words based on frequency.

However, this would result in multiple problems. For example, the phrase “cats and dogs” would turn into the meaningless phrase “cats dogs”, due to the fact that we exclude trivial words. So, in order to solve this issue, we would need to include all trivial words when we mine for phrases rather than words. This, however, could result in other meaningless phrases such as “and this disease” and “attack of”. Furthermore, some phrases are inherently meaningless to researchers such as the phrase “here we are”, which may still occur frequently. Selecting n-grams that cross punctuation marks also generally results in meaningless phrases.

We would also have an issue with double-counting, where one word or group of words may be counted in multiple phrases. This would result in meaningless subsets of phrases

being presented as frequent phrases together with their meaningful super phrases since they would occur with the same frequency. If we were to cluster the results this would likely cause these phrases to be clustered together, and the meaningless sub phrase would provide no useful additional information to the cluster that wasn't already provided by its super phrase. As an example, consider the phrases "severe cardiovascular disease" and "severe cardiovascular", the former meaningful, the latter meaningless. Since the latter always occurs with the former, its frequency would be the same, and even if the phrase "severe cardiovascular" never occurred on its own, it would still be presented as a phrase as frequent as "severe cardiovascular disease". If the phrase "severe cardiovascular disease" were present in a cluster, then the phrase "severe cardiovascular" would appear right next to it since its commonality would be the same. Similarly, the meaningful phrase "support vector machine" would result in the meaningless phrase "vector machine" having the same frequency.

Due to these issues, when we mine for phrases, we wish to restrict our search to meaningful phrases only. We refer to those meaningful phrases as ***principal phrases***.

However, meaningful, but generally less common phrases such as "support vector" would have their frequency exaggerated since the frequencies of "support vector machine" would be added to its stand-alone occurrences. This would cause any principal sub phrase such as "support vector" to appear to be more common than its super phrase, like the phrase "support vector machine".

We developed a text mining method that avoids these problems, extracts principal phrases only from large texts, and avoids all double-counting.

In addition, we created another Shiny application that uses that method to mine and cluster abstracts using principal phrases instead of words.

5.2 Principal Phrase Mining

Here we present a new method for mining principal phrases of variable length from a vector of texts.

To start with, we no longer exclude trivial words. Then, instead of mining for words, we mine for n-grams. Recall that an n-gram is any combination of n words that exists in a text. In addition, we select only those n-grams that appear between certain punctuation marks, i.e. we never select an n-gram that crosses over a punctuation mark such as a period, comma, semicolon, etc. This ensures that the text “Dealing with disease; how to cope” does not result in n-grams such as “disease how to cope” or “dealing with disease how”.

However, when doing this, it is inevitable that words and phrases appearing in the abstracts are counted multiple times, on their own as well as inside other phrases. This is undesirable, as it leads to duplication of information and the appearance of phrases that do not make sense. For example, should the phrase "abstract phrase mining" be a frequent one, then so will "abstract phrase", and "phrase mining”. The phrase "abstract phrase" is not a meaningful one and as such will rarely occur by itself. But when mining n-grams, it is a frequent occurring phrase since its super-phrase "abstract phrase mining" is frequent. The phrase “phrase mining” on the other hand, is a meaningful phrase and may also occur by itself. However, its frequency would be exaggerated since all occurrences of the phrase “abstract phrase mining” would be added to the frequency of the stand-alone phrase.

We therefore should avoid all double-counting. For example, should "abstract phrase mining papers" be contained in a document and due to this we assign one additional frequency to "abstract phrase mining", we do not then also assign an additional frequency to "abstract phrase" and "phrase mining". In addition, we also do not assign an additional frequency to "phrase mining papers", since if we were to do that, "phrase mining" would be counted twice (in different phrases) while it only occurred once.

Finally, not all frequent phrases we discover within the publications may be useful; any phrases that start or end with stop words such as "and" and "or" are removed, and we also allow for the removal of complete phrases that are not informative to specific users.

5.2.1 Proposal

The challenges related to mining phrases rather than words are described by (Liu, Shang, & Han, 2017). We adopt some of their terminology, but we will employ a method different from their suggested method of phrasal segmentation. Their quality phrases are selected using different criteria than our principal phrases; the quality of a phrase is calculated based on user-supplied sets of quality phrases and non-quality phrases. When segmenting a block of text, they use the quality of the phrases to determine the best possible segmentation. In addition, the objective of their paper is to extract meaning from large amounts of text, while we use principal phrases as features used for clustering.

A ***principal phrase*** has three properties: it needs to be popular, complete, and informative. A phrase is popular if it is frequent. It is complete if it is a complete

semantic unit, which is obtained by avoiding the double counting explained previously, and a phrase is informative if it is meaningful to the user.

