A FRAMEWORK FOR TREND MINING WITH APPLICATION TO MEDICAL DATA

VASSILIKI SOMARAKI

A thesis submitted to the University of Huddersfield in partial fulfilment of the requirements for the degree of Doctor of Philosophy

The University of Huddersfield in collaboration with Eye and Vision Science Department, University of Liverpool & Saint' Paul Eye Unit, Royal Liverpool University Hospital

June 28 2013

Copyright statement

Abstract

This thesis presents research work conducted in the field of knowledge discovery. It presents an integrated trend-mining framework and SOMA, which is the application of the trend-mining framework in diabetic retinopathy data. Trend mining is the process of identifying and analysing trends in the context of the variation of support of the association/classification rules that have been extracted from longitudinal datasets.

The integrated framework concerns all major processes from data preparation to the extraction of knowledge. At the pre-process stage, data are cleaned, transformed if necessary, and sorted into time-stamped datasets using logic rules. At the next stage, time-stamp datasets are passed through the main processing, in which the ARM technique of matrix algorithm is applied to identify frequent rules with acceptable confidence. Mathematical conditions are applied to classify the sequences of support values into trends. Afterwards, interestingness criteria are applied to obtain interesting knowledge, and a visualization technique is proposed that maps how objects are moving from the previous to the next time stamp.

A validation and verification (external and internal validation) framework is described that aims to ensure that the results at the intermediate stages of the framework are correct and that the framework as a whole can yield results that demonstrate causality. To evaluate the thesis, SOMA was developed.

The dataset is, in itself, also of interest, as it is very noisy (in common with other similar medical datasets) and does not feature a clear association between specific time stamps and subsets of the data. The Royal Liverpool University Hospital has been a major centre for retinopathy research since 1991. Retinopathy is a generic term used to describe damage to the retina of the eye, which can, in the long term, lead to visual loss.

Diabetic retinopathy is used to evaluate the framework, to determine whether SOMA can extract knowledge that is already known to the medics. The results show that those datasets can be used to extract knowledge that can show causality between patients' characteristics such as the age of patient at diagnosis, type of diabetes, duration of diabetes, and diabetic retinopathy.

Table of Contents

List of Figures

List of Tables

Dedications and Acknowledgements

I would like to express my deep and profound gratitude to my principal supervisor, Professor Lee Mc Cluskey, who has always supported my ideas and encouraged their pursuit, and also for his invaluable advice, guidance, and extensive revision of the thesis. He was endlessly patient with me, and I am deeply grateful to him.

I also would like to express my appreciation to my second supervisor, Professor Simon P. Harding, Head of the Department of Eye and Vision Science, for his support, suggestions, and valuable comments, and for his support with the honorary contract with the NHS St Paul's Eye Unit of the Royal Liverpool University Hospital. Special thanks also to all the staff in the Department of Computer and Engineering at The University of Huddersfield and the St. Paul's Eye Unit at the Royal Liverpool University Hospital, who have been helpful whenever needed. I am also very thankful to Professor Deborah Broadbent, Director of Diabetes in Department of Eye and Vision Science, Royal Liverpool Hospital, who has been helpful in providing advice many times during my work on databases. Special thanks to other members of staff of the University of the Huddersfield especially to Professor Joan Lu and Dr Di Cai for their valuable help and advice during my studies. I would like also to thank deeply Dimitrios Tsaltas and Thomas Politis with their support and valuable advices during my studies.

Last, but not least, love and gratitude to my beloved family, especially to my parents and my partner, Nikos, for their understanding and constant support; without them, the completion of my Ph.D. would not have been possible.

Chapter 1 Introduction

Ignorance is the curse of God, knowledge the wing wherewith we fly to heaven—
William Shakespeare

We now live in the information age. "Data owners" such as scientists, businesses, and medical researchers, are able to gather, store, and manage previously unimaginable quantities of data owing to technological advances and economic sciences in sensors, digital memory, and data-management techniques. In 1991, it was proposed that the amount of data stored in the world doubles every 20 months (Piatetsky-Shapiro and Frawley, 1991). At the same time, there is a growing realization and expectation that data, intelligently analysed and presented, will be a valuable resource to gain a competitive advantage.

Knowledge Discovery (KD) is a non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns from large collections of data (Fayyad et al., 1996). One of the steps KD is Data Mining (DM). DM is concerned with the actual extraction of knowledge from data, in contrast to the KD process, which is concerned with many other things such as understanding and preparation of the data, verification, and application of the discovered knowledge. In practice, however, people use terms DM, KD, and DMKD synonymously (Cios et al., 2002). The design of a framework for a KD process is an important issue. Several researchers have described a series of steps that constitute the KD process, ranging from very simple models, incorporating few steps that usually include data collection and understanding, DM, and implementation, to more sophisticated models such as the nine-step model proposed by Fayyad et al. (1996) or the six-step DMKD process model proposed by Cios et al. (2000) and Cios and Moore (2000). Cios et al. (2000) applied the model to several medical problem domains (Sacha et al., 2000; Kurgan et al., 2001, 2003).

To bridge the growing gap between data generation and data understanding, there is an urgent need for new computational theories and tools to assist humans in extracting useful knowledge from the huge volumes of data. These theories and tools are the subject of the emerging field of Knowledge Discovery in Databases (KDD), or DM, which sits at the common frontiers of several attributes including Database Management, Artificial Intelligence, Machine Learning, Pattern Recognition, and Data Visualization (Hand, 1994).

DM is a multidisciplinary field, drawing work from areas including database technology, machine learning, statistics, pattern recognition, information retrieval, neural networks, knowledge-based systems, artificial intelligence, high performance computing, and data visualization" (Han and Kamber, 2006).

In the past decade, DM techniques have been widely applied in bioinformatics (Wang et al., 2005), e-commerce (Raghavan, 2005), financial studies (Kovalerchun and Vityaev, 2000), geography (Miller and Han, 2001), marketing and sales studies (Berry and Linoff, 1997; Rypielski et al., 2002), etc.

Most DM applications routinely require datasets that are considerably larger than those that have been addressed by traditional statistical procedures. The size of the datasets often means that traditional statistical algorithms are too slow for DM problems, and alternatives have to be devised. The volume of the data is probably not very important: the number of variables or attributes often is much more important. The analysis of the way in which data change with time is an important mechanism for providing information for decision-makers, policy-makers and other "stakeholders". One way of conducting such an analysis is by considering data trends.

Trends can be defined and generated in a number of ways. One mechanism, and the focus of the work to be undertaken here, is to define trends in terms of the way that the frequency of occurrence of patterns changes with time and to employ DM techniques to identify such trends. In this work, the term "trend mining" has been adopted to describe this discovery process.

Trend mining is the process of identifying and analysing trends in the context of the variation of the support of the association rules that have been extracted from longitudinal datasets. The proposed trend-mining mechanism is founded on an Association Rule Mining (ARM) approach whereby an ARM technique is applied to a sequence of time-stamped data sets. This approach is both efficient and effective in finding trends.

The temporal data to which trend mining can be applied can take many forms; one common form of data is longitudinal data. One application domain that features data sets that are both large and temporal is medical records or, more specifically, patient records. Many branches of medicine have collected large longitudinal data sets spanning many years. These data sets in themselves constitute a wealth of information.

This type of data is of particular interest, in the context of trend mining, as the "time stamps" are defined in terms of patient visit number, as opposed to more traditional forms of temporal data. The data are also extremely noisy.

The field of medical informatics has evolved around structuring, processing, storing, and transmitting medical information for a variety of purposes (Shortliffe, 1990). One of these purposes is to develop decision-support systems that enhance the human ability to diagnose, treat, and assess prognoses of pathological conditions. Even if disease processes were fully understood, population variability would still make individualized diagnosis, treatment, and prognosis— all essential parts of good health care—difficult classification tasks. The reality is, however, that diseases are not fully understood; nor is population variability fully taken into account in many decision-making situations. Sometimes it is not possible for a clinician to employ the principles learned in the basic and clinical sciences to determine whether a patient has a given disease, whether he or she should be given a certain treatment, and how long he or she will survive.

Trends across time-stamped data sets can therefore be identified by observing the change in the support values of items sets across the data set. Trend mining is a branch of DM that focuses on the process of identifying and analysing hidden trends in temporal data. The study described in this work is directed towards longitudinal patient data (longitudinal data are data collected using the same set of attributes at a series of points over time), more specifically diabetic retinopathy screening data collected by St. Paul’s Eye Unit, Royal Liverpool University. Diabetic retinopathy (DR) is a common complication of diabetes, the most common cause of blindness in working-age people in the UK. DR is a chronic disease affecting patients with diabetes mellitus and causes damage to the retina (Kanski, 2007). Over 3,000,000 people suffer from diabetes, at least 750,000 people are registered blind or partially sighted in the UK, and the remainder are at risk of blindness. Consequently, it is important that DR be diagnosed at an early stage, and accurately.

The research objective of the work is to investigate and identify a mechanism or mechanisms, whereby longitudinal data trends can be mined and the results presented in such a way that informed decisions can be made by policy-makers, etc. Broadly, this entails a number of issues:

- The mechanism for the pre-processing of the longitudinal data required to permit the desired trend mining.
- The nature of the trend mining mechanisms to be employed.
- The identification of the process to be used to present the results in a meaningful way. Longitudinal data thus provide a record of the “progress” of some set of features associated with the subjects. Medical longitudinal data, such as DR data, typically plot the progress of a medical condition.
- Longitudinal data thus implicitly contain information concerning trends.

This work has resulted in a novel trend mining framework along with a validation and verification framework, and also resulted with an evaluation application called "SOMA", which not only enables trend mining but also supports the validation of discovered trends. This validation is based upon the selection of certain attributes for which there are known associations. Having known these associations as well as the patterns they change, trends can be identified using mathematical conditions. Hence, the main purpose of this research is to develop a novel trend-mining framework for extracting trends from longitudinal data while emphasizing the validation of those trends.

This thesis introduces the described method as a general framework for trend-mining validation and verification that can be applied generally to most trend procedures and types of data. The work also introduces trend mining, provides a description of the validation framework, and includes the experimental evaluation of the application.

1.1 Contribution and research questions

Trend mining remains an open challenge in the field of knowledge discovery in data (KDD). This can be attributed partly to the lack of a clear definition of what we mean by a "trend" (it is very much an application-dependent definition), and partly to issues associated with the modelling of time-stamped (longitudinal) data. This research addresses the development of a trend-mining framework for knowledge discovery from large databases, the development of a validation framework for trend mining, and the application of trend mining in medical data (SOMA). The trend-mining framework is an integrated platform which can be used for knowledge discovery in databases starting from the pre-processing of data and ending with the extraction of useful information.

The research questions which arise from this research work are:

- What is the most appropriate mechanism to identify, analyse and validate trends in real, noisy and longitudinal data and in particular when the input time stamped data denote patients' progress?
- Can this mechanism produce trends that may be employed for prediction purposes?
- Can trend mining be applied on medical applications?

More detailed, this thesis examines the following issues:

- How can frequent patterns and trends be discovered to facilitate the desired trend mining?
- How can changes be detected in the identified trends?
- How the large numbers of generated trends are handled? How can the interestingness of these trends be measured?

- Can be applied constraints to the data in order to anticipate interesting, desirable and useful trends?
- How can different types of trends be interpreted to the users?
- What criteria can be used to validate and verify the framework?

In the pre-processing stage the framework directed to solve issues that they arise after bringing together data from various sources such as missing values, heterogeneity of data (combination of numerical with discrete or continuous values and or categorical data) and creation of time stamps. The creation of time stamped data is very important issue for the framework. The data do not feature a clear association between specific time stamps and subsets of data. These include any time-stamped subset of data comprising data collected at different dates and stored in different locations. As a result, the creation of time-stamped subsets for analysis is not straightforward and they form quasi-longitudinal data.

At the processing stage, the framework, through the combination of association rule mining (ARM) and prototype mathematical conditions, deals with the following challenges:

- identifying temporal patterns (associations) that commonly occur in the input data;
- working on distinguish interesting knowledge through a large amount of temporal patterns
- identifying change points of state changes in temporal sequences or, alternatively, the lack of such state changes;
- the grouping (clustering) of data according to some temporal change;
- the classification of temporal data sequences.
- The knowledge that is extracted from trend mining depicts how the initial conditions (that describe a situation) of a group (e.g. patients) change over time. This type of change is called a state change. To visualize any possible state stage, a colourful representation scheme is used to interpret the results.

This thesis also aims to produce a consistent framework for validation of trend mining. The trend-mining method essentially performs "learning by discovery", and so it cannot be trained; rather, the user has to have confidence in the results it gives, that is, it should be validated. To perform validation and verification of the trend-mining framework, two complementary approaches are advocated here:
Validation: This method tests the outputs of the framework and also checks the consistency in the application that experts already know and expect. The methods include: confirmation of the framework that uncovers known causal connections in the application and confirmation of the framework that uncovers known trends in the application.
Verification: This type of validation tests whether the intermediate results and/or outputs of the framework are self-consistent. At each of the intermediate stages, a language for a set

of declarative validation rules are set up, and a systematic process of validation of each set of input data is created.

For the development and application of both the trend-mining framework and its validation framework, real world medical data are used in this research. Data came from the Diabetic Retinopathy databases maintained by the Royal Liverpool University Hospital, and these data are an example of an irregular database in that they contain 150,000 records comprising 450 attributes distributed over two databases, each composed of a number of tables.

SOMA, the application of the framework over those data, consists of three steps:

- Pre-processing: data from different sources are brought together after applying logic rules to deal with problems arising from the nature of data and to create a time-stamped subset for analysis.
- ARM stage where, for each time-stamped subset, the matrix algorithm technique is used to identify the rules, which are determined by the user specifying which are "variable attributes", or the left-hand side of the rule, and which are the key attributes, or the right-hand side of the rule, whose support and confidence exceeds the user's specified threshold values. Matrix algorithm (Yuan and Huang, 2005) it is a novel algorithm for the identification of frequent item sets based on the creation of a matrix with binary entries and its main advantage is that only one passing it is needed.
- The trend mining taking information from the ARM stage creates trends using prototypes (mathematical conditions), which show the attitude of the rules over time. Beyond this, trend mining creates a colourful representation that shows how a group of patients moving from one time-stamp to another either remain with the same rule or may move to another rule owing to changes in some of their characteristics.

A novel algorithm was created to implement the above trend mining framework. The algorithm consists of 3 parts: the first part implements pre-processing, the 2nd part is the main processing and the last part is the post-processing and outputting the results.

In order to minimize as much as possible the interaction between the user and the process, the novelty of the script is the transformation of the input attribute names and their values into a numerical language which is recognized from all parts of the algorithm without the need of any action required by the user when the algorithm proceeds from one stage to the next. The problem with existing algorithms found in the literature was that each one has its own way of reading data and as a result more work was required to prepare the data, especially to go from the pre-processing stage to the main processing stage.

1.2 Thesis Structure

The rest of this thesis is organised as follows:

Chapter 2 presents the literature review which describes the background of current KDD research with respect to a variety of methodologies in both data mining in general and trend mining in particular. Also Review of Association Rule Mining, review of trend mining (similar approaches for identifying change, such as Emerging Patterns and Jumping EPs, to which the work can be compared), review of the nature of longitudinal data.

Chapter 3 presents the Medical overview of Diabetic retinopathy and review of the data used and the challenging aspects of these data, the warehouse, logic rules, and the pre-processing. Draw out the fact that the data being used is different to more standard temporal data sets in terms of the concept of episodes. Include description of methods including definitions and schemas.

Chapter 4 introduces the Trend mining framework and description. An approach to trend mining is to use the concept of user defined temporal prototypes to define the nature of the trends of interests. The trends are defined in terms of sequences of support values associated with identified frequent patterns. The prototypes are defined mathematically so that they can be mapped onto the temporal patterns. A process to validate the intermediate data sets and the results of the trend mining process is presented. This is about how to deal with the main challenge of the framework which is how to evaluate the results of trend mining. The primary information that is required for the validation stage is a set of "expected" associations between features, given that all these features have been represented as inputs. These associations represent actual known relationships between features. The purpose is to produce a consistent framework for validation:

at each of the intermediate stages; a language for a set of declarative validation rules will be set up, and a systematic process of validation for each set of input data will be created.

at the end of the data mining pipeline, a process for testing for known associations (the expected outputs) will be created. Thus, it may use invariants and characteristics of the data to prune and/or synthesise output rules to fit the associations being looked for. Such complex "pipelines" of processes are fraught with various kinds of errors and biases that can creep into the process at each stage. When the process is for use as a research aid applied to sets of patient data, then the integrity of the data and the reliability of the results are particularly important. To promote the quality of the data mining process, and the outputs of the whole process, we propose to extend the current framework to one that incorporates a systematic method for validation: that is to check that the results obtained accord with the domain (in this case the domain of diabetes retinopathy). At the same time we will investigate the scope and value of incorporating verification checks: that is to check that the sequence of processes are working correctly. The trend mining framework application SOMA, and Aretaeus, the associated trend mining algorithm have been developed. The application is used to detect different kinds of trends across longitudinal medical datasets.

Chapter 5 details the evaluation of research work on trend mining. The aim is to evaluate the approach for the development of the advocated trend mining framework. The goal of evaluation process described here is to judge the usefulness of the discovered knowledge and the process of trend mining itself. On the one hand the evaluation of the produced rules is straightforward by using criteria evaluating novelty action ability unexpectedness reliability etc, on the other hand evaluating the processes of the framework is based on quantitative criteria which measure the performance. The evaluation by applying the framework to the DR data examines if the validation and verification are effective as part of the framework.

Chapter 6 conclude the thesis and present a summary of research work along with main findings and future work.

Finally, the Appendices present information on the data used (schemas), tables and figures from evaluation experiments.

1.3 Publications

The following papers were produced as part of the research described in this thesis:

Somaraki V., Broadbent D., Harding P.S., Coenen F. (2010). Finding temporal patterns in noisy longitudinal data: A study in diabetic retinopathy. Perner, Petra (ed.), Advances in Data Mining. Applications and Theoretical Aspects. 10th Industrial Conference, ICDM 2010, Berlin, Germany, July 12-14, 2010. Proceedings. Berlin: Springer. Lecture Notes in Computer Science 6171. Lecture Notes in Artificial Intelligence, 418–431.

This paper describes an approach to temporal pattern mining using the concept of user defined temporal prototypes to define the nature of the trends of interests. The temporal patterns are defined in terms of sequences of support values associated with identified frequent patterns. The prototypes are defined mathematically so that they can be mapped onto the temporal patterns. The focus for the advocated temporal pattern mining process is a large longitudinal patient database collected as part of a diabetic retinopathy screening programme, The data set is, in itself, also of interest as it is very noisy (in common with other similar medical datasets) and does not feature a clear association between specific time stamps and subsets of the data. The diabetic retinopathy application, the data warehousing and cleaning process, and the frequent pattern mining procedure (together with the application of the prototype concept) are all described in the paper. An evaluation of the frequent pattern mining process is also presented.

Somaraki V., Harding P.S., Broadbent D., Coenen F. (2010).SOMA: A Proposed Framework for Trend Mining in Large UK Diabetic Retinopathy Temporal Databases. Research and Development in Intelligent Systems XXVII Proceedings of AI-2010, The Thirtieth SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence Bramer, Max; Petridis, Miltos; Hopgood, Adrian (Eds.)

This paper is a continuation and extension of the previous paper and how the proposed framework is able to detect different kinds of trends within the SOMA application and how the proposed framework is able to detect different kinds of trends within longitudinal datasets. To evaluate the proposed framework the process was applied to a large collection of medical records, forming part of the diabetic retinopathy screening programme at the Royal Liverpool University Hospital.

Somaraki V., McCluskey L. (2012). Robust Validation Framework for Trend Mining. Diamond Jubilee Annual Researchers' Conference, University of Huddersfield.

An extended framework for Validation of trend mining framework is described in this paper. To validate the framework in the analysis of the generated trends a mechanism is also proposed. The framework is evaluated using longitudinal Diabetic Retinopathy screening data.

Chapter 2 Literature review

2.1 Introduction

This thesis describes an approach to finding temporal patterns in noisy longitudinal patient data and an extended internal and external validation (validation and verification) process to validate the framework. The identification of patterns in such data has many applications. One common example is the analysis of questionnaire returns collated over a number of years, for example Kimm et al.,(2000) studied the nature of physical activity in groups of adolescents and Skinner et al. studied children's food eating habits (Skinner et.al,2002).Another example of the application of longitudinal studies is in the analysis of statistical trends; an early reported example is that of Wagner (1992),who performed an extensive longitudinal study of children with special educational needs". Longitudinal studies particularly lend themselves to the analysis of patient data in medical environments where records of a series of "consultations" are available. For example Yamaguchi et. al., (2001) studied the effect of treatments for shoulder injuries, and (Levy et, al., 1996) studied the long term effects of Alzheimer's disease.

In this chapter a literature review is presented on topics that are related to the development of the trend mining framework. Firstly, is given an overview of data mining and also the following aspects are covered:

- Association Rule Mining (ARM): they are presented as algorithms for the discovery of association rules in datasets and a set of criteria for the definition of what is an interesting rule.
- Associative Classification (AC): how ARM can be used to build a classifier.
- Data mining in medical applications or Medical Data Mining(MDM): here it is presente~~d~~ how data mining techniques are applied to extract knowledge from medical data
- Trend Mining(TM): here is presented the definition of trend mining and the work on emerging and jumping patterns which are the cornerstone on which a new trend mining algorithm is built.
- Verification and validation: in this part the difference between internal and external validation is given what other researchers did in that field and what strategy will be implemented here.

2.2 Knowledge discovery in databases process

Knowledge discovery in databases (KDD) has been attracting a huge amount of research, for business, media, social network, health care, etc. As data volumes have grown dramatically, manual analysis and interpretation of data have become impractical for many domains. KDD is the overall process of discovery of novel, potentially useful, and ultimately understandable patterns in data (Fayyad, Piatetsky-Shapiro, and Smyth 1996a).

The KDD process consists of several steps (Fayyad, Piatetsky-Shapiro, and Smyth 1996b), which are shown in Figure 2.1.

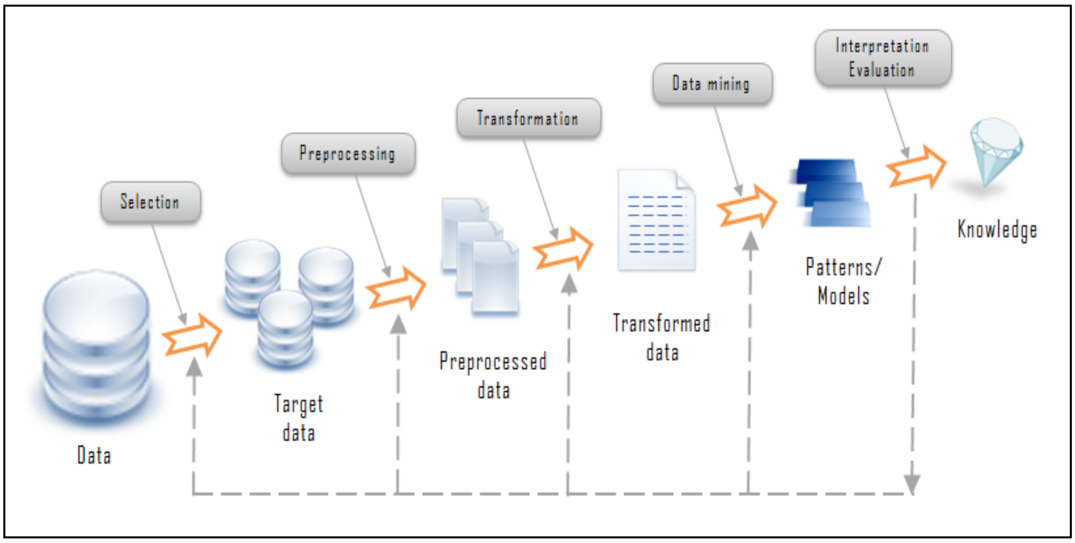


Figure 2.1 :Overview of the KDD process (Fayyad, Piatetsky-Shapiro, and Smyth 1996a)

1. Developing an understanding of the application domain and the relevant prior knowledge and identifying the goal of the KDD process from the user's viewpoint.
2. Creating a target data set: selecting a data set, or focusing on a subset of variables or data samples, on which discovery is to be performed.
3. Cleaning and pre-processing. Basic operations include removing noise if appropriate, collecting the necessary information to model or account for noise, deciding on strategies for handling missing data attributes, and accounting for time-sequence information and known changes.
4. Data reduction and projection: finding useful features to represent the data depending on the goal of the task. With dimensionality reduction or transformation.

5. Matching the goals of the KDD process (step 1) to a particular data-mining method. For example, summarization, classification, regression, clustering, and so on.
6. Exploratory analysis and model and hypothesis selection: choosing the data mining algorithm(s) and selecting method(s) to be used for searching for data patterns. This process includes deciding which models and parameters might be appropriate and matching a particular data-mining method with the overall criteria of the KDD process (for example, the end user might be more interested in understanding the model than its predictive capabilities).
7. Data mining: searching for patterns of interest in a particular representational form or a set of such representations, including classification rules or trees, regression, and clustering. The user can significantly aid the data-mining method by correctly performing the preceding steps.
8. Interpreting mined patterns: possibly returning to any of steps 1 through 7 for further iteration. This step can also involve visualization of the extracted patterns and models or visualization of the data given the extracted models.
9. Acting on the discovered knowledge: using the knowledge directly, incorporating the knowledge into another system for further action, or simply documenting it and reporting it to interested parties. This process also includes checking for and resolving potential conflicts with previously believed (or extracted) knowledge.

2.3 Data Mining methods

As noted above, data mining (DM) is part of the KDD process, and it is the stage where knowledge discovery takes place. As a highly application driven-domain, DM has incorporated many techniques from other domains such as statistics, machine learning, pattern recognition, visualization, algorithms, and many more (Figure 2). The overall goal of the DM process is to extract information from a data set and transform it into an understandable structure for further use. As a general technology, DM can be applied in many forms of data such as: database data, warehouse data, transactional data, medical data, data streams, sequence data, multimedia data, text data, spatial data, and web data.

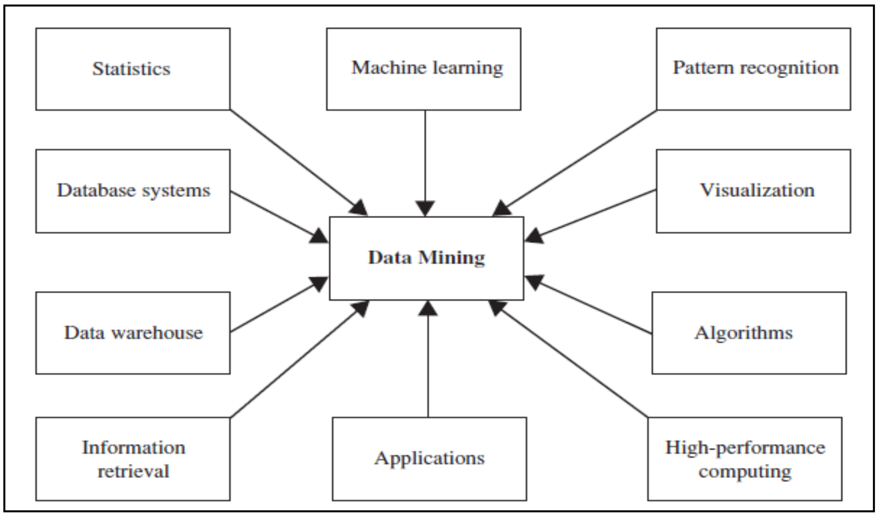


Figure 2.2 : Techniques for DM (Han et al., 2011)

Two high-level primary goals of DM in practice tend to be prediction and description. Prediction involves using some variables or attributes in the database to predict unknown or future values of other variables of interest, and description focuses on finding human-interpretable patterns describing the data. Although the boundaries between prediction and description are not always distinct (some of the predictive models can be descriptive to the extent that they are understandable, and vice versa), the distinction is useful for understanding the overall discovery goal. The relative importance of prediction and description for particular DM applications can vary considerably. The goals of prediction and description can be achieved using a variety of particular DM methods. Some of the methods for DM are described below (Witten and Frank, 2005; Han et al., 2011):

- Regression is learning a function that maps a data item to a real-valued prediction variable. There are many regression applications, such as predicting the amount of biomass present in a forest given remotely sensed microwave measurements, estimating the probability that a patient will survive given the results of a set of diagnostic tests, predicting consumer demand for a new product as a function of advertising expenditure, and predicting time series where the input variables can be time-lagged versions of the prediction variable.
- Classification is learning a function that maps (classifies) a data item into one of several predefined classes (Weiss and Kulikowski, 1991; Hand, 1981). Examples of classification methods used as part of knowledge discovery applications include the classifying of trends in financial markets (Apte and Hong, 1996) and the automated identification of objects of interest in large image databases (Fayyad, Djorgovski, and Weir, 1996).

The classifiers used are generated using what are called supervised learning methods in that they require pre-labelled training data.

- Clustering is a common descriptive task where one seeks to identify a finite set of categories or clusters to describe the data (Jain and Dubes, 1988; Titterington, Smith, and Makov, 1985). The categories can be mutually exclusive and exhaustive, or consist of a richer representation, such as hierarchical or overlapping categories. Examples of clustering applications in a knowledge discovery context include discovering homogeneous subpopulations for consumers in marketing databases and identifying subcategories of spectra from infrared sky measurements. Closely related to clustering is the task of probability-density estimation, which consists of techniques for estimating from data the joint multivariate probability density function of all the variables or attributes in the database (Silverman, 1986).
- Summarization involves methods for finding a compact description for a subset of data. A simple example would be tabulating the mean and standard deviations for all attributes. More sophisticated methods involve the derivation of summary rules (Agrawal et al., 1996), multivariate visualization techniques, and the discovery of functional relationships between variables (Zembowicz and Zytkow, 1996). Summarization techniques are often applied to interactive exploratory data analysis and automated report generation.
- Dependency modelling consists of finding a model that describes significant dependencies between variables. Dependency models exist at two levels: (1) the structural level of the model specifies (often in graphic form) which variables are locally dependent on each other, and (2) the quantitative level of the model specifies the strengths of the dependencies using a numeric scale. For example, probabilistic dependency networks use conditional independence to specify the structural aspect of the model and probabilities or correlations to specify the strengths of the dependencies (Glymour et al., 1987; Heckerman, 1996). Probabilistic dependency networks are increasingly finding applications in areas as diverse as the development of probabilistic medical expert systems from databases, information retrieval, and modelling of the human genome.
- Decision trees and rules that use univariate splits have a simple representational form, making the inferred model relatively easy for the user to comprehend. However, the restriction to a particular tree or rule representation can significantly restrict the functional form (and, thus, the

approximation power) of the model. A large number of decision tree and rule-induction algorithms are described in the machine learning and applied statistics literature (Quinlan, 1992; Breiman et al., 1984). To a large extent, they depend on likelihood-based model-evaluation methods, with varying degrees of sophistication in terms of penalizing model complexity. Greedy search methods, which involve growing and pruning rule and tree structures, are typically used to explore the super-exponential space of possible models. Trees and rules are primarily used for predictive modelling, both for classification (Apte and Hong, 1996; Fayyad, Djorgovski, and Weir, 1996) and for regression, although they can also be applied to summary descriptive modelling (Agrawal et al., 1996).

2.4 Association Rule Mining

Association Rule Mining (ARM) consists of first finding frequent item sets (set of items A and B) from which strong association rules in the form A=>B are generated. ARM was first proposed by Agrawal, Imielinski and Swami (1993). It is an important task in DM that finds correlations between items in a database. ARM is an unsupervised DM method. The classic application for ARM is market basket analysis (Agrawal et al., 1993; Agrawal and Srikant, 1994), in which business experts aim to investigate the shopping behaviour of customers in an attempt to discover regularities. In finding association rules, one tries to find groups of items that are frequently sold together in order to infer certain items from the presence of other items in the customer's shopping cart.

Agrawal and Srikant (1994) defined the task of association rule discovery as follows: Let D be a database of sales transactions, and I = {i_1, i_2, …, i_m} be a set of binary literals called items. A transaction T in D contains a set of non-empty items called an item set, such that T $\subseteq$ I. The support of an item set is defined as the proportion of transactions in D that contain that item set. An association rule is an expression $X \rightarrow Y$, where X, Y $\subseteq$ I and $X \cap Y = \theta$. The confidence of an association rule is defined as the probability that a transaction contains Y given that it contains X, and given as support (X$\cup$Y)/support(X). Given a transactional database D, the association rule problem is to find all rules that have supports and confidences greater than certain user-specified thresholds, denoted by minsupp and minconf, respectively.

The problem of producing all association rules from a transactional database can be decomposed into two sub-problems according to (Agrawal and Srikant, 1994):

- Step 1. The generation of all item sets with support greater than the minsupp threshold. These item sets are called frequent item sets. All other item sets are called infrequent.
- Step 2. For each frequent item set generated in Step1, produce all rules that pass the minconf threshold. For example if item XYZ is frequent, then we might evaluate the confidence of rules $XY \to Z$, $XZ \to Y$ and $YZ \to X$.

While the second step that involves generating the rules from the set of discovered frequent item sets is straightforward, given that frequent item sets and their supports are known (Han et al., 2000; Lim et al., 2000), the first step of finding frequent item sets is a relatively harder problem that requires extensive computation and storage (Zaki et al., 1997; Cheung et al., 1997; Lin and Dunham, 1998; Lim et al., 2000). If we consider a database that contains 1500 different distinct items, there are 2^{1500} possible different combinations of candidate item sets, most of which do not appear even once in the database. Only a small subset of this large number of candidate item sets are frequent. Many researchers have extensively investigated the problem of finding frequent item sets in association rule discovery in the last decade for the purpose of improving its efficiency (Park et al., 1995; Liu et al., 1999; Li et al., 1999; Zaki, 2000; Bayardo and Agrawal, 1999; Baralis et al., 2004).

2.4.1 Apriori algorithm

Apriori is an algorithm that has been proposed in Agrawal and Srikant (1994), and its name is based on the fact that it uses prior knowledge of frequent item sets. As mentioned earlier, the discovery of frequent item sets is accomplished in a stepwise fashion, where, in each iteration, a full pass over the training data is required to generate new candidate item sets from frequent item sets already found in the previous step. Apriori uses the "downward-closure" property, aiming to improve the efficiency of the search process by reducing the size of the candidate item sets list during each iteration.

The Apriori algorithm for finding frequent item sets is shown in Figure 2-3, where the generate candidate function shown in Figure 2-4, is used to produce C_n from F_{n-1} by merging F_{n-1} with F_{n-1}, and discarding all item sets in C_n that do not pass the support threshold.

```
1.  DB : Transactional database
2.  **F**$_{n:}$ Set of n-items that pass the *minsupp* threshold (frequent item sets)
3.  C$_{n:}$ Set of n-candidate item sets that are possibly frequent
4.  F$_1$={frequent 1-item sets};
5.  for (n=2; F$_{n-1}$≠Ø; n++) Do
6.  C$_n$=generate_candidates(F$_{n-1}$);
7.  for each transaction T∈DB  Do
8.  P$_t$ = subset(C$_n$, t)
9.  for each candidate c∈P$_t$
10. c.count++;
11. end //for
12. F$_n$={c∈C$_n$| c.count≥ *minsupp*}
13. 10. end// for
14. output = $\cup_n F_n$
```

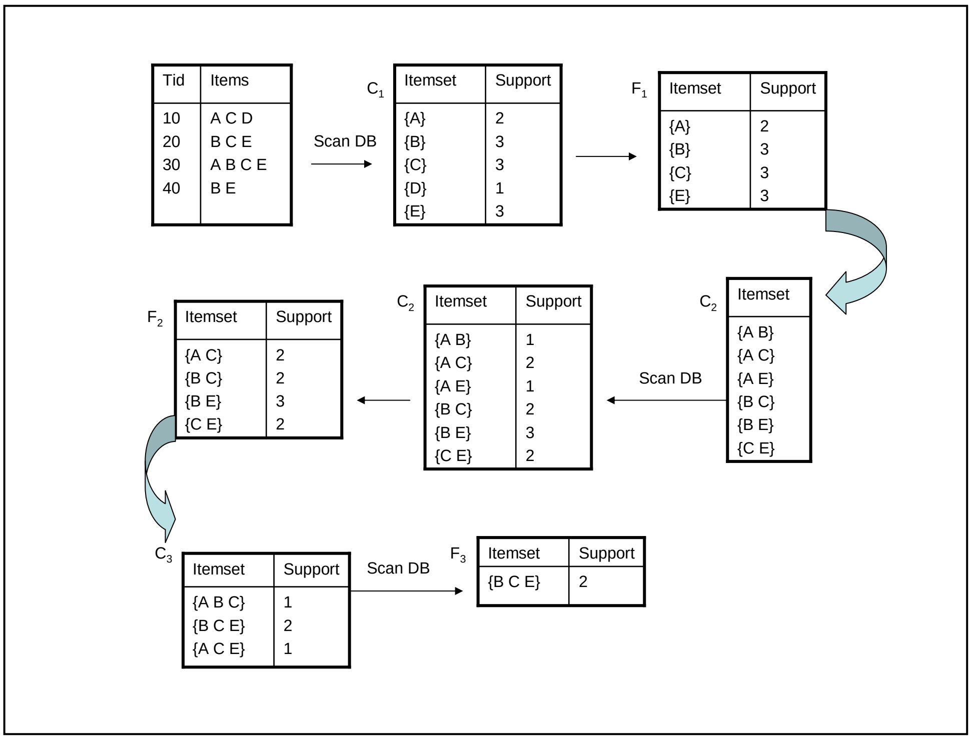
Figure 2.3 Apriori Algorithm

The subset function (line 5) finds the subset of candidate item sets contained in the current database transactional (t). Once these candidate item sets are identified from C$_n$, their supports are incremented (line 6-7). The algorithm terminates whenever there are no frequent item sets Fn in the nth iteration.

```
1.  Function Generate_candidate(F$_k$)
2.  begin
3.  C :=0;
4.  for all $f_1, f_2 \in F_k$
5.  with $f_1 = \{i_1,..., i_{k-1}, i_k\}$
6.  and $f_2 = \{i_1,..., i_{k-1}, i'_k\}$
7.  and $i_k < i'_k$ Do
8.  $f := f_1 \cup f_2 = \{i_1, i_2,..., i_{k-1}, i_k, i$
9.  if $\forall_i \in f : f - \{i\} \in F_k$
10. $C := C \cup \{f\};$
11. end if
12. end
13. return $C$
```

Figure 2.4 Apriori Generate_Candidate function

To illustrate the discovery of frequent item sets in Apriori, consider Figure 2-5, which shows the steps of Apriori's candidate generation described in Figure 2-4 on a database using a *minsupp* of 2. As shown in Figure 2-5, Apriori scans the database to find candidate item sets of length 1, and from which it determines those that pass the support threshold (F_1). In the second level, the algorithm generates candidate item set of size two 2 (C_2) and scans the database to determine which subset of them is frequent (F_2). The algorithm finally terminates after discovering frequent item sets of length three (F_3). For the database shown in Figure 2-5, Apriori requires three passes over the database in order to discover the complete set of frequent item sets.



Tid	Items
10	A C D
20	B C E
30	A B C E
40	B E

C_1

Itemset	Support
{A}	2
{B}	3
{C}	3
{D}	1
{E}	3

F_1

Itemset	Support
{A}	2
{B}	3
{C}	3
{E}	3

C_2

Itemset
{A B}
{A C}
{A E}
{B C}
{B E}
{C E}

C_2

Itemset	Support
{A B}	1
{A C}	2
{A E}	1
{B C}	2
{B E}	3
{C E}	2

F_2

Itemset	Support
{A C}	2
{B C}	2
{B E}	3
{C E}	2

C_3

Itemset	Support
{A B C}	1
{B C E}	2
{A C E}	1

F_3

Itemset	Support
{B C E}	2

Figure 2.5 :Apriori candidate generation example

2.4.2 Dynamic item-set counting

To speed up the discovery of frequent item sets in a database, a new ARM algorithm called Dynamic Item set Counting (DIC) was developed in Brin et al. (1997). DIC splits the database into several partitions marked by start points. Then, it calculates the supports of all item sets counted so far, dynamically adding new candidate item sets whenever their subsets are determined to be frequent, even if their subsets have not yet been seen at all transactions. The main difference between DIC and Apriori is that whenever a candidate item set reaches the support during a particular scan, DIC starts producing additional candidate item sets based on it, without waiting to complete the scan as Apriori does.

To accomplish the dynamic candidate item sets generation, DIC employs a prefix tree where each item counted so far is associated with a node. One of the drawbacks of DIC algorithm is its sensitivity to how homogeneous the data are. Particularly, if the database to be mined is correlated, DIC cannot recognnise that an item set is frequent unless it has been seen in most transactions.
Experimental results using census and synthetic data sets (Agrawal and Srikant, 1994) indicated that DIC is faster by 30.00% than Apriori at a support threshold of 0.5% on the synthetic database. On the large and highly correlated census database, DIC outperformed Apriori at a support threshold of 36.00%. Both algorithms require a long period of training when the support is lowered, since the items in the census database occur frequently 95% of the time and thus yielding a very large number of candidate item sets.

2.4.3 Partition

An ARM approach that minimizes the I/O time by reducing the number of database scans to two has been proposed in Savasere et al. (1995). The algorithm divides the database into small partitions such that each partition can fit in the main memory and discovers frequent item sets locally using a stepwise approach, e.g. Apriori, in the first pass. A tid-list structure for each item set in a partition is then constructed. The tid-list of an item set identifies rows in a partition that contain that item set. The cardinality of an item set tid-list divided by the total number of the transactions in a partition gives the support of that item set.

In the second pass, the algorithm performs union operations on local frequent item sets found in each partition to discover frequent item sets in the database as whole. One of the drawbacks of the partitioning algorithm is that it prefers a uniform data distribution in which, if the count of an item set is evenly distributed in each part, the vast majority of the item sets to be counted in the second pass are frequent. However, for an unevenly distributed database, the majority of item sets in the second pass may be infrequent, causing extra I/O overhead (Lin and Dunham, 2000). Furthermore, when the number of partitions increases, the number of local frequent item sets also increases, consuming processing time and increasing redundant computation, especially when these partitions overlap in several frequent item sets (Zaki et al., 1997).

A comparison of performance between Apriori and the partitioning algorithm using six market basket analysis data sets (Agrawal and Srikant, 1994) revealed that the execution times of both algorithms increase when the support is reduced. A comparison using different number of partitions against the six benchmark problems indicates that the execution time decreases when fewer partitions are used, because the candidate set normally becomes smaller.

2.4.4 Frequent pattern growth

Apriori-like techniques use a candidate generation step to find frequent item sets during each iteration, and so these techniques require significant processing time and memory. Han et al. (2000) presented a new ARM approach, called FP-growth, that generates a highly condensed frequent pattern tree (FP-tree) representation of the transactional database. Each database transaction is represented in the tree by at most one path, and the length of each path is equal to the number of frequent items in the transaction representing that path. The FP-tree is a useful data representation because (1) all of the frequent item sets in each transaction of the original database are given by the FP-tree, and since there is a lot of sharing between frequent items, the FP-tree is smaller in size than the original database; and (2) the FP-tree construction requires only two database scans, whereby, in the first scan, frequent item sets along with their support in each transaction are produced, and in the second scan, the FP-tree is constructed.

Once the FP-tree is built, a pattern growth method is used to mine association rules by using patterns of length 1 in the FP-tree. For each frequent pattern, all possible other frequent patterns co-occurring with it in the FP-tree (using the pattern links) are generated and stored in a conditional FP-tree. The mining process is performed by concatenating the pattern with those produced from the conditional FP-tree. The mining process used by the FP-growth algorithm is not Apriori-like in that there is no candidate rule generation. One primary weakness of the FP-growth method is that there is no guarantee that the FP-tree will always fit in main memory, especially in cases where the mined database is dimensionally large.

A performance comparison between FP-growth and Apriori on two 10,000 record data sets (Han et al., 2000) indicates that FP-growth is at least an order of magnitude faster than Apriori, since the candidate sets that Apriori must maintain become extremely large. Also, the searching process through the database transactions to update candidate item set support counts at any level becomes very expensive for Apriori, especially when the support threshold is set to a small value. As the number of transactions grows, the difference in processing time between the two techniques increases further. Yildiz et.al.

(2010) compared matrix algorithm with FP growth algorithm using two case studies, the first one with 10000 items and 30000 transactions and the second with 30000 items and 30000 transactions. They concluded that the performances of the two algorithms are related to the characteristics of the given datasets and the minimum support threshold. Also, they concluded that the matrix algorithm performs better than the FP-Growth and their difference in the performance is more noticeable as the minimum support threshold decreases. For minimum threshold less than 10% the matrix algorithm is more efficient by up to 230%.

2.4.5 Confidence-based approach

Another possible solution to the problem of discarding rules with high confidence and low support, which abandons the support threshold and mines only top confidence rules, has been proposed (Li et al., 1999). Given a database, the end-user has to set an item set target, which represents the consequent of the desired outcome (rules). The problem of mining high confidence rules is to find all association rules where the target is the consequent. In doing that, the algorithm divides the problem of mining confidence rules into two steps. Step 1 involves splitting the original database into two sets, one set that holds transactions containing the target item set, T1, and the other holds the rest of the transactions, T2. The algorithm discards all items of the target from transactions in T1 and T2, therefore, the set of items in the original database I, becomes $I' = I$ – target. In the second step, all item sets, X, which appear in T1 but do not appear in T2 are discovered, and rules such as $X \rightarrow tg$, is produced, where tg is the target consequent. These item sets have a zero support in T2 but non-zero support in T1 and are called Jumping Emerging Patterns (JEP). The authors of (Li et al., 1999) have adopted two border methods from (Dong, 1999) to discover item sets whose support is zero in one sub-set, but non-zero in the other sub-set. The first border algorithm finds all item sets with non-zero support in a data set and names them horizontal borders. When taking two horizontal borders produced from two sets of data, as an input, the second border algorithm can derive all item sets whose support in one is zero, but non-zero in the other one.

Experimental studies using three data sets showed that this confidence-based approach can produce high confidence rules that cannot be found by traditional association rule approaches. However, the candidate item sets generated are much larger than in the original database. Therefore, a disk-based implementation is often preferred when pruning the search space using only the confidence threshold (Wang et al., 2001).

2.4.6 Matrix algorithm

In 2005, Yuan and Huang presented a novel algorithm for generation of association rules. The algorithm is called a matrix algorithm, and it creates a binary matrix with entries 0, 1 passing over the database only once creating a set of candidate items from which association rules are produced. The process of generating the matrix is the following: first the items in I are set as columns and the transactions D as rows in the matrix.

Let I={i_1,i_2,...,i_n} be the set of items and D = {t_1,t_2,...,t_m} be the set of transactions. Then, the matrix G={g_{ij}} for i=1,...n and j=1,...m is generated using the following rule:

$$g_{ij} = \begin{cases} 1, if \ i_j \in t_i \\ 0, if \ i_j \notin t_i \end{cases}$$

Using this generated matrix association rules are produced using the matrix algorithm: The 1-item set C_1 consists of the sets which are subsets of single item in I , that is, C_1 = {{i_1},{i_2},…,{i_n}}. In order to compute the support number for each set in C_1, we express every set in C_1 as a row vector in R^n, that is, we express {i_1} as $S_1^1 = \{1, 0, ..., 0\}$ and {i_k} as:

$$S_1^k = \{0, 0, ..., 1, ...0\}$$

where the kth element is 1 and others are 0. Then the support number of the set {ik} is calculated by:

$$supp\left(\{i_k\}\right) = \sum_{j=1}^{m} \left\langle g_j, S_k^1 \right\rangle$$

Where <,> is the dot product of two row vectors and g_j j=1,…,m are the rows of matrix G.

Then the set of all the frequent 1-item sets, L_1, is generated from C_1. If the support number of {i_k} is beyond the user-specified support threshold Minsupport, that is,

$$supp\left(\{i_k\}\right) \geq Minsupp$$

Then {i_k} $\in$ L_1 .

The set of candidate 2-item sets C_2 is the joint set of L_1 with itself. Each subset in C_2 consists of two items and has the form {i_k, i_j}, k < j. Similarly, we specify each set in C_2 a row vector in R_n. For example, for the set {i_k, i_j}, the specified vector is:

$$S_{k,j}^2 = \{0, ..., 0, 0, 1, ..., 0, 0, 1, ..., 0\}$$

Where the kth and jth elements are 1 and others are 0. The support number of the set {ik, ij} is :

$$supp\left(\{i_k\}\right) = \sum_{j=1}^{m} \text{int}\left[\frac{\left\langle g_s, S_{i,k}^2 \right\rangle}{2}\right]$$

where int[·] is the integrating function that changes a real number to integer by discarding the number after decimal point.

The frequent item set L_2 is generated from C_2 with the set whose support number is beyond the user specified support threshold Minsupport, that is:

$$supp\left(\left\{i_k, i_j\right\}\right) \geq Minsupp$$

then $\{i_k, i_j\} \in$ L2.

After the frequent 2-item sets L2 is obtained, it can be used to generate C_3.

The process is repeated with successively increasing number k until either C_k or L_k is empty, where each subset in Ck has the form $\{i_{l1}, i_{l2}, \cdot \cdot \cdot, i_{lk}\}$ including k items, and is generated from the frequent (k − 1)−item sets Lk−1, and L_k is the frequent k−item sets generated from C_k with the set whose support number is beyond the user specified threshold.

At the end of procedure, we can get the all frequent item sets by the following formula. Let the procedure is terminated after step k, then:

$$L = \bigcup_{i=1}^{k-1} L_i$$

2.4.7 Multiple supports Apriori

The support constraint is the most important factor that controls the number of association rules produced (Agrawal et al., 1993; Bayardo and Agrawal, 1999; Zaki, 2000). Almost all current ARM algorithms use a single support, but setting the support to a high value results in disposal of some useful, rare items in the database. Furthermore, to capture such rare items, one has to set the support to a very small value, which can lead to the generation of many useless rules (Liu et al., 1999, Li et al., 1999).

To overcome such a problem, Liu et al. (1999) proposed a multiple-support Apriori-like approach, called MSapriori, which assigns different support values for each item in the database. This enables users to express different support requirements for different rules. The support for a particular rule in MSapriori is the lowest *minsupp* value among the items in that rule. The candidate generation step in MSapriori is similar to the generate function in the Apriori algorithm.

An evaluation study comparing the MSapriori against real data from Agrawal and Srikant (1994) reveals that MSapriori generates a smaller number of candidate item sets than that of Apriori for real-world data sets. In particular, when the support threshold is set to 0.2%, the number of frequent item sets found by MSapriori is 61% lower than that of

Apriori. However, the execution time spent to find frequent item sets for both algorithms is roughly the same.

2.4.8 Hash-based technique and pruning

Generally, the computational cost of ARM is largely determined by the speed of the discovery of frequent one- and two-item sets. Empirical results from Agrawal and Srikant (1994) suggest that the computational cost in the initial iterations dominates most of the execution time for the candidate generation phase. When the number of frequent item sets during iteration 1 is large, the expected number of candidate item sets at iteration 2 is also large, and so reducing the size of the candidate item sets in the early iterations may result in huge savings in processing time and memory. A hash-based technique, called Direct Hashing and Pruning (DHP), has been proposed in Park et al. (1995) to efficiently reduce the size of candidate item sets in early iterations.

DHP works as follows. While scanning the database to find frequent one-item sets, a hash tree, H_1, is built for candidate one-item sets to facilitate the search. The algorithm evaluates during the scan whether an item exists in the hash table, and if so, the count of the item is incremented by 1; otherwise, the item is inserted into the hash table and is given a count of 1. Also, when the occurrences of all one-item sets are counted for each transaction, all two-item sets are produced and hashed into another hash table, H_2, where a count is initialized to 1 for each item set. Once the database is scanned, we can obtain the possible candidate two-item sets from H_2.

Pruning occurs to reduce the database size during the scan in which not only a transaction is trimmed but also some of the transactions are removed. DHP trims an item in a transaction *t* if it does not have a certain number of occurrences in *t's* candidate item sets. For example, If the support is set to 2, *t* = XYZWP and four two-subsets, (XZ, XW, XP, WP), exist in the hash tree constructed for candidate two-item sets, H_2, the number of frequencies according to each item in *t* is 3, 0, 1, 2, 2, respectively. For frequent three-item sets, only three items in *t*, e.g. (X, W, P), have occurrences above the support threshold. Consequently, these three items are kept in *t* and items Y and Z are removed.

Empirical studies indicate that DHP reduces the execution times not only in the second iteration, when the hash table is employed by DHP to facilitate the production of candidate two-item sets, but also in later iterations (Park et al., 1995). In particular, the execution time required to produce candidate two-item sets by DHP is several orders of magnitude smaller than that of Apriori, but the execution time of DHP is slightly larger than that of Apriori in the first iteration, owing to time required for building the hash table for candidate two-item sets.

2.4.9 Eclat algorithm

To minimize the number of passes over the input database, the Eclat algorithm was presented in Zaki et al. (1997). It requires only one database scan, thus addressing the question of whether all frequent item sets can be derived in a single pass. Eclat uses a vertical database transaction layout, where frequent item sets are obtained by applying simple tid-list intersections, without the need for complex data structures.

A recent variation of the Eclat algorithm, called dEclat, has been proposed in (Zaki and Gouda, 2003). The dEclat algorithm uses a new vertical layout representation approach called a diffset, which only stores the differences in the transactions identifiers (tids) of a candidate item set from its generating frequent item sets. This considerably reduces the size of the memory required to store the tids. The diffset approach avoids storing the complete tids of each item set; rather the difference between the class and its member item sets are stored. Two item sets share the same class if they share a common prefix. A class represents items that the prefix can be extended with to obtain new class. For instance, for a class of item sets with prefix x, [x] = {a_1, a_2, a_3, a_4}, one can perform the intersection of xa_i with all xa_j with j>i to get the new classes. From [x], we can obtain classes [xa_1] = {a_2, a_3, a_4}, [xa_2] = {a_3, a_4}, [xa_3] = {a_4}.
Experimental results on real world data and synthetic data (Zaki and Gouda, 2003) revealed that dEclat and other vertical techniques like Eclat usually outperform horizontal algorithms like Apriori and FP-growth with regards to processing time and memory usage. Furthermore, dEclat outperforms Eclat on dense data, whereas the size of the data stored by dEclat for sparse databases grows faster than that of Eclat. Consequently, the authors concluded that for dense databases, it is better to start with a diffset representation, but for sparse databases, it is better to start with a tid-list representation and then switch to a diffset at later iterations.

2.4.10 Sampling technique

Another technique to solve Apriori’s slow counting and Eclat’s large memory requirements is to use sampling as proposed by Toivonen (1996). The presented sampling algorithm picks a random sample from the database, finds all relatively frequent patterns in that sample, and then verifies the results with the rest of the database. In cases where the sampling method does not produce all frequent sets, the missing sets can be found by generating all remaining potentially frequent sets and verifying their supports during a second pass through the database. The probability of such a failure can be kept small by using a lower support threshold than the minimum support value.

2.4.11 Measuring interestingness of rules

ARM has the potential to produce a large number of patterns as the size and dimensionality of databases increase. Most ARM algorithms employ a support–confidence threshold framework.

Let I = $\{i_{1,}i_2,\ldots,i_n\}$ be a set of items and D = $\{t_1,t_2,\ldots,t_m\}$ be a set of transactions. A rule is defined as an implication X=> Y where X,Y ⊆ I and X∩Y=0.

Item set X is called antecedent of the rule and Y is called the consequent of the rule. The support of the rule support(X=>Y) is the number of occurrences of X and Y, P(X∪Y).

The confidence of the rule confidence(X=>Y) is defined as conditional probability P(Y|X) which is the percentage of transactions on D containing X that also containing Y:

$$\mathbf{P(Y|X)} = \frac{\mathbf{support(X \to Y)}}{\mathbf{support(X)}}$$

The candidate rules must have both confidence and support at least equal with the threshold values that are given by the users.

The issue that arises is that using the framework of support and confidence only cannot guarantee that a strong rule is necessarily interesting. The pitfall of confidence can be traced to the fact that its definition ignores the support of the item set Y, support(Y). Han et al. (2011) and Pong-Ning et al. (2000) described other measures that are used to measure the interestingness of a rule.

One measure of interestingness is the lift. Lift is defined as the ratio of the confidence of the rule X=> Y over the support of Y. Lift is used to filter misleading strong association:

$$\mathbf{Lift(X \to Y)} = \frac{\mathbf{Conf(X \cup Y)}}{\mathbf{support(Y)}}$$

- If lift equals 1 X,Y are independent
- If lift is greater than 1 X,Y are positively correlated

- If lift is less that 1 X,Y are negatively correlated

The following four measures:

- all confidence,
- max_confidence,
- Kulczynski and
- Cosine, have the following property: each value is only influenced by the supports of X, Y and XUY or more exactly, by the conditional probabilities of P(X|Y) and P(Y|X), but not by the total number of transactions.

Another common property of these four measures is that the value they take ranges from 0 to 1, and the higher the value, the higher is the relationship between X, Y.

The definitions of these measures are:

All confidence:

$$all_{confidence(X,Y)} = \frac{support(X \cup Y)}{max\{support(X), support(Y)\}}$$

Max confidence:

$$\max{}_{\downarrow} confidence(X,Y) = max\{P(X \mid Y), P(Y|X)\}$$

Kulczynski:

$$\mathbf{Kulczynski(X,Y)} = \frac{1}{2}\left(P(X|Y) + P(X|Y)\right)$$

Cosine:

$$Cosine(X,Y) = \sqrt{P(X|Y) \times P(Y|X)}$$

2.5 Associative classification

Associative classification has been proposed from Liu et al., 1998. In associative classification the right side of a rule $X \rightarrow Y$ is considered to be class attribute. Associative classification builds rules based on conjunctions of attribute–value pairs that occur frequently in data. The steps of Associative Classification are:

- Mine the data for frequent item sets, that is, find commonly occurring attribute–value pairs in the data.
- Analyze the frequent item sets to generate association rules per class, which satisfy confidence and support criteria.
- Organize the rules to form a rule-based classifier.

In order to rank the strength of each rule parameters such as support of the rule, confidence of the rule, length of the antecedent part of the rule and the generation time of the rule.

In (Liu et al., 1998) (Yin et al., 2003) (Thabtah et al., 2005), (Thabtah et al., 2010) it is stated that removing redundant and misleading rules often lead to wrong classification might enhance model efficiency as well as effectiveness.
Also one of the main drawbacks of AC mining is that it often generates large number of rules since AC extract all the correlations among the items and the class are discovered as rules. Finally, the use of large number of rules necessitates high computation cost and often degrades the accuracy rates.

2.6 Temporal Pattern Mining

The objective of Temporal Pattern Mining(TPM) is to discover temporal patterns in time stamped data. However, the identification of known patterns is also seen as important as this would provide a means of validating the adopted approach. The process of frequent pattern mining in static data tables is well established within the Knowledge Discovery in Data (KDD) community and can be traced back to early work on Association Rule Mining (ARM) as first espoused by Agrawal and Srikant (1994).Less attention has been applied to temporal pattern mining. There has been reported work on Temporal ARM (TARM) where association rules are mined from time stamped data.
The TPM process described in this paper operates on binary value data sets (thus, where necessary, data must be transformed into this format using a process of normalisation and discretisation). The research described in this work also borrows from the field of Jumping and Emerging Patten (JEP) mining as first introduced by Dong and Li (1999). The distinction between the work on JEPs, and that described in this paper, is that JEPs are patterns whose frequency increases (typically) between two data sets (although some work has been done on identifying JEPs across multiple data sets, for example Khan et al. (2010). JEP mining is usually also conducted in the context of classification (see for example Fan and Kotagiri, 2003). The distinction between JEPs and the work described here is that the work is directed at patterns that change in a variety of pre-described ways over a sequence of data sets. To the best knowledge of the authors there is little reported work on temporal pattern mining or trend mining as defined above.
Zhu et al. [18], in the context of data stream mining, identify three processing models for temporal pattern mining:

- Landmark
- Damped and
- Sliding Windows.

The Landmark model discovers all frequent patterns over the entire history of the data from a particular point in time called the "landmark". The Damped model, also known as the Time-Fading model, finds frequent patterns in which each time stamp is assigned a weight that decreases with "age" so that older records contribute less than more recent records. In the Sliding Window model the data is mined by sliding a "window" through the temporal dimension. A similar categorisation may be adopted with respect to temporal pattern mining. The work described in this work adopts the Landmark model.

This thesis addresses a number of aspects of the field of temporal data mining (TDM) (Lin et al., 2002; Roddick and Spiliopoulou, 2002), with the central goal being the identification of interesting temporal rules between patterns in time-stamped data. Several authors have previously identified a number of such temporal rules. Agrawal and Srikant (1995) originally used an a priori-like technique to extract sequential patterns; this was then extended by Mannila and Toivonen (1996) to address the existence of frequent episodes and episode rules. Subsequently, a number of authors published extensions to the extraction of temporal association rules and inter-transactional association rules (Chen et al., 1998; Li et al., 2003; Tung et al., 2003).

Several authors have mined temporal rules in interval-based data. Having provided a definition of temporal patterns based on temporal relationships between interval-based events, Kam and Fu (2000) proposed an a priori-like strategy for the efficient detection of such patterns. A general methodology for the process of knowledge discovery in time series databases, addressing both the pre-processing and the rule mining step, was presented by Last et al. (2001). Cohen (2001) introduced the theory of fluent learning in order to extract common patterns in time series data and described the 'shape' of episodes using a statistical technique; this was well suited for multivariate time series data with binary variables.

A mining technique to discover containment relationships in series of interval events was proposed by Villafane et al. (2000) (Sacchi et al., 2004); such events are derived from numerical time series through a quantisation step.

The newly discovered rules between temporal patterns were applied in unsupervised neural networks to detect complex temporal patterns and to generate temporal grammatical rules for a symbolic knowledge representation (Guimaraes and Ultsch, 1999; Guimarães et al., 2001), thus highlighting the benefits of using prior knowledge to improve algorithm performance.

Höppner and Klawonn (2002b) and Höppner (2003) developed informative temporal rules on a given sequence of labelled intervals, thus improving the flexibility of the temporal pattern previously defined by Kam and Fu (2000). Using the work of Höppner and an algorithm for the discovery of temporal patterns from interval-based data proposed by Lin and Lee (2005), Winarko and Roddick recently proposed a new method to extract frequent temporal patterns and then to infer temporal rules from such patterns (Winarko and Roddick, 2005). Papapetrou et al. (2005) developed a novel formalisation of the problem of mining frequent arrangements of temporal intervals, wherein the method acts on a database of sequences of events each of which occur during a defined time interval. Of the aforementioned publications, the method that is closest to the one that will be described in this thesis is that of Höppner (2003), who suggests a formulation of the problem of extracting rules from temporal. In particular, Höppner proposes qualitative features by which the time series can be divided into segments, as well as a method for mining temporal patterns from which informative rules are derived. Höppner (2003) provides an introduction to how to learn qualitative labels (usually trends) from time stamped data, mentioning techniques such as clustering and smoothing, and wavelets.

Following the ideas of Bellazzi et al. (2005), Sacchi et al. (2007) presented raw time series introducing a step for the extraction of an interval-based representation based on the formalism of TAs . In previous proposals, even when a qualitative representation of the time series is suggested (Höppner and Klawonn, 2002b) or achieved through TAs (Bellazzi et al., 2005), the representation that is considered is always of a basic nature (e.g., intervals of increasing, decreasing or stationary trends for a single time series) and the temporal rules are always extracted between such simple patterns.

Bellazzi et al. (2005) aimed at interpreting and performing data analysis in real time, with the difference that they wished to evaluate knowledge discovery of data in batch mode. Thus it can be seen that the field of temporal data mining is rich and varied.

2.7 Temporal Logic

Temporal logic aims on extracting knowledge on sequential and complicatedly changeable. This method is an extension of the classical propositional logic where an event is true (1) or false (0). Temporal logic tries to answer what to do with the fact that true or false value of a statement changes from time to time. Temporal logic associates each point of a given flow time with a separate evaluation about an event.

The basic idea of temporal logic is to make the evaluations (true or false) time dependent. The second fundamental idea of the temporal logic is the use of distinct time point, e.g. future, past which is interpreted for a case φ as: at some time in the future the case is φ or at some time in the past φ holds.

In the work presented here the time is not used in the form of distinct points but time is used in form of time stamps which are synthesised from distinct points. The total number of time stamps forms a time interval where the knowledge is extracted not as event (true or false) but as an association between attributes which may or may not be strong and interesting within this time interval.

2.8 Longitudinal data mining

Longitudinal data are information comprising values for a set of data attributes that are repeatedly collected for the same object over a sequence of sample points, and as such it can be said to track the progress of the object in some context (Singer and Willett, 2003).

The exemplar longitudinal data set is patient data, where information concerning a patient's condition is repeatedly collected so as to track a patient's progress.

Longitudinal data may be categorized in a number of ways: one suggested categorization is that of (Singer and Willet, 2003) who identified person-level and person-period data sets. In a person-level data set, each person (subject) has one record and multiple variables containing the data from each sampling. In a person-period data set, each person (subject) has multiple records, one for each measurement occasion. Thus, a person-level data set has as many records as there are subjects in the sample, while a person-period data set has many more records (one for each subject sampling event). The former is sometimes referred to as a broad data structure, and the latter as a long data structure (Twisk, 2003). Longitudinal studies vary with regard to sample size, number of variables, and number of time stamps. Broadly speaking, there are five main types of longitudinal study based on these characteristics (Kamp and Bijleveld, 1988):

- simultaneous cross-sectional studies,
- trend studies,
- times series studies,
- intervention studies and
- Panel studies.

2.9 Data mining for medical applications

Modern medicine generates a great deal of information stored in medical databases, and it has become increasingly necessary to extract useful knowledge and provide scientific decision-making for the diagnosis and treatment of disease from the database. Because the medical information is characteristic of redundancy, multi-attribution, incompletion and closely related with time, the medical DM differs from others.

Many factors affect the success of DM on medical datasets, such as the quality of the data. If the information is irrelevant or redundant, or the data are noisy and unreliable, knowledge discovery during training is more difficult. Zhao and Wang (2010) refer to the four characteristics of medical data:

- Redundancy: The medical database is a huge data resource, and a large number of records are stored in the database every day. It may contain repeated, irrelevant, and even contradictory records. For example, for one disease, patients' symptoms, test results, and treatment measures may be the same. In addition, the medical data are also a feature of time.
- Complexity: Complexity is a remarkable feature of medical data. As the medical data obtained from medical imaging, laboratory data and the exchange between doctors and patients, they are in various forms. These include images (SPECT), signals (ECG), pure data (the signs of parameters, test results), and text (such as the identity of the patient records, descriptions of the symptoms, detection and diagnosis of the textual representation).
- Privacy: Privacy is different from security and confidentiality, in that when individuals or organizations access private information without authorization, this creates a safety issue. While researchers share private information with unauthorized individuals or institutions, this exposes the issue of confidentiality. Medical DM scientists are obliged to carry out research on the premise to protect patients' privacy.
- Missing values: Medical data collection is always out of line with the stage of processing. The main purpose of medical data collection is to cure sickness and save patients' lives. However, the purpose of medical data processing is to determine regular patterns in certain diseases. In this case, the collected data may not meet the need to cover all the information. In addition, human factors may lead to errors and incomplete information in patients' records and the expression of many medical data is uncertain and fuzzy.

In their work, Robu and Hora (2012) described the main DM techniques used in medical applications:

- Classification – in order to predict a nominal value
- Regression - estimation of an output value based on input values
- Time series analysis - is the value of an attribute examined over a time period usually at evenly spaced time intervals.
- Clustering – is a descriptive technique which consists of identifying classes or groups in sets of unclassified data. Clustering is often one of the first steps in data mining analysis. It identifies groups of related records that can be used as a starting point for exploring further relationships.

- Association rules – discovering that a set of symptoms often occur together with another set of symptoms

Delen et al. (2005) used two popular DM algorithms (artificial neural networks and decision trees) along with the most commonly used statistical method (logistic regression) to develop prediction models using a large dataset to predict breast-cancer survivability (more than 200 000 cases). They also used 10-fold cross-validation methods to measure the unbiased estimate of the three prediction models for performance-comparison purposes.

Srinivas et al. (2010), in their study in health care and prediction of heart attacks, examined the potential use of classification based DM techniques such as rule-based, decision tree, Naïve Bayes, and artificial neural networks for massive volumes of data. The health-care industry collects huge amounts of health-care data, which, unfortunately, are not “mined” to discover hidden information. They are used for data preprocessing and effective decision-making for the One Dependency Augmented Naïve Bayes classifier (ODANB) and Naïve Credal Classifier 2 (NCC2). This is an extension of Naïve Bayes to imprecise probabilities aimed at delivering robust classifications when dealing with small or incomplete data sets.

Lavrac (1999), in his paper, reviewed several DM methods for intelligent data analysis in medicine, in particular machine-learning methods. Machine-learning methods can be classified into three major groups: inductive learning of symbolic rules (such as induction of rules, decision trees, and logic programs), statistical or pattern-recognition methods (such as k-nearest neighbours or instance-based learning, discriminate analysis, and Bayesian classifiers), and artificial neural networks (such as networks with back-propagation learning, Kohonen’s self-organizing network, and Hopfield’s associative memory).

Breault et al. (2002), in their study on diabetic data warehouse, used a classification tree approach as standardized in the CART software by Salford Systems.

Li et al. (2004), in their study on DM techniques for cancer detection using serum proteomic profiling, used a support vector machine-based method as applied in this study, in which statistical testing and genetic algorithm-based methods are used for feature selection respectively. Leave-one-out cross-validation with a receiver operating characteristic (ROC) curve is used to evaluate and compare cancer-detection performance.

Silva et al. (2008), in their study on rating organ failure via adverse events, compared two DM methods: multinomial logistic regression (MLR) and artificial neural networks (ANNs). These methods were tested in the R statistical environment, using 20 runs of a fivefold cross-validation scheme. The area under the ROC curve and Brier score were used as the discrimination and calibration measures.

Srimani and Koti (2011), in their study on difference medical databases, examined the performance of different classification methods:

- Decision Trees: Decision trees are a way of representing a series of rules that lead to a class or value. Therefore, they are used for directed DM, particularly classification. One of the important advantages of decision trees is that the model is quite explainable, since it takes the form of explicit rules.
- Bayesian Network: A Bayesian Network (BN) is a graphical model for probability relationships among a set of variable features. The most interesting feature of BNs, compared with decision trees or neural networks, is the possibility of taking into account the prior information about a given problem, in terms of structural relationships among its features.
- Naïve Bayes: The Naïve Bayes classifier [8] uses the Bayes rule to compute the conditional probability of each possible class by assuming the input features to be conditionally independent, given the target feature.
- Ripper: Ripper is a rule-based learner that builds a set of rules to identify the classes while minimizing the amount of error. The error is defined by the number of training examples misclassified by the rules.
- Nearest Neighbour: Nearest Neighbour, instead of determining the tables global majority, based on the same set of features, determines the class for each instance that is not covered by a decision table entry.
- Bagging: Bagging bags a classifier to reduce variance. This works for both classification and regression, depending on the base learner. In the case of classification, predictions are generated only by averaging the probability estimates, not by voting.
- Decision stump: Decision stump builds one-level binary decision trees for datasets with a categorical or numeric class, by dealing with missing values and by treating them as a separate value and extending a third branch from the stump.
- Dagging: Dagging creates a number of disjointed, stratified folds out of the data and feeds each chunk of data to a copy of the supplied base

classifier. Predictions are made by majority vote, since all the generated base classifiers are put into the vote meta classifier.

The rapid growth of digitalized medical records presents new opportunities for mining large (terra-bytes) amounts of data, when the structure of the text record is very loose without any rules. Term mining technique can led to the identification of keywords with a significant value within the narratives of medical records. Term mining can help to convert unstructured text into structured content. Term mining can be applied in textual data (literature, admission notes, reports and summaries) and yields precise knowledge nuggets from a sea of information.
Claster et.al., (2008) presented an unsupervised neural network text-mining technique for the analysis of computed tomography scanning. Using the notes of physicians identified keywords and correlated them with either negative or positive outcome.
Ananiadou et.al.(2012) applied semantic text mining techniques in diabetes databases. Semantic text mining techniques can be customized to extract semantic types, relations and associations with multifactorial diseases such as diabetes. Currently, such extraction is being manually conducted by a large group of scientists, and therefore it is anticipated that text mining will contribute to the automation of this work.
Wu et.al. (2007) they used text mining on clinical records for cancer diagnosis. In this work, they proposed a framework for discovering the relationships between cancer diseases and potential patterns from clinical medical records. They applied a text mining process on a corpus of clinical records to extract the potential patterns. First, they utilized to the Cancer Ontology & Thesaurus for extracting and weighting the key terms from clinical records and tag various cancer-specific concepts. Second, they applied the SOM algorithm to perform a clustering process and extracting the relationships between cancer diseases and potential factors from clinical medical records. Third, they developed an approach applying a SVM method to supporting acquisition of relatedness among texts and clinical records. The results show that the integration of cancer ontology and this approach can extract the potential patterns and re-categorize clinical records. Furthermore, the system also allow combine microarray data-mining methods into the framework to find the relationships between cancer diseases and specific genes.

2.10 Trend mining

Trend mining is the process of identifying and analysing trends in the context of the variation of the support of the association rules that have been extracted from longitudinal datasets.

The identification and the analysis of trends are performed using mathematical conditions (prototypes). The aim of these identities is to separate into groups the large amount of knowledge that is hidden in the datasets. A longitudinal data is a set of finite variables that are revealingly measured. The repeated measurements create a class of multidimensional time series. The application of trend mining for the discovery of trends requires the following steps/procedure:

- Understanding the domain: In real-life applications, data are complex, and often the user must deal with problems such as missing records, conflicting values, and double records. The success of trend mining is highly related to the clarity of the input data.
- Association rule mining: After data preparation the next step is to discover useful knowledge through the application of ARM. This process is repeated as many times as the number of datasets with different time stamps.
- Trend mining algorithm: The mathematical conditions are applied in order to determine the evolution of the support of the interesting rules and create different categories of trends, depicted using a colourful representation.

Dong and Li (1999) introduced the idea of emerging patterns in order to describe the change of the support of a frequent item set from a Dataset D_1 to a dataset D_2. Let I be an item set $\{i_1, i_2, ..., i_N\}$and X be a subset of I. A transaction T contains X if $X \subseteq T$. The support of item set X in a dataset D is denoted as $supp_D(X)$ which is the number of transactions in D that contain X. Assume two datasets D_1 and D_2 and $supp_1(X)$, $supp_2(X)$ the support of X in D_1 and D_2 respectively, then the Growth Rate(X) is defined as:

$$\begin{cases} 0 \; if \; supp_{\mathbf{1}}(X) = 0 \; and \; supp_{\mathbf{2}}(X) = \mathbf{0} \\ \infty \; if \; supp_{\mathbf{1}}(X) = 0 \; and \; supp_{\mathbf{2}}(X) \neq \mathbf{0} \\ \qquad \dfrac{\mathbf{supp_2(X)}}{\mathbf{supp_1(X)}} \; otherwise \end{cases}$$

Given ρ>1 as a growth rate threshold an item set X is said to be an ρ-emerging pattern from database D_1 to D_2 if GrowthRate(X)≥ρ.

Li et.al, (2000 and 2001) presented the concept of jumping emerging patterns (JEP).
A jumping emerging pattern from D_1 to D_2 is an item set X that satisfies $supp_{D_1}(X) = 0 \; and \; supp_{D_1}(X) \geq \xi$ with î a minimum support threshold.

Dong et al., (1999) proposed new classifier called CAEP (classification by aggregating emerging patterns) based on the definitions of emerging patterns proposed by Dong and Li (1999) using the following fundamental ideas:

i) Each emerging patterns can sharply differentiate the class membership of a fraction of instances containing the emerging patterns due to the big difference between its supports in the opposing classes. They defined the differentiating power of the emerging patterns in terms of the support and their ratio on instances containing the emerging patterns.
ii) for each instance t, by aggregating the differencing power of a fixed automatically selected set of EPs a score is obtained for each class. The scores of all classes are normalized and the largest score determines t’s classes.

Terlecki and Walczak (2007) proposed the concept of JEP with negation (JEPNs) based on the concept of JEPs. They defined negation as a transaction that does not contain an item but it contains the respective negated item.

Soulet et al. (2004) proposed a new kind of emerging pattern that they termed “strong emerging patterns” (SEPs) as the emerging patterns with the best possible growth rates. In order to calculate the growth rate, they divided the database D into as many datasets as the number of different values of an item C, where C_1, C_2 represent two different classes.

Fan and Ramamohanarao (2006) proposed the generalised noise-tolerant emerging patterns (GNEPs). They defined the generalised growth rate of an item set from dataset D_1 to D_2 as: GrowthRate(X)=

$$\begin{cases} 0\ if\ supp_{D_1} = 0\ and\ supp_{D_2} > \mathbf{0} \\ \infty\ if\ supp_{D_1} = 0\ and\ supp_{D_2} = \mathbf{0} \\ \dfrac{f_2\left(supp_{D_2}(X)\right)}{f_1\left(supp_{D_1}(X)\right)}\ otherwise \end{cases}$$

Where $f_1(x)$ and $f_2(x)$ are two monotone function and $\forall$ $x \geq 0$, $f_1(x) \geq 0$, $f_2(x) \geq 0$.

Given two thresholds $\delta_1 > 0$ and $\delta_2 > 0$ with $\delta_2 \gg \delta_1$.

An item set X is GNEP from D_2 and D_1 if:

$$\frac{f_2\left(supp_{D_2}(X)\right)}{f_1\left(supp_{D_1}(X)\right)} \geq \frac{f_2(\delta_2)}{f_1(\delta_1)}$$

$$f_2\left(supp_{D_2}(X)\right) \geq f_2(\delta_2)$$

$$any\ proper\ subset\ of\ X\ does\ not\ satisfy\ conditions\ 1\ and\ \mathbf{3}$$

Zhu et al. (2002), in the context of data stream mining, identify three processing models temporal pattern mining:

- Landmark,

- Damped and
- Sliding Windows.

The Landmark model discovers all frequent patterns over the entire history of the data from a particular point in time called the “landmark”. The Damped model, also known as the Time-Fading model, finds frequent patterns in which each time stamp is assigned a weight that decreases with “age” so that older records contribute less than more recent records. In the Sliding Window model, the data are mined by sliding a "window" through the temporal dimension. A similar categorization may be adopted with respect to temporal pattern mining.

Kohavi et al. (2002) defined trend-mining techniques to extract trends from time-stamped data collections, and Nohuddin et al. (2012) used SOM to identify trends using cattle-movement data. Other related work by Streibel (2008) used quantitative numeric financial data, and qualitative text corpi data extracted from business news articles, to forecast financial market trends. Google provides Google Trends, a public web facility that supports the identification of trends associated with keyword search volume. Raza and Liyanage (2008) introduced a trend-analysis approach to mine and monitor data for abnormalities and faults in industrial production processes.

The major difference of the proposed trend mining algorithm is that it examines any number of datasets, and the identities that determine the trends must be valid across all datasets.

2.11 Validation and verification

Verifying and validating a system are very important processes in the development of a knowledge-based system. Verification tests examine whether the system is built correctly. Thus, verification examines the internal procedures and that is why here it is called “internal”. Validation tests are aimed at building the right system, and so validation needs to ensure that the system produces the right output.

Verification was defined by Branstad and Cherniavsky (1982) as "the demonstration of the consistency, completeness and correctness of the software".

O’Leary (1993) presented a review of case-based systems and concluded that the validation of each system has used the comparison of the system with human experts or machine learning. Murrel and Plant (1997) examined 33 tools of validation and verification with 145 testing techniques, and these techniques were categorized into three categories: requirements/design methods, static testing, and dynamic testing.

O'Keele and Preece (1996) noted that the verification of a system can be achieved using three measures: conflict, redundancy, and deficiency. Conflict refers to the ability of the system to arrive at logically inconsistent conclusions from consistent input; redundancy refers to the presence within the system of logically unnecessary structures that never affect the relationship between the input and output of the system; and deficiency refers to the absence of structures that should be present, logically, for the system to arrive at conclusions for all valid input cases. They worked on verifying their system attempting to detect anomalies. Thus, an anomaly could indicate any of the above measures. Anomaly detection is focused on the usage of rules. In terms of validation, O'Keele and Preece presented a list of methods used for calibration of knowledge-based systems (rules, heuristic, case testing) and also suggested that a strategy should be implemented for validation and verification of a knowledge-based system. They suggested the following guidelines for the development of a strategy:

- choice and specification of the criteria of validation and verification;
- development of a list of methods for validation and verification;
- mapping of validation and verification methods into the life cycle of the system.

Liu et al. (2010) presented a comparison of 11 validation measures for five clustering aspects: monotonicity, noise, density, subclusters, and skewed distribution. They defined internal validation as the process that relies on information in data, and external validation as the process based on external information not contained in data. External validation measures know the "true" cluster number in advance, so they are mainly used to choose an optimal clustering algorithm on a specific data set. On the other hand, internal validation measures can be used to choose the best clustering algorithm as well as the optimal cluster number without any additional information.

Theodoridis and Koutroubas (1999) identified three approaches to validate clustering results. The first approach is based on external criteria. This implies that we evaluate the results of a clustering algorithm based on a pre-specified structure, which is imposed on a data set and reflects our intuition about the clustering structure of the data set. The second approach is based on internal criteria. We may evaluate the results of a clustering algorithm in terms of quantities that involve the vectors of the data set themselves (e.g. proximity matrix). The third approach of clustering validity is based on relative criteria. Here, the basic idea is to evaluate a clustering structure by comparing it with other clustering schemes, resulting in the same algorithm but with different parameter values.

There are existing examples of DM process models that incorporate the notion of validation, such as CRISP-DM, but these tend to be very general models and do not relate to a specific framework. The central idea of validation in the DM community is one of checking the results by "cross-validation", where the data are split into a training set and a testing set, and the results of the data mining applied to the testing set are compared with the training set. This kind of statistical validation, of course, is not possible when the goal of the system is to discover trends. In this case, a more application-specific method has to be adopted, so the following strategy has been developed here for the verification and validation of the trend-mining framework:

- At the beginning and end of the constituent processes, a language for a set of declarative validation rules will be established, and a systematic process of validation of each set of input data will be created.
- At the end of the DM pipeline, a process for testing for known associations (the expected outputs) will be created. This involves using invariants and characteristics of the data to prune and/or synthesize output rules to fit the associations being sought.

2.12 Conclusion

An overview of various aspects of DM has been presented in this chapter. ARM plays a vital role in the trend-mining framework, which is presented in this research. The matrix algorithm was selected, owing to its ability to scan a dataset only once, which is very important when a large number of time stamps are present. The trend-mining definition has been given in order to clarify the term trend, and the strategy of the validation has been given, as it is crucial to prove that the framework has been built appropriately and provides the right output.

Chapter 3 Medical Overview and Data Description

3.1 Introduction

In the UK, about 3,000,000 people are diabetic and one third of them have signs of diabetic retinopathy. This disease has many side effects, such as a higher risk of eye disease, a higher risk of kidney failure, and other complications. However, early detection of the disease and proper care management can make a difference. To combat this disease, a national scheme in England introduced a regular screening programme for diabetic patients. The application domain, with respect to this study, is diabetic data.

Patient information, clinical symptoms, eye-disease diagnosis, and treatments are routinely recorded in these databases, and medical longitudinal data are used to plot the progress of a medical condition and implicitly provide information about various trends. After 20 years of data collection from the diabetic-retinopathy (DR) screening process, a wealth of information has been gathered, and this naturally this has led to the application of knowledge-discovery and data-mining techniques to discover interesting patterns in the data.

DR is the most common cause of blindness in working-age people in the UK. It is a chronic multifactorial disease affecting patients with diabetes mellitus and causes damage to the retina. About 750,000 are registered blind or partially sighted in the UK, and the remainder are at risk of blindness.

The RLUH screening programme currently deals with some 17,000 people with diabetes registered with GP[1] within the Liverpool Primary Care Trust[2] per year. Consequently, a substantial amount of data is available for analysis, and further details on the data collection are presented in the next subsection.

The objective is to find rules that can be used by medical doctors to improve their daily tasks, that is, to understand more about diabetes or to discover something special about the treatment, patient management, and also to stop the progress of DR. Although knowledge discovery in databases has reportedly been very successful in domains such

[1] In the UK GP stands for "General Practitioner", essentially a family doctor

[2] Primary Care Trusts are organizational units established to manage local health services in the UK.

as fraud detection, targeted marketing, etc., we found in comparison that there have been relatively few applications of data-mining techniques to the health sector. This is important for two reasons. First, the data obtained by health clinics are typically very noisy. Many of the patient records contain typographical errors, missing values, or incorrect data on details such as street names date of birth, etc.; worse, many records are in fact duplicate records. Cleaning these data takes a tremendous amount of effort and time. In addition, many of the data collected are not in the forms that are suitable for data mining. They need to be transformed to more meaningful attributes before mining can proceed. Second, health doctors are usually too busy to see patients every day, and medics cannot afford the time or energy to sieve through the thousands of rules generated by state-of-the-art mining techniques in the diabetic patient database. Thus, it is important to present the discovered rules in an easy-to-understand way for interpretation.

These concerns are addressed in the validation chapter. To overcome the problem of noisy data, a semi-automatic data-cleaning system based on logic rules has been developed. The system reconciles database format differences by allowing doctors to specify the mapping between attributes in different format styles and in the encoding schemes used. To resolve the problem of too many rules being generated by the state-of-the-art mining techniques, a user-orientated approach is applied to provide a step-by-step exploration of the data to better understand the discovered patterns.

3.2 Diabetes overview

Diabetes mellitus is the most common metabolic disease worldwide. Quality and Outcomes Framework data suggest that there are 1,766,391 patients registered as diabetic in England, a prevalence of 3.55%. DR is a frequent complication of both types of diabetes and represents the most common cause of blind registration in the working-age population in the Western world (Harding S.P., Broadband B.D., 2009).