Our method by design only selects frequent phrases. It employs a rectification process that ensures the phrases are complete. Non-informative phrases are excluded using a user-defined selection of start-stop-words and end-stop-words, where phrases are excluded if they respectively start or end with those words.

Finally, a selection of user-defined phrases that should always be excluded since they are not considered informative, may also be supplied.

5.2.2 Method

The method usually employed using text mining of words involves processing each document word for word and keeping track of the number of times each word appears in each document, excluding so-called "stop-words". See section 4.3, the section on abstract mining, which describes the process using the “tm” package in R (Feinerer, An Introduction to Text Mining in R, 2008). A similar method can be used for n-grams; we can run through each document and keep track of the number of times each n-gram appears in each document.

Unfortunately, for our current problem, n-grams will not do; we would like to obtain principal phrases instead. In order to obtain these, we wish to have access to the positions of the n-grams in the various documents. Since position information is not available

through the functions supplied in the "tm" package, we have developed our own. Using these functions, we can obtain all principal phrases from a collection of texts.

Instead of the TermDocumentMatrix construct, we created a phraseDoc construct, with a function of the same name that creates a phraseDoc object from a vector of texts.

We no longer create a corpus, and cleaning it is also no longer necessary as cleaning is a part of the phraseDoc function. A function to create a (sparse) matrix with phrases (terms) as rows and document numbers (indices to the vector of texts) is provided to aid applications such as our Shiny application that will need the information in this form in order to cluster the documents. Furthermore, the function removePhrases is provided which will remove a set of phrases from a phraseDoc object.

5.2.2.1 The phraseDoc Object

The phraseDoc is a construct central to the selection of principal phrases. The function with the same name takes as input a vector with texts, inspects the contents and outputs a phraseDoc object. The object contains several fields, see Table 5.

Table 5: Fields of a phraseDoc Object

phrase	An integer vector with indices of the phrases as they appear in phrases$phrase	
doc	An integer vector containing document numbers for the corresponding phrase	
block	An integer vector containing the block number within the document where the corresponding phrase appears	
pos	An integer vector containing the position within the block where the corresponding phrase appears	
phrases	phrase	A character vector with unique principal phrases found in the collection of texts
	pwrds	An integer vector containing the number of words in the corresponding phrase
	freq	An integer vector containing the number of times the corresponding phrase occurs in the collection of texts

Note that the fields phrase, doc, block, and pos all have the same length; they indicate that the phrase indicated by the index in phrase[k] appears in document doc[k], block block[k], at position pos[k].

Also, the fields phrases$phrase, phrases$pwrds, and phrases$freq all have the same length; they indicate that the phrase in phrases$phrase[i] has phrases$pwrds[i] words and appears in the entire collection of texts/documents phrases$freq[i] times.

5.2.2.2 Selecting Principal Phrases

The process performed by the phraseDoc function allows for a variety of user input in addition to the vector of texts, all of which have default values. See Table 6 for a description of the input fields and their default settings.

Table 6: Parameters of the phraseDoc Function

Input Field	***Default***	***Description***
mn	2	Minimum number of words per phrase
mx	8	Maximum number of words per phrase
ssw	Output of the function stopStartWords()	A character vector containing all words a phrase should not start with
sew	Output of the function stopEndWords()	A character vector containing all words a phrase should not end with
sp	Output of the function stopPhrases()	A character vector containing all phrases that should be excluded
min.freq	2	The minimum number of times a phrase should appear in the collection of texts in order to be included
qp	A function that returns FALSE if the parameter freq is less than min.freq, and TRUE otherwise	A function with 2 parameters, phrase and frequency, that returns TRUE when the phrase is considered principal, and FALSE if it is not.
max.phrases	1500	The maximum number of principal phrases to be collected from the collection of texts
shiny	FALSE	Should be set to TRUE if called from a shiny application, FALSE otherwise. If TRUE, the function will output progress information to the Shiny application.

We process each text by breaking it into ***blocks***, which are identified by any of the following punctuation marks: [].!(),:;?|{}. For each block, we transform its text to lower case, then we inspect it and identify within it all suitable n-grams of length between the minimum and maximum supplied. An n-gram is suitable if it does not start with a word in the supplied list of stop-start-words, it does not end with a word in the supplied list of stop-end-words, and also does not appear in the list of stop-phrases. Then we record the block and the position within the block as well as the document for each of these suitable n-gram phrases we encounter.

When all suitable n-gram phrases and their positions have been obtained, we determine frequencies for each n-gram phrase, the so-called ***raw frequencies***. We check to see if the total number of n-gram phrases exceeds the maximum supplied. if it does, we note the minimum frequency of the set of most frequent phrases of size equal to the supplied maximum. If this minimum exceeds the supplied minimum frequency, it will replace it. We then remove all phrases with frequencies less than the minimum. Each phrase removed will also have its positions removed.