Diabetes is a chronic disease that occurs either when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin it produces. Insulin is a hormone that regulates blood sugar. Hyperglycaemia, or raised blood sugar, is a common effect of uncontrolled diabetes and over time leads to serious damage to many of the body's systems, especially the nerves and blood vessels. Worldwide, 347 million people have diabetes. In 2004, an estimated 3.4 million people died from consequences of fasting high blood sugar. A similar number of deaths has been estimated for 2010.

More than 80% of diabetes deaths occur in low- and middle-income countries (WHO, 2013).

Type 1 diabetes (previously known as insulin-dependent, juvenile, or childhood-onset diabetes) is characterized by deficient insulin production and requires daily administration of insulin. The cause of type 1 diabetes is not known, and it is not preventable with current knowledge. Symptoms include excessive excretion of urine, thirst, constant hunger, weight loss, vision changes, and fatigue. These symptoms may occur suddenly.

Type 2 diabetes (formerly non-insulin-dependent or adult-onset diabetes) results from the body's ineffective use of insulin. Type 2 diabetes comprises 90% of people with diabetes around the world (World Health Organization, 2011) and is primarily the result of excess body weight and physical inactivity. Symptoms may be similar to those of type 1 diabetes, but are often less marked, and so the disease may be diagnosed several years after the onset, once complications have already arisen. Until recently, this type of diabetes was seen only in adults, but it is now also occurring in children.
The most common effects of diabetes are as follows:
Over time, diabetes can damage the heart, blood vessels, eyes, kidneys, and nerves.
Diabetes increases the risk of heart disease and stroke: 50% of people with diabetes die of cardiovascular disease (primarily heart disease and stroke) (Morrish et al., 2001)
Combined with reduced blood flow, neuropathy (nerve damage) in the feet increases the chance of foot ulcers, infection, and eventual need for limb amputation.
Diabetes is among the leading causes of kidney failure (World Health Organization, 2011).
The overall risk of dying among people with diabetes is at least double the risk of their peers without diabetes (Roglic, et al.,2005)
The DR focused on in this study is an important cause of blindness, occurring as a result of long-term accumulated damage to the small blood vessels in the retina. One per cent of blindness globally can be attributed to diabetes (World Health Organization, 2012).

3.3 Diabetic Retinopathy overview

DR is the leading cause of blindness in people of working age in industrialized countries and accounts for 1.8 million of the 37 million cases of blindness throughout the world. The total number of people with diabetes is projected to rise from 285 million in 2010 to 439 million in 2030. DR is a chronic multi-factorial disease affecting patients with diabetes mellitus and causes damage to the retina (Stangos, 2009). Patients with diabetes are more likely to develop eye problems such as cataracts and glaucoma, but the disease's effect on the retina is the main threat to vision. It occurs when high blood sugar damages small blood vessels at the back of the eye, called the retina. All people with diabetes are at risk from this disease. There are things that can be done to reduce one's risk and prevent or slow vision loss. DR can affect both eyes, and there may not be any signs at first. As the condition worsens, blood vessels weaken and leak blood and fluid. As new blood vessels grow, they also leak, causing blocks in vision.

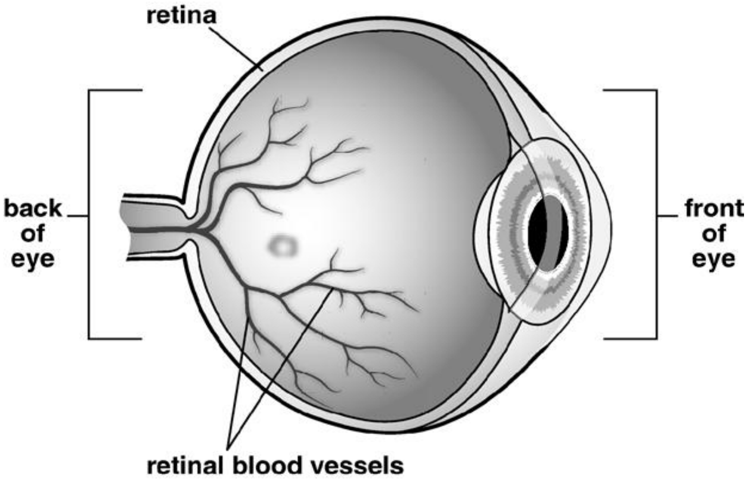


Figure 3.1: Circulatory system of the retina

Over time, diabetes affects the circulatory system of the retina. The earliest phase of the disease is known as background DR. In this phase, the arteries in the retina become weakened and leak, forming small, dot-like hemorrhages. These leaking vessels often lead to swelling, or oedema, in the retina and decreased vision (Figure 3.1). The next stage is known as proliferative DR. In this stage, circulation problems cause areas of the retina to become oxygen-deprived or ischemic. New, fragile vessels develop as the circulatory system attempts to maintain adequate oxygen levels within the retina. This process is called neovascularization. Unfortunately, these delicate vessels hemorrhage easily, and blood may leak into the retina and vitreous, causing spots or floaters, along with decreased vision. In the later phases of the disease, continued abnormal vessel growth and scar tissue may cause serious problems such as retinal detachment and glaucoma.

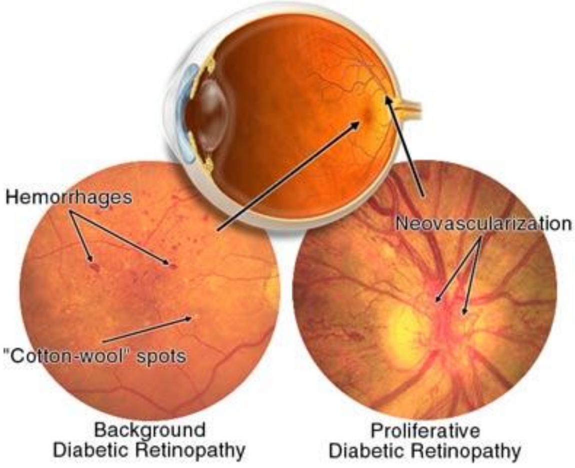
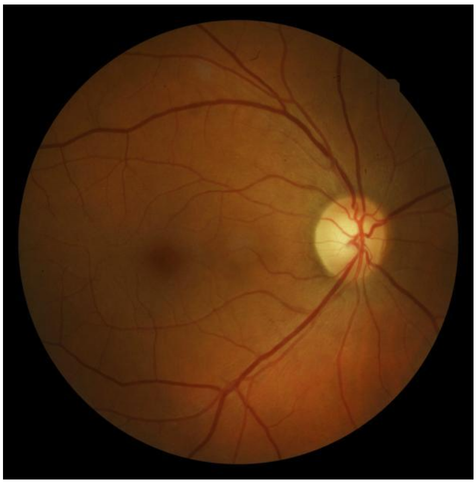


Figure 3.2 :Mechanism of DR development

The effect of DR on vision varies widely, depending on the stage of the disease (Figure 3.2). Some of the common symptoms of DR are listed below, but diabetes may cause other eye symptoms such as blurred vision (this is often linked to blood sugar level, floaters, flashes, and sudden loss of vision).Diabetic patients require routine eye examinations so that related eye problems can be detected and treated as early as possible. Most diabetic patients are frequently examined by an internist or endocrinologist, who in turn works closely with the ophthalmologist. The diagnosis of DR is made following a detailed examination of the retina with an ophthalmoscope, and the prognosis for visual recovery is dependent on the severity of the detachment. Researchers have found that diabetic patients who are able to maintain appropriate blood sugar levels have fewer eye problems than those with poor control. Diet and exercise also play important roles in the overall health of those with diabetes.



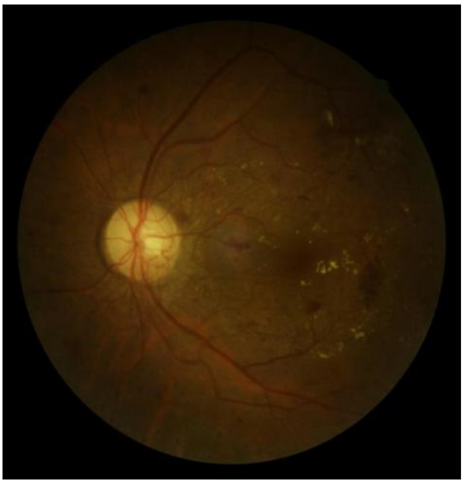


Figure 3.3: Photograph of a normal eye Figure 3.4 : Photograph of a DR eye

Diabetic macular oedema is the leading cause of legal blindness in diabetics and can be present at any stage of the disease but is more common in patients with proliferative DR (Figures 3.3 and 3.4)

3.4 Epidemiology

The best predictor of DR is the duration of the disease. After 20 years of diabetes, nearly 99% of patients with type 1 diabetes and 60% with type 2 have some stage of DR, and 33% of patients with diabetes have signs of DR. People with diabetes are 25 times more likely to become blind than the general population (Yanoff and Duker, 2008).

Some important points in Research strategy for Diabetes:

- Diabetes occurs in men, women, the young and old and in all races. No group is spared.
- There is no known cure for diabetes and available treatments have limited success in controlling the devastating consequences of the condition.
- Diabetes affects 5 per cent of the world's population and its prevalence is doubling every generation.
- The International Diabetes Federation estimates that in 2005 around 333 million1 people in the world aged 20–79 had diabetes.
- More than two million people in the UK have been diagnosed with diabetes. This number is predicted to reach three million by 2010.

- It is estimated that around another 750,000 people in the UK have diabetes but do not know they have it.
- There are over 250,000 people in the UK with Type 1 diabetes. This is caused by an absolute lack of the hormone insulin, resulting from autoimmune destruction of the body's pancreatic islet beta cells.
- Around 1.8 million people have Type 2 diabetes, 2, 3 representing about 90 per cent of diabetes cases. Type 2diabetes is due to varying combinations of insulin deficiency and insulin resistance.
- The incidence of Type 1 diabetes in children is rising at a rate of 3–4 per cent a year. 4 We do not know why.
- The increase in Type 2 diabetes is closely linked to an aging population and rapidly rising numbers of obese or overweight people.
- Diabetes is the leading cause of kidney failure, blindness in adults, and amputations. It can lead to impotence, can affect mental health and wellbeing, and is a major risk factor for heart disease, stroke and birth defects.
- On average, life expectancy is reduced by 20 years in people with Type 1 diabetes and by 10 years in people with Type 2 diabetes. In the next 10 years there will be a 25 per cent increase in the number of diabetes-related deaths.
- The clinical supervision and care of people with diabetes currently consumes 5 per cent of the NHS budget (about £10 million a day) and 10 per cent of hospital in-patient resources. The NHS spend on diabetes will rise to around 10 per cent of the NHS budget by 2011.
- Various national plans outlining the standards of care that should be expected by people with diabetes exists but its full implementation across the NHS has yet to be achieved.
- People with diabetes (or their carers) are responsible for the day-to-day management of their condition.

3.5 Symptoms

DR is asymptomatic in the early stages of the disease, but as the disease progresses, the symptoms may include: blurred vision, floaters, fluctuating vision, distorted vision, dark areas in the vision, poor night vision, impaired color vision, and partial or total loss of vision. Known risk factors include: duration of diabetes, poor blood sugar control, HTN and hyperlipidaemia. The effect of intensive diabetes treatment on the progression of DR in insulin-dependent diabetes mellitus reduced the risk of developing retinopathy by 76% and slowed the progression of retinopathy by 54%.

3.6 Data collection and pre-processing

The ophthalmologist is in a unique position to collect information on sight loss. Without the collection and analysis of this type of data, it is not possible to understand the changing epidemiology of DR as well as other important conditions leading to sight loss and blindness. Collection of outcome data is essential for a screening programme to:

- undertake audit to ensure that systems are working effectively;
- demonstrate cost-effectiveness of screening as an intervention;
- understand inequalities in access to services;
- use the information to improve services for the future.

The Royal Liverpool University Hospital (RLUH) has been a major centre for retinopathy research since 1991. Retinopathy is a generic term used to describe damage to the retina of the eye which can, in the long term, lead to visual loss. Retinopathy can result from a number of causes, for example: diabetes, age- related macular degeneration (AMD), high blood pressure and various genetic conditions. In diabetes the retinopathy progresses over a number of years through well characterized stages. Treatment comprises the application of laser to the retina and is most effective during the stages before vision is affected. Screening Programmes for people with diabetes have recently been established in all four UK nations to detect retinopathy and refer for prompt treatment.

RLUH has collected a substantial amount of data over a considerable period of time as part of its diabetic retinopathy research and screening programme.

Screening takes place within the community and is conducted by technicians who perform photography and record data images on “lap-tops" which are then down-loaded (typically) at the end of each day. Retinal images are graded at a central grading facility at a separate time within a few weeks with results recorded onto a database. If disease is detected on the retinal photographs worse than a predetermined level or if photographs are upgradable or unobtainable then patients are invited to a dedicated hospital outpatient clinic for further examination by an ophthalmologist using more specialized slit lamp biomicroscopy (A high intensity light source instrument to facilitate examination of the human eye). Data on retinal findings are entered into the database. This clinical assessment can occur several months after the initial photographic screening.

Four types of data associated with a single screening sequence are collected:

1. General demographic data.
2. Data on visual acuity (clarity of vision).
3. Data from grading of retinal images.
4. Data from biomicroscopy of the retina.

The full screening sequence is referred to as a “screening episode".

People with diabetes are usually screened once a year with the option to rescreen early (typically 6 months) depending on presence of intermediate levels of disease indicating greater risk of progression. The RLUH screening programme currently deals with some 17,000 people with diabetes registered with general practices within Liverpool Primary Care Trust per year. Overall details of some 20,000 patients have been recorded. Consequently a substantial amount of data is available for analysis. Some further details of the data collection are presented in the following sub-section.

3.6.1 Diabetic retinopathy Databases

Longitudinal data is data that is repeatedly sampled and collected over a period of time with respect to some set of subjects. Typically values for the same set of attributes are collected at each sample points. The sample points are not necessarily evenly spaced. Similarly the data collection process for each subject need not necessarily be commenced at the same time. A regular longitudinal data set is one where data at each sample point is collected simultaneously for all subjects. Most longitudinal data sets are not regular. The most common example of irregular longitudinal data sets are patient medical records where patients enter and leave the “system” continuously and data is collected during consultations which occur at irregular intervals (episodes/time stamps). One example of an irregular longitudinal database, and the focus of the research described here, is the Diabetic Retinopathy screening dataset maintained by The Royal Liverpool University Hospital (RLUH).

Longitudinal data thus provides a record of the “progress” of some set of features associated with the subjects. Medical longitudinal data, such as the Diabetic Retinopathy data, typical plots the progress of some medical condition.

St Paul’s Eye Unit is a major referral centre for patients with diabetic retinopathy. Recruitment will be run from established NHS services and referrals.

Baseline screening takes place in diabetic retinopathy assessment clinic, medical retina clinic and general ophthalmology clinic of St. Paul’s Eye Unit at Royal Liverpool University Hospital. Data are collected and stored on the encrypted disk in accordance with Data Protection Act and Caldicott guidelines.

The data which are collected for a warehouse with 22,000 patients, 150,000 episodes, at least 1200 values of attributes, include demographic details, visual acuity data,

photographic grading results, data from biomicroscopy of the retina and results from biochemistry investigations. One challenging task is the application of logic rules either to address the missing value problem or to retrieve knowledge from existing information. One of the major challenge of the work described, is that the data collection is extremely large and complex; comprising about 450 attributes (of various types: categorical, quantitative, text, etc.), distributed over two databases each composed of a number of datasets. Another challenge represented by the data, was that unlike more standard longitudinal data sets, there was no clear association between specific time stamps and subsets of the data. The data warehousing process established to prepare the data for mining is therefore also described.
The datasets, contained in the RLUH database and used to construct the warehouse in this research work, are:
1. Patient Details. Table containing background information regarding individual patients.
2. General. Demographic patient details and visual acuity data.
3. Photodetails. Results from the photographic grading.
4. Biomicroscopy. Results from the slit lamp biomicroscopy in cases where this has been conducted.
5. Risk Factors. Results from blood pressure and biochemistry investigations known to be associated with an increased risk of progression of retinopathy.

The Diab database contains paper work and film photos from 1991 – 2005. This database consists of several datasets:

- TbDiabEyePatient (8437×17)
- Dead (Yes or No)
- Sex (M or F)
- Age (number)
- Personal Details
- tbDiabRetinPhotDetails (63110×141)
- Examination Details for both eyes. The vast majority of the columns are numerical data. There are few columns with string data.
- tbDiabRetinBiomicDetails (15070×171)
- Details of exam date,grading for both eyes. Both string and numerical data.
- tbEyeGeneral (62875×52)
- Age of exam, eye conditions (cataract, glaucoma, family history glaucoma, weak/lazy eye etc). Both string and numerical data.

The RAD is the risk assessment database. It consists of two tables: one contains the risk factors and the other demographic data. The risk factors table contains medical information regarding the patient e.g. diastolic pressure, systolic pressure, range random cholesterol, urea, creatine etc. The table with the demographic data contains personal details of the patients and details of diabetes (type, year of diagnosis, diabetes carer)

- All databases contain both numerical and string values. Also, all of them have missing values.
- Some medical characteristics have numerical values. The following schemas show the maximum and the minimum values that these characteristics can take (Appendix 1).

Data collected from the diabetic retinopathy screening process described above is stored in a number of databases. The structure (tables) of Diab database reflects the mechanism whereby patients are processed and includes historical changes in the process. Screening commenced in 1991 when data was recorded in a bespoke database system called Epi-Info. The number of records in the Epi-Info database is small and for this reason it is not considered appropriate with respect to the intended temporal pattern mining study described here. Epi-Info was replaced with a more sophisticated system, Diab, in 1991, which describes the data used in this study. Diab, in turn, was replaced with a national database system, Orion, in 2005. The design and implementation of Orion does not lend itself to simple extraction of data for temporal pattern mining purposes and thus the data contained in this latest database system also does not form part of the current study. Thus the study described here deals with data collected from 1995 to 2005.

The RLUH, as opposed to the screening programme, also maintains a clinical investigations database called Ice. This database includes information about biochemical "risk factors" that are known to be associated with progression of diabetic retinopathy. Not all patients included in the screening programme have records on ICE. The screening programme has its own Risk Factors database, maintained by the programme team, containing data mostly extracted from ICE.

An additional complication was that the data, in common with similar patient datasets, was very noisy in that it contained much missing and anomalous data.

This issue was addressed by defining a set of logic rules. In the context of missing data the logic rules were used to derive appropriate values. In the case of anomalous data, the logic rules were used to derive additional attributes to formulate consensus values.

The nature of the longitudinal data is of interest because it does not fit into any standard categorization of such data, in that the "time stamp" used is the sequential patient consultation event number. The duration between consultations is also variable.

3.6.2 Data warehousing and cleaning

For the study described in this thesis, before any investigation of trend mining could commence the five database tables identified in the above subsection (Patient, General, Photodetails, Biomicroscopy and Risk factors) were combined into a single warehouse (i.e. a static data repository specifically intended for the application data mining and data analysis tools). The creation of the data warehouse required data anonymization and data cleaning.

The anonymisation of the data tables was initiated by removing patient names. Although this was straightforward, this presented a second problem as in many cases the patient name was the common "key" linking database tables.

An obvious candidate for a universal common key was patient NHS (National Health Service) numbers; however this was missing with respect to some 8000 records and consequently had to be added manually. The NHS number was then used for the construction of the data warehouse; on completion the NHS number was replaced by a sequential record number so that individual records could not be traced back to individual patients.

The next step after anonymisation was data cleaning. There were three principal issues to be addressed:

1. Missing values
2. Contradictory values (conflict)
3. Duplicate records

The first two issues were addressed by developing a set of logic rules.

The problem of missing attribute values is well established in the context of data mining. In any large database, we encounter a problem of missing values. A missing value may have been accidentally not entered, or purposely not obtained for technical, economic, or ethical reasons.

The missing value problem is widely encountered in medical databases, since most medical data are collected as a by-product of patient-care activities, rather than for organized research protocols, where exhaustive data collection can be enforced. In the emerging federal paradigm of minimal risk investigations, there is preference for data mining solely from by-product data. Thus, in a large medical database, almost every patient-record is lacking values for some feature, and almost every feature is lacking values for some patient-record.

One approach to address this problem is to substitute missing values with most likely values; another approach is to replace the missing value with all possible values for that attribute. Still another approach is intermediate: specify a likely range of values, instead of only one most likely. The difficulty is how to specify the range in an unbiased manner. The generally agreed view is that removing records with missing data is the least favoured option as this may introduce bias. The reduction of the overall data set size, by removing records that contain missing values, is not considered to be critical. There is significant scientific work to support this view. Approaches to the imputation of missing values have been extensively researched from a statistical perspective (Kalton, Kasprzyk,1986),(Little, Rubin, 2002)(Mumoz, Rueda,2009) .Example imputation methods include: nearest neighbour imputation, mean imputation, ratio imputation and regression imputation.

The approach to missing data advocated in this study is to define and implement a set of logical rules to address the missing value problem; this is discussed further in the following section, 3.6.3.

In this study we define 2 different types of missing data: the data that is not entered because there is no meaning in some cases (not applicable) and the data that is accidentally not entered (not recordable).

With respect to missing values the evidence of such a missing value could be interpreted in three ways:

- The value was either unknown or mistakenly omitted at time of collection.
- The missing value indicated a negative response to a question suggested by the field.
- The clinician considered the field to be inapplicable for the given case.

For example some attributes indicated responses to question such as "does the patient have one weak eye", to which, in many cases, the clinician had inserted a "yes" if the answer to the question was an affirmative and left the field blank otherwise (the latter can thus be interpreted as either a "no", or a "don't know".

A set of "if . . . then . . . "logical rules were therefore developed to address this issue. The logic rules were written in such a way that they could also be used for data validation purposes. The operation of these rules is best illustrated using some examples (See Appendix 1 for all rules).

Consider the field SeeGPRegulary featured in the Diab General dataset.

This field can have three possible values: 1 ("No"), 2 ("Yes") and 9 ("Don't know").

In the event of a missing value for this field it can be derived from another field, in the set of database tables, LastSeeGP; asking when the patient last saw their GP for anything.

The LastSeeGP field can have the following values:

- 1 ("Within last 6 months")

- 2 ("Within last 6 to 12 months")
- 3 ("More than a year ago")
- 9 ("Don't know")

The logic rule is then as shown below (the null value indicates a missing field). The rule states that if the value for SeeGPRegulary is missing and the value for LastSeeGP is also missing, or set to 9 ("Don't know"), we set the value for SeeGPRegulary to 9. If the patient has seen their GP with the last 12 months (LastSeeGP field set to 1 or 2) we set the value for SeeGPRegulary to 9 ("Yes").

Otherwise we set the value of SeeGPRegulary to 1:

- if (SeeGPRegulary == null)
- if (LastSeeGP == 9) or (LastSeeGP == null) then (SeeGPRegulary = 9)
- if (LastSeeGP == 1) or (LastSeeGP == 2) then (SeeGPRegulary = 2)
- if (LastSeeGP == 3) then (SeeGPRegulary = 1)

With respect to contradictory/anomalous values this issue can be exemplified by the diAgeDiag field, the age of the patient when diabetes was first diagnosed. Within the application domain this has been recognised as a question patients find very difficult to answer, and consequently clinicians responsible for gathering data often leave this field blank if they feel that a patient is unable to give a definitive answer. In addition it was found that patients may give a different answer over different consultations, hence it was believed to get less accurate with the passing of time. The rule adopted in this case was to take the first recorded value of the field as this was likely to be the most accurate.

3.6.3 Issues and challenges of medical data

The field of medical informatics has evolved around structuring, processing, storing and transmitting medical information for a variety of purposes (Shortliffe, 1990). One such purposes is to develop decision-support systems that enhance the clinician's ability to diagnose, treat and assess prognoses of pathological conditions. Even if disease processes were fully understood, population variability would still make individualised diagnosis, treatment and prognosis, all essential parts of good health care, difficult classification tasks. The reality is, however, that diseases are not fully understood, nor is population variability fully taken into account in many decision-making situations. Sometimes it is not possible for a clinician to employ the principles learned in the basic and clinical sciences to determine whether a patient has a given disease, whether the patient should be given a certain treatment or how long the patient will survive.

Medical informatics has been an important area for the application of computing and database technology for at least four decades and this thesis presents a number of new research challenges in this area. These include the need for complex-data modelling features, advanced temporal support, advanced classification structures, continuously valued data, dimensionally reduced data and the integration of very complex data. In addition, the support for clinical treatment protocols and medical research is an interesting area for research.

It is extremely important to have a good understanding of the data when embarking on a data mining project and this is facilitated by considering the following questions:

- What data is available?
- What available data is actually relevant or useful?
- Can the data be enriched from other sources?
- Are there historical datasets available?
- Who is the real expert on the data to whom questions can be addressed?
- Are the results at all sensible?

Cios et al. (2002) refer to a number of important issues and challenges that were also encountered in the work of this thesis, and which will now be described.

Medical datasets often contain insignificant, redundant or inconsistent data objects or attributes that present a number of issues and challenges such as:

- Dimensionality reduction. The large volume and heterogeneity of medical databases makes it unlikely that any data-mining tool can succeed with raw data (Cios and Moore, 2000). The tools may require that a sample is extracted from the database in the hope that results obtained in this manner are representative of the entire database. Dimensionality reduction can be achieved in two ways:
- Updating. Medical databases are constantly updated by, say, adding records (for an existing or new patient), or by replacement of the existing records. This requires methods that are able to incrementally update the knowledge learned so far.
- Missing values. The medical information collected in a database is often incomplete and it is very difficult to avoid the problem of missing values. This happens either because some values were accidentally not entered or not obtained for technical or ethical reasons. Sometimes the patients themselves are unsure of the answers to some of the questions they are asked. Therefore, databases typically contain significant levels of noise. For data mining purposes it is important to eliminate this noise in order to achieve accuracy in the results. There are several ways to address the problem of missing values. For example, it might be possible to substitute

missing values with the most likely values; another approach is to replace the missing value with all possible values (Cios et al., 1998). Another approach is to use the experience of clinicians in order to create logic rules to replace missing values. One of the major concerns in large longitudinal medical datasets is how to find natural groupings (clusters); objects are clustered together if they are similar to one another and at the same time are not similar with objects from other groups. Without at least partial human supervision (Cios, 2001), it is easy to end up with results that do not make sense. This thesis will use a variety of methods to investigate and treat the missing values problem: (i) to assign cell averages, (ii) to use the frequency distribution of every field to decide whether or not to include that field in the analysis, (iii) to replace empty attributes with a global constant value, (iv) to use logic rules based on the experience and human knowledge of the domain.

- Data ownership. Data ownership is a topic of debate in the field of medical data mining. Legally, ownership is determined by who is entitled to sell a particular item of property (Moore and Berman, 2000). The corpus of human medical data potentially available for data mining is enormous, with thousands of terabytes being generated annually in UK.
- Privacy and security of human data. Privacy and security are areas of concern with medical data. UK law includes guidelines for the concealment of individual patient identifiers. At stake is not only a potential breach of patient confidentiality, with the possibility of ensuing legal action, but also erosion of the physician–patient relationship, in which the patient is often candid with the physician in the expectation that such private information will never be made public. Under some guidelines concealment of identifiers must be irreversible.
- While it is possible for these special requirements to be managed by appropriate regulatory agencies, this is not possible in the case of totally anonymised data. There are four forms of patient data:
- Anonymous data. Data where the patient identification was removed at the time the data was collected. For example, a block of tissue may be taken from an autopsy on a patient with a certain disease to serve as a control tissue block in the histology laboratory. The patient's identifiers are not recorded at the time of specimen collection and thus can never be recovered.
- Anonymised data. Data that are collected initially with the patient identifiers, which are subsequently and irrevocably removed. That is, there

can never be a possibility of returning to the patient's record and obtaining additional information. While this practice was common in the past, it is no longer used as standard since accidental duplication is possible and such data is thus difficult to verify for corrections or additional data.

- De-identified data. Data that are collected initially with the patient-identifiers and are subsequently encoded or encrypted.
- Identified data. Fully identifiable data which can only be collected under significant review by the institution, federal guidelines, etc. with the patient giving written informed consent.

If one employs only data that are collected as part of the ordinary diagnosis and treatment of patients, so that there is no change in patient management (course of treatment) as a result of the research, such as pressure on the patient to accept or refuse certain management or call-back for additional data that might upset the patient or next of kin, then the only risk of using such data is the loss of confidentiality to the patient.

- Administrative issues. Emerging guidelines for patient privacy specify a number of administrative policies and procedures that would not ordinarily be required for non-medical data mining (Saul, 2000). Such policies are required to evaluate and certify that appropriate security measures are in place in the place of research. There must be legal contracts between the organisation and any outside parties given access to individually identifiable health information that require the outside parties to protect the data.
- Security issues. There must be security training for all staff accessing computer-based databases, including awareness training for all personnel, periodic security reminders, user education concerning virus protection, user education in the importance of monitoring login failures, password management, and how to report discrepancies. These and many other rules impose constraints upon medical data miners that other academic researchers may regard as burdensome and stifling to the creativity of scientific research. Researchers must carefully assess the perceived need for information such as postcodes (which might be necessary for epidemiological studies), that have the potential to also render the data re-identifiable in combination with other information (Sweeney, 2001).
- Statistical philosophy. There is an emerging doctrine that data mining methods themselves, especially statistical methods, and the basic assumptions underlying these methods, may be fundamentally different for medical data. Human medicine is primarily a patient-care activity and

has only a secondary role as a research resource. Generally, the only justification for collecting data in medicine, or the refusal to collect certain data, is to benefit the individual patient. Some patients might consent to be involved in research projects that do not benefit them directly, but such data collection is typically very small-scale, narrowly focused, and highly regulated by legal and ethical considerations. The major points of statistical philosophy in medicine may be organised under these general headings:

  - Ambush in statistics
  - Data mining as a superset of statistics
  - Data mining and knowledge discovery process

- Importance of physician's interpretation: The physician's interpretation of images, signals or any other clinical data is normally written in unstructured free-text English that is very difficult to standardise and thus difficult to mine. Even specialists from the same discipline very often cannot agree on unambiguous terms to be used in describing a patient's condition.
- Volume and complexity of medical data: Raw medical data are voluminous and heterogeneous. Medical data may be collected from various sources including images, patient interviews and physician's notes. All data-elements may influence diagnosis, prognosis and treatment plan and must be taken into account in data mining research.

3.7 Conclusion

This chapter gives a quick overview of diabetic retinopathy and continues with the issues of pre-processing and post-processing (before and after rule generation) that have largely been ignored by the data mining research community. Yet these issues are critical to the success of any real-life applications to deal with these issues, we have proposed the use of a semi-automatic data cleaning system for cleaning the noisy data and an exploration mining strategy for easy understanding of the rules generated by the state-of-the-art data mining techniques.

Chapter 4 Trend-mining framework description

4.1 Introduction

In modern applications data from different sources that vary after some time must come together and be transformed into a dataset from which the user can extract knowledge. That kind of knowledge should show causality between the data attributes. This means that any discovered rule of the form $X \rightarrow Y$ should show a relationship between the attributes in X and the attributes in Y. However, when the amount of data is huge the data must be filtered out from noise, identify frequent item sets and distinguish which of them are interesting or not. Moreover, when data change in time it is important to know how relation between attribute changes and how causality is affected. The trend mining framework that is proposed in this thesis forms a part of the investigation of how to deal with large noisy and time varying data.

This chapter provides a description of the trend-mining framework, detailing all major steps of the framework: pre-processing, main processing association rule mining (ARM) and trend generation, and output production. Moreover, this chapter provides the main aspects of verification and validation. The fundamental idea of verification is to test whether the intermediate results and/or outputs of the framework are self-consistent and the main idea of validation is to test the outputs of the framework and also check the consistency of them in respect of what experts already know.

The chapter contains the following parts: firstly it provides a generic description of the trend-mining framework, and continues with the aspects of verifying and validating the framework and finally provides a description of SOMA which is the application of trend mining in the area of diabetic retinopathy.

SOMA starts with the pre-processing of data, continues with the Association Rule Mining algorithm, continues with the new trend generation algorithm Aretaeus, and ends with the (colourful) visualization of trends producing a mosaic-based representation of knowledge. Pre-processing includes the preparation of time-stamped datasets through the application of logic rules both for the creation of time stamps and for the band creation of continuous variables, for the transformation of the continuous variables into categorical using bands: for example the age of patient is a continues variable and it's become categorical using bands; so if age is between 0 and 12 the categorical value is 1, if age is between 12 and 20 then the categorical value is 2 and so on, as well as to correct errors in the datasets. The Association Rule Mining algorithm, that is described here, is used for the discovery of 'interesting' rules. The trend-mining algorithm, Aretaeus, uses mathematical conditions to produce trends, based on the changes of the support count of an association rule and then it classifies them into groups based on how the support count of a rule changes in each time stamp, and finally a description of the concept behind the visualization technique is given.

4.2 Trend mining framework

As stated earlier, this section describes the trend mining framework. More specifically, it outlines the fundamental principles behind each element of the trend mining.

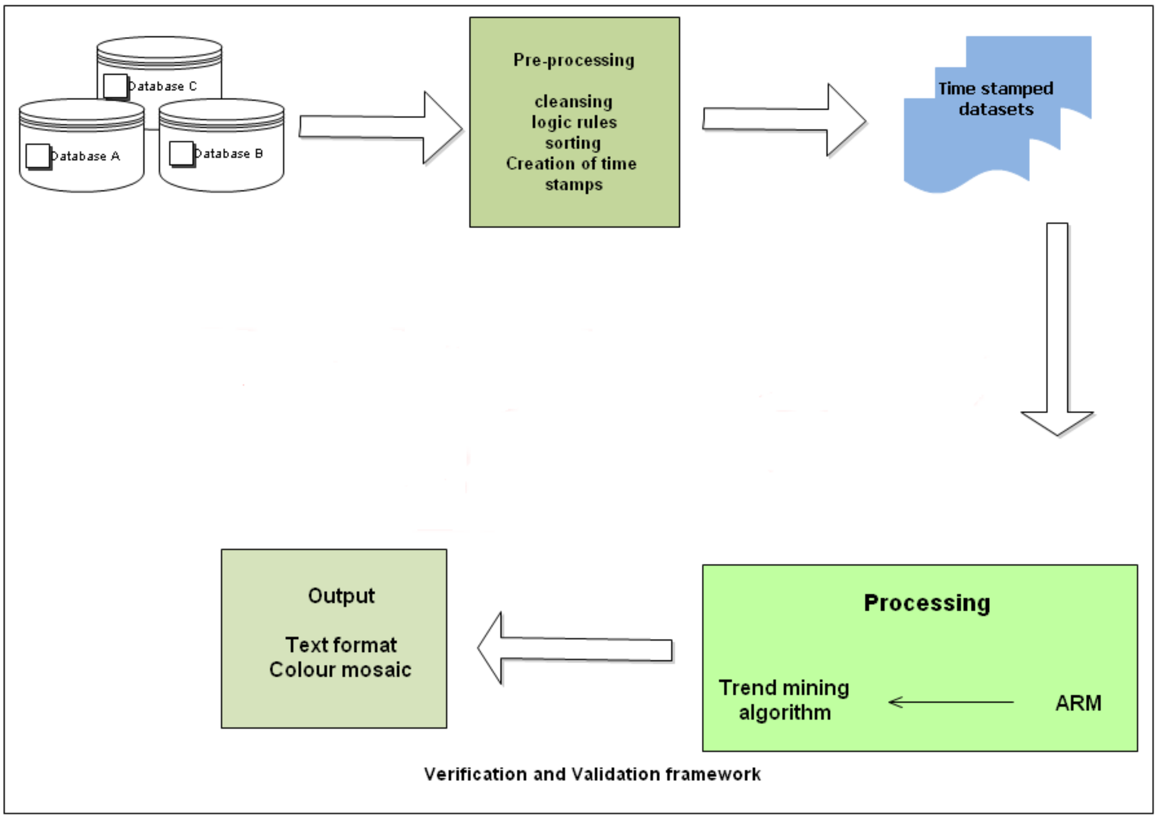


Figure 4.1: Trend mining framework representation

Figure 4.1 shows the stages of trend mining framework. Data from different sources bring into pre-processing and they are transformed into time – stamped datasets which then enter the main processing stage where ARM first and trend mining algorithm subsequently aims to identify frequent and interesting rules and then to create trends. The final stage is to output the discovered knowledge. The output has two forms; one is text which reports analytically the trends and how their characteristics change at every time stamp (support, confidence, and lift) and the second form aims to represent using colours how a group of objects ,which initially have the same attributes values, appear from one trend to another at each time stamp.

4.2.1 Pre-processing

Pre-processing is the first step of the framework where data from different sources come together. Data may have several forms, discrete, continues, numerical, text or combinations of these. Real data come with several problems such as missing values, duplicated records, values entered by mistake and so on. Another issue, which is related to discovery of frequent item sets at the next stage of the framework, is how to treat continuous values. The reason is that the presence of continuous values makes it difficult to identify frequent item set and therefore the amount of extracted knowledge might become huge. The framework deals with this problem by sorting the continuous values into bands. Depending on the data, bands represent equal intervals or not.

Another important function of the pre-processing stage is the application of logic rules to reduce noise from data and to ensure that data are consistent with the domain that they describe. To replace missing values several methods can be used either the averaged observed value or the most frequent observed value or to use the knowledge of an expert who for example can combine values of other attributes to determine how to replace the missing value.

After the application of logic rules, sorting and cleansing, pre-processing performs the task of creating time stamped datasets. Each dataset contains data under certain time conditions and thus all datasets show how data evolve with time. However, when there are data from different sources and data are not collected with the same frequency or when different data collection takes place on different dates then it is complicated to define a clear association of how to define a time stamp. The solution of this problem lies solely on the knowledge of the domain in order to define the time window between each time stamp.

4.2.2 Main processing

The main processing stage of the framework consists of two processes:

- The association rule mining (ARM) process.
- The trend generation and categorization process.

The mechanisms behind each process are totally different and a detailed description is given in the following subsections. Another difference is that the ARM – process is

repeated for every time stamp while the trend generation and categorization takes place only once after all time stamped datasets have passed through the ARM – process.
The ARM –process performs the following tasks: firstly to indentify the frequent item sets and create rules of the following form $X \rightarrow Y$. Y is the consequent of the rule and it is the subset that contains attribute(s) that describe a class or a transaction. X is the antecedent of the rule and it is a subset that contains attributes which may or not be related with subset Y. Then it keeps the frequent item sets and calculates characteristics that measure interestingness confidence, lift and confidence of the inverse rule. Thus the filtering process is twofold; one part is to measure frequency of existence and the other is to measure interestingness.

4.2.2.1 Association rule mining

The ARM process is the stage where the data are filtered by the identification of rules X-> Y which are frequent and interesting. The support threshold and the confidence threshold are determined by the user of the framework and they are used in the filtering process. The number of occurrences of the item that contains subsets X and Y must be greater or equal to the support threshold. In addition, the confidence of the rule X->Y must be greater or equal to the confidence threshold. The threshold is chosen by the user depending on the amount of knowledge (number of discovered rules) the user wants to reveal.

In large datasets time is an important factor and in this thesis where the ARM – process is repeated in every time stamp the mechanism of identifying frequent and interesting item sets must be capable to perform those tasks with the least passes through data. The matrix algorithm has been chosen for its ability to identify which item sets are frequent with one pass through of the dataset. Below, the major steps of the algorithm are described.

The first thing that the algorithm does is to look for all the single items (attribute values) from all datasets. These items form I, where:
$I = \{i_1, i_2, i_3... i_N\}$.

The next step is the formation of the generating matrix $G = \{g_{ij}\}$. If M is the number of patients, then i = 1, 2... M and j = 1, 2... N. If D is the set of transactions, then $D = \{t_1, t_2, t_3 ,..., t_M\}$. In this study, the term transaction refers to each line of each time stamped dataset which is created from pre-processing. Therefore, the generating matrix G is a M × N, where:

$$g_{ij} = \begin{cases} 1, \mathit{if}\ i_j \in t_i \\ 0, \mathit{if}\ i_j \notin t_i \end{cases}$$

Then, using I, the algorithm produces all candidate k-item sets, using combinations of items in I, C.

C= {{C_1}, {C_2}... {C_N}},

Where C_1 contains all candidate one-item sets, C_2 contains all candidates two-item sets, and so on.
Then, for each candidate item set in Ci (i=1, 2, 3... N), the vector S is produced.
Vector S is a binary vector and has space equal to the number N.
Let C be a candidate item set:

$\mathit{if}\ c \in C_1\ \mathit{then}\ S\ \mathit{has\ only\ one\ unite\ element}, S^1$
$\mathit{if}\ c \in C_2\ \mathit{then}\ S\ \mathit{has\ only\ two\ unite\ elements}, S^2$
$\mathit{if}\ c \in C_3\ \mathit{then}\ S\ \mathit{has\ only\ three\ unite\ elements}, S^3$

and so on.
Then, for each candidate, c creates its feature vector from the following equation for every time stamp (episode):

$\mathit{if}\ c \in C_1$

$$\sup(\{c\}) = \sum_{j=1}^{M} \langle g_j, S^1 \rangle$$

$\mathit{if}\ c \in C_2$

$$\sup(\{c\}) = \sum_{j=1}^{M} \mathrm{int}\left[\frac{\langle g_j, S^2 \rangle}{2}\right]$$

Generally, if l=1, 2, 3, ..., N, the support of each candidate l-item set would be given by

$$\sup(\{c\}) = \sum_{j=1}^{M} \mathrm{int}\left[\frac{\langle g_j, S^l \rangle}{l}\right]$$

The sign <> denotes the inner product of two vectors, and the sign int[.] denotes a function that changes a real number into an integer, e.g. int[0.8] =0.0.

The minimum support value MinSup works as a threshold that determines which item set is frequent or not. If sup({c}) is less than the MinSup, then it is assigned the value 0.

Let $\overrightarrow{SUPP}$ be the feature vector of a candidate item set; if the sum of the vector components equals zero, this means that across all datasets, this item set has support count less than the threshold, and so it is discarded from the model for further analysis.

The candidate item sets whose vector $\overrightarrow{SUPP}$ has at least one non-zero element are those that undergo further analysis. The next pseudo-code describes how the association rule mining algorithm works.

```
Set E to the total number of time stamps
Set P to the number of objects
Set I to the number of all 1-item sets
Set C to the number of all candidates
For k=1 to E
    Set G^k to the matrix algorithm for time stamp k
    For kk=1 to C
        For ii=1 to P
            $\text{sup} = \text{sup} + \text{int}\left[\frac{\left\langle G^k(i,:), S^l(kk)\right\rangle}{l}\right]$
End
```

```
For kk=1 to C
    If sum { $\overrightarrow{SUPP}$ (kk, :)} = 0
        discard
    Else
        Keep
End
```

From the above description to can be understood the importance of the prepossessing stage which takes as impute heterogeneous data and transforms them into files written with the same manner, binary datasets.

After the discovery of frequent item sets is completed the ARM- processes is looking for rules X->Y whose confidences equal or exceed the confidence threshold. However, unlike some Apriori algorithms the ARM process here knows from the user which subset of items forms the left hand side of the rule X and which subset of items forms the right hand side of the rule Y.

The next step is to keep the rules that have satisfied both the support and confidence threshold, and calculate supplementary measures of interestingness, lift and the confidence of the inverse rule to avoid any misleading conclusion from the sole use of confidence.

After, repeating the procedure above for all time stamps the next step is to create vectors for each rule. The element of each vector is the support value of the rule for every time stamp, so the first element is the support at the first time stamp, the second element is the support value at the second time stamp and so on. In the case where a rule X->Y does not satisfy the support threshold condition in some time stamps then its vector has zero elements at those time stamps. Those vectors are the input of the second part of the main process, the trend mining algorithm.

The matrix algorithm principles were taken from the work of Yuan and Huang (2005). For the need of this research work a script has been created to be incorporated with the pre-processing and trend generation scripts. The reason for that was to have a script with the least interaction from the user . During the pre-processing stage the algorithm uses an internal language in order to recognise the attributes and their values. So it was decided to write a novel script for matrix algorithm which will be able to understand the pre-processing. The whole script was created in MATLAB.

4.2.2.2 Trend mining

Trend mining is implemented using mathematical prototypes on the vectors of support in order to show how the support for each rule changes at every time stamp and thus helps the visualization tool to identify how the changes on the support may be linked or not with changes to the values of attributes either at the left or the right hand side of the rule.

The following categories of trend have been identified:

- Increasing: the support increases with every time stamp, and the growth rate is greater than or equal to the growth rate threshold GR, which is defined as: Let I be a frequent item set in D1, D2, …, D_n with support S1, S2, ..., S_n, where n is the number of timestamps; the growth rate GR is then:
- $$GR = \sum_{i=1}^{n-1} \frac{S_{i+1} - S_i}{S_i} + 1$$

- Decreasing: the support decreases at each time stamp but never becomes 0.
- Constant: the support either remains constant or does not change above or below a tolerance threshold.
- Jumping: initially, the support is zero, at some point becomes non-zero, and then remains non-zero.
- Disappearing: the support from non-zero becomes zero and stays zero for the rest of the time stamp.
- Fluctuating: the support changes without falling into any of the other classes, described above.

The table below describes how, mathematically, the trends are categorized:

Table 4.1: Mathematical conditions for trend categorization

Type	Mathematical conditions
Increasing	$\frac{S_{i+1}}{S_i} > 1, \forall\, i \in [1, n-1]$, GR>ρ
Decreasing	$\frac{S_{i+1}}{S_i} < 1, \forall\, i \in [1, n-1]$
Constant	$\frac{S_{i+1}}{S_i} = 1 \pm k, \forall\, i \in [1, n-1]$, k : tolerance threshold
Fluctuating	$\frac{S_{i+1}}{S_i} = 1 \pm k, \forall\, i \in [1, n-1] \quad and \quad \frac{S_{j+1}}{S_j} > 1, \forall\, i \in [1, n-1], j \neq i$ $\frac{S_{i+1}}{S_i} = 1 \pm k, \forall\, i \in [1, n-1] \quad and \quad \frac{S_{j+1}}{S_j} < 1, \forall\, i \in [1, n-1], j \neq i$ $\frac{S_{i+1}}{S_i} > 1, \forall\, i \in [1, n-1] \quad and \quad \frac{S_{j+1}}{S_j} < 1, \forall\, i \in [1, n-1], j \neq i$ $\frac{S_{i+1}}{S_i} > 1, \forall\, i \in [1, n-1] \quad and \quad \frac{S_{j+1}}{S_j} < 1, \forall\, i \in [1, n-1], j \neq i \quad and \quad \frac{S_{l+1}}{S_l} = 1 \pm k, \forall\, l \in [1, n-1]\ l \neq j, l \neq i$
Jumping	$for\ m < n: \quad S_i = 0, \forall\, i \in [1, m] \quad and \quad S_i > 0\ \forall\, i \in [m+1, n]$
Disappearing	$for\ m < n: \quad S_i > 0, \forall\, i \in [1, m] \quad and \quad S_i = 0\ \forall\, i \in [m+1, n]$

4.2.3 Representation of the trends

The last task of representation is to output the discovered knowledge, and this is achieved in two ways. One way is the generation of a text output where the outcome is recorded as the name of the rule and what the values of certain parameters are for every time stamp. These parameters are support, confidence, lift, and the criteria of interestingness measures, described in an earlier chapter.

Another way is the creation of a colourful representation by creating groups of objects. The idea behind this type of representation is as follows. At the first time stamp, each trend's objects represent a different group of objects. Therefore each group of objects consist of objects that have the same values of attributes at the first time stamp. Each group is allocated a unique colour. At the next time stamps, the number of objects of each of the initial groups present at each time stamp is examined. For example, if, initially, there are g different object groups {G_1, G_2, G_3 ... G_g}, and each has the following number of objects {N_1, N_2, N_3 ... N_g}, let us assume that for a rule at the M^{th} time stamp, there are K objects. Then, if any of the groups is a subset of K, the M^{th} square is filled with as many colours as there are numbers of different groups that are subsets. The percentage of each group is coloured with the colour of that group. For example, if K consists of objects of the groups G_1 and G_2, then the square is filled with the colours of groups G_1 and as $\frac{N_1}{K} \times L$ and$\frac{N_2}{K} \times L$, where L is the length of the square. With this allocation technique, colours initially for the group of patients, the user is able to see how objects are moving through rules in every time stamp.

4.3 SOMA: An application of trend mining in diabetic retinopathy

SOMA is the framework for the application of trend mining in the field of diabetic retinopathy. Diabetic retinopathy screening data are collected by The Royal Liverpool University Hospital (RLUH), a major centre for retinopathy research. The nature of the longitudinal data is of interest because it does not fit into any standard categorization of such data, in that the "time stamp" used is the sequential patient-consultation event number. The duration between consultations is also variable (Somaraki et al., 2010). For the temporal pattern identification process, the annual sequence was taken as the "time stamp". The number of screening episodes per patient that have been recorded varies between one and 20, with an average of five consultations. It should also be noted that in some cases, a patient did not complete an annual screening episode (in which case there was no record for that episode), although this did not adversely affect the temporal pattern mining process. In some other cases, the sequence of episodes was terminated because the patient “dropped” out of the screening programme (was referred to the Hospital Eye Service, moved away, or died).

The data associated with a single episode, as also noted above, may actually be recorded over several months. In some cases, it was not clear whether a particular set of data entries belonged to a single episode or not. Some empirical evaluations indicated that the elapsed time between logging the initial screening data and (where appropriate) the results of biomicroscopy were less than 91 days. This was used as a working threshold to identify episode boundaries. For the research described here, a window of 91 days was therefore used to collate data into a single screening episode.

The time lapse between screening episodes is typically 12 months, although the data collection shows a great deal of variation resulting from practical considerations affecting the implementation of the screening programme (illustrated in Figure 4.2). As noted above, according to the nature of the retinopathy, additional episodes may occur, and consequently, more than one consultation can take place per year, in which case the second consultation was ignored.

The initial screening data are stored in the General dataset and the next visit, which concerns information from eyes imaging, are stored in the Photodetails dataset, and data from Biomicroscopy of retina are stored in the Biomicroscopy dataset. In order to combine data from all these datasets and form an episode, the time interval from General to Photodetails, and from Photodetails to Biomicroscopy, should be less than 91 days. To identify the next time stamp, SOMA uses the date from General and checks when the next record in General took place. If the elapsed time from the next record is within 12 ± 6 months, then this record is the start of the next episode (Figure 4.3). All these are part of an internal procedure. SOMA asks the user only which attributes and which datasets will be used.

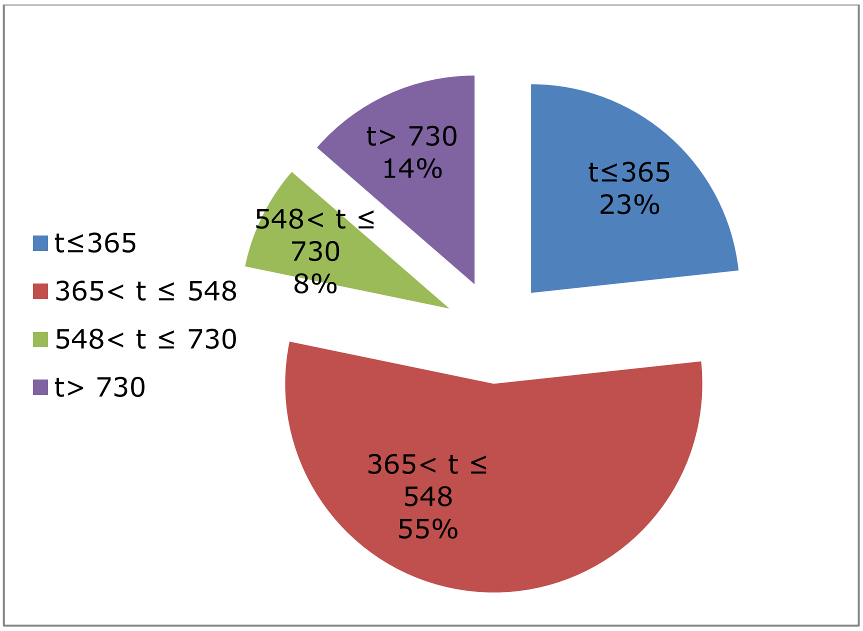


Figure 4.2: Patient distribution based on the time interval from the previous to the next episode in days.

The following figure depicts a time line showing how the episodes are generated. Let's say that a patient P_1 has episodes with G_1,G_2, … ,G_n being the visits which are registered at General dataset with G_1<G_2<…<G_n, Ph_1,Ph_2,… ,Ph_n being the visits which are registered at Photodetails dataset with Ph_1<Ph_2< … <Ph_n and B_1,B_2,…,B_n being the visits which are registered at Biomicroscopy dataset with B_1<B_2<…<B_n .

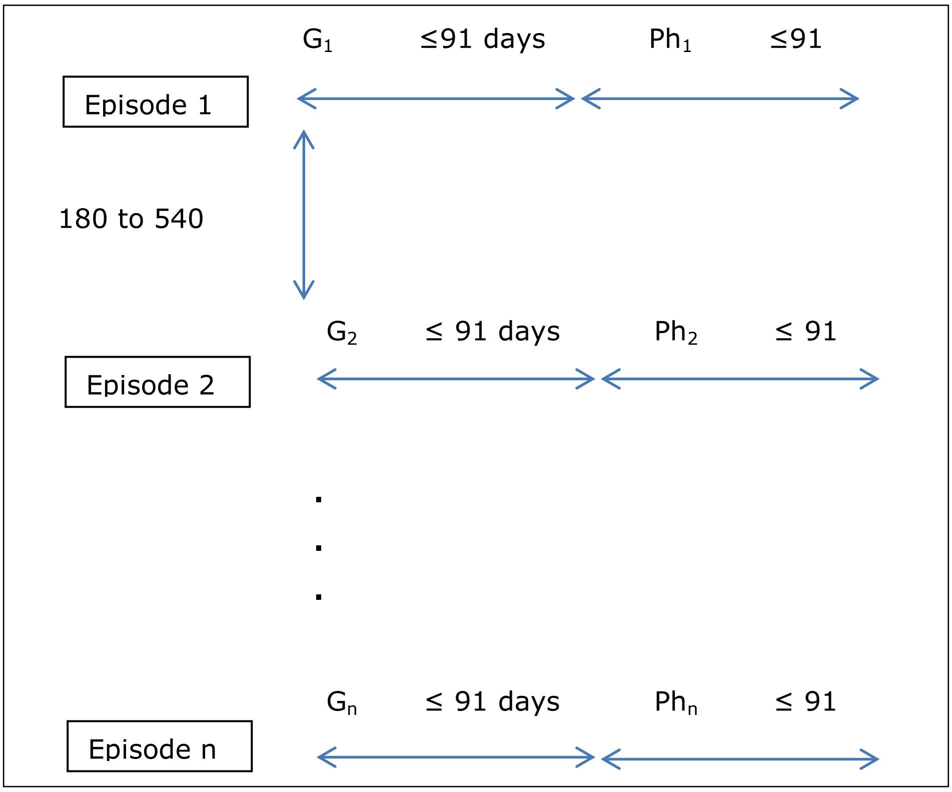


Figure 4.3: Timeline of episodes creation

The above Figure shows the procedure that is follow from the framework to form the episodes and it is repeated for each patients.

The data are stored in three different repositories and, in their raw form, cannot be formed. The pre-processing stage within SOMA is very important because, apart from the form of the episodes, data are transformed in a way in which they can be understood from the parts that follow pre-processing. The following principles are applied:

- Each attribute is allocated a unique number creating an ordered list, e.g.: <Age_at_Exam>::=1, <Visual_Acuity_Right_Best>::=2, <Visual_Acuity_Left_Best>::=3, and so on. The allocation starts from the General dataset, continues to the Photodetails dataset, and ends with the Biomicroscopy dataset. This way of labelling the attributes allows SOMA to understand from which dataset to read information about an attribute.
- The value of each attribute is treated according to its type. The datasets contain attributes that can take both continuous numerical values (e.g. the age of a patient) and discrete values (both numerical and categorical). For discrete values, the processing is straightforward, as each value is assigned a characteristic number. The key issue is to create bands for the continuous values. For each attribute that has continuous values, SOMA creates intervals that cover the range of values. Then, the range of each band is allocated a unique number. For example, the field that provides information for the age of the patients is transformed as follows: <Age_at_Exam>::= <0–12> | <12–20> | <20–30> | <30–40> | <40–50> | <50–60> | <60–70> | < 70 >. After this categorization, SOMA will use the following norm to use this attribute: <1>::=<1>|<2>|<3>|<4>|<5>|<6>|<7>|<8>. On the left-hand side, the number denotes the attribute, and the number on the right-hand side denotes the value of the attribute. The range of each band is very important, because it will affect the frequency of appearance, in every time stamp, and thus determines how frequently a certain band of an attribute appears. For example, let us take the attribute of the age of a patient. The results will be different if the values are separated into two bands, below 50 years old and above 50 years old, or if the age is separated into three bands. In addition, the attributes that characterize the level of diabetic retinopathy for the left and right eye are banded in such a way that they provide information on whether a patient has or does not have diabetic retinopathy, in the left or right eye, or both. Afterwards, these attributes are merged into a new attribute, "diabetic retinopathy", which states whether a patient suffers

from diabetic retinopathy or not. The generation and categorization of trends are both based on the support count of the value of each attribute. Each value of an attribute represents an item, and the combination of items creates item sets; if the values of an attribute are scattered, the number of item sets may become so high that it will be practically impossible to identify useful knowledge, even if the support threshold were reduced to a very low level. This can be easily understood if we take into account the attributes with continuous values such as the age of a patient, which can take any value and, moreover, will change from time stamp to time stamp. In such a case with no transformation process in every time stamp, there will be as many item sets as the number of different ages of all patients, and in order to track them, the support threshold must be less than 1/N (where N is the number of patients).

- Another feature of this arithmetical language is that it allocates each patient a unique number that is universal. A patient who can be found in all datasets is indicated by the same number, and this allows SOMA to identify each individual patient for every time stamp.

When the time-stamped datasets (episodes) are created, the logic rules are applied so as to ensure that the values of attributes are correct and correspond to true medical situations for the present study. Also, the logic rules are applied to replace missing values and thus to reduce the percentage of information that is missing. The logic rules are a set of if clauses; the if clauses state that if a certain combination of values of attributes exists, then the value of a field should take a certain value. Here is an example:

If < Present Treatment > ::=< diet and insulin > && < Calculated age at diagnosis > ::= < ≥ 30 and <40> && < diInsTab > ::=<Don't know> | <Null> && < dbPastTreat > ::= <tablets and the insulin> then <calculated diabetes type> ::= <diabetes type 1>.

This rule implies that if a patient's present treatment is diet and insulin, patient's age when diagnosed with diabetes was between 30 and 40 years old, if the patient doesn't know for how long was taking tablets before insulin then the diabetes type of the patient is diabetes type 1.

Figure 4.4 below depicts how the algorithm searches through the datasets: the algorithm moves horizontally from one dataset to another in order to form an episode. When an episode is confirmed through logic rules, the algorithm moves vertically to the following records.

Following this process, the final outcome is a number of datasets, and each represents one episode. In each dataset, every line represents a patient, and every column represents an attribute.

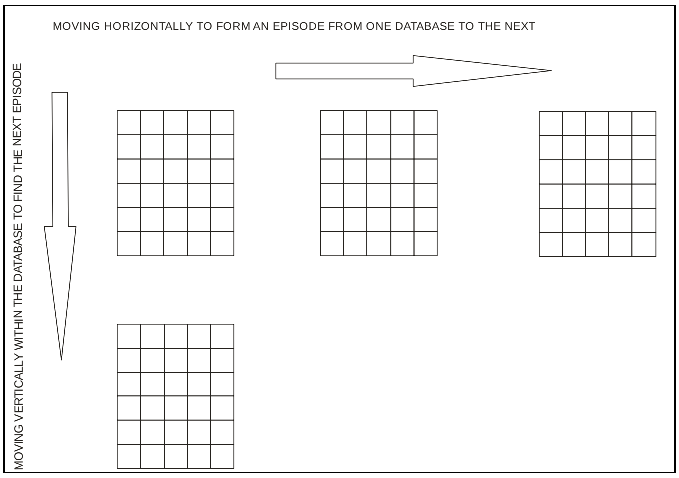

MOVING HORIZONTALLY TO FORM AN EPISODE FROM ONE DATABASE TO THE NEXT
MOVING VERTICALLY WITHIN THE DATABASE TO FIND THE NEXT EPISODE

Figure 4.4: How framework reads data

```
For i=1 to {total number of patients in general dataset}
    set j to i
    set date1 to date of visit of patient i
    set line1 to i
    set line2 to zero
    set line3 to zero
    for jj =1 to {total number of patients in photodetails}
        if jj is i
            if date1 differs from date of visit of patient jj ≤ 91 days
                set line2 to jj
            end
        end
    end
    for jj =1 to {total number of patients in biomicroscopy}
        if jj is i
            if date1 differs from date of visit of patient jj ≤  91days
                set line3 to jj
            end
        end
    end
    for ii=1 to number of attributes selected
        set jj to number of attribute ii
        if jj ≤  {total number of attributes in general}
            set array(ii) to General(line1,jj)
        end
        if jj >   {total number of attributes in general}  and ≤ { number of attributes in photodetails}
            set array(ii) to Photodetails(line2-{number of attributes in general}, jj)
        end
        if jj > {total number of attributes in photodetails} and ≤ {total number of attributes in biomicroscopy}
            set array(ii) to Photodetails(line2-{ total number of attributes in general} – {total number of attributes in   photodetails}, jj)
        end
    end
End
```

At this stage, the user must specify the following parameters:

- which attributes are to be examined;
- the number of time stamps.

At the next stage, Aretaeus[3] software is used, which incorporates the application of rule mining and trend mining. It asks the user to specify the following:

- in the context of association rules, which attributes are the variable attributes (the left side of the rule) and which are the key attributes (the right side of the rule);
- the threshold value for both the support and the confidence.

Next, the algorithm applies the matrix algorithm to identify item sets, which contain only the defined user attributes with support higher than the support threshold, and then checks if the rule has a greater confidence than the threshold, in at least one time stamp. If both thresholds have been satisfied, the implementation algorithm calculates the confidence of the inverse rule and the lift.

[3] Aretaeus (AD 130–200) was the first physician to give diabetes its proper name. In his treatise "On the causes and symptoms of chronic diseases Book II, he used the Greek word "diabetes" (meaning "siphon") to describe the disease. He stated that "Diabetes is a remarkable affection not very frequent among men being a melting down of the flesh and blood into urine." "Mellitus" was added later by others to denote the sweet taste of urine. "Mellitus" means "honey" in Greek.

The final step of the SOMA framework is the creation of a coloured representation showing how the patients are moving through the trends. The support count provides information only about the number of patients who have certain characteristics, and provides no information about who they are. On the other hand, keeping the identities of the patients anonymous is a major task in medical databases. The solution to this problem is to group patients in the first time stamp based on the values of the attributes and allocates each group a unique colour. Each trend is represented as a line with squares, as many as the number of time stamps. If a trend has 0 support count in a time stamp, the square is left white.

At the following time stamps, SOMA examines where the patients are. Let us say that a trend in the 2-second time stamp has support count S2. The following conditions may then apply:

- The patients all belong to the same group and the square will have the same colour as in the first time stamp.
- The patients all belong to the same group, but this group appears for the first time at this time stamp. Then, this group is allocated a new colour and is considered from this point as an extra group.

N1 patients belong to existing group A_1, and N_2 patients belong to existing group A_2, and so on, such that N_1+N_2+...=S_2. In such a case, the patients come from difference groups, and the square is coloured in as many colours as there are different groups. The percentage of each group Ni/S2 determines the area of the square, which is painted in the specific colour. If some of these groups have appeared for the first time, they are considered as new groups and are allocated a unique colour.

The same process is followed for the rest of the time stamp until the algorithm reaches the last one.

4.4 Validation and verification of the Framework

The trend-mining method essentially performs "learning by discovery", and hence we cannot train it; rather, we have to have confidence in the results it provides, that is, it should be validated. To perform a validation of the trend-mining framework, we advocate two complementary approaches:

i) Verification

Verification tests whether the intermediate results and/or outputs of the framework are self-consistent.

ii) Validation

This method tests the outputs of the framework and also checks the consistency of the application that experts already know and expect. The methods include:

- confirmation of the framework that reveals known causal connections in the application;
- confirmation of the framework that reveals known trends in the application.

4.4.1 Verification

Verification attempts to measure the extent to which a set of parameters affects the self-consistence of trend mining framework. In this chapter, several measures are discussed below. In order to perform verification SOMA application is used.

4.4.1.1 Number of time stamps

The first parameter for the framework is the number of time stamps. The size of the dataset is determined by the number of time stamps. Also, another factor that is very important is the time window which determines how data recorded in different times can be collocated into a time stamp.

In SOMA for every patient, each time stamp or episode consists of collated data that have been recorded under different consultations. Here, a window of 91 days is used as the threshold to create an episode. Moreover, to move to the next episode, there is a window of 365 days ± 180 days.

4.4.1.2 Completeness of dataset

This measure refers to the degree of complexity of the datasets and how much information they contain. Even using logic rules at the pre-processing stage, it is not possible to fill all the empty values.

The numerator of the ratio, of the support count of an item set over the total number of transactions of a dataset, that is used to calculate the support value, is the number of occurrence of an item set. The more complete a dataset is, the more information can be extracted from it.

4.4.1.3 Rule conflict

Sometimes, if a dataset is very dense (very large), there is the probability that an item set of "variable attributes" (the antecedent part of the rule) belongs to two or more different "key-variable" values (consequent).

If X = $\{x_1, x_2, ..., x_n\}$ is an item set, the following rules are a set of conflict rules:

$X \rightarrow Y_1$

$X \rightarrow Y_2$

Where Y_1 and Y_2 are item sets of "key attributes".

In such a case, the methodology to deal with this problem will affect the final results of trend mining. One way to deal with the conflict rules is to discard both of them from the results. Another way is to perform a comparison between the conflict rules in terms of support and confidence across all time stamps.

4.4.1.4 Banding of continuous attributes

In the dataset, there are attributes that are continuous variables and must be banded within constant intervals. If an attribute that has continuous values is not banded, each individual value will represent an individual item, and as a result a very low support threshold would be required for this attribute to appear in the final result.

Therefore, the length of the interval of the band affects the results in the context that when the length of a band covers a high percentage of values, its probability of appearing in the trend increases. Let us take an attribute, A, which is a continuous variable and separated into m bands:

B_1, B_2, B_3, ..., B_m

which have the following percentages p_1, p_2, p_3,..., p_m.

In order for all the bands to appear in the final results, the support threshold S_{thres} must follow the following rule:

$S_{thres} \leq \min[p_1, p_2, p_3, ..., p_m]$

The attributes that are time-related have more effect on the results (e.g. age at examination). The reason for this is that the trends are time-dependent. Let's take an attribute that is time-dependent and separated into bands, each of which has length m years within the interval [t, t + m].

The time from previous to next time stamp (episode) varies from 6 (0.5 year) months to 18 months (1.5 year).

Let's take a case where there are N time stamps and take three options:

- the time between time stamps is 0.5 year;
- the time between the time stamps is 1 year; and
- the time between the time stamps is 1.5 years.

In this case:

- the elapsed time T = (N – 1) * .5 years
- T = (N – 1) years
- the elapsed time is T = (N – 1) * 1.5.

Let's assume that the interval [t, t + m] has minimum value t_{min}.

If t_{min} plus the elapsed time T gives:

$t_{min} + T > t + m$

This means that between time stamp 1 and time stamp N, this attribute will change band from [t, t + m] to [t + m, t + 2m]. Therefore, the type and number of each trend can be varied because of the specific length of a band and without necessarily providing useful information about the data.

4.4.1.5 Parameterization

In the trend mining framework, there are four parameters whose values control the effectiveness of the framework:

- support threshold : the minimum support required for an item set;
- confidence threshold: the minimum confidence required for an association rule;
- growth rate : the rate that shows how the support increases across all time stamps;
- tolerance : parameter to determine a constant trend.

All the above parameters are user-specified.

The support threshold and confidence threshold control which rules from a dataset are kept and which are discarded. Also, the support threshold controls the type of trend: if support for a rule is above the threshold at all time stamps, the trend could be increasing, decreasing, or fluctuating. Otherwise it would fall into the category of jumping or disappearing.

The growth-rate threshold determines if the increase in the support of a rule is sufficient to be characterized as an increasing trend. However, tolerance is the threshold at which, when the growth rate is less than the tolerance, the trend is characterized as constant.

4.4.2 Validation

Researchers have found that diabetic patients who are able to maintain appropriate blood sugar levels have fewer eye problems than those with poor control. Diet and exercise play important roles in the overall health of those with diabetes. These are some examples of common-sense clauses that can be justified from the trends. The more trends the SOMA produces (smaller support), the more clauses that can be "exported". The confidence of the trend also plays an important role, as it measures the validity of the trend. The higher the confidence of the role, the more valid it is.

The validation of the entire SOMA framework is based on known associations between the attributes that are selected. Among the attributes that have been selected, there should be at least one that has the role of the "key attribute". In this research, a "key attribute" could be an attribute that characterizes the status of diabetic retinopathy for each patient. The other attributes play the role of variables, "variable attributes" whose values affect the “key attribute”. At the end of the SOMA framework, the rules that are produced are compared with known associations, given by the experts.

4.4.2.1 The SOMA output language

The SOMA output language consists of keywords, which represent the attributes and values that represent a certain time interval or a certain situation such as the type of diabetes or the type of treatment. This output language is used by SOMA to process input data to produce the final results. The following table reveals the keywords and their associative values that SOMA uses to create output rules.

Table 4.2: SOMA output language

Attribute	SOMA keyword	SOMA value	Interpretation
Age at the date of exam	Age_at_Exam	1	<12
		2	12 – 20 years old
		3	20 – 30 years old
		4	30 – 40 years old
		5	40 – 50 years old
		6	50 – 60 years old
		7	60 – 70 years old
		8	>70 years old
Age at diagnosis	calculated_age_at_diagnosis	1	<12
		2	12 – 20 years old
		3	20 – 30 years old
		4	30 – 40 years old
		5	40 – 50 years old
		6	50 – 60 years old
		7	60 – 70 years old
		8	>70 years old
Treatment of diabetes	Present_Treatment	1	Diet alone
		2	Diet and tablet
		3	Diet and insulin
		4	Tablets and insulin
Type of diabetes	calculated_diabetes_type	1	Type 1
		2	Type 2 diet controlled
		3	Type 2 oral controlled
		4	Type 2 insulin required
Duration of diabetes	calculated_diabetes_duration	1	<5 years
		2	5-10 years
		3	10 – 15 years
		4	15 – 20 years
		5	>20 years

4.4.2.2 How SOMA reveals knowledge

The output of SOMA framework is a combination of textual, numerical and graphical representation. For every association rule which meets the conditions of support threshold and confidence threshold SOMA provides the association rule in text format. It outputs the rule as pairs of attributes and the associative value, and with an arrow the left-hand side of the rule is separated from the right-hand side of the rule. The following is an example of what the textual output looks like:

```
Age_at_Exam=8  Present_Treatment=4  calculated_age_at_diagnosis=7
calculated_diabetes_type=4  calculated_diabetes_duration=3   --> DR=0
```

The SOMA output is written in the same way as for reading input data. The values of the attributes represent either an interval, e.g. "Age_at_Exam = 8", meaning that the age of a patient at the examination is higher than 70 years old, or a certain condition, e.g. "calculated_diabetes_type = 4", meaning that the type of diabetes of a patient is of type 2, and insulin needs to be taken by the patient.

SOMA after the associative classification process outputs in textual format information about the kind of trends depending on how the support of a rules varies across all time stamps. The kinds of trends the SOMA produces are:

- Increasing
- Decreasing
- Jumping
- Disappearing
- Constant
- Fluctuating

The numerical output of SOMA concerns the support value of the rule, the minimum and maximum confidence, and the lift for each time stamp. Moreover, it outputs the growth rate of a trend if it is increasing.

In particularly the output of the framework gives information about the rule: the support count of $X \rightarrow Y$ for each time stamp, the kind of trend , the maximum and minimum confidence of the rule $X \rightarrow Y$ and finally the lift of the rule. The conclusion from this output is that although that the rule has high confidence, the fact that lift is less than one lead s to the conclusion that the attributes of X are negatively correlated with the attributes of Y.

```
Have the following support counts:
1   5

This trend is increasing with growth rate 5.00

This rule has P(Y|X) :
 maximum confidence : 71.4286 %
 minimum confidence : 25.0000 %

This rule lift is :
   0.30007     0.94446
```

If, in any time stamp, the support count is less than the support threshold, the confidence, lift, and support are set to 0.

Another form of textual output that SOMA produces is how patients move from rule to rule at each time stamp. Here is an example:

```
Age_at_Exam=8  Present_Treatment=4  calculated_age_at_diagnosis=7
calculated_diabetes_type=4  calculated_diabetes_duration=3   --> DR=0

 From the 1 timestamp to the 2 timestamp there are more 4 patients

From the previous time stamp continue 0 patients

 PATIENT # at this time stamp came from the trend :
Age_at_Exam=8  Present_Treatment=2  calculated_age_at_diagnosis=7
calculated_diabetes_type=3  calculated_diabetes_duration=3   --> DR=0
 PATIENT # at this time stamp came from the trend :
Age_at_Exam=8  Present_Treatment=2  calculated_age_at_diagnosis=7
calculated_diabetes_type=3  calculated_diabetes_duration=3   --> DR=0
 PATIENT # at this time stamp came from the trend :
Age_at_Exam=8  Present_Treatment=4  calculated_age_at_diagnosis=7
calculated_diabetes_type=4  calculated_diabetes_duration=3   --> DR=1
```

In the example above, this rule has one patient in the first time stamp and five patients at the second. The end user is informed that four new patients were found to have the characteristics of this rule. From the previous time stamp, no patient continues to exist, which means that this patient has a missing value, and so it is ignored by SOMA. Also, three patients out of five in time stamp 1 had different characteristics, which are also given in textual form. The # symbol is used to protect the identity of the patients. This type of output can help spot changes in the values of the attributes from one time stamp to the next.

In cases where a rule has a high support count in every time stamp, the textual output is so large that it is very difficult to check the movement of the patients from time stamp to time stamp. To solve this problem, SOMA creates a graphical representation, using the following procedure:

- Each trend is allocated a sequential number, 1, 2, 3, etc.
- A mesh is created with square boxes of equal size P × T where P is the number of trends, and T is the number of time stamps. Each line represents a trend; on the left-hand side of the mesh, the trend is written, and on the right-hand side of the mesh, the type of trend is written.
- For the first time stamp in each trend, the number of patients in that trend forms an individual group, each of which is allocated a unique colour, except white, which is allocated when the support count is zero, and black, which is allocated to a patient who appears for the first time in a time stamp different than the first.
- At the following time stamps, SOMA examines the proportion of patients in relation to the groups formed in the first time stamp, so the box may contain more than one colour depending on how many different groups of patient can be found.

SOMA has reserved two colours for special cases:

- white: when a time stamp has no patient;
- black: when in a time stamp, a patient appears but does not belong to any of the initial groups; this happens when patients have missing values at certain time stamps, and so SOMA ignores them.

4.5 Conclusion

This chapter gives a general description of the trend mining framework and also gives details of the application of the framework in Diabetic retinopathy, SOMA. The implementation of SOMA consists of 3 stages: pre-processing for cleaning the datasets and creating episodes, the processing stage using the matrix algorithm and the trend generation and the post processing which regards the categorization of trends and the implementation of visualization technique for the representation of trends. Finally, this chapter provides a framework for the verification and validation of the Trend Mining. Verification is based on checking how parameters of the framework affect how to produce the results right and on the other hand validation aims to check whether the results are right.

Chapter 5 Evaluation

5.1 Introduction

This chapter describes the evaluation of research work described in the thesis which is centred on trend mining. The aim is to validate and verify the approach for the development of the advocated trend mining framework. The goal of evaluation process described here is to judge the usefulness of the discovered knowledge and the process of trend mining itself. On the one hand the evaluation of the produced rules is straightforward by using criteria evaluating novelty action ability unexpectedness reliability etc, on the other hand evaluating the processes of the framework is based on quantitative criteria which measure its performance.

Trend mining validation and verification are aimed at identifying whether the trend mining framework:

- can determine whether the intermediate results and/or outputs of the framework are self-consistent
- produces result in the context of the domain that the data describe

The evaluation by applying the framework to the DR data, examines if the validation and verification are effective as part of the framework. The evaluation is based on an approach embodied in the SOMA framework using the diabetic retinopathy (DR) data. This approach consists of three different directions:

- evaluation of the verification techniques within the pre-processing stages;
- evaluation of the verification techniques within the processing stages;
- evaluation of the validation techniques used to determine the quality of the discovered knowledge.

For each of those directions, a set of criteria is created. The results from the evaluation are measured against the specified criteria, and a scoring system is implemented in an attempt to measure how successful or not the evaluation is.

The chapter provides:

- details on the evaluation setup, e.g. data used in the parameter setup and the type of output;
- details on the criteria for measuring the results and what we want to discover;
- details on outputs of the evaluation;
- discussion about the results of the evaluation.

5.2 Verification experiments

This section describes experiments in the context of testing how parameters given by the user such as the number of time stamps, support threshold, confidence threshold, growth rate threshold, and tolerance affect the amount of knowledge that is discovered as well as the running time, and how the number of variables affects the performance. Another set of experiments concerns how the intervals of bands of time variables affect the model.
Table 5.1 below shows the size of each temporal dataset in number of patients vs the number of time stamps/episodes.

Table 5.1: Number of patients per number of episodes

Number of episodes	Number of patients
2	10968
3	6037
4	3696
5	2328
6	1420
7	887
8	546
9	329
10	172

The number of episodes and thus the size of the datasets determine the value of the support threshold that is required in each case to extract useful information for analysis. The size of the data set is the denominator of the ratio that determines the support value of the item set.
The following tables show the conditions of the experiments. The 1st column shows the threshold values of support, confidence, growth rate, tolerance and the number of variables. The first experiments for the verification of the trend mining framework examine how the input parameters from the user affect the performance of the framework in the context of the total number of trends and the running time that is required in order to discover those trends. The parameters concerning these experiments are:

- Support threshold
- Confidence threshold
- Growth rate threshold

- Tolerance
- Number of variables

For the experiments a dual core Pentium Intel D processor was used with clock 3.0 GHz and the algorithm was developed in MATLAB® 2009a.

Tables 5.2 – 5.4 show the values of the parameters that are selected from experiment A to Y. in all experiments the same variables are used, but in experiments V, W, X and Y variables which are time related have been modified in the context of the intervals that are used. Each of the experiments from A to Y is repeated for several time stamps. Table 5.1 shows the number of patients for each case.

Tables below {5.5 – 5.28} contain the results from those experiments. There are as many tables as the different conditions. In each table the 1st column refers the kind of trends and the rest columns show the results for each time stamps. The last line shows the elapsed time for each time for each time excluding the moments when the user interacts with the algorithm in order to enter the parameters.

From the experiments can be concluded that:

- For the same conditions, in all experiments, as the number of the time stamp increases the elapsed time decreases because the size of datasets decreases.
- For the same conditions, in all experiments, the number of trends, from the experiments with 2 time stamps to the experiments of 10 time stamps, becomes at least double which indicates that as the number of time stamps increases the datasets contains more information.
- Comparing the experiments with the same parameters and variables but with the altered time intervals, experiment Q vs experiment Y, experiment G vs experiment X, experiment U vs experiment W and experiment T vs experiments V, it can be seen that in the experiments with the altered time intervals the total number of trends decreases and as the result the elapsed time decreases. In the experiments with the altered intervals Y,X,W and V the time related variables are split in only 3 intervals while in the Q,G,U and T experiments have been used 8 intervals.

Table 5.2: Experimental conditions

	A	B	C	D	E	F	G	H	I
Support threshold	0.05	0.05	0.05	0.01	0.01	0.01	0.005	0.005	0.005
Confidence threshold	50	70	90	50	70	90	50	70	90
Growth rate Threshold	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01
Tolerance	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
Number of variables	6	6	6	6	6	6	6	6	6

Table 5.3: Experimental conditions

	K	L	M	N	O	P	Q	R
Support threshold	0.05	0.05	0.05	0.01	0.01	0.01	0.005	0.005
Confidence threshold	50	70	90	50	70	90	50	70
Growth rate Threshold	1.005	1.005	1.005	1.005	1.005	1.005	1.005	1.005
Tolerance	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
Number of variables	6	6	6	6	6	6	6	6

Table 5.4: Experimental conditions

	S	T	U	V	W	X	Y
Support threshold	0.005	0.005	0.005	0.005	0.005	0.005	0.005
Confidence threshold	90	50	50	50	50	50	50
Growth rate Threshold	1.005	1.01	1.005	1.01	1.005	1.01	1.005
Tolerance	0.0001	0.001	0.0001	0.001	0.0001	0.001	0.0001
Number of variables	6	8	8	8^*	8^*	6^*	6^*

The superscript (*) denotes that the bands have been altered. For each experiment, a table has been produced that shows the number of trends and the elapsed time.