For example, if the minimum frequency equals 2, and the maximum number of n-gram phrases equals 1500 and the number of n-grams is greater than that, we select the 1500 most frequent phrases and inspect the lowest frequency. If this equals 5, then the minimum frequency will be set to 5 and all n-gram phrases with a frequency equal to 1, 2, 3, or 4 will be deleted together with their positions. This will result in somewhat more than 1500 n-gram phrases remaining.

We choose this method rather than to cut the number of phrases at the supplied maximum to make the choice of those phrases less arbitrary; also, this way we initially start with more than the maximum number of phrases, which is desirable since the rectifying process will remove many phrases, which could then possibly lead to a very small selection of phrases left. Note that the maximum number of phrases only indicates an approximate; it is possible (although unlikely) that more than that number of phrases will be supplied in the end.

At this point we are left with a reasonably small selection of frequent n-gram phrases (depending on the maximum), on which we will perform the rectification process. Once the rectification process is complete, we have a selection of principal phrases with their positions in the vector of texts, which will be provided as a phraseDoc object.

Note, that the creation of the phraseDoc object is a standalone process that can be used in circumstances other than our Shiny abstract phrase mining application.

5.2.2.2.1 Rectification Process

The rectification process will run through each phrase, starting with the most frequent n-gram phrases of the greatest length, continuing until the least frequent n-gram phrase of the shortest length has been processed.

For each n-gram phrase we find all its positions. We check to see if the phrase is a principal phrase using the supplied function. The default of this function designates the phrase as a principal phrase only when it has more positions than the minimum frequency

allows. If the phrase is not a principal phrase, it will be removed together with all its positions. If it ***is*** a principal phrase, any position for another phrase that starts or ends within the positions of this one, will be removed.

For example, say that our minimum frequency is 3, and that block 10 of document 5 equals "The authors wrote abstract phrase mining papers". We assume that "the" and "wrote" are in both the stop-start-words and the stop-end-words. See Table 7 for the frequencies assigned to the phrases for this example.

Table 7: Rectification Example

n-grams	n	frequencies			
		Raw	After removing infrequent n-grams	After processing Doc.5, Block 10	Rectified
authors wrote abstract	3	3	3	2	-
authors wrote abstract phrase	4	1	-	-	-
authors wrote abstract phrase mining	5	1	-	-	-
authors wrote abstract phrase mining papers	6	1	-	-	-
abstract phrase	2	10	10	9	-
abstract phrase mining	3	10	10	10	10
abstract phrase mining papers	4	1	-	-	-
phrase mining	2	20	20	19	6
phrase mining papers	3	5	5	4	4
mining papers	2	5	5	4	-

After removing infrequent n-grams (those with a raw frequency less than or equal to 3), we see that the largest phrases still under consideration consist of 3 words; we have “authors wrote abstract”, “abstract phrase mining” and “phrase mining papers”. In this case, "abstract phrase mining" would be processed before the other phrases since it is the most frequent 3-gram and 3-grams are processed before 2-grams. We will process all its positions in any of the documents/blocks where it occurs.

The phrase "abstract phrase mining" takes up positions 4, 5, and 6 in block 10 document 5. When we process document 5, block 10, position 4, we will remove all positions for document 5, block 10, that either start or end in positions 4, 5, or 6. The phrase "authors wrote abstract" ends in position 4, so this position will be removed. The phrases "abstract phrase", "phrase mining", "phrase mining papers", and "mining papers" start respectively in positions 4, 5, 5, and 6, and so their positions here will also be removed.

When all positions for "abstract phrase mining" have been processed, "authors wrote abstract" will have less than 3 positions left, while "abstract phrase" will have no positions left. When the phrase "authors wrote abstract" is being processed later, it will be removed together with its remaining positions since it no longer fulfills the frequency requirement. The same fate belies "abstract phrase" since it has no positions left; thus, neither of these two n-gram phrases makes it as a principal phrase.

We see that after rectification, only the phrases "abstract phrase mining", "phrase mining", and "phrase mining papers" are left and thus they are the only ones identified as principal phrases.

Note that we give priority to phrases consisting of many words; the reason for this is that these long phrases are less likely than smaller ones to be frequent, so if they DO make the minimum frequency, they are more likely to be principal phrases. Of course, there may be some of these passing this requirement that are not actually meaningful; some errors are to be expected. Note, however, that phrases that should be excluded can always be added to the `sp` parameter. Should one wish to use the original stop-phrases from the function in addition to a user-defined set, the code:

```
sp=c(<vector with phrases to be excluded>,stopPhrases())
```

can be added when calling the phraseDoc function. Note that the phrases in the vector with phrases to be excluded should be in all lower case.