Table 5.5– Experiment A

	Experiment A								
	Time stamps								
	2	3	4	5	6	7	8	9	10
Increasing	0	0	0	0	0	0	0	0	0
Decreasing	1	2	0	0	0	0	0	0	0
Fluctuating	0	2	0	0	0	0	0	3	6
Constant	0	0	0	0	0	0	0	0	0
Jumping	0	0	2	2	2	2	4	2	2
Disappearing	3	1	3	3	3	4	4	3	3
Total	4	5	5	5	5	6	8	8	11
Elapsed time (sec)	423.36	153.24	243.34	182.56	132.60	102.96	85.99	67.91	63.38

Table 5.6 – Experiment B

	Experiment B								
	Time stamps								
	2	3	4	5	6	7	8	9	10
Increasing	0	0	0	0	0	0	0	0	0
Decreasing	1	0	0	0	0	0	0	0	0
Fluctuating	0	0	0	0	0	0	0	3	6
Constant	0	0	0	0	0	0	0	0	0
Jumping	0	0	2	2	2	2	3	2	2
Disappearing	3	3	3	3	3	4	4	3	3
Total	4	3	5	5	5	6	7	8	11
Elapsed time (sec)	425.55	329.48	254.19	194.34	143.30	107.65	88.38	71.92	61.17

Table 5.7 – Experiment C

	Experiment C								
	Time stamps								
	2	3	4	5	6	7	8	9	10
Increasing	0	0	0	0	0	0	0	0	0
Decreasing	0	0	0	0	0	0	0	0	0
Fluctuating	0	0	0	0	0	0	0	2	6
Constant	0	0	0	0	0	0	0	0	0
Jumping	0	0	0	0	0	0	1	0	0
Disappearing	2	3	3	3	3	4	4	3	3
Total	2	3	3	3	3	4	5	5	9
Elapsed time (sec)	429.90	325.56	251.61	190.70	141.87	106.73	87.05	71.21	60.87

Table 5.8– Experiment D

	Experiment D								
	Time stamps								
	2	3	4	5	6	7	8	9	10
Increasing	14	7	5	1	3	1	0	0	0
Decreasing	16	8	4	0	0	0	0	0	0
Fluctuating	0	3	5	9	7	13	12	16	40
Constant	0	0	0	0	0	0	0	0	0
Jumping	1	2	5	5	6	6	8	12	15
Disappearing	0	2	6	12	14	13	12	14	13
Total	31	22	25	27	30	33	32	42	68
Elapsed time (sec)	428.44	341.27	262.60	199.98	150.38	115.44	93.06	79.35	67.24

Table 2.9 - Experiment E

	Experiment E								
	Time stamps								
	2	3	4	5	6	7	8	9	10
Increasing	6	7	5	1	3	1	0	0	0
Decreasing	10	8	4	0	0	0	0	0	0
Fluctuating	0	3	5	9	7	13	12	16	40
Constant	0	0	0	0	0	0	0	0	0
Jumping	3	2	5	5	6	6	8	12	15
Disappearing	2	2	6	12	14	13	12	14	13
Total	21	22	25	27	30	33	32	42	68
Elapsed time (sec)	431.93	334.25	261.32	199.201	150.38	115.44	93.06	79.35	67.24

Table 5.10 - Experiment F

	Experiment F								
	Time stamps								
	2	3	4	5	6	7	8	9	10
Increasing	0	1	0	0	0	0	0	0	0
Decreasing	5	8	4	0	0	0	0	0	0
Fluctuating	0	1	5	5	6	9	11	14	35
Constant	0	0	0	0	0	0	0	0	0
Jumping	0	0	1	1	2	2	4	6	14
Disappearing	2	2	6	12	14	13	12	14	13
Total	7	12	16	18	22	24	27	34	62
Elapsed time (sec)	426.19	330.24	259.41	196.54	147.45	110.75	92.25	75.74	65.97

Table 5.11 - Experiment G

	Experiment G								
	Time stamps								
	2	3	4	5	6	7	8	9	10
Increasing	14	13	8	3	3	1	0	0	0
Decreasing	16	10	5	4	0	0	0	0	0
Fluctuating	0	7	14	18	13	17	36	59	111
Constant	0	0	0	0	0	0	0	0	0
Jumping	1	3	4	9	10	12	14	28	48
Disappearing	0	2	6	6	16	19	22	18	26
Total	31	35	37	40	42	49	72	105	185
Elapsed time (sec)	439.08	341.19	261.59	202.35	155.24	117.52	96.75	82.19	74.85

Table 5.12 - Experiment H

	Experiment H								
	Time stamps								
	2	3	4	5	6	7	8	9	10
Increasing	13	13	8	3	3	1	0	0	0
Decreasing	16	10	5	4	0	0	0	0	0
Fluctuating	0	7	14	18	13	17	33	50	103
Constant	0	0	0	0	0	0	0	0	0
Jumping	1	3	4	8	10	10	13	21	44
Disappearing	0	2	6	6	16	19	22	18	25
Total	30	35	37	39	42	47	68	89	172
Elapsed time (sec)	433.00	335.08	265.20	203.21	153.67	117.41	98.11	81.90	74.07

Table 5.13 - Experiment I

	Experiment I								
	Time stamps								
	2	3	4	5	6	7	8	9	10
Increasing	3	1	0	1	0	0	0	0	0
Decreasing	8	10	5	4	0	0	0	0	0
Fluctuating	0	4	11	14	11	13	29	48	103
Constant	0	0	0	0	0	0	0	0	0
Jumping	0	2	3	2	3	6	9	19	43
Disappearing	0	2	6	6	16	18	21	18	25
Total	11	19	25	27	30	37	59	85	171
Elapsed time (sec)	429.41	331.61	258.32	198.70	150.22	114.81	97.66	81.27	74.09

Table 5.14 - Experiment K

	Experiment K								
	Time stamps								
	2	3	4	5	6	7	8	9	10
Increasing	0	0	0	0	0	0	0	0	0
Decreasing	1	0	0	0	0	0	0	0	0
Fluctuating	0	0	0	0	0	0	0	3	6
Constant	0	0	0	0	0	0	0	0	0
Jumping	0	0	2	2	2	2	4	2	2
Disappearing	3	3	3	3	3	4	4	3	3
Total	4	3	5	5	5	6	8	8	11
Elapsed time (sec)	430.21	330.40	254.54	192.67	145.01	102.54	87.07	63.08	54.19

Table 5.15 - Experiment L

	Experiment L								
	Time stamps								
	2	3	4	5	6	7	8	9	10
Increasing	0	0	0	0	0	0	0	0	0
Decreasing	1	0	0	0	0	0	0	0	0
Fluctuating	0	0	0	0	0	0	0	3	6
Constant	0	0	0	0	0	0	0	0	0
Jumping	0	0	2	2	2	2	3	2	2
Disappearing	3	3	3	3	3	4	4	3	3
Total	4	3	5	5	5	6	7	8	11
Elapsed time (sec)	422.71	324.75	249.59	185.48	143.21	106.78	87.30	71.60	60.28

Table 5.16 - Experiment M

	Experiment M								
	Time stamps								
	2	3	4	5	6	7	8	9	10
Increasing	0	0	0	0	0	0	0	0	0
Decreasing	0	0	0	0	0	0	0	0	0
Fluctuating	0	0	0	0	0	0	0	2	6
Constant	0	0	0	0	0	0	0	0	0
Jumping	0	0	0	0	0	0	1	0	0
Disappearing	2	3	3	3	3	4	4	3	3
Total	2	3	3	3	3	4	5	5	9
Elapsed time (sec)	414.26	317.90	244.35	186.27	135.85	97.24	80.88	63.06	56.62

Table 5.17 - Experiment N

	Experiment N								
	Time stamps								
	2	3	4	5	6	7	8	9	10
Increasing	6	7	5	1	3	1	0	0	6
Decreasing	10	8	4	0	0	0	0	0	0
Fluctuating	0	3	5	9	7	13	12	16	40
Constant	0	0	0	0	0	0	0	0	0
Jumping	3	2	5	5	6	6	8	12	15
Disappearing	2	2	6	12	14	13	12	14	13
Total	21	22	25	27	30	33	32	42	68
Elapsed time (sec)	417.96	320.11	249.78	186.77	140.12	103.96	88.19	69.50	63.50

Table 5.18 – Experiment O

	Experiment O								
	Time stamps								
	2	3	4	5	6	7	8	9	10
Increasing	6	7	5	1	3	1	0	0	0
Decreasing	10	8	4	0	0	0	0	0	0
Fluctuating	0	3	5	9	7	13	12	15	35
Constant	0	0	0	0	0	0	0	0	0
Jumping	3	2	5	5	6	6	8	11	15
Disappearing	2	2	6	12	14	13	12	14	13
Total	21	22	25	27	30	33	32	40	63
Elapsed time (sec)	438.04	338.09	267.29	203.23	151.90	116.02	92.83	76.96	64.89

Table 5.19- Experiment P

	Experiment P								
	Time stamps								
	2	3	4	5	6	7	8	9	10
Increasing	0	1	0	0	0	0	0	0	0
Decreasing	5	8	4	0	0	0	0	0	0
Fluctuating	0	1	5	5	6	9	11	14	35
Constant	0	0	0	0	0	0	0	0	0
Jumping	0	0	1	1	2	2	4	6	14
Disappearing	2	2	6	12	14	13	12	14	13
Total	7	12	16	18	22	24	27	34	62
Elapsed time (sec)	411.23	312.95	244.01	183.36	136.13	100.86	83.43	69.23	62.70

Table 5.20 - Experiment Q

	Experiment Q								
	Time stamps								
	2	3	4	5	6	7	8	9	10
Increasing	14	13	8	3	3	1	0	0	0
Decreasing	16	10	5	4	0	0	0	0	0
Fluctuating	0	7	14	18	13	17	36	59	111
Constant	0	0	0	0	0	0	0	0	0
Jumping	1	3	4	9	10	12	14	28	48
Disappearing	0	2	6	6	16	19	22	18	26
Total	31	35	37	40	42	49	72	105	185
Elapsed time (sec)	416.64	327.64	254.87	193.39	145.55	112.94	94.17	80.68	73.10

Table 5.21 - Experiment R

	Experiment R								
	Time stamps								
	2	3	4	5	6	7	8	9	10
Increasing	13	13	8	3	3	1	0	0	0
Decreasing	16	10	5	4	0	0	0	0	0
Fluctuating	0	7	14	18	13	17	33	50	103
Constant	0	0	0	0	0	0	0	0	0
Jumping	1	3	4	8	10	10	13	21	44
Disappearing	0	2	6	6	16	19	22	18	25
Total	30	35	37	39	42	47	68	89	172
Elapsed time (sec)	438.26	339.79	263.63	201.72	152.23	114.63	96.19	83.43	65.92

Table 5.22 - Experiment S

	Experiment S								
	Time stamps								
	2	3	4	5	6	7	8	9	10
Increasing	3	1	0	1	0	0	0	0	0
Decreasing	8	10	5	4	0	0	0	0	0
Fluctuating	0	4	11	14	11	13	29	48	103
Constant	0	0	0	0	0	0	0	0	0
Jumping	0	2	3	2	3	6	9	19	43
Disappearing	0	2	6	6	16	18	21	18	25
Total	11	19	25	27	30	37	59	85	171
Elapsed time (sec)	418.65	322.70	256.24	194.03	146.95	111.03	94.29	78.48	73.59

Table 5.23 - Experiment T

	Experiment T								
	Time stamps								
	2	3	4	5	6	7	8	9	10
Increasing	15	12	10	3	1	0	0	0	0
Decreasing	17	10	6	0	0	0	0	0	0
Fluctuating	0	6	9	16	17	26	52	71	171
Constant	1	0	0	0	0	0	0	0	0
Jumping	1	5	7	11	14	19	24	40	65
Disappearing	2	6	11	18	20	16	4	0	0
Total	36	39	43	48	52	61	80	111	236
Elapsed time (sec)	749.27	553.18	411.33	304.19	211.47	150.47	113.77	86.59	78.99

Table 5.24 – Experiment U

	Experiment U								
	Time stamps								
	2	3	4	5	6	7	8	9	10
Increasing	15	12	10	3	1	0	0	0	0
Decreasing	17	10	6	0	0	0	0	0	0
Fluctuating	0	6	9	16	17	26	52	71	171
Constant	1	0	0	0	0	0	0	0	0
Jumping	1	5	7	11	14	19	24	40	65
Disappearing	2	6	11	18	20	16	4	0	0
Total	36	39	43	48	52	61	80	111	236
Elapsed time (sec)	744.30	530.30	397.75	291.78	200.72	162.36	106.39	82.20	70.76

Table 5.25 - Experiment V

	Experiment V								
	Time stamps								
	2	3	4	5	6	7	8	9	10
Increasing	6	4	2	1	1	0	0	0	0
Decreasing	10	8	5	3	0	0	0	0	0
Fluctuating	0	5	10	10	15	20	24	29	76
Constant	0	0	0	0	0	0	0	0	0
Jumping	2	2	4	7	10	11	17	29	41
Disappearing	0	0	1	2	4	2	1	0	0
Total	18	19	22	23	30	33	42	58	117
Elapsed time (sec)	324.66	251.45	193.62	156.15	121.90	102.09	81.38	71.84	64.60

Table 5.26 – Experiment W

	Experiment W								
	Time stamps								
	2	3	4	5	6	7	8	9	10
Increasing	6	4	2	1	1	0	0	0	0
Decreasing	10	8	5	3	0	0	0	0	0
Fluctuating	0	5	10	10	15	20	24	29	76
Constant	0	0	0	0	0	0	0	0	0
Jumping	2	2	4	7	10	11	17	29	41
Disappearing	0	0	1	2	4	2	1	0	0
Total	18	19	22	23	30	33	42	58	117
Elapsed time (sec)	329.62	248.21	194.49	156.76	123.64	96.19	77.34	67.73	58.89

Table 5.27– Experiment X

	Experiment X								
	Time stamps								
	2	3	4	5	6	7	8	9	10
Increasing	4	3	4	4	1	0	0	0	0
Decreasing	10	6	4	3	3	3	2	1	1
Fluctuating	0	8	8	10	15	14	26	29	53
Constant	0	0	0	0	0	0	0	0	0
Jumping	1	1	3	6	8	12	10	15	17
Disappearing	2	0	1	1	1	1	4	5	5
Total	17	18	20	24	28	30	42	50	76
Elapsed time (sec)	194.29	158.08	127.14	105.11	87.16	73.74	67.89	63.42	56.85

Table 5.28 - Experiment Y

	Experiment Y								
	Time stamps								
	2	3	4	5	6	7	8	9	10
Increasing	4	3	4	4	1	0	0	0	0
Decreasing	10	6	4	3	3	3	2	1	1
Fluctuating	0	8	8	10	15	14	26	29	53
Constant	0	0	0	0	0	0	0	0	0
Jumping	1	1	3	6	8	12	10	15	17
Disappearing	2	0	1	1	1	1	4	5	5
Total	17	18	20	24	28	30	42	50	76
Elapsed time (sec)	195.11	157.76	133.04	111.13	92.55	78.21	70.53	63.59	57.03

From the tables with the results (5.5 – 5.28) the following conclusions can be extracted:

- for the same support threshold, as the confidence threshold increases, the number of rules that are discovered decreases;
- for the same confidence threshold, as the support threshold decreases, the number of rules that are discovered increases;
- for the same values of support and confidence threshold, the change in growth rate threshold and tolerance does not affect the amount of rules that are discovered;

- for the same confidence and support threshold and number of variables, as the number of time stamps increases, the elapsed time decreases;
- for the same number of time stamps and variables, the elapsed time does not change significantly;
- for the same confidence and support threshold and number of time stamps, the elapsed time is increased as the number of time stamps increases;
- the increase in time-band intervals results in a significant decrease in the number of rules and the elapsed time, under the same conditions.

5.3 Validation of the trend mining framework

The proposed technique for validation of the trend mining framework based on using known relations between the attributes that describe the domain. That way the user can provide the attributes that describe the left and the right hand side of the rule. The validation of the entire SOMA framework is based on known associations between the attributes that are selected.

5.3.1 Validation experiments

Here, some rules are provided to show how attributes are related to diabetic retinopathy disease and how these rules are validated by SOMA:

- Rule 1: If the duration of diabetes is longer than 15 years, it is very likely that this patient will suffer from diabetic retinopathy.
- Rule 2: If the type of diabetes is type 1, the patient is likely to suffer from diabetic retinopathy.
- Rule 3: If a patient is over 60 years old, has type 2 diabetes, and changes the treatment from insulin and oral control to insulin and tablets, this patient is likely to develop diabetic retinopathy.
- Rule 4: If the age of a patient becomes greater than 70 years old, then it is quite likely that this patient will develop diabetic retinopathy.

For rule 1, the end user should check for rules that include the attribute that SOMA recognizes as diabetes duration, “calculated_diabetes_duration”, and has the value 4 or 5, which means 15–20 years and more than 20 years, respectively. For rule 2, SOMA should reveal rules that include the right type of diabetes and present treatment values, as it is known that patients who have diabetes type 1 require to take insulin, and so SOMA should also reveals this causality.

The SOMA outputs the following rules:

```
Age_at_Exam=6  Present_Treatment=2  calculated_age_at_diagnosis=4
calculated_diabetes_type=3  calculated_diabetes_duration=4   --> DR=1
Age_at_Exam=4  Present_Treatment=3  calculated_age_at_diagnosis=2
calculated_diabetes_type=1  calculated_diabetes_duration=5   --> DR=1
Age_at_Exam=4  Present_Treatment=3  calculated_age_at_diagnosis=3
calculated_diabetes_type=1  calculated_diabetes_duration=4   --> DR=1
Age_at_Exam=4  Present_Treatment=3  calculated_age_at_diagnosis=2
calculated_diabetes_type=1  calculated_diabetes_duration=5   --> DR=1
Age_at_Exam=5  Present_Treatment=3  calculated_age_at_diagnosis=3
calculated_diabetes_type=1  calculated_diabetes_duration=5   --> DR=1
```

All of the above rules confirm that patients have a diabetes duration of 15–20 years (in SOMA output language calculated_diabetes_duration = 4) or longer than 20 years (calculated_diabetes_duration = 5). In all rules above, both supports and confidence are greater than the threshold, but what is really important are the high values of the lift.
Lift is the ratio of the confidence of the rule over the support count of the item set appearing in the rule consequent. Sometimes, the value of the confidence of the rule may be misleading because the calculation of confidence ignores the support count of the item set appearing in the rule consequent. If lift is greater than 1, the item set appearing in the rule antecedent and the item set appearing in the rule consequent are positively correlated; if it is less than 1, they are negatively correlated; and if it is 1, they are independent.

For all the above rules for all time stamps, lift is greater than 2, which means that there is a high positive correlation between the left- and the right-hand side of the rules. Also, from rules II–V, it is confirmed that in the case of a patient who suffers from diabetes type 1, their treatment contains insulin (Present_treatment = 3).
Before showing how SOMA confirms rules 3 and rule 4, let us first explain what a trend-mining algorithm should output to conclude that rules 3 and 4 are validated by SOMA. Rule 3 describes a situation where patients have two characteristics of their condition stable, i) age older than 60 years old and ii) diabetes type 2, but the third characteristic which is treatment of diabetes changes from insulin and oral control to insulin and tablets. This change in the diabetes treatment results in a change in the diabetic retinopathy condition, from not having it to an increased likelihood of developing diabetic retinopathy. This abrupt change in the treatment of diabetes can be described with a jumping trend because the final

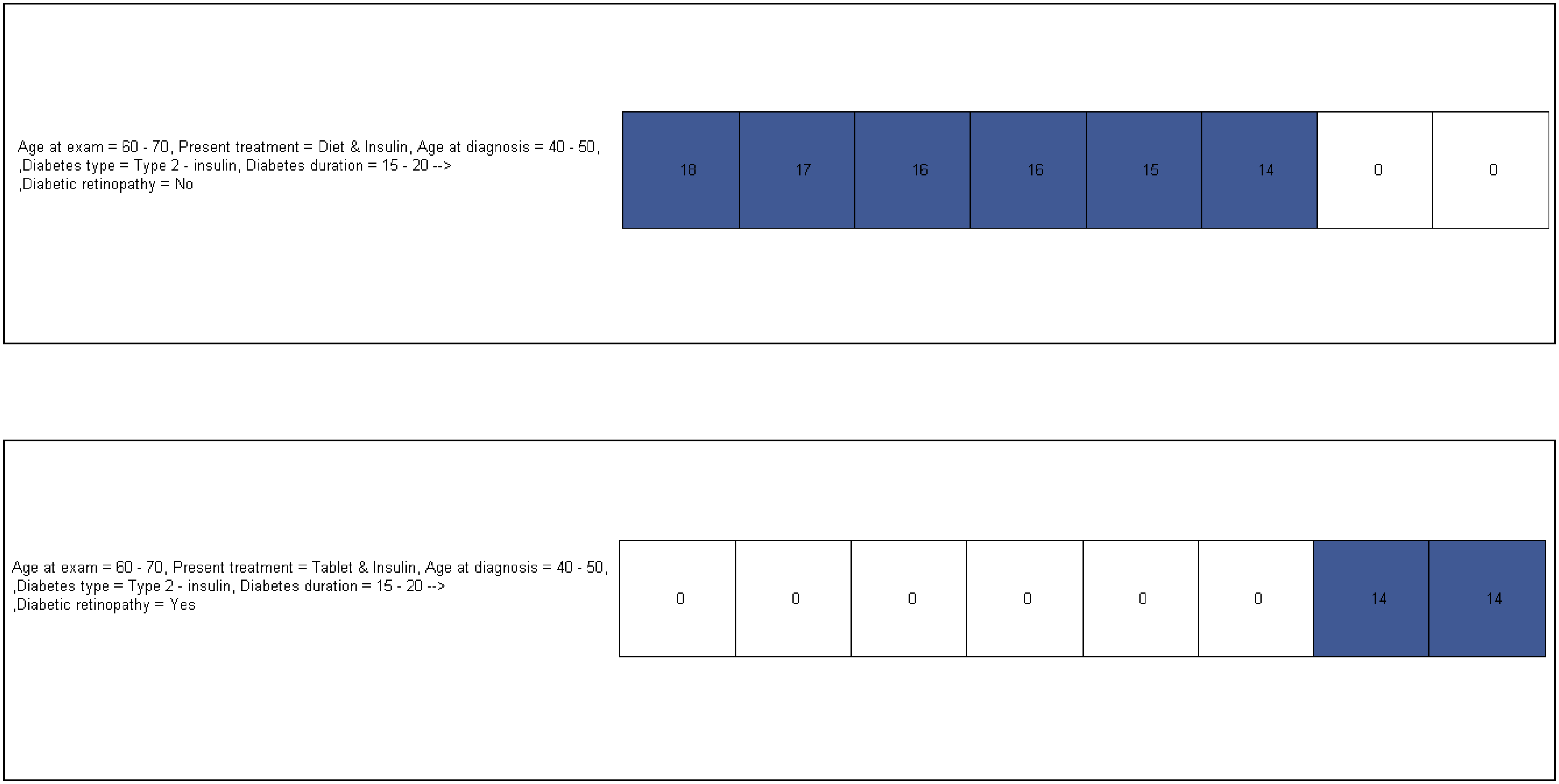
condition of those patients initially does not exist but jumps sometimes. Thus, patients move from a trend describing the initial condition to a trend describing the final condition. Rule 4 describes a situation that shows how the change in the age of patients affects their condition of diabetic retinopathy. When the age of a patient increases above 70, the value of the attribute age_at_exam (age at the time of the hospital visit) changes from 6 (age between 60 and 70 years old) to 7 (above 70 years old). However, as time passes, the age of patients increases, and as a result there will be patients whose age will go to the next age interval.

In SOMA, the rule 3 can be validated with the following pair for trends:

```
Age_at_Exam=7 Present_Treatment=3 calculated_age_at_diagnosis=5
calculated_diabetes_type=4 calculated_diabetes_duration=4   --> DR=0
```

```
Age_at_Exam=7  Present_Treatment=4  calculated_age_at_diagnosis=5
calculated_diabetes_type=4  calculated_diabetes_duration=4  --> DR=1
```

This pair of rules confirms that for patients aged 60–70 years old (Age_at_Exam = 7), their treatment changes from insulin and diet control (calculated_diabetes_type = 3) to insulin and tablet control (calculated_diabetes_type = 4).These patients from the disappearing trend appear to the jumping trend. This change can be depicted using the following coloured schema 5-1. The same background colour in the boxes denotes that the patients at each time stamp have begun from the same association, which is written in a format that the end user can understand and not in the language that SOMA uses to perform its tasks. At time stamp 7, the first association disappears and jumps the second one with 14 patients from the original 18 having diabetic retinopathy and a change in their treatment.



Age at exam = 60 - 70, Present treatment = Diet & Insulin, Age at diagnosis = 40 - 50, ,Diabetes type = Type 2 - insulin, Diabetes duration = 15 - 20 --> ,Diabetic retinopathy = No	18	17	16	16	15	14	0	0

Age at exam = 60 - 70, Present treatment = Tablet & Insulin, Age at diagnosis = 40 - 50, ,Diabetes type = Type 2 - insulin, Diabetes duration = 15 - 20 --> ,Diabetic retinopathy = Yes	0	0	0	0	0	0	14	14

Figure 5.1: Pair of Disappearing and Jumping trends

As regards rule 4 in SOMA language the following pair of decreasing and increasing trends can be used for validation:

```
Age_at_Exam=7 Present_Treatment=4 calculated_age_at_diagnosis=4
calculated_diabetes_type=4 calculated_diabetes_duration=5   --> DR=0
```

```
Age_at_Exam=8 Present_Treatment=4 calculated_age_at_diagnosis=4
calculated_diabetes_type=4 calculated_diabetes_duration=5   --> DR=1
```

The first trend is decreasing because from the previous time stamp to the next, the number of the patients’ decreases and for the same reason the second trend is increasing because the patients from the former are added to the initial number of patients. From the following coloured representation, this is clear because the second trend has two colours: the first shows the number of patients who continue from the first time stamp and the second shows the number of patients coming from the first trend. In total, the number of patients increases, which means that it is increasing trend. These two trends validated rule 4, which states that patients older than 70 years of age are likely to develop diabetic retinopathy.

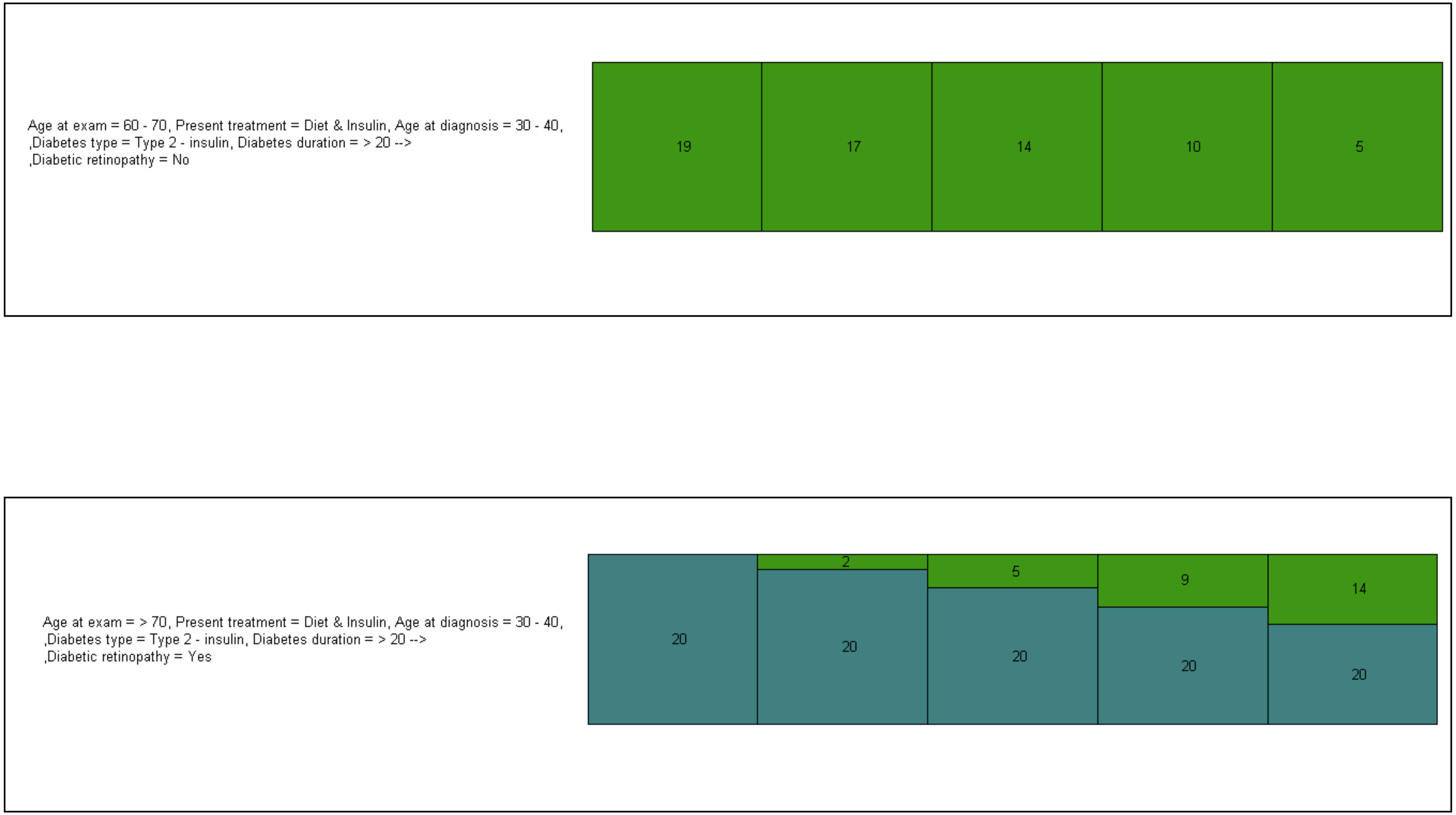


Figure 5.2: Pair of decreasing and increasing trends

5.4 Evaluation of the trend mining framework

In this research work, a set of Validation and Verification criteria was used to ensure the quality of the discovered knowledge. To evaluate this, a case study is necessary to incorporate trend mining within it. The SOMA framework is used to show the effectiveness of trend mining with DR data, and this effectiveness is measured using a set of criteria incorporating various aspects of the quality of knowledge discovery and by showing that the results score well with respect to those criteria.

The evaluation has three main aims: verification of the pre-processing; evaluation of the verification of processing; and evaluation of the validation. The evaluation of the verification of the pre-processing considers issues related to the pre-processing tasks before trend mining algorithm is applied:

- noise reduction;
- object distribution;
- Time-stamp duration.

The evaluation of the verification of main processing considers how interesting the rules are from the association mining point of view, and how changes in trends are related to objects. The final aim, evaluation of the validation, is related to the quality of the knowledge discovered. The concept behind the evaluation is to create a scoring system for each of the criteria. For each criterion, a set of targets is established, and for each target, a score is allocated. After checking the framework against all criteria, the final score is summed up; a higher score indicates a better performance and greater effectiveness of the framework.

5.4.1 Criteria discussion

In this section, each criterion is explained and linked with the application of SOMA, and a discussion is provided about what the criterion aims to achieve and how this is associated with sets of targets:

- Noise reduction: Raw data from databases are noisy with missing values, error values, etc. During the first stage of trend mining, data undergo cleansing using domain-related logic rules. The lower the percentage of missing values in the time-stamped datasets, the more likely information can be extracted from them. The aim is to achieve the greatest possible reduction in noise, expressed as a percentage by comparing the time-stamp datasets before and after use. SOMA postulates the use of specific logic rules, which have been set up from the physicians and determine the values of the attributes used.
- Object distribution: This criterion evaluates at each time stamp whether an object has enough information in order to be included or not into the trend mining process. The term enough is interpreted as completeness. If an object has no missing information has enough information. The evaluation wants to establish if this happens in all time stamps .Each line of a dataset refers to a certain object. For any number of time-stamped datasets, every line refers to the same object. If, in any line, there is a missing value, this object is omitted from the framework processes after pre-processing. The aim of this criterion is to check the percentage of objects that are not ignored across all datasets. In this case study, the objects are the patient from the DR databases.
- Time-stamp duration and interval: The time-stamp length and the interval between two consecutive time stamps criterion is related to both the process of the trend-mining framework in terms of whether data are longitudinal or not, but also related to the application of the domain. As explained previously, a time stamp may consist of a single event or may be a combination of multiple events associated with an object. In the case of the latter, one must ensure that the time interval between multiple events and the time interval between time stamps are mostly uniform. In DR databases, different visits to medics and screeners are recorded in which different types of data are collected. From one dataset to the next, the interval must be no more than 91 days so as to produce a time stamp for a patient. Additionally, for the same patient, the interval from one time stamp to the next must be 12±6 months. The target is related to the uniformity of the length of the time stamp and of the interval from one time stamp to another for the same patient.

- Interestingness of information: One of the major tasks of trend mining is to identify which trend is interesting or not. As described in chapter 4 a trend containing information of the support of a rule $X \to Y$ is calculated. If, in any time stamp, the confidence of the rule $X \to Y$ is greater than the threshold, the trend is accepted as interesting; otherwise, it is discarded. However, the use of confidence only is not adequate, and so more measures of interestingness are applied: lift, max confidence all confidence, cosine, and Kulczynski. These calculations are carried out for each time stamp. The aim is for every measure to exceed a threshold value for each time stamp. The more measures are achieved, and for more time stamps, the more interesting the rule.
- Change-point detection: The change-point detection criterion evaluates the framework in the context of the ability of the model to identify meaningful changes from one condition to another for the same group of objects. The aim is to measure how many meaningful change points are detected. SOMA examines whether the change in DR status can be depicted on the disappearing–jumping trends concerning the same group of patients.
- Quality of discovered knowledge: This criterion is simply concerned with the application domain of the trend mining. A set of rules describing a condition are used to check the quality of the knowledge. These kinds of rules have the form of an if-clause using the antecedent and the consequent parts of the discovered rules and describe a relation between them. The criterion here is to check whether trend mining can confirm those rules and how accurately that can be done. The confirmation is measured by examining if the framework can give either the same rule or the opposite rule and the accuracy is measured by examining how many attributes are included into the discovered rule, both antecedent and consequent. Here, the rules used concern how and whether certain characteristics affect DR.

5.4.2 Criteria implementation

This section describes how the evaluation criteria are implemented in order to measure the effectiveness and performance of the framework:

- **First criterion**: the interestingness of the information.

The main filter of the vast amount of information is the confidence threshold. If an association rule has a confidence greater than or equal to a threshold in any time stamp, then it is considered by the framework to be interesting information. However, the use of confidence only is not an adequate measure for the interestingness, and so the framework calculates other measures, too: lift, max confidence, all confidence, cosine, and Kulczynski.

If an association rule has a high lift (more than 1), and the measures are greater than the threshold given for confidence, then it is assumed that this association rule is valid. More particularly, for each association rule, a 5×N matrix is created where N denotes the time stamps, where the association rule has non-zero support, and the figure lines denote: lift, max confidence, all confidence, Kulczynski, and cosine measures. If the rule has exceeded the threshold in terms of the measures, it is assigned 1; otherwise 0. The maximum score that a rule can have across all time stamps is 5×N. If S is the total score of a rule across all time stamps, the interestingness of the rule is given by the following ratio:

$$\frac{S}{5 \times N} \times 100\%$$

- **Second criterion: noise reduction**

For each time stamp, the number of missing values before and after the pre-processing is measured. If the reduction in the number of missing values is above a threshold (40%) and is preserved across all datasets, then the aim has been achieved, and the score is incremented by one or not.

- **Third criterion: time-stamp length and interval**

This criterion examines whether the length of the time stamp and its interval from the next are within a given range i) for a patient across all time stamps and ii) across the records within the same dataset. If the percentage of the time stamps that lie within the given range is above a given value, the score is accredited.

- **Fourth criterion: discovery of change points**

At the end of the framework, there is a colourful representation where a group of patients is formed, based on the patients' characteristics at the first time stamp. Following each group of patients, from its unique colour, we are able to track the patients. We are particularly interested in tracking change points. A change point is defined as follows: at any time stamp T, a group of patients G is moving from rule R_1 at time stamp T–1 to rule R_2 at time stamp T.

This change point can indicate just a change in the characteristics of the group, for example, a change in any variable attribute (left-hand side of the rule), or it may indicate that any change in left-hand side of the rule, they initially belong, may result in a change in "key attribute".

Similarly with the second criterion, the score is accredited when a group of patients from a decreasing trend moves to an increasing trend. The higher the score, the better the system works.

- **Fifth criterion: object distribution**

For all time-stamped data set, each line refers to the same patient. If the line has no missing values across all datasets, it is considered a valid record. The total number of such records is calculated as a percentage against the number of all records. If the percentage exceeds a threshold value given by the user, the score is accredited.

- **Sixth criterion: Quality of discovered knowledge**

Knowledge (relations, rules, characterizations) is harvested from the experts in the application that they already know about, and that they expect the system should discover, given the data collected. The application of this criterion to the case study involved acquiring medical criteria such as:

- If a patient suffers from cataract, it is possible to develop DR.
- If a patient is diagnosed with diabetes in young age, they have an increased risk of suffering from DR.
- If a patient suffers from type 1 diabetes, they are more likely to develop DR.
- If a patient suffers from type 2 diabetes and is on insulin, they are more likely to develop DR.
- If a patient suffers from type 2 diabetes for more than 20 years, they are more likely to develop DR.
- If a patient suffers from type 1 diabetes for more than 5 years, they are more likely to develop DR.

The greater the numbers of criteria from the above that are confirmed by an interesting trend, the higher the score the system obtains. Each time a criterion is met, the score is incremented by 1.

At the end, the final score is calculated, and this is checked against the maximum score that could be measured in the framework evaluation.

5.4.3 Experimental set-up

Experiments were conducted using the DR data described in chapter 4, to evaluate the SOMA framework. In particular, the data combine information from the General and Photodetails datasets. The attributes used were as follows:

- age at exam
- present treatment of patient
- diabetes type
- diabetes duration
- age at diagnosis
- DR in the left eye
- DR in the right eye
- DR.

The last attribute is not originally included in the datasets and is synthesized by merging the attributes regarding the DR on either the left or right eye. The reason for this is that we are interested in the situation of a patient, not which eye has the disease.

The first seven attributes from the list were chosen for two reasons:

- These attributes are less noisy after the application of logic rules.
- The medics believe that these attributes are very important in terms of trying to link DR to certain characteristics.

The attributes have continuous and discrete values. Continuous values are the values from attributes that show some relation to time (age and duration). For those attributes (age at exam, age at diagnosis, and diabetes duration), the values are converted to discrete values using the following bands:

For age at exam and age at diagnosis, the following:

- 0 – 12 years old → Band 1 → Value =1
- 12 – 20 years old → Band 2 → Value =2
- 20 – 30 years old → Band 3 → Value =3
- 30 – 40 years old → Band 4 → Value =4
- 40 – 50 years old → Band 5 → Value =5
- 50 – 60 years old → Band 6 → Value =6
- 60 – 70 years old → Band 7 → Value =7
- > 70 years old → Band 8 → Value =8

For diabetes duration, the following transformation was used:

- 0 – 5 years → Band 1 → Value =1
- 5 – 10 years → Band 2 → Value =2

- 10 – 15 years → Band 3 → Value =3
- 15 – 20 years → Band 4 → Value =4
- > 20 years → Band 5 → Value =5

The experiment is repeated with 5 to 10 time stamps to examine short-term and long-term changes.

The following table shows the parameters used for all experiments:

Table 5.29: Experimental Parameters

Parameters	Value
Support threshold	0.01 %
Confidence threshold	40 %
Growth rate threshold	1.01
Tolerance	0.00001

The experiments were conducted using all patients who complied with the time-stamp rules: i) irrespective of their DR status; ii) examining those who have developed DR in every time stamp; and iii) those who have not developed DR in all time stamps. The reason for this is to examine not only how manipulation of the dataset may affect the results but also how evaluation measures are affected by that kind of manipulation.

However, a different approach is used to evaluate the quality of discovered knowledge. Based on the rules used for the medical criteria, different experiments were conducted using only the attributes described in the rules, using all patients, and using a stricter confidence threshold.

5.4.4 Experimental results and evaluation

This section presents the results from the experiments with the aim to evaluate the trend-mining framework and SOMA in particular. As stated earlier, the datasets were manipulated either to contain a certain class (condition) of the disease or to pick up all patients. Under those conditions, Table 5.30 shows the sizes of the datasets used.

Table 5.30: Dataset size

	Series 1	Series 2	Series 3
Time stamps	ALL patients	Patients with DR in all time stamps	Patients with no DR in all time stamps
5	2328	51	1303
6	1420	25	715
7	887	10	411
8	546	7	231
9	329	5	135
10	179	2	75

It can be seen that the amount of information available regarding patients with DR is very small compared with that regarding non-DR patients or those who have developed DR at a certain stage of their life. Hereafter, the experiments concerning all patients will be referred to as series 1, experiments for patients with DR in all time stamps as series 2, and the other category as series 3.

5.4.4.1 Noise reduction

Tables 5.31– 5.33 show the noise reduction as a percentage and also present the scores. This percentage concerns the values that have undergone cleansing successfully over the total number of values of a dataset in every time stamp.
The algorithm for each time stamped dataset, calculates the number of cells which potentially can undergo the cleaning process. After the cleaning process the algorithm counts the number of cells whose values have been changed from the cleaning process. The ratio of the later number of cell over the former number of cells gives the percentage of the cleaning process.
Tables 5.31 – 5.33 show the results. Each column in those tables represents the results of each experiment and each line represents the time stamp. The threshold for those experiments was set to 40% for each experiment and if the noise reduction is above that percentage the score for each experiment increases by one. For each experiment the maximum score is equal to the number of time stamps. The last two lines of the tables show the score and the maximum score for each experiment respectively. The higher score means more successful noise reduction.

It can be seen that the use of a threshold of 40% noise reduction is successful in series 1 and 3 in attaining the maximum score. The results for series 2 range from 0% to 90%

(in terms of scoring), but if another threshold had been used, e.g. 48%, the score for all series would be 0. The reason for that is that the knowledge of experts which was used is not enough to replace the missing attributes with a value. It would be more helpful to replace missing values using another method, for example, using the mean value, but even so, such a method would have to change the support of the antecedent and consequently all measures of interestingness.

Table 5.31: Noise reduction (%) for series 1

	5 time stamps	6 time stamps	7 time stamps	8 time stamps	9 time stamps	10 time stamps
1st	42.6730	42.6962	42.6156	42.6156	42.8137	42.9370
2nd	42.6055	42.6660	42.7444	42.7444	42.6835	42.8571
3rd	42.6362	42.6660	42.7444	42.7444	42.6400	42.6975
4th	42.6178	42.6559	42.6961	42.6961	42.7703	42.7773
5th	40.9732	42.6358	42.6800	42.6800	42.9006	42.7773
6th	N/A	41.0664	42.6317	42.6317	42.7269	42.9370
7th	N/A	N/A	41.1016	41.1016	42.8137	42.6975
8th	N/A	N/A	N/A	42.6156	42.7269	42.8571
9th	N/A	N/A	N/A	N/A	41.2505	42.6975
10th	N/A	N/A	N/A	N/A	N/A	41.8994
Score	5	6	7	8	9	10
Max score	5	6	7	8	9	10

Table 5.32: Noise reduction (%) for series 2

	5 time stamps	6 time stamps	7 time stamps	8 time stamps	9 time stamps	10 time stamps
1st	41.1765	38.8571	40.0000	38.7755	40.0000	42.8571
2nd	40.0560	40.0000	37.1429	38.7755	40.0000	42.8571
3rd	39.7759	40.0000	40.0000	34.6939	40.0000	42.8571
4th	40.6162	38.8571	40.0000	38.7755	34.2857	42.8571
5th	38.6555	40.0000	38.5714	38.7755	40.0000	42.8571
6th	N/A	40.0000	38.5714	36.7347	40.0000	42.8571
7th	N/A	N/A	40.0000	36.7347	37.1429	42.8571
8th	N/A	N/A	N/A	38.7755	37.1429	42.8571
9th	N/A	N/A	N/A	N/A	40.0000	35.7143
10th	N/A	N/A	N/A	N/A	N/A	42.8571
Score	3	4	4	0	6	9
Max score	5	6	7	8	9	10

Table 5.33: Noise reduction (%) for series 3

	5 time stamps	6 time stamps	7 time stamps	8 time stamps	9 time stamps	10 time stamps
1st	42.7804	42.8571	42.7876	42.9190	42.8571	42.8571
2nd	42.7585	42.7772	42.8571	42.7953	42.8571	42.8571
3rd	42.7804	42.8372	42.7876	42.9190	42.8571	42.8571
4th	42.8133	42.8172	42.8224	42.9190	42.8571	42.8571
5th	42.7365	42.8172	42.7876	42.8571	42.8571	42.8571
6th	N/A	42.7772	42.8224	42.9190	42.7513	42.8571
7th	N/A	N/A	42.8224	42.9190	42.8571	42.6667
8th	N/A	N/A	N/A	42.9190	42.8571	42.8571
9th	N/A	N/A	N/A	N/A	42.8571	42.8571
10th	N/A	N/A	N/A	N/A	N/A	42.8571
Score	5	6	7	8	9	10
Max score	5	6	7	8	9	10

Appendix 2 shows an analytical representation of the plots for noise reduction. Those plots are colour bars and each bar shows the percentage of values that are corrected for every time stamp for all experiments.

5.4.4.2 Object distribution

This measure aims to evaluate the extent to which all objects are used or not. If an object has a missing value, it is omitted from the framework. Tables 5.34-5.36 show the object distribution. At the start of an experiment the user enters the number of time stamps. Depending on the number of the time stamps there is a certain number of objects, O. Aafterwards the algorithm checks how many of the objects O have information in all time stamps. Let's say that there's one experiment with 5 time stamps. Each object, after cleaning must have values in every time stamp, if not then this object is ignored.

Table 5.6 shows the number of objects O for reach experiment, for each series. Each number in that table is the denominator for the calculation of the ratio for the object distribution.

The SOMA then for each time stamp calculates the ratio of the objects that have values in all time stamps over the number O.

Here, the threshold that is used is 90 %. Tables 5.34 – 5.36 show the distribution of objects. Again, experiments from series 1 and 3 have better scores than the experiments in series 2. This was an expected finding, since experiments from series 2 have the smallest scores for noise reduction.

Table 5.34: Patient distribution (%) for series 1

	5 time stamps	6 time stamps	7 time stamps	8 time stamps	9 time stamps	10 time stamps
1st	93.0842	93.8732	91.4318	88.0952	85.4103	74.3017
2nd	93.9433	93.0282	92.3337	90.1099	84.8024	81.0056
3rd	94.1581	94.4366	93.9121	93.0403	92.0973	86.5922
4th	93.9433	93.5915	93.4611	93.5897	93.3131	90.5028
5th	92.6546	93.1690	92.5592	91.9414	92.4012	92.7374
6th	N/A	92.6056	92.3337	92.6740	91.4894	91.0615
7th	N/A	N/A	91.6573	92.6740	92.0973	89.9441
8th	N/A	N/A	N/A	90.4762	91.1854	91.0615
9th	N/A	N/A	N/A	N/A	89.0578	88.8268
10th	N/A	N/A	N/A	N/A	N/A	88.8268
Score	5	6	7	7	6	4
Max score	5	6	7	8	9	10

Table 5.35: Patient distribution (%) for series 2

	5 time stamps	6 time stamps	7 time stamps	8 time stamps	9 time stamps	10 time stamps
1st	84.3137	84	80	85.7143	85.4103	74.3017
2nd	84.3137	88	80	85.7143	84.8024	81.0056
3rd	88.2353	88	90	85.7143	92.0973	86.5922
4th	84.3137	92	80	100.0000	93.3131	90.5028
5th	90.1961	92	80	85.7143	92.4012	92.7374
6th	N/A	92	90	85.7143	91.4894	91.0615
7th	N/A	N/A	100	85.7143	92.0973	89.9441
8th	N/A	N/A	N/A	100.0000	91.1854	91.0615
9th	N/A	N/A	N/A	N/A	89.0578	88.8268
10th	N/A	N/A	N/A	N/A	N/A	88.8268
Score	1	3	4	2	6	4
Max score	5	6	7	8	9	10

Table 5.36: Patient distribution (%) for series 3

	5 time stamps	6 time stamps	7 time stamps	8 time stamps	9 time stamps	10 time stamps
1st	94.4743	95.5245	93.6740	89.6104	83.7037	84.0000
2nd	96.3162	95.9441	94.4039	92.2078	85.9259	82.6667
3rd	95.6255	98.0420	95.8637	94.8052	94.0741	86.6667
4th	96.3929	95.8042	96.3504	95.2381	95.5556	93.3333
5th	95.3952	96.9231	95.1338	95.2381	94.0741	97.3333
6th	N/A	96.0839	95.6204	94.8052	97.0370	93.3333
7th	N/A	N/A	95.3771	95.6710	94.8148	97.3333
8th	N/A	N/A	N/A	93.9394	95.5556	97.3333
9th	N/A	N/A	N/A	N/A	93.3333	94.6667
10th	N/A	N/A	N/A	N/A	N/A	93.3333
Score	5	6	7	7	7	7
Max score	5	6	7	8	9	10

5.4.4.3 Time-stamp distribution

To evaluate the time-stamp uniformity, three rules are used:

- Rule I .For each object, the average time interval and the standard deviation are calculated.
- Rule II. For each time stamp, the average time interval and the standard deviation for all objects are calculated.
- Rule III. The maximum time interval from General to Photodetails, 91 days, is divided into three intervals: 0–29 days, 30–60 days, and more than 60 days. For each time stamp, the percentage of patients belonging to that interval is calculated. The criterion is to check the uniformity of the sum of the percentage of patients who go from the General to Photodetails within 60 days, and the average and standard deviation are then calculated for each time stamp.

For all three rules, the ratio of the standard deviation over the average should not exceed a threshold value, set here to 0.15. The aim of this rule is to examine whether the objects need the time interval to combine information from various sources.

Table 5.37: Patient distribution in intervals for every time stamp, from General to Photodetails – Series 1

	< 30 days	30 – 60 days	> 60 days	Sum of two first intervals
1st time stamp	76.29 %	20.75 %	2.96 %	97.04 %
2nd time stamp	66.88 %	30.93 %	2.19 %	97.81 %
3rd time stamp	39.6 %	58.55 %	1.85 %	98.15 %
4th time stamp	46.56 %	50.64 %	2.8 %	97.2 %
5th time stamp	29.47 %	48.5 %	22.03 %	77.97 %

Table 5.38: Patient distribution in intervals for every time stamp, from General to Photodetails – Series 2

	< 30 days	30– 60 days	> 60 days	Sum of two first intervals
1st time stamp	76.290%	20.750%	2.960%	97.040%
2nd time stamp	66.880%	30.930%	2.190%	97.810%
3rd time stamp	39.600%	58.550%	1.850%	98.150%
4th time stamp	46.560%	50.640%	2.800%	97.200%
5th time stamp	29.470%	48.500%	22.030%	77.970%

Table 5.39: Patient distribution in intervals for every time stamp, from General to Photodetails – Series 3

	< 30 days	30– 60 days	> 60 days	Sum of two first intervals
1st time stamp	78.740%	18.110%	3.150%	96.850%
2nd time stamp	68.150%	30.010%	1.840%	98.160%
3rd time stamp	38.070%	59.940%	1.990%	98.010%
4th time stamp	43.510%	53.340%	3.150%	96.850%
5th time stamp	28.550%	49.730%	21.720%	78.280%

Table 5.40: Patient distribution in intervals for every time stamp, from General to Photodetails – Series 1

	< 30 days	30– 60 days	> 60 days	Sum of two first intervals
1st time stamp	68.800%	27.460%	3.740%	96.260%
2nd time stamp	77.250%	20.420%	2.330%	97.670%
3rd time stamp	69.930%	28.590%	1.480%	98.520%
4th time stamp	40.990%	57.820%	1.190%	98.810%
5th time stamp	45.350%	52.180%	2.470%	97.530%
6th time stamp	26.970%	51.200%	21.830%	78.170%

Table 5.41: Patient distribution in intervals for every time stamp, from General to Photodetails – Series 2

	< 30 days	30– 60 days	> 60 days	Sum of two first intervals
1st time stamp	60.000%	36.000%	4.000%	96.000%
2nd time stamp	64.000%	36.000%	0.000%	100.000%
3rd time stamp	72.000%	28.000%	0.000%	100.000%
4th time stamp	48.000%	52.000%	0.000%	100.000%
5th time stamp	40.000%	56.000%	4.000%	96.000%
6th time stamp	44.000%	52.000%	4.000%	96.000%

Table 5.42: Patient distribution in intervals for every time stamp, from General to Photodetails – Series 3

	< 30 days	30– 60 days	> 60 days	Sum of two first intervals
1st time stamp	68.800%	27.460%	3.740%	96.260%
2nd time stamp	77.250%	20.420%	2.330%	97.670%
3rd time stamp	69.930%	28.590%	1.480%	98.520%
4th time stamp	40.990%	57.820%	1.190%	98.810%
5th time stamp	45.350%	52.980%	1.670%	98.330%
6th time stamp	26.970%	51.200%	21.830%	78.170%

Table 5.43: Patient distribution in intervals for every time stamp, from General to Photodetails – Series 1

	< 30 days	30– 60 days	> 60 days	Sum of two first intervals
1st time stamp	60.880%	34.610%	4.510%	95.490%
2nd time stamp	70.350%	26.270%	3.380%	96.620%
3rd time stamp	79.590%	18.830%	1.580%	98.420%
4th time stamp	71.590%	27.960%	0.450%	99.550%
5th time stamp	38.780%	60.770%	0.450%	99.550%
6th time stamp	42.390%	54.900%	2.710%	97.290%
7th time stamp	23.110%	53.440%	23.450%	76.550%

Table 5.44: Patient distribution in intervals for every time stamp, from General to Photodetails – Series 2

	< 30 days	30– 60 days	> 60 days	Sum of two first intervals
1st time stamp	70.000%	10.000%	20.000%	80.000%
2nd time stamp	60.000%	40.000%	0.000%	100.000%
3rd time stamp	70.000%	30.000%	0.000%	100.000%
4th time stamp	80.000%	20.000%	0.000%	100.000%
5th time stamp	50.000%	50.000%	0.000%	100.000%
6th time stamp	30.000%	60.000%	10.000%	90.000%
7th time stamp	50.000%	50.000%	0.000%	100.000%

Table 5.45: Patient distribution in intervals for every time stamp, from General to Photodetails – Series 3

	< 30 days	30– 60 days	> 60 days	Sum of two first intervals
1st time stamp	63.260%	33.820%	2.920%	97.080%
2nd time stamp	72.510%	24.820%	2.670%	97.330%
3rd time stamp	82.970%	15.330%	1.700%	98.300%
4th time stamp	73.480%	26.280%	0.240%	99.760%
5th time stamp	37.710%	61.800%	0.490%	99.510%
6th time stamp	39.420%	57.420%	3.160%	96.840%
7th time stamp	21.900%	53.770%	24.330%	75.670%

Table 5.46: Patient distribution in intervals for every time stamp, from General to Photodetails – Series 1

	< 30 days	30– 60 days	> 60 days	Sum of two first intervals
1st time stamp	45.790%	45.050%	9.160%	90.840%
2nd time stamp	63.190%	34.070%	2.740%	97.260%
3rd time stamp	74.180%	23.990%	1.830%	98.170%
4th time stamp	83.330%	15.750%	0.920%	99.080%
5th time stamp	72.340%	27.290%	0.370%	99.630%
6th time stamp	32.780%	66.850%	0.370%	99.630%
7th time stamp	38.460%	59.160%	2.380%	97.620%
8th time stamp	22.160%	52.010%	25.830%	74.170%

Table 5.47: Patient distribution in intervals for every time stamp, from General to Photodetails – Series 2

	< 30 days	30– 60 days	> 60 days	Sum of two first intervals
1st time stamp	57.470%	14.290%	28.240%	71.760%
2nd time stamp	85.710%	0.000%	14.290%	85.710%
3rd time stamp	42.860%	57.140%	0.000%	100.000%
4th time stamp	57.140%	42.860%	0.000%	100.000%
5th time stamp	85.710%	14.290%	0.000%	100.000%
6th time stamp	57.140%	42.860%	0.000%	100.000%
7th time stamp	28.570%	71.430%	0.000%	100.000%
8th time stamp	42.860%	57.140%	0.000%	100.000%

Table 5.48: Patient distribution in intervals for every time stamp, from General to Photodetails – Series 3

	< 30 days	30– 60 days	> 60 days	Sum of two first intervals
1st time stamp	48.920%	42.290%	8.790%	91.210%
2nd time stamp	64.940%	32.030%	3.030%	96.970%
3rd time stamp	74.460%	23.380%	2.160%	97.840%
4th time stamp	84.850%	13.850%	1.300%	98.700%
5th time stamp	74.030%	25.970%	0.000%	100.000%
6th time stamp	33.330%	66.670%	0.000%	100.000%
7th time stamp	36.360%	60.170%	3.470%	96.530%
8th time stamp	21.650%	51.520%	26.830%	73.170%

Table 5.49: Patient distribution in intervals for every time stamp, from General to Photodetails – Series 1

	< 30 days	30– 60 days	> 60 days	Sum of two first intervals
1st time stamp	38.300%	38.300%	23.400%	76.600%
2nd time stamp	45.590%	45.590%	8.820%	91.180%
3rd time stamp	65.350%	33.430%	1.220%	98.780%
4th time stamp	75.380%	23.400%	1.220%	98.780%
5th time stamp	84.800%	15.200%	0.000%	100.000%
6th time stamp	73.560%	26.140%	0.300%	99.700%
7th time stamp	28.270%	71.430%	0.300%	99.700%
8th time stamp	34.040%	52.310%	13.650%	86.350%
9th time stamp	22.190%	52.280%	25.530%	74.470%

Table 5.50: Patient distribution in intervals for every time stamp, from General to Photodetails – Series 2

	< 30 days	30– 60 days	> 60 days	Sum of two first intervals
1st time stamp	60.000%	0.000%	40.000%	60.000%
2nd time stamp	60.000%	40.000%	0.000%	100.000%
3rd time stamp	100.000%	0.000%	0.000%	100.000%
4th time stamp	60.000%	40.000%	0.000%	100.000%
5th time stamp	40.000%	60.000%	0.000%	100.000%
6th time stamp	80.000%	20.000%	0.000%	100.000%
7th time stamp	40.000%	60.000%	0.000%	100.000%
8th time stamp	20.000%	80.000%	0.000%	100.000%
9th time stamp	60.000%	40.000%	0.000%	100.000%

Table 5.51: Patient distribution in intervals for every time stamp, from General to Photodetails – Series 3

	< 30 days	30– 60 days	> 60 days	Sum of two first intervals
1st time stamp	37.780%	37.780%	24.440%	75.560%
2nd time stamp	48.150%	42.220%	9.630%	90.370%
3rd time stamp	66.670%	31.850%	1.480%	98.520%
4th time stamp	77.040%	20.740%	2.220%	97.780%
5th time stamp	87.410%	12.590%	0.000%	100.000%
6th time stamp	74.810%	25.190%	0.000%	100.000%
7th time stamp	26.670%	73.330%	0.000%	100.000%
8th time stamp	31.110%	63.700%	5.190%	94.810%
9th time stamp	22.220%	49.630%	28.150%	71.850%

Table 5.52: Patient distribution in intervals for every time stamp, from General to Photodetails – Series 1

	< 30 days	30– 60 days	> 60 days	Sum of two first intervals
1st time stamp	46.930%	32.400%	20.670%	79.330%
2nd time stamp	38.550%	32.960%	28.490%	71.510%
3rd time stamp	49.720%	45.810%	4.470%	95.530%
4th time stamp	64.800%	34.640%	0.560%	99.440%
5th time stamp	78.210%	20.670%	1.120%	98.880%
6th time stamp	84.360%	15.640%	0.000%	100.000%
7th time stamp	72.630%	26.820%	0.550%	99.450%
8th time stamp	26.260%	73.180%	0.560%	99.440%
9th time stamp	26.910%	65.920%	7.170%	92.830%
10th time stamp	19.550%	51.960%	28.490%	71.510%

Table 5.53: Patient distribution in intervals for every time stamp, from General to Photodetails – Series 2

	< 30 days	30– 60 days	> 60 days	Sum of two first intervals
1st time stamp	0.000%	50.000%	50.000%	50.000%
2nd time stamp	50.000%	0.000%	50.000%	50.000%
3rd time stamp	100.000%	0.000%	0.000%	100.000%
4th time stamp	100.000%	0.000%	0.000%	100.000%
5th time stamp	50.000%	50.000%	0.000%	100.000%
6th time stamp	50.000%	50.000%	0.000%	100.000%
7th time stamp	100.000%	0.000%	0.000%	100.000%
8th time stamp	100.000%	0.000%	0.000%	100.000%
9th time stamp	50.000%	50.000%	0.000%	100.000%
10th time stamp	100.000%	0.000%	0.000%	100.000%

Table 5.54: Patient distribution in intervals for every time stamp, from General to Photodetails – Series 3

	< 30 days	30– 60 days	> 60 days	Sum of two first intervals
1st time stamp	46.670%	32.000%	21.330%	78.670%
2nd time stamp	36.000%	37.330%	26.670%	73.330%
3rd time stamp	53.330%	40.000%	6.670%	93.330%
4th time stamp	69.330%	30.670%	0.000%	100.000%
5th time stamp	82.670%	14.640%	2.690%	97.310%
6th time stamp	89.330%	10.670%	0.000%	100.000%
7th time stamp	76.000%	24.000%	0.000%	100.000%
8th time stamp	22.670%	77.330%	0.000%	100.000%
9th time stamp	28.000%	66.670%	5.330%	94.670%
10th time stamp	17.330%	48.000%	34.670%	65.330%

Tables 5.37 – 5.54 shows the results for Rule III of the time needed to combine information for patients with 5 to 10 time stamps for series 1, 2 and 3. In those tables, the 1st column shows the time stamp, the second column is the percentage of patients who need less than 30 days to combine information from different sources the third column shows the percentage of patients that need 30 to 60 days and the fourth the percentage of patients who need more that 60 days.

The last column shows the percentage of patients for each time stamp who need less than 60 days to combine information from General and Photodetails. For those tables the procedure that is followed is that for the values of the last are averaged and their standard deviation is calculated. If the ratio of the standard deviation over the average value is smaller than a threshold value, which here is set to 0.15, the test is successful and it is interpreted that the majority of the patients in all time stamps are examined by medics in less than 60 days and the score for this experiment is accredited with 1. However, by applying Rule I and Rule II, the outcome is that only 6 out of 2328 patients have a uniform time interval for linking data from different sources, and 0 out of 5 time stamps have uniformity over the elapsed time in days from General to Photodetails. The same approach is used for all experiments, and Tables 5.14–5.16 below summarize the scores of the 3 rules for series 1, 2, and 3. Appendix 3 provides graphs regarding the time intervals for all series of experiments.

Table 5.55: Score summary for series 1; from General to Photodetails.

	Rule I	Rule II	Rule III
5 time stamps	6 / 2328	0 / 5	1 / 1
6 time stamps	0 /1420	0 /6	1 / 1
7 time stamps	0 / 887	0 / 7	1 / 1
8 time stamps	0 / 546	0 / 8	1 / 1
9 time stamps	0 / 329	0 / 9	1 / 1
10 time stamps	0 / 179	0 /10	1 / 1

Table 5.56: Score summary for series 2; from General to Photodetails

	Rule i	Rule ii	Rule iii
5 time stamps	0 / 51	0 / 5	1 / 1
6 time stamps	0 / 25	0 /6	1 / 1
7 time stamps	0 / 10	0 / 7	1 / 1
8 time stamps	0 / 7	0 / 8	1 / 1
9 time stamps	0 / 5	0 / 9	1 / 1
10 time stamps	0 / 2	0 /10	1 / 1

Table 5.57: Score summary for series 3; from General to Photodetails

	Rule i	Rule ii	Rule iii
5 time stamps	2 / 1303	0 / 5	1 / 1
6 time stamps	0 / 715	0 /6	1 / 1
7 time stamps	0 / 411	0 / 7	1 / 1
8 time stamps	0 / 231	0 / 8	1 / 1
9 time stamps	0 / 135	0 / 9	1 / 1
10 time stamps	0 / 75	0 /10	1 / 1

In Tables 5.55–5.57, the first column shows whether, for the same object in every time stamp, the link from general to Photodetails, in terms of time interval, is uniform; the second column shows whether, in every time stamp, there is uniformity in objects regarding the link from general to Photodetails. The third column shows whether the majority of the objects go from general to Photodetails in a predefined time interval.
It can be concluded, from columns 1 and 2, that:

- Each object has its own time pattern from time stamp to time stamp, to link general to Photodetails.
- In every time stamp, there is no uniformity in patients regarding the elapsed time from general to Photodetails.

Column 3 shows that in all experiments, the majority of objects go from general to Photodetails within a time interval of 60 days.
The following tables 5.58 – 5.75 show analytically the results for Rule III examining the distribution of patients from time-stamp to time-stamp. The interval of 180 to 540 days that intervenes between time-stamps is broken into smaller intervals , as it can been seen in the tables first line.

Table 5.58: Patient distribution in intervals from previous to next time stamp – Series 1

	180-252 days	253-324 days	325-396 days	397-468 days	469-540 days	Sum of 180 -468 days
1st to 2nd time stamp	1.890%	7.131%	54.300%	29.300%	7.379%	92.621%
2nd to 3rd time stamp	0.902%	5.713%	48.320%	30.460%	14.605%	85.395%
3rd to 4th time stamp	0.473%	4.983%	45.790%	29.770%	18.985%	81.016%
4th to 5th time stamp	0.773%	5.069%	37.970%	31.010%	25.178%	74.822%

Table 5.59: Patient distribution in intervals from previous to next time stamp – Series 2

	180-252 days	253-324 days	325-396 days	397-468 days	469-540 days	Sum of 180 -468 days
1st to 2nd time stamp	1.961%	5.882%	50.980%	33.330%	7.847%	92.153%
2nd to 3rd time stamp	0.000%	11.760%	56.860%	17.650%	13.730%	86.270%
3rd to 4th time stamp	0.000%	5.882%	47.060%	29.410%	17.648%	82.352%
4th to 5th time stamp	0.000%	3.922%	41.180%	31.370%	23.528%	76.472%

Table 5.60: Patient distribution in intervals from previous to next time stamp – Series 3

	180-252 days	253-324 days	325-396 days	397-468 days	469-540 days	Sum of 180 -468 days
1st to 2nd time stamp	1.919%	8.289%	54.410%	27.860%	7.522%	92.478%
2nd to 3rd time stamp	0.537%	5.219%	46.740%	32.390%	15.114%	84.886%
3rd to 4th time stamp	0.461%	5.526%	44.900%	30.700%	18.414%	81.587%
4th to 5th time stamp	1.074%	5.679%	36.070%	31.160%	26.017%	73.983%

Table 5.61: Patient distribution in intervals from previous to next time stamp – Series 1

	180-252 days	253-324 days	325-396 days	397-468 days	469-540 days	Sum of 180 -468 days
1st to 2nd time stamp	1.408%	7.606%	53.940%	26.480%	10.566%	89.434%
2nd to 3rd time stamp	0.845%	6.638%	56.340%	29.440%	6.737%	93.263%
3rd to 4th time stamp	0.916%	6.408%	49.080%	28.800%	14.797%	85.204%
4th to 5th time stamp	0.493%	5.423%	45.490%	30.350%	18.244%	81.756%
5th to 6th time stamp	0.634%	5.282%	39.370%	28.730%	25.984%	74.016%

Table 5.62: Patient distribution in intervals from previous to next time stamp – Series 2

	180-252 days	253-324 days	325-396 days	397-468 days	469-540 days	Sum of 180 -468 days
1st to 2nd time stamp	4.000%	4.000%	64.000%	16.000%	12.000%	88.000%
2nd to 3rd time stamp	0.000%	4.000%	52.000%	36.000%	8.000%	92.000%
3rd to 4th time stamp	0.000%	8.000%	64.000%	20.000%	8.000%	92.000%
4th to 5th time stamp	0.000%	0.000%	48.000%	36.000%	16.000%	84.000%
5th to 6th time stamp	0.000%	0.000%	52.000%	28.000%	20.000%	80.000%

Table 5.63: Patient distribution in intervals from previous to next time stamp – Series 3

	180-252 days	253-324 days	325-396 days	397-468 days	469-540 days	Sum of 180 -468 days
1st to 2nd time stamp	1.399%	8.531%	55.240%	23.640%	11.190%	88.810%
2nd to 3rd time stamp	0.839%	6.993%	56.500%	28.530%	7.138%	92.862%
3rd to 4th time stamp	0.559%	6.434%	47.550%	30.490%	14.967%	85.033%
4th to 5th time stamp	0.420%	6.294%	43.500%	32.170%	17.616%	82.384%
5th to 6th time stamp	0.979%	6.014%	37.760%	29.230%	26.017%	73.983%

Table 5.64: Patient distribution in intervals from previous to next time stamp – Series 1

	180-252 days	253-324 days	325-396 days	397-468 days	469-540 days	Sum of 180 -468 days
1st to 2nd time stamp	1.127%	6.877%	59.410%	28.180%	4.406%	95.594%
2nd to 3rd time stamp	0.451%	6.877%	55.020%	27.620%	10.032%	89.968%
3rd to 4th time stamp	1.015%	6.990%	57.500%	28.520%	5.975%	94.025%
4th to 5th time stamp	0.789%	5.862%	51.070%	27.960%	14.319%	85.681%
5th to 6th time stamp	0.451%	5.186%	46.450%	31.340%	16.573%	83.427%
6th to 7th time stamp	0.902%	6.426%	38.560%	27.510%	26.602%	73.398%

Table 5.65: Patient distribution in intervals from previous to next time stamp – Series 2

	180-252 days	253-324 days	325-396 days	397-468 days	469-540 days	Sum of 180 -468 days
1st to 2nd time stamp	10.000%	10.000%	50.000%	20.000%	10.000%	90.000%
2nd to 3rd time stamp	10.000%	0.000%	60.000%	20.000%	10.000%	90.000%
3rd to 4th time stamp	0.000%	0.000%	70.000%	30.000%	0.000%	100.000%
4th to 5th time stamp	0.000%	10.000%	70.000%	20.000%	0.000%	100.000%
5th to 6th time stamp	0.000%	0.000%	60.000%	20.000%	20.000%	80.000%
6th to 7th time stamp	0.000%	0.000%	60.000%	30.000%	10.000%	90.000%

Table 5.66: Patient distribution in intervals from previous to next time stamp – Series 3

	180-252 days	253-324 days	325-396 days	397-468 days	469-540 days	Sum of 180 -468 days
1st to 2nd time stamp	1.217%	7.299%	58.390%	29.680%	3.414%	96.586%
2nd to 3rd time stamp	0.243%	9.002%	54.740%	25.050%	10.965%	89.035%
3rd to 4th time stamp	0.973%	7.299%	58.640%	28.470%	4.618%	95.382%
4th to 5th time stamp	0.243%	6.326%	50.850%	28.220%	14.361%	85.639%
5th to 6th time stamp	0.730%	5.596%	45.500%	32.360%	15.814%	84.186%
6th to 7th time stamp	1.460%	7.299%	37.470%	28.950%	24.821%	75.179%

Table 5.67: Patient distribution in intervals from previous to next time stamp – Series 1

	180-252 days	253-324 days	325-396 days	397-468 days	469-540 days	Sum of 180 -468 days
1st to 2nd time stamp	1.282%	5.128%	68.680%	20.330%	4.580%	95.420%
2nd to 3rd time stamp	0.550%	5.861%	60.070%	30.400%	3.119%	96.881%
3rd to 4th time stamp	0.366%	7.692%	56.960%	26.370%	8.612%	91.388%
4th to 5th time stamp	0.916%	6.044%	61.360%	26.190%	5.490%	94.510%
5th to 6th time stamp	0.916%	6.044%	51.100%	27.660%	14.280%	85.720%
6th to 7th time stamp	0.366%	4.762%	46.150%	31.320%	17.402%	82.598%
7th to 8th time stamp	1.282%	7.326%	40.290%	25.460%	25.642%	74.358%

Table 5.68: Patient distribution in intervals from previous to next time stamp – Series 2

	180-252 days	253-324 days	325-396 days	397-468 days	469-540 days	Sum of 180 -468 days
1st to 2nd time stamp	0.000%	14.290%	57.140%	28.570%	0.000%	100.000%
2nd to 3rd time stamp	14.290%	14.290%	42.860%	14.290%	14.270%	85.730%
3rd to 4th time stamp	14.290%	0.000%	71.420%	14.290%	0.000%	100.000%
4th to 5th time stamp	0.000%	0.000%	71.430%	28.570%	0.000%	100.000%
5th to 6th time stamp	0.000%	14.290%	71.420%	14.290%	0.000%	100.000%
6th to 7th time stamp	0.000%	0.000%	57.140%	14.290%	28.570%	71.430%
7th to 8th time stamp	0.000%	0.000%	71.430%	28.570%	0.000%	100.000%

Table 5.69: Patient distribution in intervals from previous to next time stamp – Series 3

	180-252 days	253-324 days	325-396 days	397-468 days	469-540 days	Sum of 180 -468 days
1st to 2nd time stamp	2.165%	6.061%	67.970%	18.610%	5.194%	94.806%
2nd to 3rd time stamp	0.000%	5.195%	60.610%	31.170%	3.025%	96.975%
3rd to 4th time stamp	0.433%	9.957%	58.870%	21.650%	9.090%	90.910%
4th to 5th time stamp	1.299%	4.762%	62.770%	28.570%	2.599%	97.401%
5th to 6th time stamp	0.433%	6.494%	51.520%	26.840%	14.713%	85.287%
6th to 7th time stamp	0.433%	4.329%	42.420%	36.360%	16.458%	83.542%
7th to 8th time stamp	2.165%	7.792%	39.830%	25.970%	24.243%	75.757%

Table 5.70: Patient distribution in intervals from previous to next time stamp – Series 1

	180-252 days	253-324 days	325-396 days	397-468 days	469-540 days	Sum of 180 -468 days
1st to 2nd time stamp	0.912%	4.863%	61.090%	28.270%	4.865%	95.135%
2nd to 3rd time stamp	0.000%	4.863%	75.080%	17.230%	2.827%	97.173%
3rd to 4th time stamp	0.912%	8.815%	60.490%	26.750%	3.033%	96.967%
4th to 5th time stamp	0.608%	7.903%	54.410%	28.270%	8.809%	91.191%
5th to 6th time stamp	1.520%	8.207%	56.230%	28.270%	5.773%	94.227%
6th to 7th time stamp	0.912%	5.471%	50.150%	27.050%	16.417%	83.583%
7th to 8th time stamp	0.304%	4.863%	46.200%	29.480%	19.153%	80.847%
8th to 9th time stamp	0.912%	7.903%	37.390%	25.230%	28.565%	71.435%

Table 5.71: Patient distribution in intervals from previous to next time stamp – Series 2

	180-252 days	253-324 days	325-396 days	397-468 days	469-540 days	Sum of 180 -468 days
1st to 2nd time stamp	0.000%	20.000%	60.000%	20.000%	0.000%	100.000%
2nd to 3rd time stamp	0.000%	20.000%	60.000%	20.000%	0.000%	100.000%
3rd to 4th time stamp	20.000%	20.000%	60.000%	0.000%	0.000%	100.000%
4th to 5th time stamp	0.000%	0.000%	80.000%	20.000%	0.000%	100.000%
5th to 6th time stamp	0.000%	0.000%	100.000%	0.000%	0.000%	100.000%
6th to 7th time stamp	0.000%	20.000%	80.000%	0.000%	0.000%	100.000%
7th to 8th time stamp	0.000%	0.000%	80.000%	0.000%	20.000%	80.000%
8th to 9th time stamp	0.000%	0.000%	80.000%	20.000%	0.000%	100.000%

Table 5.72: Patient distribution in intervals from previous to next time stamp – Series 3

	180-252 days	253-324 days	325-396 days	397-468 days	469-540 days	Sum of 180 -468 days
1st to 2nd time stamp	0.741%	4.444%	57.040%	33.330%	4.445%	95.555%
2nd to 3rd time stamp	0.000%	2.963%	75.560%	18.520%	2.957%	97.043%
3rd to 4th time stamp	0.000%	7.407%	62.220%	28.150%	2.223%	97.777%
4th to 5th time stamp	0.741%	9.630%	56.300%	27.700%	5.629%	94.371%
5th to 6th time stamp	2.222%	6.667%	57.040%	31.110%	2.961%	97.039%
6th to 7th time stamp	0.000%	5.926%	48.150%	28.890%	17.034%	82.966%
7th to 8th time stamp	0.000%	5.185%	42.220%	33.330%	19.265%	80.735%
8th to 9th time stamp	0.741%	10.370%	31.810%	28.890%	28.189%	71.811%

Table 5.73: Patient distribution in intervals from previous to next time stamp – Series 1

	180-252 days	253-324 days	325-396 days	397-468 days	469-540 days	Sum of 180 -468 days
1st to 2nd time stamp	0.000%	4.469%	56.420%	32.400%	6.711%	93.289%
2nd to 3rd time stamp	0.000%	3.911%	70.390%	24.580%	1.119%	98.881%
3rd to 4th time stamp	0.000%	5.028%	78.770%	14.530%	1.672%	98.328%
4th to 5th time stamp	0.559%	12.850%	58.100%	26.820%	1.671%	98.329%
5th to 6th time stamp	0.559%	10.610%	52.510%	26.820%	9.501%	90.499%
6th to 7th time stamp	2.235%	11.730%	57.540%	22.910%	5.585%	94.415%
7th to 8th time stamp	1.117%	6.145%	58.100%	24.020%	10.618%	89.382%
8th to 9th time stamp	0.559%	3.911%	51.960%	27.370%	16.200%	83.800%
9th to 10th time stamp	1.117%	2.821%	37.430%	25.140%	33.492%	66.508%

Table 5.74: Patient distribution in intervals from previous to next time stamp – Series 2

	180-252 days	253-324 days	325-396 days	397-468 days	469-540 days	Sum of 180 -468 days
1st to 2nd time stamp	0.000%	0.000%	0.000%	50.000%	50.000%	50.000%
2nd to 3rd time stamp	0.000%	0.000%	100.000%	0.000%	0.000%	100.000%
3rd to 4th time stamp	0.000%	50.000%	50.000%	0.000%	0.000%	100.000%
4th to 5th time stamp	50.000%	50.000%	0.000%	0.000%	0.000%	100.000%
5th to 6th time stamp	0.000%	0.000%	50.000%	50.000%	0.000%	100.000%
6th to 7th time stamp	0.000%	0.000%	100.000%	0.000%	0.000%	100.000%
7th to 8th time stamp	50.000%	50.000%	0.000%	0.000%	0.000%	100.000%
8th to 9th time stamp	0.000%	0.000%	50.000%	0.000%	50.000%	50.000%
9th to 10th time stamp	0.000%	0.000%	50.000%	50.000%	0.000%	100.000%

Table 5.75: Patient distribution in intervals from previous to next time stamp – Series 3

	180-252 days	253-324 days	325-396 days	397-468 days	469-540 days	Sum of 180 -468 days
1st to 2nd time stamp	0.000%	5.333%	52.000%	37.330%	5.337%	94.663%
2nd to 3rd time stamp	0.000%	5.333%	62.670%	30.670%	1.327%	98.673%
3rd to 4th time stamp	0.000%	1.333%	81.330%	14.670%	2.667%	97.333%
4th to 5th time stamp	0.000%	12.000%	53.330%	33.330%	1.340%	98.660%
5th to 6th time stamp	1.333%	12.000%	54.670%	21.330%	10.667%	89.333%
6th to 7th time stamp	4.000%	8.000%	61.330%	24.000%	2.670%	97.330%
7th to 8th time stamp	0.000%	6.667%	53.330%	30.670%	9.333%	90.667%
8th to 9th time stamp	0.000%	5.333%	49.330%	29.330%	16.007%	83.993%
9th to 10th time stamp	0.000%	9.333%	38.670%	32.000%	19.997%	80.003%

Tables 5.76 - 5.78 detail the scores, this time examining the time interval from time stamp to time stamp for each patient.

Table 5.76: Score summary for series 1; from General to Photodetails

	Rule I	Rule II	Rule III
5 time stamps	1762 / 2328	4 / 4	1 / 1
6 time stamps	1093 / 1420	5 / 5	1 / 1
7 time stamps	697 / 887	5 / 6	1 / 1
8 time stamps	0 / 546	6 / 7	1 / 1
9 time stamps	0 / 329	6 / 8	1 / 1
10 time stamps	0 / 179	8 /9	1 / 1

Table 5.77: Score summary for series 2; from General to Photodetails

	Rule I	Rule II	Rule III
5 time stamps	41 / 51	4 / 4	1 / 1
6 time stamps	21/25	4 / 5	1 / 1
7 time stamps	7 / 10	3 / 6	1 / 1
8 time stamps	4 / 7	4 / 7	1 / 1
9 time stamps	3 / 5	6 / 8	1 / 1
10 time stamps	0 / 2	6 /9	1 / 1

Table 5.78: Score summary for series 3; from General to Photodetails

	Rule I	Rule II	Rule III
5 time stamps	968 / 1303	3 / 4	1 / 1
6 time stamps	538 / 715	4 / 5	1 / 1
7 time stamps	323 / 411	5 / 6	1 / 1
8 time stamps	185 / 231	6 / 7	1 / 1
9 time stamps	113 / 135	7 / 8	1 / 1
10 time stamps	66 / 75	9 / 9	1 / 1

Tables 5.76–5.78 show that there is uniformity in the creation of episodes not only for the same object in every time stamp but also from the previous to the next time stamp where the time interval is similar. As regards the third column in the tables above, the time interval from episode to episode ranges from 180 to 540 days and is broken down into five intervals: 180–252 days, 253–324 days, 325–396 days, 397–468 days, and 469–540 days. It is calculated from the average of the sum of the patients that their time interval from time stamp to time stamp is a maximum of 468 days. The score 1/1 shows that there is uniformity. The approach followed here is the same followed before

when it was examined the formation of a time stamp. The only thing that changes is the number of intervals.

5.4.4.4 Interestingness

This section presents the results of the experiments designed to evaluate the trend-mining framework and SOMA in particular. As stated earlier, the datasets were manipulated either to contain a certain class (condition) of the disease or to pick up all patients.

Interestingness criterion aims to identify whether:

- The discovered rules are interesting
- At which time stamp the discovered rules are interesting; in other words the criteria want to examine whether the interestingness duration is in all time stamp or not.

Here interestingness should not been confused with the interestingness as a measure whether or the information is useful or not. The interestingness here is not used as a qualification criterion but as quantification.

In large databases the amount of knowledge can be huge, hence the algorithm should be able to distinguish which rule is interesting or not. Using just confidence and support we can discover strong rules but this doesn't mean that they are indeed interesting.

In this thesis we are using 5 criteria to evaluate the interestingness of the discovered rule:

- Lift
- All confidence
- Max confidence
- Kulzusnki
- Cosine.

In chapter 2 is given details about those criteria. The following procedure is followed:

For each of the above criteria the user sets a target which the discovered rules have to achieve. For lift the threshold must be above 1 which means that for any rule $X \rightarrow Y$ X,Y are positively correlated. For the rest 4 criteria the threshold must be a percentage from 0 to 100 %. For each discovered rule, $\rightarrow Y$, the 5 criteria is calculated. If the target is achieved the score increases by 1. So for each rule the maximum score that can be achieved is 5 and the minimum and then this final score is calculated as a percentage. If the final score expressed as a percentage, is greater than the threshold which the user gives then the rule is interesting at a specific time stamp. That procedure is repeated for the rule $X \rightarrow Y$ for all time stamps.

For the experiments in this chapter, for lift the threshold is set above 1 and for the rest 4 criteria the threshold is set to 50 %.

The tables 5.20, 5.21 and 5.22 summarize the scores for the experiments. The line with the name "Total" shows the total number of trends. The following lines show the number of trends with the interestingness score above 50% below 50% the percentage of those trends and the maximum and minimum score of all trends, respectively.

Tables5.82 to 5.84 show analytically the results for the experiments on the interestingness. Those tables provide the score for each of the 5 criteria for each time stamp, showing the number of trends that exceed the threshold, the number of trends that don't exceed the threshold, the maximum and the minimum score.

In most case the score is ranged around 60% and this is due to the fact that the criteria are affected from the confidence of the inverse rule $Y \rightarrow X$ and this can be seen from tables 5.82 to 5.84 and in particularly from the score of all confidence which is the mean of the confidences of the rule $X \rightarrow Y$ and $Y \rightarrow X$ and from the score of cosine which is the square root of the product of the confidence of a rule and its inverse.

It can be seen that the amount of information available regarding patients with DR is very small compared with that regarding non-DR patients or those who have developed DR at a certain stage of their life. Hereinafter, experiments concerning all patients will be referred to as series 1, experiments for patients with DR in all time stamps as series 2, and the other category as series 3.

The first measure of evaluation presented here is the interestingness of the rules not only at a certain point but in all time stamps. By applying a scoring system (described previously) using certain criteria apart from confidence, there is a maximum score that can be achieved, and the results are compared against that maximum value in Tables 5.79 - 5.81.

Table 5.79: Summary of interestingness score for series 1

Time stamps / Trends	5	6	7	8	9	10
Total	423	383	338	308	259	188
interestingness score more than 50%	228	231	214	224	197	153
interestingness score less than 50%	195	152	124	84	62	35
% interestingness score more than 50%	53.9	60.31	63.31	72.73	76.06	81.38
% interestingness score less than 50%	46.1	39.69	36.39	27.27	23.94	18.62
Maximum score %	60	60	60	60	60	60
Minimum score%	8	10	6.66	11.42	12	12.5

Table 5.80: Summary of interestingness score for series 2

Time stamps / Trends	5	6	7	8	9	10
Total	62	40	19	16	12	6
interestingness score more than 50%	0	1	3	1	12	6
interestingness score less than 50%	62	39	16	15	0	0
% interestingness score more than 50%	0	2.5	15.79	6.25	100	100
% interestingness score less than 50%	100	97.5	84.21	93.75	0	0
Maximum score %	40	52	60	56	67.5	80
Minimum score %	40	40	40	40	60	80

Table 5.81: Summary of interestingness score for series 3

Time stamps / Trends	5	6	7	8	9	10
Total	290	226	172	146	127	100
interestingness score more than 50%	0	0	0	0	0	0
interestingness score less than 50%	290	226	172	146	127	100
% interestingness score more than 50%	0	0	0	0	0	0
% interestingness score less than 50%	100	100	100	100	100	100
Maximum score %	40	40	40	40	40	47.5
Minimum score %	40	40	40	40	40	40

In all these tables, the first line shows the total line of trends under examination; the second line shows the number of trends for which their score exceeds 50%; and the third line shows the number of trends with a score below 50%. Lines 4 and 5 show, respectively, the percentage of trends scoring greater than and less than 50%, and the last two lines show the maximum and minimum scores. Each column represents an experiment with its number of time stamps.

The next three tables, represent analytically the scoring for each series of experiments showing the results for each measure for all patients, for patients that have no DR and for patients with DR respectively. They present the number of occurrences which each criterion has exceeded, or not, a certain threshold, which is set to be equal to the confidence threshold, and this rule regards all confidence, max confidence, Kulczynski, and cosine. As regards lift, it is measured if it is greater than, equal to, or less than the unity. Regarding counting the number of occurrences, the trends under examination are those that have exceeded the confidence threshold in any time stamp. All Confidence, max confidence, Kulczynski, and cosine have a range of values from 0 to 100%, while lift is a positive real number.