5.2.2.3 Create a Matrix from a phraseDoc Object

A phraseDoc object may be converted to a matrix that provides frequency information per document for each phrase using the `as.matrix.phraseDoc` function, to be called as `as.matrix(pd)` where `pd` is a phraseDoc object. This function creates the matrix by converting the phraseDoc object to a simple_triplet_matrix object first, which has a column for each document and a row for each phrase. Each element of the matrix has the number of instances of the phrase in the row for the document in the column. Many elements in this matrix will be zero, making the matrix sparse.

5.2.2.4 Remove a Collection of Phrases from a phraseDoc Object

A collection of phrases may be removed from a phraseDoc object using the `removePhrases.phraseDoc` function, to be called as `removePhrases(pd,phrs)` where `pd` is a phraseDoc object and `phrs` a character vector containing the phrases to be removed. This function will remove all references to any of the phrases in `phrs` from the phraseDoc while keeping its structure intact.

5.2.3 Performance

The functions for this process have been written in R. As such, performance of these functions can be improved by creating them in a compiled programming language instead, to be called by an R function.

We executed the process on 10 different sized output files from PubMed. The results are displayed in Table 8.

Table 8: File Size and Number of Publications vs. Processing Time

Size (in KB)	Time (in seconds)					# Publications
	Read file	Create Raw Frequencies	Remove Infrequent	Rectify	Total	
46	0	0	0	0	1	11
368	0	4	0	5	9	157
562	1	6	2	11	20	368
1838	1	7	3	5	16	419
9024	4	31	12	11	54	3748
18524	6	69	28	19	122	3513
31804	17	188	68	54	327	10443
34582	17	204	80	60	361	9523
79830	40	505	217	118	880	23830
97873	51	640	243	187	1121	29298

Note that these times are dependent on the power of the computer the process is run on and will vary significantly dependent on which computer is used. As such, the times in this table should be read only for their relative values between the different input files.

We plotted the total time versus the size of the file and ran a linear model on the data. See Figure 31.

The adjusted R-square for linearity is .9929, indicating it is likely that the relationship between total processing time and size is linear.

We also found that for the 6 largest file sizes, creating the n-grams took 57% of the total time, while the rectification took around 16% of the total time.

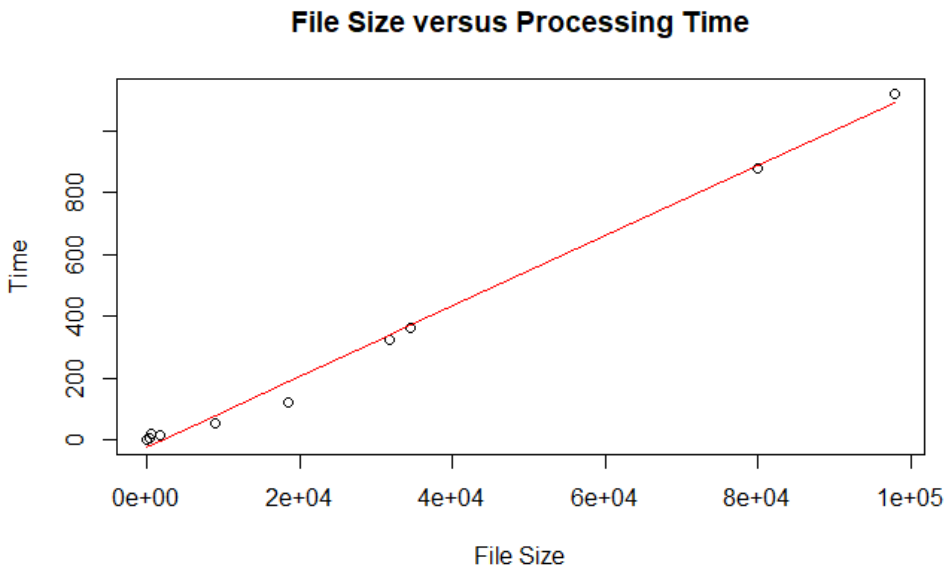

Figure 31: File Size vs. Processing Time

5.3 Cluster the Phrase/Document Matrix

Once the sparse phrase/document matrix has been created from the PhraseDoc object, we can cluster the documents (texts). We have a sparse matrix with principal phrases as rows, documents as columns, and the number of instances of a phrase within a document at the intersection of the rows and columns.

We then cluster the columns of this matrix using k-means clustering; columns that are most similar will be put in the same cluster. The number of clusters can be varied by the user. See section 4.3.5.1 for a description on how to perform k-means clustering for such a matrix. Each cluster will contain a subset of the total set of texts/documents, and a set of principal phrases associated with those texts. The most frequent principal phrases in each cluster will provide useful information about the texts in that cluster.

5.4 Shiny Application

Upon entering the Shiny application, the user will arrive at a screen hat will show three tabs: “Main”, “Abstracts”, and “Titles”, with the Main tab being current. See Figure 32.

5.4.1 The Main Tab

Figure 32: The Abstract Phrase Mining Application

The application at this point will require a MEDLINE input from PubMed, which may be obtained as described in section 4.4.1. Once this file is uploaded to the Abstract Phrase Mining Shiny application, a phraseDoc object is created using the process described in section 5.2.2.2. After that, the information from the phraseDoc object is transformed to a sparse matrix containing phrase/document frequencies, and clustered using k-means as described in section 5.3. This clustering is then presented on the screen as in Figure 33 where we have once again selected the takotsubo file, but this time chosen 20 clusters.

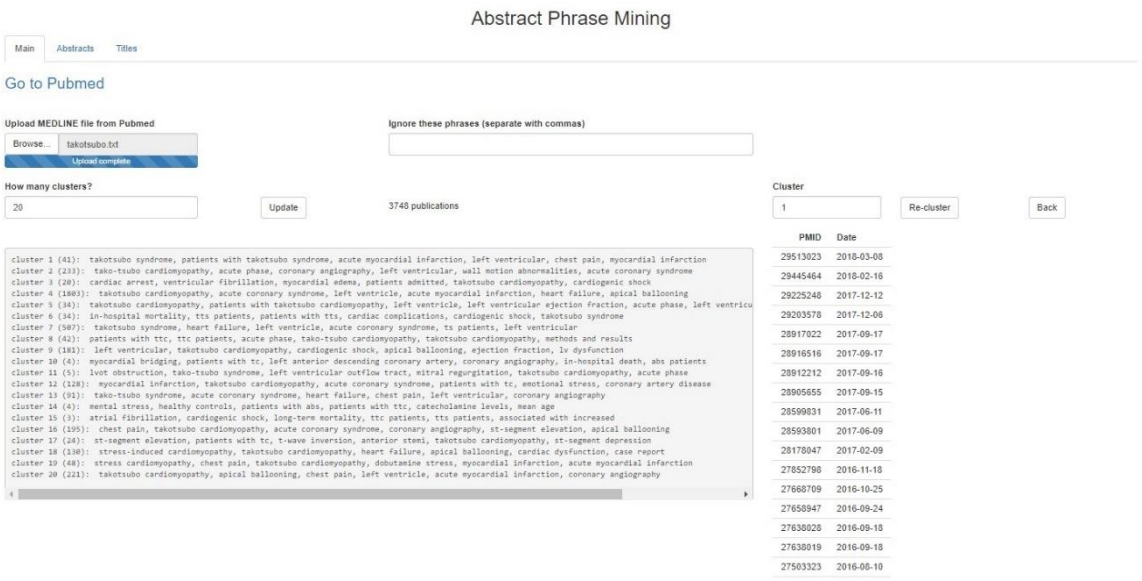

Figure 33: Abstract Phrase Mining on the Takotsubo File

Note that mining for principal phrases is more time consuming than mining for words; a progress bar in the bottom right corner displays the progress of the process.

The total number of publications in the file is displayed as well as the clusters. For each cluster, the number of publications in the cluster is shown in parentheses after the cluster number, followed by the most frequent principal phrases in the cluster. The user may select the number of clusters or may leave the default of 6 clusters in place. The field “Cluster” is by default set to 1 but can be changed. Underneath this field, the PMID and the date of all publications in this cluster are displayed (here cluster 1).

To change the number of clusters or to change the ignored phrases, the user should press the "Update" button. The “Re-cluster” button may be pressed to re-cluster all

publications for the cluster in the “Cluster” field (here cluster 1). The “Back” button can be used to return to a previous clustering.

Specific phrases can be removed from the clustering. For example, in the takotsubo example, the search term for the PubMed file was “takotsubo”, so the phrases “takotsubo cardiomyopathy” and “takotsubo syndrome” are too frequent to be useful. To remove them, they should be entered in the “Ignore these phrases (separate with commas) field. Commas should be placed between every two phrases in this field. They may be accompanied by spaces (but don’t need to be). See Figure 34 where these two fields have been removed.