Table 5.82: Interestingness score for series 1

ALL Confidence	5 t.s.	6 t.s.	7 t.s.	8 t.s.	9 t.s.	10 t.s.
Number of trends above threshold	0	0	0	0	0	0
Number of trends below threshold	1228	1177	1041	946	810	609
Maximum value	10.22 %	10.92 %	12.26 %	13.22 %	12.69 %	18.75 %
Minimum value	0.047 %	0.078 %	0.12%	0.20 %	0.33 %	0.60 %
Cosine	5 t.s.	6 t.s.	7 t.s.	8 t.s.	9 t.s.	10 t.s.
Number of trends above threshold	0	0	0	0	0	0
Number of trends below threshold	1228	1177	1041	946	810	609
Maximum value	28.14 %	31.72 %	31.90 %	34.68 %	30.72 %	36.25 %
Minimum value	1.13 %	1.37 %	1.54 %	2.25 %	3.37 %	4.57 %
Max Confidence	5 t.s.	6 t.s.	7 t.s.	8 t.s.	9 t.s.	10 t.s.
Number of trends above threshold	1104	1065	968	879	760	573
Number of trends below threshold	124	112	73	67	50	36
Maximum value	100 %	100 %	100 %	100 %	100 %	100 %
Minimum value	3.17 %	4 %	2.38 %	3.70 %	4.55 %	9.09 %
Kulczynski	5 t.s.	6 t.s.	7 t.s.	8 t.s.	9 t.s.	10 t.s.
Number of trends above threshold	750	776	743	705	631	502
Number of trends below threshold	148	401	298	241	179	107
Maximum value	51.17 %	52.53 %	53.83 %	54.62 %	54.12 %	55.42 %
Minimum value	1.91 %	2.23 %	1.69 %	2.76 %	3.62 %	6.71 %
Lift	5 t.s.	6 t.s.	7 t.s.	8 t.s.	9 t.s.	10 t.s.
Number of trends >1	980	923	822	773	672	524
Number of trends =1	0	0	0	0	0	0
Number of trends <1	300	254	219	173	138	85
Maximum value	10.87	10.14	11.2278	11.14	10.96	13.76
Minimum value	0.24	0.26	0.2112	0.27	0.37	0.37

Table 5.83: Interestingness score for series 2

ALL Confidence	5 t.s.	6 t.s.	7 t.s.	8 t.s.	9 t.s.	10 t.s.
Number of trends above threshold	0	0	0	0	4	12
Number of trends below threshold	139	93	47	45	30	0
Maximum value	13.72 %	20 %	30 %	28.57 %	40 %	50 %
Minimum value	1.96 %	4 %	10 %	14.28 %	20 %	50 %
Cosine	5 t.s.	6 t.s.	7 t.s.	8 t.s.	9 t.s.	10 t.s.
Number of trends above threshold	0	5	36	40	34	12
Number of trends below threshold	139	88	11	5	0	0
Maximum value	37.04 %	44.72 %	54.77 %	53.45 %	63.24 %	70.71 %
Minimum value	14.0 %	20 %	31.62 %	37.79 %	44.72 %	70.71 %
Max Confidence	5 t.s.	6 t.s.	7 t.s.	8 t.s.	9 t.s.	10 t.s.
Number of trends above threshold	139	93	47	45	34	12
Number of trends below threshold	0	0	0	0	0	0
Maximum value	100 %	100 %	100 %	100 %	100 %	100 %
Minimum value	100 %	100 %	100 %	100 %	100 %	100 %
Kulczynski	5 t.s.	6 t.s.	7 t.s.	8 t.s.	9 t.s.	10 t.s.
Number of trends above threshold	139	93	47	45	34	12
Number of trends below threshold	0	0	0	0	0	0
Maximum value	56.86 %	60 %	65 %	64.28 %	70 %	75 %
Minimum value	50.98 %	52 %	55 %	57.11 %	60 %	75 %
Lift	5 t.s.	6 t.s.	7 t.s.	8 t.s.	9 t.s.	10 t.s.
Number of trends >1	0	0	0	0	0	0
Number of trends =1	139	93	47	45	34	12
Number of trends <1	0	0	0	0	0	0
Maximum value	1	1	1	1	1	1
Minimum value	1	1	1	1	1	1

Table 5.84: Interestingness score for series 3

ALL Confidence	5 t.s.	6 t.s.	7 t.s.	8 t.s.	9 t.s.	10 t.s.
Number of trends above threshold	0	0	0	0	0	0
Number of trends below threshold	819	707	595	505	454	335
Maximum value	10.36 %	11.88 %	12.65 %	15.15 %	11.85 %	22.67 %
Minimum value	0.07 %	0.14 %	0.24 %	0.43 %	0.74 %	1.33 %
Cosine	5 t.s.	6 t.s.	7 t.s.	8 t.s.	9 t.s.	10 t.s.
Number of trends above threshold	0	0	0	0	0	0
Number of trends below threshold	819	707	595	505	454	335
Maximum value	32.18 %	34.47 %	35.56 %	38.92 %	34.42 %	47.61 %
Minimum value	2.77 %	3.73 %	4.93 %	6.57 %	8.50 %	11.54 %
Max Confidence	5 t.s.	6 t.s.	7 t.s.	8 t.s.	9 t.s.	10 t.s.
Number of trends above threshold	819	707	595	505	454	335
Number of trends below threshold	0	0	0	0	0	0
Maximum value	100 %	100 %	100 %	100 %	100 %	100 %
Minimum value	100 %	100 %	100 %	100 %	100 %	100 %
Kulczynski	5 t.s.	6 t.s.	7 t.s.	8 t.s.	9 t.s.	10 t.s.
Number of trends above threshold	819	707	595	505	454	335
Number of trends below threshold	0	0	0	0	0	0
Maximum value	55.18 %	55.94 %	56.32 %	57.57 %	55.92 %	50.67 %
Minimum value	50.03 %	50.06%	50.12 %	50.21 %	50.37 %	61.33 %
Lift	5 t.s.	6 t.s.	7 t.s.	8 t.s.	9 t.s.	10 t.s.
Number of trends >1	0	0	0	0	0	0
Number of trends =1	819	707	595	505	454	335
Number of trends <1	0	0	0	0	0	0
Maximum value	1	1	1	1	1	1
Minimum value	1	1	1	1	1	1

From the tables 5.79-5.81, it can be seen that:

- In all cases the maximum scoring has not been achieved. The highest scoring that has been achieved was 80%, and this refers to a trend from series 2 for 10 time stamps.
- In series 1 and series 3, the all confidence criterion remained below the threshold. Also, in the series 2 experiments for five time stamps to eight time stamps, the all confidence remains below the threshold. This means that only the $X \rightarrow Y$ has confidence above the threshold, while the $Y \rightarrow X$ does not have confidence above the threshold. Practically, this can be interpreted as follows: characteristics are linked to DR, but DR cannot be linked to the same characteristics.
- In series 2 and series 3, the lift is always equal to 1. This can be explained by the definition of lift. As mentioned earlier, lift is defined as the ratio of $P(A \cup B)$ over the product of$P(A) \cdot P(B)$. In those two series, $P(B) = 1$ because in those datasets, all lines contain the same B, and $P(A) = P(A \cup B)$. Therefore, in cases where the dataset is manipulated in such a way that B has one value only, lift will always be equal to 1. Therefore, the use of lift as a measure of interestingness cannot be reliable in such cases.

5.4.4.5 Quality of the knowledge discovered

To check the quality of knowledge discovered, specific rules were used that describe the relation of some attributes that exist in the databases with DR. For this reason, experiments were conducted using exclusively the attributes dictated by the rules. Moreover, the confidence threshold used in those experiments is higher than the confidence threshold that was used in the experiments above. At the following lines, the rules are presented along with the results from the experiments, which show whether the rules are confirmed or not:

- If a patient suffers from cataracts, it is possible to develop DR: in the experiment with six time stamps (episodes), it has been shown that if a patient does not suffer from cataracts, then they are unlikely to suffer from DR. There is a maximum confidence of 90.64% and a minimum of 49.8%. The lift is well above 1, with a minimum value of 1264.9. A very important finding from this experiment is that the inverse rule $Y \rightarrow X$ has a minimum confidence of 87.2% and maximum confidence of 92.85%.

- The younger a patient is when diagnosed with diabetes, the more likely this patient will develop DR. In the experiments with six time stamps, the following rule occurs: if a diabetic patient is diagnosed at the age of 60–70 years, they are unlikely to suffer from DR. This rule has a maximum confidence of 92.34% and a minimum of 53.17%, and the lift is very high, with a minimum value of 489.44%. However, the inverse rule has less confidence, with a minimum of 33% and maximum of 34%, and in this rule, the DR is not strongly linked to age of diabetes diagnosis.
- If a patient has suffered from type 2 diabetes for more than 20 years, it is very likely that this patient will develop DR. An experiment with seven time stamps has shown the following rule: if a patient suffers from type 2 diabetes, and the duration of diabetes is less than 5 years, is unlikely to develop DR. This rule has a confidence of 84–96.4%, and the minimum lift value is 20.2. Although the inverse rule has a small confidence, ranging from 1.83 to 30%, it has a lift greater than 1. Also, another rule from the same experiment is that if a patient has suffered from type 2 diabetes for more than 20 years, they are likely to develop DR. This rule has a confidence of 50–100% and minimum lift 1.1, and although the confidence of the inverse rule may be low, again the lift of the inverse rule is greater than 1, (1.4).
- If a patient suffers from type 1 diabetes, they are likely to develop DR. This rule has been confirmed by experiments using different numbers of time stamps. In experiments with nine time stamps, a high value of confidence at 83.87% has been recorded. Both in this experiment and in others, where the maximum confidence reached 55%, the lift remained above 1. However, the inverse rule has a very low confidence (<15%) but also the lift is well above 1 with a minimum lift of 3.
- If a patient suffers from type 2 diabetes and is on insulin treatment, this patient is likely to suffer from DR. This rule is confirmed by many experiments, with a confidence ranging from 9% to 61%, and the lift varying from 1.17 to 226.9.
- If a patient suffers from type 2 diabetes, and the duration of diabetes is longer than 20 years then this patient is likely to develop DR. In an experiment with five time stamps, it has been confirmed that patient who suffers from diabetes type 2 for more than 25 years is likely to suffer from DR. This rule has a confidence ranging from 75% to 100% and a lift ranging from 3.30 to 16.07.

According to the clinicians of Saint Paul Eye Unit the first two clauses appear to provide new evidence while the other 4 fit with the accepted thinking and therefore provide validation to the approach described in this work.

5.4.5 Discussion

One of the important problems in KDD is the evaluation of the discovered knowledge. In real-life applications, the number of discovered rules is huge, and it is difficult for the end user to identify interesting ones. In this thesis, evaluation concerns not the only the discovered knowledge from the proposed trend-mining framework, but also the procedure that the framework follows to discover the knowledge. Similar work has been proposed and described in Nohuddin (2012). In that thesis the author described an approach which is designed to support "end-to-end" social network data and named it predictive trend mining framework. The author divides the framework into two parts; one part is the trend discovery with the use of the FTP Apriori algorithm and in the second part SOM is used to group and visualise the produced trends. Finally the evaluation approach concentrates on the final result of the framework ignoring the intermediate stages.

Geng and Hamillton (2006) used the term "interesting measures" to facilitate a general approach to automatically identifying interesting patterns. They used this term in three ways, or roles, to use their terminology. First, the measures can be used to prune uninteresting patterns during the mining process. Second, measures can be used to rank the patterns according to their interestingness, and finally they are used during post-processing to select interesting patterns.

In this research work, a similar approach is used. First, the measures are used to create time-stamped datasets with the least noise. Second, they are used to discover interesting patterns in data. Third, measures are used to rank the interestingness and filter the rules. Fourth, the measures are used to evaluate the interestingness of the discovered knowledge.

Geng and Hamilton (2006) categorized the "interesting measures" as follows:

- objective measures;
- subjective measures;
- semantic measures.

Objective measures are based only on the raw data, and no knowledge about the user or application is required. A subjective measure takes into account both the data and the user of these data. To define a subjective measure, access to the user's domain or background knowledge about the data is required. A semantic measure considers the semantics and explanations of the patterns. Because semantic measures involve domain knowledge from the user, Yao et al. (2006) considered them to be a special type of subjective measure.

The measures used for the evaluation of this thesis come from the categories of subjective and objective measures. The measures used in the pre-processing are subjective measures because both the creation of the time-stamped datasets and the development of rules for cleaning require knowledge of the domain.

Measures for ARM and ranking are objective, since no knowledge of the domain is needed. ARM is based on the support threshold and confidence threshold, and ranking is achieved using the support values in every time stamp by using mathematical conditions and also by calculating the following properties of the rules: lift, cosine, all confidence, max confidence, and Kulc. Filtering, on the other hand, is subjective because the user uses knowledge of the domain to select the antecedent and consequent of the rule.

Measures to evaluate the knowledge fall into the category of subjective measures because they are based on the knowledge of the domain. The measure that is used in this thesis to evaluate the results is based on the user's existing knowledge and follows an approach similar to the work presented by Liu et al. (1997). They introduced the concept of general impressions. General impressions are if–then clauses that describe the relation between a condition variable and a class value, and reflect the user's knowledge of the domain.

In Liu et al. (1997), the criterion they use to compare a discovered rule r against a set of general impressions G:{G1, G2, …, G_N} is that the rule r and any general impression from the set G must have the same consequent. In this thesis, this principle is modified not only by accepting a general impression with the same consequent but also by accepting general impressions with the opposite consequent if and only if the antecedent in r is also the opposite of the antecedent of a general impression.

Liu et al. (1999) extended their approach to the evaluation by proposing another technique to rank rules according to knowledge background. Based on the user's knowledge, the discovery rules can be classified into the three following categories:

- An unexpected rule is a rule that is unexpected or previously unknown to the user if it has an unexpected condition, an unexpected consequent, or both.
- A confirming rule is a rule that partially or completely matches a user's existing knowledge.
- An actionable rule is a rule that a user can use to do something to their advantage.

The evaluation of the knowledge discovered by the advocated trend-mining framework concentrates on confirming rules to validate the knowledge discovery of the framework.

The rules discovered by knowledge-discovery methods must be of interest to end users to be considered as useful. Therefore, evaluation both of the interestingness of the rules and of the method used to produce those rules is an active and very important task in knowledge discovery, but there are no single measures that can be applied everywhere because KDD can be used in different application domains. Thus, the evaluation approach advocated is adjusted to the proposed trend-mining framework of this thesis.

From the evaluation of the thesis using the SOMA framework, the following conclusions arise. The noise reduction barely reaches levels above 50%, mainly owing to the fact that in the specific example, the expert's knowledge was used instead of a more generic method such as using the average value or the value that is observed more frequently. On the other hand, even though the percentage of noise reduction is low, the object distribution from time stamp to time stamp is very good, since at least 90 objects from the experiments are used throughout all the time stamps. As regards the interestingness of the produced rules, the score rarely exceeds 60% because two of the measures are based on the confidence of the inverse rule. On the other hand, if the dataset is manipulated into a specific characteristic, regarding the consequent of the rule, owing to the definition of confidence, some criteria are nullified.

The figures in appendix 4 show colourful representations, how patients are moving from time stamp to time stamp and from trend to trend. The concept behind this method is based on the definition of trends. As it was referred previously the trend shows how the support changes from time stamp to time stamp. Thus, in each time stamp the support value shows the number of objects which the framework identifies as having certain attribute values, given by the trend. The representation passes through the following steps:

- At the first time stamp the framework allocates with a unique colour every group of objects (patients in the case of SOMA).Each group fills a square with its colour. The number of different groups at this point is equal to the number of trends. If a trend has 0 support value then it is given the white colour at the square that represent this trend at this time stamp.
- At the next time stamp, the framework examines whether or not the group changes in terms of number of objects and of the trend where the objects are located. In other words, the framework tracks the objects and adjusts the square colours according to what was described earlier in chapter 4.

- The orientation of this mosaic of colours is :from bottom to top each line is a trend, at the bottom is the first trend at the top the last one; form left to right is every time stamp.

The aim of this kind of representation is to help the end users to follow the trends along with the text outcome of the framework.

Regarding the evaluation of the discovered knowledge, although the use of trend mining makes it possible to predict what experts expect, the confidence ranges from very low to very high values. Also, the inverse rule has very rarely been justified with a high confidence value. By contrast, lift has always shown very good values above 1. In this specific case using the diabetic retinopathy databases, the attributes showing diabetic retinopathy are rather fewer in number than the attributes showing no diabetic retinopathy and so it has been accepted that there are ranges in the lift and confidence values.

Chapter 6 Conclusion & Future Work

In this research work it is presented a Trend Mining framework with aim to extract hidden trend from longitudinal datasets. The proposed trend mining mechanism is founded on an Association Rule Mining (ARM) approach whereby an ARM technique is applied to a sequence of time stamped data sets. This approach is both efficient and effective at finding trends. The disadvantage, given appropriate input parameters, is that a great many trends may be discovered; the number of identified trends can of course be reduced by adjusting the parameters, but at the risk of losing potentially valuable knowledge.

The application of this novel framework consists of 3 steps:

- Pre-processing of the data: applying cleansing techniques to reduce noise and preparation of time-stamped datasets.
- Main process: creation of trends.
- Evaluation of the discovered knowledge.

SOMA is the application of trend mining framework on the diabetic retinopathy datasets which contain data collected from St Paul's Eye clinic of Royal University Liverpool Hospital. This data is itself of particular interest, in the context of trend mining, as the "time stamps" are defined in terms of patient visit number, as opposed to more traditional forms of temporal data. The data is also extremely noisy. SOMA represents trends as constraints on parameters over intervals that correspond to phases of a process. This representation is based on how expert diagnosticians verbally report their knowledge of trends. For this reason trend templates may be useful for knowledge acquisition and explanation of trends.

SOMA monitors process data and matches them to hypotheses which include a trend template and a chronology of how the data fall into different stages of the trend. Our prototype application to growth chart monitoring produces plausible hypotheses on a real patient. The Aretaeus trend discovery algorithm provides a useful mechanism for trend classification.

Within this research work a generic framework has been proposed for the Validation and Verification of trend mining. The verification examines if the intermediate results are self-consistent and the validation tries to uncover known causal connections in the application.

In order to evaluate the trend mining a set of criteria has been established which covers all stages of trend mining. Those criteria were adjusted to the application of diabetic retinopathy data.

As a result of this application it can be concluded that noise reduction based on the expert knowledge of the domain is not pretty much effective. Even though, the noise reduction didn’t exceed the 50 % the object distribution was quite good. As regards the interestingness of the results this is measured using max_confidence, all_confidence, cosine, Kulc and lift. The first 4 criteria depend not only from the confidence of the rule $X \rightarrow Y$, but also they depend on the confidence of the inverse rule $Y \rightarrow X$. The evaluation showed that using the data from the specific diabetic retinopathy databases, very rarely there was an inverse rule with high confidence.

Another conclusion is that using a mosaic of colours to describe the trends, it makes the visualization a very difficult task because the amount of trends is huge.

To examine the quality of the discovered knowledge, specific rules were used. Those rules are related with the domain of diabetic retinopathy and describe the relation of certain characteristics, such as diabetes type, duration, treatment etc. The trend mining was able to predict those rules but not always with high confidence values. Even though, lift values were well above 1 which shows that the attributes are highly related.

6.1 Future work

In order to improve the trend mining framework the following tasks are suggested for future studies:

- The framework should be tested using different kinds of databases, for several large scale experiments.
- Improvement of the pre-processing stage. In terms of noise reduction, a novel way should be implemented to deal with this task. A Bayesian network could be implemented to fill the missing values using the existing data. Thus the noise will be cleared and the datasets that will be produced will give more accurate results. Another challenge in the pre-processing is the formation of the time stamp; events, whose combination creates an episode, do not occur periodically, so it would be very interesting to have time stamps with different time interval for each stamp.
- Another issue is the use of threshold values. In the work presented in this thesis, the framework targets knowledge with a frequency above a threshold. However, in some domains what is of great importance may be events that do not occur frequently (infrequent patterns / trends). Therefore it would be very challenging to make the framework work on events that occur rarely.

- Techniques to predict the interval between change points/state changes in records. For example to act as a guide for establishing safe disease screening intervals in medical records.
- Techniques to identify the key attributes in longitudinal data sets that influence a particular classification (i.e. the features that influence patient progress in the context of a given disease/condition) and reduce the interaction with the end use on that issue, making the framework more autonomous.
- So far the framework uses one dimension to produce trends; the time dimension. A field to expand the present work would be the use of more dimensions, where a dimension will represent a specific attribute and time stamps will be defined by the intervals of the values of this chosen attribute. For example in SOMA the user could use the attribute that shows the kind of treatment and assumes that different values of that attributes are the new stamps, the attribute-stamps. The time that so far was used to define an episode, with this approach could be converted into another attribute.
- Visualization techniques. Although trend mining produces a colourful representation the difficulty increases when the amount of the produced knowledge is huge. Therefore several filters may be needed to allow the user ,via an interface ,to choose what should be visualized and to clear the amount of information based on some values like lift, confidence or the kind of trend.

Appendices

Appendix 1 – Schemas for logic rules

Table A1.1 Schema for DiabRetinaPhotodetails dataset

Liverpool Diabetes Eye Study Schema Description:

Database DIAB

Dataset DiabRetinaPhotodetails Date: 03/03/10

Data Label	Description	Data type	Value	Narrative	Logic Rules
StudyIDNo	ID number	Number	Xxxxx		
NHS No	ID number (key patient identifier for Study Tables)	Number	10 digits		
ExamDate	Date	Date/ Time	dd/mm/yy		
REField1NAS	Field position nasal field			Ignore warehouse	
REField2UTQ	Upper temporal quadrant			Ignore warehouse	
REField3LTQ	Lower temporal quadrant			Ignore warehouse	

REQual	Clarity and focus			Ignore warehouse	
REHMA	Haemorrhages and/or micro aneurysms	nominal	None=0, Quest=1, <2A=2, ≥2A=3, CG=90		If null and REOUTCOME8=90 set as 90; else set as =NR
RENVD	New Vessels Disc	nominal	None=0, Quest=1, <10A=2, ≥10A=3, CG=90		If null and REOUTCOME8=90 set as 90; else set as =NR
RECWS8A	Cotton wool spot	nominal	None=0, Quest=1, <six=2, ≥six=3, CG=90		If null and REOUTCOME8=90 set as 90; else set as =NR
RENVE	New vessels elsewhere	nominal	None=0, Quest=1, <1/2DA=2, ≥1/2DA=3, CG=90		If null and REOUTCOME8=90 set as 90; else set as =NR
REVBVRVL6A	Venous Beading and /or Venous	nominal	None=0, Quest=1,		If null and REOUTCOME8=90 set as 90; else set as =NR

	Reduplication and/or Venous Loop		1Quad=2 2Quads=3 3Quads= 4 4Quads=5 CG=90		
REFVP	Fibrovascular proliferation	nominal	None=0, Quest=1, FPE=2, FPD=3, FPE+FPD=4, TRD=5, CG=90		If null and REOUTCOME8=90 set as 90; else set as =NR
REIRMA	Intraretinal microvascular abnormality	nominal	None=0, Quest=1, <8A=2, ≥8A=3, CG=90		If null and REOUTCOME8=90 set as 90; else set as =NR
REPRH VH	Pre-Retinal Haemorrhage Vitreous Hemorrhage	nominal	None=0, Quest=1, PRH=2, VH=3, PRH+VH=4, CG=90		If null and REOUTCOME8=90 set as 90; else set as =NR
RERET	Retinopathy	nominal	None=10,	calculated from	Calculated from above 8

	level		Quest=12, HMA<2A=20, HMA≥2A and /or CWS<six=30, CWS≥six and/or IRMA<8A and/or VB/VR/VL (1 quad only)=40, IRMA≥8A and/or VB/VR/VL≥2 quads=50, FVP and/or PDR and /or PRP=60, PDR+HRC=70, PDR+TRD=71, CG-total VH=72, CG=90	above 8 attributes	attributes See appendix for full rule If null and LEOUTCOME8=90 set as 90; else set as NR
RELASER	Any laser; PRP, focal or grid	nominal	None=0, Quest=1, PRP=2, Focal /Mac Grid=3, CG=90		If null and REOUTCOME8=90 set as 90; else set as =NR
REMAC-EX	Presence of macular exudates	nominal	None=0, Quest=1, Present>1DD=2 , Circinate=3, Present.≤1DD and /or Laser=4,		If null and REOUTCOME8=90 set as 90; else set as =NR

			Other (non-DR) =8, CG=90		
RECUPDISC	Cup disc ratio ≥0.7	nominal	No=0, Quest=1, Yes=2, CG=90	If yes = sign of glaucoma ignore analysis	
REOTHER1	None-OTHER1	nominal	Yes=0 (no other disease) No=-1 (other disease present) CG=90	This field records absence of other eye disease Ignore analysis	Where REOTHER2-23 = not 0 or empty field, then enter 0
REOTHER2	Drusen /ARMD-OTHER2	nominal	Yes=1 disease present) No =0	ignore analysis	Where null set as 0
REOTHER3	CNVM-OTHER4	nominal	Yes =2 No=0	ignore analysis	Where null set as 0
REOTHER4	Naevus-OTHER5	nominal	Yes=3 No=0	ignore analysis	Where null set as 0
REOTHER5	Epiretinal Membrane-stopped after 1998	nominal	Yes=4 No=0	ignore analysis	Where null set as 0
REOTHER6	C/BRAO-stopped after 1998	nominal	Yes=5 No=0	ignore analysis	Where null set as 0

REOTHER7	CRVO-OTHER9	nominal	Yes=6 No=0	ignore analysis	Where null set as 0
REOTHER8	BRVO-OTHER10	nominal	Yes=7 No=0	ignore analysis	Where null set as 0
REOTHER9	Other Disc-stopped after 1998	nominal	Yes=8 No=0	ignore analysis	Where null set as 0
REOTHER10	Rhematogenous RD - stopped after 1998	nominal	Yes=9 No=0	ignore analysis	Where null set as 0
REOTHER11	Vitreous Opacity-stopped after 1998	nominal	Yes=10 No=0	ignore analysis	Where null set as 0
REOTHER12	Couldn't Grade throughout -OTHER17jjk[[/	nominal	Yes=90 No=0	ignore analysis	Where null set as 0
REOTHER13	Other-OTHER18	nominal	Yes =11 No = 0	ignore analysis	Where null set as 0
REOTHER14	Age-related Macular Degeneration /Retinal Pigment Epithelial	nominal	Yes =20 No = 0	ignore analysis	Where null set as 0

	Change-OTHER3				
REOTHER15	OTHER6	nominal	21	ignore analysis	Where null set as 0
REOTHER16	Central Retinal Artery Occlusion -OTHER7	nominal	Yes = 22 No = 0	ignore analysis	Where null set as 0
REOTHER17	Retinal Artery Occlusion - OTHER8	nominal	Yes=23 No=0	ignore analysis	Where null set as 0
REOTHER18	Rhematogenous Retinal Detachment-OTHER11	nominal	Yes=24 No = 0	ignore analysis	Where null set as 0
REOTHER19	Myelinated Nerve Fibres - OTHER12	nominal	Yes=25 No=0	ignore analysis	Where null set as 0
REOTHER20	Myopic Degeneration-OTHER13	nominal	Yes=26 No=0	ignore analysis	Where null set as 0
REOTHER21	Tited Disc-OTHER14	nominal	Yes=27 No=0	ignore analysis	Where null set as 0
REOTHER22	Asteroid Hyalosis-OTHER15	nominal	Yes=28 No=0	ignore analysis	Where null set as 0

REOTHER23	Hollenhorst Plaque-OTHER16	nominal	Yes=29 No=0	ignore analysis	Where null set as 0
REDISCSP	Specify Disc	text	Comments	Ignore warehouse	
REVITOPSP	Vitreous Opacity	text	Comments	Ignore warehouse	
RESTEREO	Was stereo photography performed		Yes/No	Ignore analysis	
REMACOED	Assessment of macular oedema	nominal	None=0, Quest=1, present, not CSMO=2, Circinate=3, Present CSMO=4, other=8, CG=90	Ignore analysis Only recorded if stereo present on photos.	
REOUTCOME1	Screen negative	nominal	0	Ignore analysis	
REOUTCOME2	Screen pos – retinopathy level 30 and above	nominal	1	Ignore analysis	
REOUTCOME3	Screen pos-Maculopathy level 3 and above	nominal	2	Ignore analysis	
REOUTCOME5	Screen pos-	nominal	4	Ignore analysis	

	Disc				
REOUTCOME6	Screen pos-diabetic other	nominal	5	Ignore analysis	
REOUTCOME7	Non-diabetic STED	nominal	6	Ignore analysis	
REOUTCOME8	Couldn’t Grade throughout	nominal	90	Ignore analysis	
REOUTCOME9	No photos	nominal	99	Ignore analysis	
RE CALCULATED OUTCOME	composite score for grading outcome for RE	nominal	screen –ve 0, screen +ve retinopathy =1, screen +ve maculopathy =2, screen +ve retinopathy and screen +ve maculopathy =3 ungradable = 90 null field = NR	This field records the outcome of the screening episode If retinopathy or macular exudates is ungradable then eye is ungradable If either retinopathy or maculopathy attributes are null (empty) then eye	calculated from RERET and REMACEX If <RERET> = 10, 12, 20 and <REMACEX> = 0,1,2,8 then set =0 if <RERET> = 30,40,50,60,70,71,72 and <REMACEX> =0,1,2 ,8,90,NR then set as =1 If <RERET> = 10, 12, 20,90,NR and <REMACEX> =3,4 then set as =2

				cannot be categorized for this field and is set as NR	if <RERET> = 30,40,50,60,70,71,72 and <REMACEX> =3,4 then set as =3 Ungradable If <RERET> = 90 or <REMACEX> = 90 then set = 90 Null attributes If <RERET> = NR or <REMACEX> = NR then set = NR
LEField1NAS	Field position nasal field			Ignore warehouse	
LEField2UTQ	Upper temporal quadrant			Ignore warehouse	
LEField3LTQ	Lower temporal quadrant			Ignore warehouse	
LEQual	Clarity and focus			Ignore warehouse	

LEHMA	Hemorrhages and/or Micro Aneurisms	nominal	None=0, Quest=1, <2A=2, ≥2A=3, CG=90		If null and LEOUTCOME8=90 set as 90; else set as NR
LENVD	New vessels Disc	nominal	None=0, Quest=1, <10A=2, ≥10=3, CG=90		If null and LEOUTCOME8=90 set as 90; else set as NR
LECWS8A	Cotton wool Spot	nominal	None=0, Quest=1, <six=2, ≥six=3, CG=90		If null and LEOUTCOME8=90 set as 90; else set as NR
LENVE	New vessels Disc Elsewhere	nominal	None=0, Quest=1, <1/2DA=2, ≥1/2DA(5)=3, CG=90		If null and LEOUTCOME8=90 set as 90; else set as NR
LEVB/VR/VL6A	Venous Beading and /or Venous Reduplication and/or Venous	nominal	None=0, Quest=1, 1Quad=2 2Quads=3 ,		If null and LEOUTCOME8=90 set as 90; else set as NR

	Loop		3Quads= 4 , 4Quads=5 , CG=90		
LEFVP	Fibro vascular Proliferation	nominal	None=0, Quest=1, FPE=2, FPD=3, FPE+FPD=4, TRD=5, CG=90		If null and LEOUTCOME8=90 set as 90; else set as NR
LEIRMA	IntraRetinal Micro vascular Anomaly	nominal	None=0, Quest=1, <8A=2, ≥8A=3, CG=90		If null and LEOUTCOME8=90 set as 90; else set as NR
LEPRHVH	Pre-Retinal Hemorrhage Vitreous Hemorrhage	nominal	None=0, Quest=1, PRH=2, VH=3, PRH+VH=4, CG=90		If null and LEOUTCOME8=90 set as 90; else set as NR
LERET	Retinopathy level	nominal	None=10, Quest=12, HMA<2A=20,	calculated from above 8 attributes	Calculated from above 8 attributes See appendix for full rule

			HMA≥2,CWS<six=30, CWS≥six,IRMA,VB/VR/VI=40, IRMA≥8,VB/VR/VL≥2 quads=50, FVP,PDR±PRP=60, PDR+HRC=70,PDR+TRD=71, CG-total VH=72, CG=90		If null and LEOUTCOME8=90 set as 90; else set as NR
LELASER	Any macular laser; focal or grid	nominal	None=0, Quest=1, PRP=2, Mac Grid=3, CG=90		If null and LEOUTCOME8=90 set as 90; else set as NR
LEMACEX	Presence of macular exudates	nominal	None=0, Quest=1, Present,>1DD=2 , Circinate=3, Present.≤1DD±Laser=4, Other=5, CG=90		If null and LEOUTCOME8=90 set as 90; else set as NR
LECUPDISC	Cup disc ratio ≥0.7	nominal	No=0, Quest=1,	If yes = sign of glaucoma	

			Yes=2, CG=90	ignore analysis	
LEOTHER1	None-OTHER1	nominal	0	This field records absence of other eye disease	Where LEOTHER2-23 = not 0 or empty field, then enter 0
LEOTHER2	Drusen/ARMD-OTHER2	nominal	1	Ignore analysis	Where null set as 0
LEOTHER3	CNVM-OTHER4	nominal	2	Ignore analysis	Where null set as 0
LEOTHER4	Naevus-OTHER5	nominal	3	Ignore analysis	Where null set as 0
LEOTHER5	Epiretinal Membrane-stopped after 1998	nominal	4	Ignore analysis	Where null set as 0
LEOTHER6	C/BRAO-stopped after 1998	nominal	5	Ignore analysis	Where null set as 0
LEOTHER7	CRVO-OTHER9	nominal	6	Ignore analysis	Where null set as 0
LEOTHER8	BRVO-OTHER10	nominal	7	Ignore analysis	Where null set as 0
LEOTHER9	Other Disc-stopped after 1998	nominal	8	Ignore analysis	Where null set as 0
LEOTHER10	Rhematogenous RD - stopped after 1998	nominal	9	Ignore analysis	Where null set as 0

LEOTHER11	Vitreous Opacity-stopped after 1998	nominal	10	Ignore analysis	Where null set as 0
LEOTHER12	CG-OTHER17	nominal	90	Ignore analysis	Where null set as 0
LEOTHER13	Other-OTHER18	nominal	11	Ignore analysis	Where null set as 0
LEOTHER14	Age-related Macular Degeneration /Retinal Pigment Epithelial defect Change-OTHER3	nominal	20	Ignore analysis	Where null set as 0
LEOTHER15	OTHER6	nominal	21	Ignore analysis	Where null set as 0
LEOTHER16	Central Retinal Artery Occlusion-OTHER7	nominal	22	Ignore analysis	Where null set as 0
LEOTHER17	Branch Retinal Artery Occlusion-OTHER8	nominal	23	Ignore analysis	Where null set as 0
LEOTHER18	Rhegmatogenous Retinal	nominal	24	Ignore analysis	Where null set as 0

	Detachment - OTHER11				
LEOTHER19	Myelinated Nerve Fibres - OTHER12	nominal	25	Ignore analysis	Where null set as 0
LEOTHER20	Myopic Degeneration-OTHER13	nominal	26	Ignore analysis	Where null set as 0
LEOTHER21	Tited Disc-OTHER14	nominal	27	Ignore analysis	Where null set as 0
LEOTHER22	Asteroid Hyalosis-OTHER15	nominal	28	Ignore analysis	Where null set as 0
LEOTHER23	Hollenhorst Plaque-OTHER16	nominal	29	Ignore analysis	Where null set as 0
LEDISCSP	Specify Disc	text	Comments	Ignore warehouse	
LEVITOPSP	Vitreous Opacity	text	Comments	Ignore warehouse	
LESTEREO	Was stereo photography performed	nominal	Yes/No	Ignore analysis	
LEMACOED	Assessment of macular oedema	nominal	None=0, Quest=1, present, not CSM=2,	Only recorded if stereo present on photos.	

			Circinate=3, Present CSMO=4, other=8, CG=90		
LEOUTCOME1	Screen negative	nominal	0	Ignore analysis	
LEOUTCOME2	Screen positive -retinopathy	nominal	1	Ignore analysis	
LEOUTCOME3	Screen positive-maculopathy	nominal	2	Ignore analysis	
LEOUTCOME4	Screen pos -VA	nominal	3	Ignore analysis	
LEOUTCOME5	Screen pos-disc	nominal	4	Ignore analysis	
LEOUTCOME6	Screen pos-diabetic other	nominal	5	Ignore analysis	
LEOUTCOME7	Non-diabetic Sight Threatening E Disease	nominal	6	Ignore analysis	
LEOUTCOME8	Couldn't Grade throughout	nominal	90	Ignore analysis	
LEOUTCOME9	No photos	nominal	99	Ignore analysis	
LE CALCULATED OUTCOME	composite score for grading	nominal	screen –ve 0, screen +ve retinopathy	This field records the outcome of	calculated from LERET and LEMACEX

	outcome for RE		=1, screen +ve maculopathy =2, screen +ve retinopathy and screen +ve maculopathy =3 ungradable = 90 null field = NR	the screening episode If retinopathy or macular exudates is ungradable then eye is ungradable If either retinopathy or maculopathy attributes are null (empty) then eye cannot be categorized for this field and is set as NR	If <LERET> = 10, 12, 20 and <LEMACEX> = 0,1,2,8 then set =0 if <LERET> = 30,40,50,60,70,71,72 and <LEMACEX> =0,1,2 ,8,90,NR then set as =1 If <LERET> = 10, 12, 20,90,NR and <LEMACEX> =3,4 then set as =2 if <LERET> = 30,40,50,60,70,71,72 set = 1 and <LEMACEX> =3,4 then set as =3 Ungradable If <LERET> = 90 or <LEMACEX> = 90 then set = 90

					Null attributes If <LERET> = NR or <LEMACEX> = NR then set = NR
BEOUTCOME1	Screen neg	nominal	0	Ignore analysis	
BEOUTCOME2	Screen pos-retinopathy	nominal	1	Ignore analysis	
BEOUTCOME3	Screen pos-maculopathy	nominal	2	Ignore analysis	
BEOUTCOME4	Screen pos-VA	nominal	3	Ignore analysis	
BEOUTCOME5	Screen pos-disc	nominal	4	Ignore analysis	
BEOUTCOME6	Screen pos-Diabetic other	nominal	5	Ignore analysis	
BEOUTCOME7	Non –diabetic STD	nominal	6	Ignore analysis	
BEOUTCOME8	Couldn’t Grade throughout	nominal	90	Ignore analysis	
BEOUTCOME9	No photos	nominal	99	Ignore analysis	
BE CALCULATED OUTCOME	composite score for grading outcome for RE	nominal	screen –ve 0, screen +ve retinopathy =1, screen +ve maculopathy	This field records the outcome of the screening episode by patient	calculated from RE CALCULATED OUTCOME and LE CALCULATED OUTCOME

			=2, screen +ve retinopathy and screen +ve maculopathy =3 ungradable = 90 null field = NR	Highest grade in either eye takes precedence If maculopathy and retinopathy exist in either or both eyes then set as 3 Ungradable Where both attributes = 90 set as 90 Where one field = 90: i) and other is 0 set as 90 ii) and other is 1,2, or 3 set as 1,2,3 respectively Null attributes Where both	If < RE CALCULATED OUTCOME > = 0 and < LE CALCULATED OUTCOME > =0 then set = 0 If (< RE CALCULATED OUTCOME > or < LE CALCULATED OUTCOME > = 1) and (< RE CALCULATED OUTCOME > or < LE CALCULATED OUTCOME > = 0,1) then set = 1 If (< RE CALCULATED OUTCOME > or < LE CALCULATED OUTCOME > = 2) and (< RE CALCULATED OUTCOME > or < LE CALCULATED OUTCOME > = 0,2) then set = 2 If (< RE CALCULATED OUTCOME > or < LE CALCULATED OUTCOME > = 1)

				attributes = null set as not recorded (NR) Where one field = null : i) and other is 0 set as NR ii) and other is 1,2, or 3 set as 1,2,3 respectively	and (< RE CALCULATED OUTCOME > or < LE CALCULATED OUTCOME > = 2) then set = 3 If either < RE CALCULATED OUTCOME > = 3 or < LE CALCULATED OUTCOME > =3 then set = 3 Ungradable If <RE CALCULATED OUTCOME> = 90 and <LE CALCULATED OUTCOME > = 90 then set = 90 If (<RE CALCULATED OUTCOME> or <LE CALCULATED OUTCOME > = 90) and (<RE CALCULATED OUTCOME> or <LE CALCULATED OUTCOME > = 0) then set = 90

					If (<RE CALCULATED OUTCOME> or <LE CALCULATED OUTCOME > = 90) and (<RE CALCULATED OUTCOME> or <LE CALCULATED OUTCOME > = 1) then set = 1 If (<RE CALCULATED OUTCOME> or <LE CALCULATED OUTCOME > = 90) and (<RE CALCULATED OUTCOME> or <LE CALCULATED OUTCOME > = 2) then set = 2 If (<RE CALCULATED OUTCOME> or <LE CALCULATED OUTCOME > = 90) and (<RE CALCULATED OUTCOME> or <LE CALCULATED OUTCOME > = 3) then set = 3

					Null attributes If (<RE CALCULATED OUTCOME> or <LE CALCULATED OUTCOME > = null) and (<RE CALCULATED OUTCOME> or <LE CALCULATED OUTCOME > = 0) then set = NR If (<RE CALCULATED OUTCOME> or <LE CALCULATED OUTCOME > = null) and (<RE CALCULATED OUTCOME> or <LE CALCULATED OUTCOME > = 1) then set = 1 If (<RE CALCULATED OUTCOME> or <LE CALCULATED OUTCOME > = null) and (<RE CALCULATED OUTCOME> or <LE CALCULATED OUTCOME > = 2) then set = 2

					If (<RE CALCULATED OUTCOME> or <LE CALCULATED OUTCOME > = null) and (<RE CALCULATED OUTCOME> or <LE CALCULATED OUTCOME > = 3) then set = 3
DiabStedEye	Auto generated- by rule .Not on Photo form-6/98	nominal	Yes=Y, No=N	Replace autogenerated with value calculated from elsewhere As in BE OUTCOME but level 40 or above for retinopathy and no account of presence of maculopathy and retinopathy	calculated from < BE CALCULATED OUTCOME> and <RERET> and <LERET> If <BE CALCULATED OUTCOME> = 0, 90 or NR then set as N If <BE CALCULATED OUTCOME> = 2,3 then set as Y If <BE CALCULATED OUTCOME> = 1 and (<RERET> = 30 and <LERET>

					= 30) then set as N else set as Y
Action	Recall	nominal	Yes=Y, No=N	Ignore warehouse	

	Inclusive rule									
		HMA	NVD	CWS8A	NVE	VB/VR/VL6A	FVP	IRMA	PRHVH	Laser
	variables	0,1,2,3,90,NR	0,1,2,3,90,NR	0,1,2,3,90,NR	0,1,2,3,90,NR	0,1,2,3,4,5,90,NR	0,1,2,3,4,5,90,NR	0,1,2,3,90,NR	0,1,2,3,4,90,NR	0,1,2,3,90,NR
10		is not = 2 or 3	is not = 2 or 3	is not = 2 or 3	is not = 2 or 3	is not = 2,3,4 or 5	is not = 2,3,4 or 5	is not = 2 or 3	is not = 2,3 or 4	is not = 2,3
12	any field = 1	is not = 2 or 3	is not = 2 or 3	is not = 2 or 3	is not = 2 or 3	is not = 2,3,4 or 5	is not = 2,3,4 or 5	is not = 2 or 3	is not = 2,3 or 4	is not = 2,3
20	HMA =2		is not = 2 or 3	is not = 2 or 3	is not = 2 or 3	is not = 2,3,4 or 5	is not = 2,3,4 or 5	is not = 2 or 3	is not = 2,3 or 4	is not = 2,3
3	HMA =2 and/or		is not		is not =	is not =	is not = 2,3,4	is not = 2	is not = 2,3	is not = 2,3

0	CWS8A =2		= 2 or 3		2 or 3	2,3,4 or 5	or 5	or 3	or 4	
40	CWS8A = 3 and/or VB/VR/VL6A =2 and/or IRMA =2		is not = 2 or 3		is not = 2 or 3		is not = 2,3,4 or 5		is not = 2,3 or 4	is not = 2,3
50	VB/VR/VL6A =3,4,5 and/or IRMA =3		is not = 2 or 3		is not = 2 or 3		is not = 2,3,4 or 5		is not = 2,3 or 4	is not = 2,3
60	NVD =2 and/or NVE =2 and/or FVP = 2,3,4 and/or laser = 2						is not = 5		is not = 2,3 or 4	
70	either NVD =3; or PRHVH =2,3,4 and (NVD =2 and/or NVE =3)						is not = 5			
71	FVP = 5 and (NVD = 2,3 and/or NVE= 2,3 and/or PRHVH = 2,3,4)									
72	PRHVH = 3,4 and all other attributes are = 90									
9	all attributes = 90									

0										
	Inclusive rule									
		HMA	NVD	CWS8A	NVE	VB/VR/VL6A	FVP	IRMA	PRHVH	Laser
	variables	0,1,2,3,90,NR	0,1,2,3,90,NR	0,1,2,3,90,NR	0,1,2,3,90,NR	0,1,2,3,4,5,90,NR	0,1,2,3,4,5,90,NR	0,1,2,3,90,NR	0,1,2,3,4,90,NR	0,1,2,3,90,NR
10		is not = 2 or 3	is not = 2 or 3	is not = 2 or 3	is not = 2 or 3	is not = 2,3,4 or 5	is not = 2,3,4 or 5	is not = 2 or 3	is not = 2,3 or 4	is not = 2,3
12	any field = 1	is not = 2 or 3	is not = 2 or 3	is not = 2 or 3	is not = 2 or 3	is not = 2,3,4 or 5	is not = 2,3,4 or 5	is not = 2 or 3	is not = 2,3 or 4	is not = 2,3
20	HMA =2		is not = 2 or 3	is not = 2 or 3	is not = 2 or 3	is not = 2,3,4 or 5	is not = 2,3,4 or 5	is not = 2 or 3	is not = 2,3 or 4	is not = 2,3
30	HMA =2 and/or CWS8A =2		is not = 2 or 3		is not = 2 or 3	is not = 2,3,4 or 5	is not = 2,3,4 or 5	is not = 2 or 3	is not = 2,3 or 4	is not = 2,3
40	CWS8A = 3 and/or VB/VR/VL6A =2 and/or IRMA =2		is not = 2 or 3		is not = 2 or 3		is not = 2,3,4 or 5		is not = 2,3 or 4	is not = 2,3
50	VB/VR/VL6A =3,4,5 and/or IRMA =3		is not = 2		is not = 2 or 3		is not = 2,3,4 or 5		is not = 2,3 or 4	is not = 2,3

			or 3							
60	NVD =2 and/or NVE =2 and/or FVP = 2,3,4 and/or laser = 2						is not = 5		is not = 2,3 or 4	
70	either NVD =3; or PRHVH =2,3,4 and (NVD =2 and/or NVE =3)						is not = 5			
71	FVP = 5 and (NVD = 2,3 and/or NVE= 2,3 and/or PRHVH = 2,3,4)									
72	PRHVH = 3,4 and all other attributes are = 90									
90	all attributes = 90									