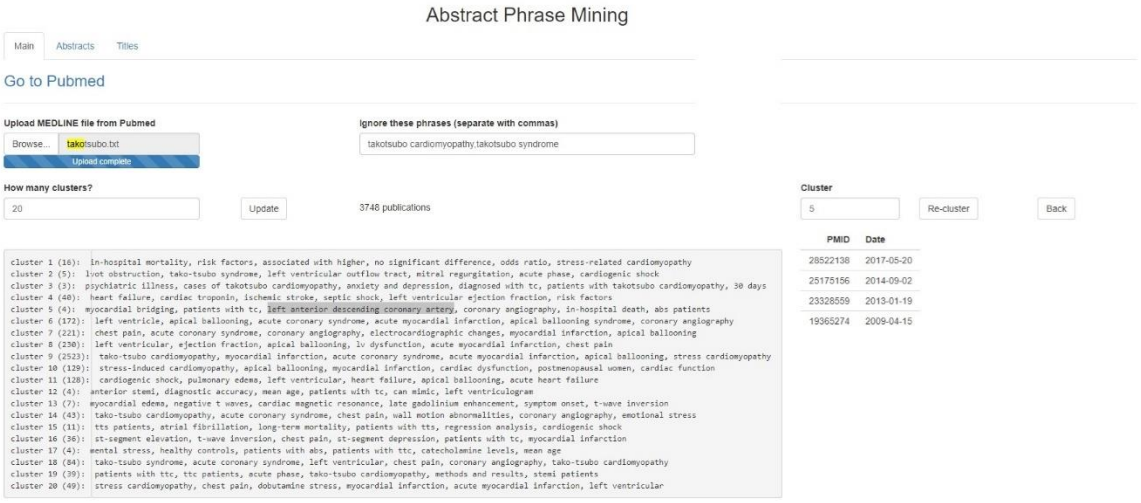

Figure 34: Ignoring Phrases

The cluster output can be examined in order to identify the cluster with a list of common principal phrases that have somewhat unexpected associations. In the above example, the phrase “left anterior descending coronary artery” is considered of interest. The cluster of

interest, which is in this case number 5, should be entered in the “Cluster” field. The user may then select the "Abstracts" tab to examine the abstracts for the publications in the selected cluster, or the "Titles" tab to see the titles.

After the initial clustering is completed, the user may select a specific cluster (which will automatically update the list of PMIDs and dates as described above), and then select "Re-cluster". This will extract all publications in the selected cluster and re-cluster them according to the number of clusters requested. A new set of clusters will then be created that will replace the existing ones.

This process may be repeated indefinitely. If at any time the user wishes to return to a previous state, the "Back" button may be selected. Repeated use of the "Back" button will eventually return to the original set of publications from the MEDLINE file.

5.4.2 The Abstracts Tab

The "Abstracts" tab will display the PMID, date, title and Abstract of the first publication for the cluster selected on the "Main" tab. The user can scroll through all the publications in the cluster using the "Previous" and "Next" buttons underneath the abstract. In addition, the "Download Cluster" button may be selected which will create a html file with all PMIDs, dates, titles, and abstracts in the current cluster. See Figure 35 where PMID 19365274 is selected from cluster 5 of Figure 34. The phrase of interest that resulted in the investigation of this cluster and this publication, “left anterior descending coronary artery”, has been highlighted. We see that it appears twice in this abstract.

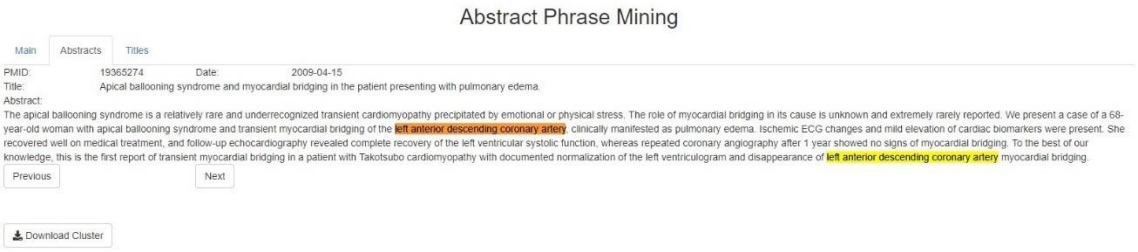

Figure 35: The Abstracts Tab

5.4.3 The Titles Tab

The "Titles" tab will display the PMID, date, and title for all publications in the selected cluster. See Figure 36 where we display the titles for cluster 5 of Figure 34. It appears that our phrase of interest, “left anterior descending coronary artery”, is present in one of the titles of the publications.

Figure 36: The Titles Tab

5.5 Discussion

We described a new method and a set of R-functions to select principal phrases from a vector of texts, as well as determine frequencies of those principal phrases in each of the provided texts. This is a new approach to text mining for which the objective is to provide principal phrases as features for modeling/clustering rather than to provide meaning to large volumes of text.

In addition, we describe a new method to identify new research ideas based on mining principal phrases from bibliographic databases such as PubMed. This may assist investigators in the health sciences. This method allows researchers to start with general search terms and find publications with unusual, unexpected findings of interest for further investigation and potential inclusion in new clinical trials.

This abstract mining application is helpful for creating a clustering structure of principal phrases in abstracts in order to identify interesting publications in a very short time, especially when the search criteria result in a fairly large number of publications. The method can be expanded to other research-oriented websites like Google Scholar, ResearchGate, etc.

Other text mining applications were reported by (Loughran & McDonald, 2011) in finance using text mining to study liabilities, and by (Dalianis, 2018) who used text mining of electronic medical records. More popular is the usage of text mining for analyzing tweets for the purpose of market research and for detecting social network sentiments about a topic such as politics, the economy etc. (Gupta & Bhathal, 2018). The

application reported in this paper is qualitatively different from other text mining applications since instead of describing and analyzing existing information or structures it leads to the development of new research ideas.

5.6 Limitations and Strengths

Like for the abstract mining algorithm, the abstract phrase mining algorithm is also operator dependent in the choice of the search criteria, the number of clusters, re-clustering and other details.

Also, since PubMed is dynamic with publications added daily, findings may not always be reproducible. This issue may be alleviated by keeping (and preferably renaming) the file obtained from PubMed.

This method has significant strengths since it provides fast access to information across many abstracts and may lead to identification of new ideas for research.

5.7 Future Direction

We plan to provide an R package with the functionality described in this dissertation, that provides the capability to extract principal phrases and their frequencies from bodies of texts while removing all double counting of those frequencies.

We aim to install our Shiny application on the Shiny platform in order to provide access to it to the public.

The principal phrase mining process has the potential to be useful for many other applications. This should be investigated and, if shown to be desirable, implemented. It could, for example, be used to perform principal phrase mining on tweets, source code, and many other collections of texts.

The abstract phrase mining process can be expanded to obtain abstracts from other sources, in particular other bibliographical databases.

It may also be worthwhile to investigate methods that will allow us to obtain information directly from the bibliographical databases without an individual having to save a MEDLINE file and load it into the Shiny application manually.

We will sit down with medical experts in order to determine an appropriate list of stop-start words, stop-end words, and stop-phrases. These lists are passed to the principal phrase mining procedure as described in section 5.2.2.2.

Should we find that too many meaningless phrases are included in the clustering, we may utilize the `qp()` function such as described in section 5.2.2.2, to be passed to the

principal phrase mining procedure. This function determines whether a given phrase is principal or not. By default, it will determine this based on frequency alone, but we have the option to replace the default function. We could, for example, add functionality in addition to the frequency requirement. When a phrase passes the frequency requirement, we could calculate features on the phrase and compare those to corresponding features calculated on the list of excluded phrases. If they are comparable, we would exclude the phrase, whereas if the features are not comparable, we would designate the phrase as a principal phrase and include it in the clustering. Alternatively, we could ask our medical experts for a specific list of principal phrases so we can cluster new phrases either with the principal phrases or excluded phrases in order to determine whether to designate a phrase as a principal phrase or not.

References

Abroug, F., Souheil, E., Ouanes, I., Dachraoui, F., Fekih-Hassen, M., & Ouanes Besbes, L. (2015, July). Scorpion-Related Cardiomyopathy: Clinical Characteristics, Pathophysiology, and Treatment. *Clin Toxicol (Phila), 53*(6), 511-8. doi:10.3109/15563650.2015.1030676

Adams, J. U. (2007, Nov 23). Interdisciplinary Research: Building Bridges, Finding Solutions. *Science, 318*(5854), 1315-18. doi:10.1126/science.318.5854.1315

Anter, E., Jessup, M., David, J., & Callans, D. J. (2009). Atrial Fibrillation and Heart Failure. *Circulation, 119*, 2516-2525.

Chen, Y. T., Chen, K. S., Chen, J. S., Lin, W. W., Hu, w. H., Chang, M. K., . . . Chiang, B. N. (1990, Sep). Aortic and Pulmonary Input Impedance in Patients with Cor Pulmonale. *Jpn Heart J., 31*(5), 619-29.

Dalianis, H. (2018). *Clinical Text Mining: Secondary Use of Electronic Patient Records.* Springer.

D'Orio, Lambermont, B., Detry, O., Kolh, P., Potty, P., Gerard, P., & Marcelle, R. (1998, May). Pulmonary Impedance and Right Ventricular-Vascular Coupling in Endotoxin Shock. *Cardiovasc Res., 38*(2), 375-82.

Drazner, M. H. (2011). The Progression of Hypertensive Heart Disease. *Circulation, 123*, 327-334.

Durante, D., & Dunson, D. B. (2018). Bayesian Inference and Testing of Group Differences in Brain Networks. *Bayesian Analysis, 13*(1), 29-58.

Efron, B. (1979). Bootstrap Methods: Another Look at Jackknife. *Ann. Stat., 7*, 1-26.

Feinerer, I. (2008, October). An Introduction to Text Mining in R. *R News, 8*(2), 19-22.

Feinerer, I. (2017). Introduction to the tm Package. Text Mining in R. cran.r-project.org. Retrieved from https://cran.r-project.org/web/packages/tm/vignettes/tm.pdf

Feinerer, I., Hornik, K., & Meyer, D. (2008). Text Mining Infrastructure in R. *Journal of Statistical Software, 25*(5), 1-54.

Gupta, G., & Bhathal, G. S. (2018). *Sentiment Analysis of English Tweets Using Data Mining: Data Mining, Sentiment Analysis.* BookRix.

Hua, W., Mond, H. G., & Strathmore, N. (1997, Jan). Chronic Steroid-Eluting Lead Performance: A Comparison of Atrial and Ventricular Pacing. *Pacing Clin Electrophysiol, 20*(1), 17-24.

Kanazawa, M., Kawamura, K., Takahashi, T., Miura, M., Tanaka, Y., Koyama, M., . . . Shimohata, T. (2015, July). Multiple Therapeutic Effects of Progranulin on Experimental Acute Ischaemic Stroke. *Brain, 138*(7), 1932-48. doi:10.1093/brain/awv079.

Koski, T., & Noble, J. (2011). *Bayesian Networks An Introduction.* John Wiley & Sons.

Kostis, W. J., Deng, Y., Pantazopoulos, J. S., Moreyra, A. E., & Kostis, J. B. (2010, Nov). Myocardial Infarction Data Acquisition System (MIDAS14) Study Group. Trends in Mortality of Acute Myocardial Infarction after Discharge from the Hospital. *Circ Cardiovasc Qual Outcomes, 3(6)*, 581-9.

Kotecha, D., & Piccini, J. P. (2015, Dec 7). Atrial Fibrillation in Heart Failure: What Should We Do? *Eur Heart J, 36(46)*, 3250-7. doi:10.1093/earheartj/ehv513

Larson, R., & Edwards, B. H. (2010). *Calculus of a Single Variable* (9 ed.). Brooks/Cole Cengage Learning.

Liu, J., Shang, J., & Han, J. (2017). *Phrase Mining from Massive Text and Its Applications.* Morgan & Claypool.

Loughran, T., & McDonald, W. (2011, January 6). When is a Liability not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance, 66*(1), 35-65.

Popper, K. R. (1963). *Conjectures and Refutations: The Growth of Scientific Knowledge.* New York: Routledge & Kegan Paul.

PubMed. (1996, Jan). Retrieved from www.ncbi.nlm.nih.gov: https://www.ncbi.nlm.nih.gov/pubmed/

Schemper, M., Kaider, A., Wakounig, S., & Heinze, G. (2013, June 17). Estimating the Correlation of Bivariate Failure Times under Censoring. *Stat Med, 32(27)*, 4781-4790. doi:10.1002/sim.5874

Shiny. (1996, Jan). Retrieved from https://shiny.rstudio.com/

Sugimachi, M., Shishido, T., & Sunagawa, K. (2001, February). Low Compliance Rather than High Reflection of ArterialSystem DecreasesStroke Volume in Arteriosclerosis: A Simulation. *Jpn J Physiol, 51*(1), 43-51.

Swerdel, J. N., Janevic, T. M., Cabrera, J., Cosgrove, N. M., Sedjro, J. E., Pressel, S. L., . . . Kostis, J. B. (2014, Aug). Rapid Decreases in Blood Pressure from Antihypertensive Treatment were Associated with Increased Cancer Mortality in the Systolic Hypertension in the Elderly Program. *Cancer Epidemiology, Biomarkers & Prevention, 23(8)*, 1589-97.

Wystrychowski, G., Kolonko, A., Chudek, J., Zudowska-Szczechowska, E., Wiecek, A., & Grzeszczak, W. (2011, October). Systemic Vascular Hemodynnamics and Transplanted Kidney Survival. *Transplant Proc., 43*(8), 2922-5. doi:10.1016/j.transproceed.2011.08.014

Yue-Cheng, H., Zuo-Cheng, L., Xi-Ming, L., Yuan, D. Z., Dong-Xia, J., Ying-Yi, Z., . . . Hong-Liang, C. (2013, Jan). Long-Term Follow-Up Impact of Dual-Chamber Pacing on Patients with Hypertrophic Obstructive Cardiomyopathy. *Pacing Clin Electrophysiol., 36*(1), 86-93. doi:10.1111/pace.12016

Zhu, H., & Wang, M. (2014, May 16). Nonparametric Inference on Bivariate Survival Data with Interval Sampling: Association Estimation and Testing. *Biometrika, 101*, 519-533. doi:10.1093/biomet/asu005