Table A1.2 : Schema for General Dataset

Liverpool Diabetes Eye Study Schema Description: Database DIAB Dataset DiabEyeGeneral Version date 03/03/10						
Data Label	Description	Data type	Value	Unit	Narrative	Logic Rules
StudyIDNo	ID number	Number	Xxxxx			
NHS No	ID number (key patient identifier for Study Tables)	Number	10 digits no gaps			
Examination Date	Date	Date/ Time	dd/mm/yy			
Age at Exam	Age	integer	Xxx	Years		Calculated from <PatDOB> in DiabPatientDetails and <Examination Date>. Ignore entered data
Visual Acuity Right Best	Visual acuity recorded on Bailey-Lovie chart	nominal	6/5=0,6/6=1,6/9=2,6/12=3,6/60=7 <6/60=8,NPL=9 – now uses Bailey Lovie	logMAR	-0.20-+1.00 (+2.00,+3.00,+4.00,+5.00)	Where VA also exists in DiabBiomicroscopy <VARE> in visit related to this episode replace by the data from DiabBiomicroscopy <VARE>

Visual Acuity Left Best	Visual acuity recorded on Bailey-Lovie chart	nominal	6/5=0,6/6=1,6/9=2,6/12=3,6/60=7 <6/60=8,NPL=9	logMAR	-0.20-+1.00 (+2.00,+3.00,+4.00,+5.00)	Where VA also exists in DiabBiomicroscopy <VALE> in visit related to this episode replace by the data from DiabBiomicroscopy <VARE
See GP Regularly	Do you see your GP regularly for diabetes care?	nominal	no=1,yes=2,Don't know=9		Is the patient currently under active review by the GP? If <see GP regularly> is yes and <lastSeeGP> is within the last 12 months then implies under current GP care. If <see GP regularly> = yes and <lastSeeGP> is more than 12 months then implies not under current GP care.	Calculated from <See GP Regularly> and <LastSeeGP> If <See GP Regularly>=1 and <Last see GP>=1 or 2 then set <See GP Regularly>=2 If <See GP Regularly>=2 and <Last see GP>=1 or 2 then set <See GP Regularly>=2 If <Last see GP>=3

						then set <See GP Regularly>=1
						Assignment of missing data rules
						if <See GP Regularly>= null and <Last See GP>=1 or 2 then <see GP regularly>=2.
						if <See GP Regularly>= null and <Last See GP>=3 then <see GP regularly>=1.
						if <See GP Regularly>= null and <Last See GP>=null then <see GP regularly>=NR
						Data cleansing completed 21.10.09

						DMB
Attended Diabetes Clinic	Have you attended a hospital clinic for diabetes in the last 2 years?	nominal	No=1,yes=2		Is the patient currently under the care of a diabetologist? If <att diab clinic> is no implies not under care of a diabetologist regardless of data in <further diab appt> If <att diab clinic> is yes and <further diab appt> = yes then implies under current diabetologist care If <att diab clinic> is yes and <further diab appt> is no then implies not under current diabetologist care	Calculated from <Attended Diabetes Clinic> and <Further diab appointment> If <Attended Diabetes Clinic> =1 the set <att diab clinic> as = 1 If <Attended Diabetes Clinic> =2 and <further diab appt> = 2 then set <att diab clinic> =2 If <Attended Diabetes Clinic> =2 and <further diab appt> =1 then set <att diab clinic> =1 Assignment of empty attributes If <Attended Diabetes Clinic> = null then set

						<Attended Diabetes Clinic> = NR Data cleansing completed 21.10.09 DMB
Further Diab Appointment	If yes , have you another appointment to be seen	nominal	No=1,yes=2			If <further diab appointment> = null and <Attended Diabetes Clinic> = null then set <further diab appointment> = NR If <further diab appointment> = null and <Attended Diabetes Clinic> = 1 then set <further diab appointment> = NA If <further diab appointment> = null and <Attended Diabetes

						Clinic> = 2 then set <further diab appointment> = NR Data cleansing completed 19.10.09 DMB
Diab Doctor Name	Doctor Name	text			Ignore warehouse	
Diab Hospital Address	Hospital Address	text			Ignore warehouse	
Attended Eye Dept Att<2yrs?	Have you attended St. Paul eye Hospital or any other eye department in the last 2 years?	nominal	No=1,yes=2,Don't know=9		Is the patient currently under an ophthalmologist? If <attended eye dept> is no implies not under ophthalmic care regardless of data in <further eye appointment> (will be either 1 – no or not applicable code) If <attended eye dept> is Yes, and patient has a further eye appt implies	Calculated from <Attended Eye Dept> and <Further Eye Appointment> If <Attended Eye Dept> =1 the set <Attended Eye Dept> as = 1 If <Attended Eye Dept> =2 and <Further Eye Appointment> = 2 then set <Attended Eye Dept> =2

					currently under an ophthalmologist. If <attended eye dept> is Yes, and patient has no further eye appt implies not currently under an ophthalmologist.	If <Attended Eye Dept> =2 and <Further Eye Appointment > =1 then set <Attended Eye Dept>=1 Assignment of empty attributes If <Attended Eye Dept> = null and <Further Eye Appointment > = null then set <Attended Eye Dept> = NR Data cleansing completed 21.10.09
Further Eye Appointment	If yes do you have another appointment to be seen?	nominal	No=1,yes=2,Don't know=9		Used as a support for < Attended Eye Dept>	If <Further Eye Appointment> = null and <Attended Eye Dept> = null then set <Further Eye Appointment> = NR

						If <Further Eye Appointment> = null and <Attended Eye Dept > = 1 then set <Further Eye Appointment> = NA If <Further Eye Appointment> = null and <Attended Eye Dept > = 2 then set <Further Eye Appointment> = NR Data cleansing completed 19.10.09 DMB
diEyeDeptDocName	Department Doctor’s name	text			Ignore warehouse	
diEyeDeptDocAddr	Department Doctor’s Address	text			Ignore warehouse	
WeakEye	Have you ever been told that you	nominal	no=1,yes=2,null =0		Ignore analysis	

	have a weak or lazy eye?					
WeakEyeLeft	Have you ever been told that you have a weak Eye left ?	nominal	Yes=1,no=0		Ignore analysis	
WeakEyeRight	Have you ever been told that you have a weak eye right?	nominal	Yes=1,no=0		Ignore analysis	
Cataracts	Have you ever been told that you have Cataract eye?	nominal	no=1,yes=2,null =0		Ignore analysis	
CataractEyeLeft	Have you ever been told that you have a Cataract left eye?	nominal	Yes=1,no=0		Ignore analysis	If = null set as no = 0
CataractEyeRight	Have you ever been told that you have a Cataract right eye?	nominal	Yes=1,no=0		Ignore analysis	If = null set as no = 0
Glaucoma	Have you ever been told that you have Glaucoma ?	nominal	no=1,yes=2,null =0		Ignore analysis	

GlaucomaEyeLeft	Have you ever been told that you have a Glaucoma left eye?	nominal	Yes=1,no=0		Ignore analysis	
GlaucomaEyeRight	Have you ever been told that you have a Glaucoma right eye?	nominal	Yes=1,no=0		Ignore analysis	
Other Problem	Have you ever been told that you have a Other eye problem?	Nominal	No=1,yes=2,Don't know=9		Ignore analysis	
OtherEyeProbLeft	Have you ever been told that you have Other eye problem left?	Nominal	Yes=1,no=0		Ignore analysis	
OtherEyeProbright	Have you ever been told that you have Other eye problem right?	Nominal	Yes=1,no=0		Ignore analysis	
OtherEyeSpecify	Describe any other eye problems	text			Ignore warehouse	
Years Diabetic(months)	How long have you been diabetic?	integer	Holds years Diabetic(06/98)-	months	Ignore warehouse	Ignore entered data

			diYears Band converted into this field at 06/98			
Calculated diabetes duration					Duration of diabetes calculated for each episode in years	Calculate from <age at exam> and <Calculated age at diagnosis> If missing data use <Duration: years> in Risk factor table.
Calculated age at diagnosis					The age at diagnosis taken from the first visit with available data Used to calculate type of diabetes	Calculate from <diAgeDiag> at first visit with available data
diYearsSource	How that data was calculated	Nominal	A=Accurate, D=Derived		Ignore analysis	
Years Diabetic Band		Nominal	<1 yr=0, 1-5 yr=1, 6-10 yr=2, 11-15 yr=3, 16-20 yr =4, 21+ yr =5, Don’t know =9,		Banded by screener based on years diabetic Ignore analysis	

			fieldname previously diYears – 06/98			
Present Treatment	What is your present treatment?	Nominal	Diet alone =1, diet and tablets =2, diet and insulin =3, tablets and insulin =4		Aims to determine whether diet, tablet or insulin controlled to determine whether patient has type 1 or type 2 DM if insulin requiring to determine point of treatment change Used to calculate type of diabetes	Where null use <diCurrTreat> in related visit in DiabBiomicroscopy within same episode (≤91 days). Else use last available observation from DiabEyeGeneral. If still null set as NR.
diInsTab	If you are on insulin, did you have a period of time on tablets before starting insulin	Nominal	No=1, yes<1 year=2, yes ≥1year=3, Don’t know=9 – if on insulin was patient on tablets – 06/98		Ignore analysis Used to calculate type of diabetes	
Calculated Diabetes Type			Type 1 = 1 Type 2 diet controlled = 2 Type 2 oral		1 = <30 years old and currently on insulin; ≥ 30 < 40 years old on insulin and <12 months tablets	Calculated from <diInsTab>, <Present Treatment>, <dbPastTreatment> ,

			controlled = 3 Type 2 insulin requiring = 4 Unclassifiable =9 Empty field = not recorded		2 = ≥ 40 years old currently on diet control 3 = ≥ 40 years old on oral control 4 = ≥ 30 currently on insulin and ≥ 12 months tabs 9 = <30 not on insulin; ≥ 40 on insulin and <12 months tablets <diInsTab> takes precedence over <dbPastTreatment>; not completed prior to 02/98 Insulin taked precedence over tablets and diet Tablets takes precedence over diet	<calculated age at diagnosis> Appendix 1 to DiabEyeGeneral schema sets out logic rule in full Where missing data use <Calculated Diabetes Type> from Risk Factors table
LastSeeGP	When did you last see your GP in your practice for anything(any GP within practice allowed)?	Nominal	<6/12=1, 6-12/12=2, >12/12=3 , 0=missing data		Ignore analysis Field required for input into <see GP regularly> Data cleansing completed 19.10.09 DMB	If <LastSeeGP> = null then set <LastSeeGP> = NR

diAgeDiag	How old were you when your diabetes was diagnosed ?	Integer	0=unknown, else recorded in years	years	Ignore analysis Used to calculate the age at diagnosis	Where there is more than one patient episode take duration data from earliest visit
WhyVisit		Nominal	No=1, yes=2, Don’t know=9		Ignore warehouse	
dbHPB1	Have you ever been told that you have High blood pressure?	Nominal	No=1,yes=2		Ignore warehouse	
dbHPB2	Have you ever been told that you have Foot Ulcers ?	Nominal	No=1,yes=2		Ignore warehouse	
dbHPB3	Have you ever been told that you have Circulatory problems?	Nominal	No=1,yes=2		Ignore warehouse	
dbHPB4	Have you ever been told that you have Nerve problems?	Nominal	No=1,yes=2		Ignore warehouse	
dbHPB5	Have you ever	Nominal	No=1,yes=2		Ignore warehouse	

	been told that you have Kidney problems?					
ndbOthPrb	Do you have any of the following diabetic problem: High blood pressure/nerve problem/foot ulcers /kidney problems/circulatory problems	Nominal	No=1,yes=2,Don't know=9		Ignore warehouse	
ndbOthPrbS	Describe any other diabetic problems	text			Ignore warehouse	
dbPastTreat	What was the Past treatment?	nominal	Diet=1, diet then tablets =2,Diet then insulin=3,tablets =4,tables<yr then Insulin=5,Tablets >1yr then insulin=6,tablets then		Ignore analysis Used to calculate type of diabetes	

			diet=7,insulin then tablets=8,insulin =9,Don’t know=10			
diSmoke	Do you smoke or have you smoked any time in the last 10 years?	Nominal	No=1, Yes=2, Don’t know=9 smoke during last 10 years 06/98		Started being collected in 1998 Not applicable before 01/06/1998	if date = <01/06/1998 set empty field = NA if no then set as no if yes set as yes if don’t know set as no if date = ≥01/06/1998 then set empty field as NR
diFamGlau	Is there any family history of Glaucoma?	Nominal	No=1,yes=2,don’t know=9-Family history of Glaucoma – 06/98		Ignore warehouse	
New Patient	New Patient	Nominal	Yes=Y, No=N		Ignore warehouse	

Table A1.3 – Schema for DiabBiomicroscopy dataset

Liverpool Diabetes Eye Study Schema Description: Database DIAB Dataset: DiabBiomicroscopy Version date: 03/03/10					
Data Label	Description	Data type	Value	Narrative	Logic Rules
StudyIDNo	ID number (key patient identifier for Study Tables)	Number	Xxxxx		
ExamDate	Date	Date/ Time	dd/mm/yy		
VARE	Visual acuity recorded on Bailey-Lovie chart	nominal	6/5=0,6/6=1,6/9=2,6/12=3,6/60=7 <6/60=8,NPL=9 – now uses Bailey Lovie	LogMAR -0.20-+1.00 (+2.00,+3.00,+4.00, +5.00) Where VA exists in one episode in Diab Gen AND bio assume the bio VA likely to be more accurate	Where VA data also exists in DiabEyeGeneral visit related to this episode, replace DiabEyeGeneral <Visual Acuity Right Best> with data from this field
VALE	Visual acuity recorded on Bailey-Lovie chart	nominal	6/5=0,6/6=1,6/9=2,6/12=3,6/60=7 <6/60=8,NPL=9	LogMAR -0.20-+1.00 (+2.00,+3.00,+4.00, +5.00)	Where VA data also exists in DiabEyeGeneral visit related to this episode, replace DiabEyeGeneral <Visual Acuity Left Best> with data from this field
B.P.	Blood pressure	number			

Pulse Rate	Pulse rate	number		Ignore warehouse	
IOPRE	Intraocular pressure IOP R	Number		<21 is normal and >21 is abnormal Ignore analysis	
IOPLE	Intraocular pressure IOP L	Number		<21 is normal and >21 is abnormal Ignore analysis	
diYears	How long have you been diabetic?	Nominal	<1 yr=0, 1-5 yr=1, 6-10 yr=2, 11-15 yr=3, 16-20 yr =4, 21+ yr =5, Don't know =9, fieldname previously diYears – 06/98	Ignore analysis	
diCurrTreat	What is your Present treatment?	Nominal	Diet alone =1, diet and tablets =2, diet and insulin =3, tablets and insulin =4	Use this field to populate missing data in DiabGen. Some patients attend for bio without having had photo	Where <Present Treatment> in Daib Gen is null or 0 use this field to populate if visit is ≤91 days previously

				attendance	
diHospClinAtt	Have you attended a hospital clinic for diabetes in the last 2 years?	nominal	No=1,yes=2,Don’t know=9	Ignore warehouse	
diEyeDeptAtt	Have you attended St.Paul eye Hospital or any other eye department in the last 2 years?	nominal	No=1,yes=2,Don’t know=9	Ignore warehouse	
FamHistGlaucoma	Is there any family history of Glaucoma?	Nominal	No=1,yes=2,don’t know=9-Family history of Glaucoma – 06/98	Ignore warehouse	
CornealOpacRE	Corneal opacity(right)	nominal	no=1,yes=2		
CornealOpacLE	Corneal opacity(left)	nominal	no=1,yes=2		
CataractRE	Cataract (Right	nominal	no=1,yes=2		

	eye)				
CataractLE	Cataract (Left eye)	nominal	no=1,yes=2		
CatRENO	Cataract Nuclear opacity(right eye)	nominal	1,2,3,4,5,6		If null and <CataractRE=1 then set as NA; else set as NR
CatRENC	Cataract Nuclear colour (right eye)	nominal	1,2,3,4,5,6		If null and <CataractRE=1 then set as NA; else set as NR
CatREC	Cataract cortical(right eye)	nominal	1,2,3,4,5,6		If null and <CataractRE=1 then set as NA; else set as NR
CatREP	Cataract Nuclear opacity(right eye)	nominal	1,2,3,4,5,6		If null and <CataractRE=1 then set as NA; else set as NR
CatLENO	Cataract Nuclear opacity(left eye)	nominal	1,2,3,4,5,6		If null and <CataractLE=1 then set as NA; else set as NR
CatLENC	Cataract Nuclear colour (left eye)	nominal	1,2,3,4,5,6		If null and <CataractLE=1 then set as NA; else set as NR
CatLEC	Cataract cortical(left eye)	nominal	1,2,3,4,5,6		If null and <CataractLE=1 then set as NA; else set as NR
CatLEP	Cataract Nuclear opacity(left eye)	nominal	1,2,3,4,5,6		If null and <CataractLE=1 then set as NA; else set as NR
Calculated cataract sufficient to interfere with photography Right eye			1 = yes 2 = no	calculated from <CatRENO>, <CatRENC>, <CatREC>, <CatREP>	Cataract is sufficient to interfere with photography if <CatRENO> = 4, 5 or 6; and / or <CatRENC> = 4, 5 or 6; and / or <CatREC> = 4,5 or 6; and / or <CatREP> = 1,2,3,4,5 or 6.

Calculated cataract sufficient to interfere with photography Left eye			1 = yes 2 = no	calculated from <CatLENO>, <CatLENC>, <CatLEC>, <CatLEP>	Cataract is sufficient to interfere with photography if <CatLENO> = 4, 5 or 6; and / or <CatLENC> = 4, 5 or 6; and / or <CatLEC> = 4,5 or 6; and / or <CatLEP> = 1,2,3,4,5 or 6.
VitreousOpacityRE	Vitreous Opacity Right Eye	nominal	no=1,yes=2		
VitreousOpacityLE	Vitreus Opacity Left Eye	nominal	no=1,yes=2		
CuppedDiscRE	Clinician opinion	nominal	no=1, yes=2	Ignore analysis	
CuppedDiscLE	Clinician opinion	nominal	no=1,yes=2	Ignore analysis	
PseudophakiaRE	Had cataract removed IOL(Right)	nominal	no=1,yes=2	Ignore analysis	
PseudophakiaLE	Had cataract removed IOL(left)	nominal	no=1,yes=2	Ignore analysis	
PostSynechiaeRE	Posterior Synechiae-adhesion between lens and iris (right)	nominal	no=1,yes=2	Ignore analysis	
PostSynechiaeLE	Posterior Synechiae-adhesion between	nominal	no=1,yes=2	Ignore analysis	

	lens and iris(left)				
SmallPupilRE	Small Pupil e.g. autonomic neuropathy(right)	nominal	no=1,yes=2	Ignore analysis	
SmallPupilLE	Small Pupil e.g. autonomic neuropathy(left)	nominal	no=1,yes=2	Ignore analysis	
OtherRE	An unspecified reason exists(right eye)	nominal	no=1,yes=2	Ignore analysis	
OtherLE	An unspecified reason exists(left eye)	nominal	no=1,yes=2	Ignore analysis	
OtherRESpecify	Other right eye specify	nominal	Comments	Ignore analysis	
OtherLESpecify	Other left eye specify	nominal	Comments	Ignore analysis	
ReturnRe	Other reason for return	nominal	No = 1 Yes = 2	Ignore warehouse	
Return Sp	Specify other reason for return	text	Comments	Ignore warehouse	
REHMA	Haemorrhages and/or micro aneurysms	nominal	None=0, Quest=1, <2A=2,		If null and REOUTCOME8=90 set as 90; else set as NR

			≥2A=3, CG=90		
RENVD	New Vessels Disc	nominal	None=0, Quest=1, <10A=2, ≥10A=3, CG=90		If null and REOUTCOME8=90 set as 90; else set as NR
RECWS8A	Cotton wool spot	nominal	None=0, Quest=1, <six=2, ≥six=3, CG=90		If null and REOUTCOME8=90 set as 90; else set as NR
RENVE	New vessels elsewhere	nominal	None=0, Quest=1, <1/2DA=2, ≥1/2DA=3, CG=90		If null and REOUTCOME8=90 set as 90; else set as NR
REVBVRVL6A	Venous Beading and /or Venous Reduplication and/or Venous Loop	nominal	None=0, Quest=1, 1Quad=2 2Quads=3 3Quads= 4 4Quads=5 CG=90		If null and REOUTCOME8=90 set as 90; else set as NR

REFVP	Fibrovascular proliferation	nominal	None=0, Quest=1, FPE=2, FPD=3, FPE+FPD=4, TRD=5, CG=90		If null and REOUTCOME8=90 set as 90; else set as NR
REIRMA	Intraretinal microvascular abnormality	nominal	None=0, Quest=1, <8A=2, ≥8A=3, CG=90		If null and REOUTCOME8=90 set as 90; else set as NR
REPRH VH	Pre-Retinal Haemorrhage Vitreous Hemorrhage	nominal	None=0, Quest=1, PRH=2, VH=3, PRH+VH=4, CG=90		If null and REOUTCOME8=90 set as 90; else set as NR
RERET	Retinopathy level	nominal	None=10, Quest=12, HMA<2A=20, HMA≥2A and /or CWS<six=30, CWS≥six and/or	calculated from above 8 attributes	Calculated from above 8 attributes See appendix for full rule If null and REOUTCOME8=90 set as 90; else set as NR

			IRMA<8A and/or VB/VR/VL (1 quad only)=40, IRMA≥8A and/or VB/VR/VL≥2 quads=50, FVP and/or PDR and /or PRP=60, PDR+HRC=70, PDR+TRD=71, CG-total VH=72, CG=90		
RELASER	Any laser; PRP, focal or grid	nominal	None=0, Quest=1, PRP=2, Focal /Mac Grid=3, CG=90		If null and REOUTCOME8=90 set as 90; else set as NR
REMAC-EX	Presence of macular exudates	nominal	None=0, Quest=1, Present>1DD=2 , Circinate=3, Present.≤1DD and /or Laser=4, Other (non-DR) =8,		If null and REOUTCOME8=90 set as 90; else set as NR

			CG=90		
RECUPDISC	Cup disc ratio ≥0.7	nominal	No=0, Quest=1, Yes=2, CG=90		If null and REOUTCOME8=90 set as 90; else set as NR
REOTHER1	None-OTHER1	nominal	0	ignore analysis	Where REOTHER2-23 = not 0 or empty field, then enter 0
REOTHER2	Drusen /ARMD-OTHER2	nominal	1	ignore analysis	Where empty field set as 0
REOTHER3	CNVM-OTHER4	nominal	2	ignore analysis	Where empty field set as 0
REOTHER4	Naevus-OTHER5	nominal	3	ignore analysis	Where empty field set as 0
REOTHER5	Epiretinal Membrane-stopped	nominal	4	ignore analysis	Where empty field set as 0
REOTHER6	C/BRAO-stopped	nominal	5	ignore analysis	Where empty field set as 0
REOTHER7	CRVO-OTHER9	nominal	6	ignore analysis	Where empty field set as 0
REOTHER8	BRVO-OTHER10	nominal	7	ignore analysis	Where empty field set as 0
REOTHER9	Other Disc-stopped	nominal	8	ignore analysis	Where empty field set as 0
REOTHER10	Rhematogenous RD –stopped	nominal	9	ignore analysis	Where empty field set as 0
REOTHER11	Vitreous Opacity-stopped	nominal	10	ignore analysis	Where empty field set as 0
REOTHER12	Couldn’t Grade	nominal	90	ignore analysis	Where empty field set as 0

	throughout - OTHER17				
REOTHER13	Other-OTHER18	nominal	11	ignore analysis	Where empty field set as 0
REOTHER14	Age-related Macular Degeneration /Retinal Pigment Epithelial Change-OTHER3	nominal	20	ignore analysis	Where empty field set as 0
REOTHER15	OTHER6	nominal	21	ignore analysis	Where empty field set as 0
REOTHER16	Central Retinal Artery Occlusion - OTHER7	nominal	22	ignore analysis	Where empty field set as 0
REOTHER17	Retinal Artery Occlusion - OTHER8	nominal	23	ignore analysis	Where empty field set as 0
REOTHER18	Rhematogenous Retinal Detachment-OTHER11	nominal	24	ignore analysis	Where empty field set as 0
REOTHER19	Myelinated Nerve Fibres -OTHER12	nominal	25	ignore analysis	Where empty field set as 0
REOTHER20	Myopic Degeneration-	nominal	26	ignore analysis	Where empty field set as 0

	OTHER13				
REOTHER21	Tited Disc-OTHER14	nominal	27	ignore analysis	Where empty field set as 0
REOTHER22	Asteroid Hyalosis-OTHER15	nominal	28	ignore analysis	Where empty field set as 0
REOTHER23	Hollenhorst Plaque-OTHER16	nominal	29	ignore analysis	Where empty field set as 0
REDISCSP	Specify Disc	text	Comments	Ignore analysis	
REVITOPSP	Vitreous Opacity	text	comments	Ignore analysis	
REMACOED	Assessment of macular oedema	nominal	None=0, Quest=1, present, not CSMO=2, Circinate=3, Present CSMO=4, other=8, CG=90		If null and REOUTCOME8=90 set as 90; else set as NR
REOUTCOME1	Screen negative	nominal	0	Ignore analysis	
REOUTCOME2	Screen pos – retinopathy level 30 and above	nominal	1	Ignore analysis	
REOUTCOME3	Screen pos-Maculopathy level 3 and above	nominal	2	Ignore analysis	

REOUTCOME4	Screen pos –VA	nominal	3	Ignore analysis	
REOUTCOME5	Screen pos- disc	nominal	4	Ignore analysis	
REOUTCOME6	Screen pos- diabetic other	nominal	5	Ignore analysis	
REOUTCOME7	N-diabetic STED	nominal	6	Ignore analysis	
REOUTCOME8	Couldn’t Grade throughout	nominal	90	Ignore analysis	
REOUTCOME9	No photos	nominal	99	Ignore analysis	
RE CALCULATED OUTCOME	composite score for grading outcome for RE	nominal	screen –ve 0, screen +ve retinopathy =1, screen +ve maculopathy =2, screen +ve retinopathy and screen +ve maculopathy =3 ungradable = 90 null field = NR	This field records the outcome of the screening episode If retinopathy or macular exudates or macular oedema is ungradable then eye is ungradable If either retinopathy or maculopathy attributes are null (empty) then eye cannot be categorized for this	calculated from RERET and REMACEX and REMACOED If <RERET> = 10, 12, 20 and <REMACEX> = 0,1,2,8 and <REMACOED> = 0,1,2,8 then set =0 if <RERET> = 30,40,50,60,70,71,72 and <REMACEX> =0,1,2 ,8,90,NR and <REMACOED> =0,1,2 ,8,90,NR then set as =1 If <RERET> = 10, 12, 20,90,NR and (<REMACEX> =3,4 or <REMACOED> =3,4) then set as =2

				field and is set as NR	if <RERET> = 30,40,50,60,70,71,72 and (<REMACEX> =3,4 or <REMACOED> =3,4) then set as =3 Ungradable If <RERET> = 90 or <REMACEX> = 90 or <REMACOED> =90 then set = 90 Null attributes If <RERET> = NR or <REMACEX> = NR or <REMACOED> =NR then set = NR
LEHMA	Hemorrhages and/or Micro Aneurisms	nominal	None=0, Quest=1, <2A=2, ≥2A=3, CG=90		If null and LEOUTCOME8=90 set as 90; else set as NR
LENVD	New vessels Disc	nominal	None=0, Quest=1, <10A=2, ≥10=3,		If null and LEOUTCOME8=90 set as 90; else set as NR

			CG=90		
LECWS8A	Cotton wool Spot	nominal	None=0, Quest=1, <six=2, ≥six=3, CG=90		If null and LEOUTCOME8=90 set as 90; else set as NR
LENVE	New vessels Disc Elsewhere	nominal	None=0, Quest=1, <1/2DA=2, ≥1/2DA(5)=3, CG=90		If null and LEOUTCOME8=90 set as 90; else set as NR
LEVB/VR/VL6A	Venous Beading and /or Venous Reduplication and/or Venous Loop	nominal	None=0, Quest=1, 1Quad=2 2Quads=3 , 3Quads= 4 , 4Quads=5 , CG=90		If null and LEOUTCOME8=90 set as 90; else set as NR
LEFVP	Fibro vascular Proliferation	nominal	None=0, Quest=1, FPE=2, FPD=3, FPE+FPD=4, TRD=5,		If null and LEOUTCOME8=90 set as 90; else set as NR

			CG=90		
LEIRMA	IntraRetinal Micro vascular Anomaly	nominal	None=0, Quest=1, <8A=2, ≥8A=3, CG=90		If null and LEOUTCOME8=90 set as 90; else set as NR
LEPRHVH	Pre-Retinal Hemorrhage Vitreous Hemorrhage	nominal	None=0, Quest=1, PRH=2, VH=3, PRH+VH=4, CG=90		If null and LEOUTCOME8=90 set as 90; else set as NR
LERET	Retinopathy level	nominal	None=10, Quest=12, HMA<2A=20, HMA≥2,CWS<six=30, CWS≥six,IRMA,VB/VR/VI=40, IRMA≥8,VB/VR/VL≥2 quads=50, FVP,PDR±PRP=60, PDR+HRC=70,PDR+TRD=71,	calculated from above 8 attributes	Calculated from above 8 attributes See appendix for full rule If null and LEOUTCOME8=90 set as 90; else set as NR

			CG-total VH=72, CG=90		
LELASER	Any macular laser; focal or grid	nominal	None=0, Quest=1, PRP=2, Mac Grid=3, CG=90		If null and LEOUTCOME8=90 set as 90; else set as NR
LEMACEX	Presence of macular exudates	nominal	None=0, Quest=1, Present,>1DD=2 , Circinate=3, Present.≤1DD±Laser=4, Other=8, CG=90		If null and LEOUTCOME8=90 set as 90; else set as NR
LECUPDISC	Cup disc ratio ≥0.7	nominal	No=0, Quest=1, Yes=2, CG=90	If yes = sign of glaucoma ignore analysis	
LEOTHER1	None-OTHER1	nominal	0	Ignore analysis This field records absence of other eye disease	Where LEOTHER2-23 = not 0 or empty field, then enter 0
LEOTHER2	Drusen/ARMD-	nominal	1	ignore analysis	Where empty field set as 0

	OTHER2				
LEOTHER3	CNVM-OTHER4	nominal	2	ignore analysis	Where empty field set as 0
LEOTHER4	Naevus-OTHER5	nominal	3	ignore analysis	Where empty field set as 0
LEOTHER5	Epiretinal Membrane-stopped	nominal	4	ignore analysis	Where empty field set as 0
LEOTHER6	C/BRAO-stopped	nominal	5	ignore analysis	Where empty field set as 0
LEOTHER7	CRVO-OTHER9	nominal	6	ignore analysis	Where empty field set as 0
LEOTHER8	BRVO-OTHER10	nominal	7	ignore analysis	Where empty field set as 0
LEOTHER9	Other Disc-stopped	nominal	8	ignore analysis	Where empty field set as 0
LEOTHER10	Rhematogenous RD –stopped	nominal	9	ignore analysis	Where empty field set as 0
LEOTHER11	Vitreous Opacity-stopped	nominal	10	ignore analysis	Where empty field set as 0
LEOTHER12	CG-OTHER17	nominal	90	ignore analysis	Where empty field set as 0
LEOTHER13	Other-OTHER18	nominal	11	ignore analysis	Where empty field set as 0
LEOTHER14	Age-related Macular Degeneration /Retinal Pigment Epithelial defect Change-OTHER3	nominal	20	ignore analysis	Where empty field set as 0

LEOTHER15	OTHER6	nominal	21	ignore analysis	Where empty field set as 0
LEOTHER16	Central Retinal Artery Occlusion-OTHER7	nominal	22	ignore analysis	Where empty field set as 0
LEOTHER17	Branch Retinal Artery Occlusion-OTHER8	nominal	23	ignore analysis	Where empty field set as 0
LEOTHER18	Rhegmatogenous Retinal Detachment -OTHER11	nominal	24	ignore analysis	Where empty field set as 0
LEOTHER19	Myelinated Nerve Fibres -OTHER12	nominal	25	ignore analysis	Where empty field set as 0
LEOTHER20	Myopic Degeneration-OTHER13	nominal	26	ignore analysis	Where empty field set as 0
LEOTHER21	Tited Disc-OTHER14	nominal	27	ignore analysis	Where empty field set as 0
LEOTHER22	Asteroid Hyalosis-OTHER15	nominal	28	ignore analysis	Where empty field set as 0
LEOTHER23	Hollenhorst Plaque-OTHER16	nominal	29	ignore analysis	Where empty field set as 0
LEDISCSP	Specify Disc	text	comments	Ignore warehouse	
LEVITOPSP	Vitreous Opacity	text	comments	Ignore warehouse	

LEMACOED	Assessment of macular oedema	nominal	None=0, Quest=1, present, not CSMO=2, Circinate=3, Present CSMO=4, other=8, CG=90		If null and LEOUTCOME8=90 set as 90; else set as NR
LE CALCULATED OUTCOME	composite score for grading outcome for RE	nominal	screen –ve 0, screen +ve retinopathy =1, screen +ve maculopathy =2, screen +ve retinopathy and screen +ve maculopathy =3 ungradable = 90 null field = NR	This field records the outcome of the screening episode If retinopathy or macular exudates or macular oedema is ungradable then eye is ungradable If either retinopathy or maculopathy attributes are null (empty) then eye cannot be categorized for this	calculated from LERET and LEMACEX and LEMACOED If <LERET> = 10, 12, 20 and <LEMACEX> = 0,1,2,8 and <LEMACOED> = 0,1,2,8 then set =0 if <LERET> = 30,40,50,60,70,71,72 and <LEMACEX> =0,1,2 ,8,90,NR and <LEMACOED> =0,1,2 ,8,90,NR then set as =1 If <LERET> = 10, 12, 20,90,NR and (<LEMACEX> =3,4 or <LEMACOED> =3,4) then set as =2

				field and is set as NR	if <LERET> = 30,40,50,60,70,71,72 and (<LEMACEX> =3,4 or <LEMACOED> =3,4) then set as =3 Ungradable If <LERET> = 90 or <LEMACEX> = 90 or <LEMACOED> =90 then set = 90 Null attributes If <LERET> = NR or <LEMACEX> = NR or <LEMACOED> =NR then set = NR
LEOUTCOME1	Screen negative	nominal	0	Ignore analysis	
LEOUTCOME2	Screen positive – retinopathy	nominal	1	Ignore analysis	
LEOUTCOME3	Screen positive-Maculopathy	nominal	2	Ignore analysis	
LEOUTCOME4	Screen pos -VA	nominal	3	Ignore analysis	
LEOUTCOME5	Screen pos-Disc	nominal	4	Ignore analysis	
LEOUTCOME6	Screen pos-	nominal	5	Ignore analysis	

	diabetic other				
LEOUTCOME7	Non-diabetic Sight Threatening E Disease	nominal	6	Ignore analysis	
LEOUTCOME8	Couldn’t Grade throughout	nominal	90	Ignore analysis	
LEOUTCOME9	No photos	nominal	99	Ignore analysis	
BEOUTCOME1	Screen neg	nominal	0	Ignore analysis	
BEOUTCOME2	Screen pos-retinopathy	nominal	1	Ignore analysis	
BEOUTCOME3	Screen pos-maculopathy	nominal	2	Ignore analysis	
BEOUTCOME4	Screen pos-VA	nominal	3	Ignore analysis	
BEOUTCOME5	Screen pos-Disc	nominal	4	Ignore analysis	
BEOUTCOME6	Screen pos-Diabetic other	nominal	5	Ignore analysis	
BEOUTCOME7	Non –diabetic STD	nominal	6	Ignore analysis	
BEOUTCOME8	Couldn’t Grade throughout	nominal	90	Ignore analysis	
BEOUTCOME9	No photos	nominal	99	Ignore analysis	
BE CALCULATED OUTCOME	composite score for grading outcome for RE	nominal	screen –ve 0, screen +ve retinopathy =1,	This field records the outcome of the screening episode by	calculated from RE CALCULATED OUTCOME and LE CALCULATED OUTCOME

			screen +ve maculopathy =2, screen +ve retinopathy and screen +ve maculopathy =3 ungradable = 90 null field = NR	patient Highest grade in either eye takes precedence If maculopathy and retinopathy exist in either or both eyes then set as 3 Ungradable Where both attributes = 90 set as 90 Where one field = 90: i) and other is 0 set as 90 ii) and other is 1,2, or 3 set as 1,2,3 respectively Null attributes Where both	If < RE CALCULATED OUTCOME > = 0 and < LE CALCULATED OUTCOME > =0 then set = 0 If (< RE CALCULATED OUTCOME > or < LE CALCULATED OUTCOME > = 1) and (< RE CALCULATED OUTCOME > or < LE CALCULATED OUTCOME > = 0,1) then set = 1 If (< RE CALCULATED OUTCOME > or < LE CALCULATED OUTCOME > = 2) and (< RE CALCULATED OUTCOME > or < LE CALCULATED OUTCOME > = 0,2) then set = 2 If (< RE CALCULATED OUTCOME > or < LE CALCULATED OUTCOME > = 1) and (< RE CALCULATED OUTCOME > or < LE CALCULATED OUTCOME > = 2) then set = 3 If either < RE CALCULATED

				attributes = null set as not recorded (NR) Where one field = null : i) and other is 0 set as NR ii) and other is 1,2, or 3 set as 1,2,3 respectively	OUTCOME > = 3 or < LE CALCULATED OUTCOME > =3 then set = 3 Ungradable If <RE CALCULATED OUTCOME> = 90 and <LE CALCULATED OUTCOME > = 90 then set = 90 If (<RE CALCULATED OUTCOME> or <LE CALCULATED OUTCOME > = 90) and (<RE CALCULATED OUTCOME> or <LE CALCULATED OUTCOME > = 0) then set = 90 If (<RE CALCULATED OUTCOME> or <LE CALCULATED OUTCOME > = 90) and (<RE CALCULATED OUTCOME> or <LE CALCULATED OUTCOME > = 1) then set = 1 If (<RE CALCULATED OUTCOME> or <LE CALCULATED OUTCOME > = 90) and (<RE CALCULATED

					OUTCOME> or <LE CALCULATED OUTCOME > = 2) then set = 2 If (<RE CALCULATED OUTCOME> or <LE CALCULATED OUTCOME > = 90) and (<RE CALCULATED OUTCOME> or <LE CALCULATED OUTCOME > = 3) then set = 3 Null attributes If (<RE CALCULATED OUTCOME> or <LE CALCULATED OUTCOME > = null) and (<RE CALCULATED OUTCOME> or <LE CALCULATED OUTCOME > = 0) then set = NR If (<RE CALCULATED OUTCOME> or <LE CALCULATED OUTCOME > = null) and (<RE CALCULATED OUTCOME> or <LE CALCULATED OUTCOME > = 1) then set = 1 If (<RE CALCULATED OUTCOME> or <LE CALCULATED OUTCOME > =

					null) and (<RE CALCULATED OUTCOME> or <LE CALCULATED OUTCOME > = 2) then set = 2 If (<RE CALCULATED OUTCOME> or <LE CALCULATED OUTCOME > = null) and (<RE CALCULATED OUTCOME> or <LE CALCULATED OUTCOME > = 3) then set = 3
DiabStedEye	Auto generated-by rule .Not on Photo form-6/98	nominal	Yes=Y, No=N	Replace autogenerated with value calculated from elsewhere As in BE OUTCOME but level 40 or above for retinopathy and no account of presence of maculopathy and retinopathy	calculated from < BE CALCULATED OUTCOME> and <RERET> and <LERET> If <BE CALCULATED OUTCOME> = 0, 90 or NR then set as N If <BE CALCULATED OUTCOME> = 2,3 then set as Y If <BE CALCULATED OUTCOME> = 1 and (<RERET> = 30 and <LERET> = 30) then set as N else set as Y

Action1	Annual Review		1	Ignore warehouse	
Action2	Glaucoma Suspect		2	Ignore warehouse	
Action3	Retinal Clinic		3	Ignore warehouse	
Action4	General Clinic		4	Ignore warehouse	
Action5	Continued Ophthalmology F/U		5	Ignore warehouse	
Action6	6 month assessment clinic		6	Ignore warehouse	
Action7	W/L Laser		7	Ignore warehouse	
Action8	W/L Cataract Extraction		8	Ignore warehouse	
Action9	For referral elsewhere		9	Ignore warehouse	
FIELDCHANGE			Y/N	Ignore warehouse	

Logic rules for RERET and LERET (example given is RERET)

	Inclusive rule									
		HMA	NVD	CWS8A	NVE	VB/VR/VL6A	FVP	IRMA	PRHVH	Laser
	variables	0,1,2,3,90,NR	0,1,2,3,90,NR	0,1,2,3,90,NR	0,1,2,3,90,NR	0,1,2,3,4,5,90,NR	0,1,2,3,4,5,90,NR	0,1,2,3,90,NR	0,1,2,3,4,90,NR	0,1,2,3,90,NR
1		is not	is not	is not	is not =	is not =	is not = 2,3,4	is not = 2	is not = 2,3	is not = 2,3

0		= 2 or 3	= 2 or 3	= 2 or 3	2 or 3	2,3,4 or 5	or 5	or 3	or 4	
12	any field = 1	is not = 2 or 3	is not = 2 or 3	is not = 2 or 3	is not = 2 or 3	is not = 2,3,4 or 5	is not = 2,3,4 or 5	is not = 2 or 3	is not = 2,3 or 4	is not = 2,3
20	HMA =2		is not = 2 or 3	is not = 2 or 3	is not = 2 or 3	is not = 2,3,4 or 5	is not = 2,3,4 or 5	is not = 2 or 3	is not = 2,3 or 4	is not = 2,3
30	HMA =2 and/or CWS8A =2		is not = 2 or 3		is not = 2 or 3	is not = 2,3,4 or 5	is not = 2,3,4 or 5	is not = 2 or 3	is not = 2,3 or 4	is not = 2,3
40	CWS8A = 3 and/or VB/VR/VL6A =2 and/or IRMA =2		is not = 2 or 3		is not = 2 or 3		is not = 2,3,4 or 5		is not = 2,3 or 4	is not = 2,3
50	VB/VR/VL6A =3,4,5 and/or IRMA =3		is not = 2 or 3		is not = 2 or 3		is not = 2,3,4 or 5		is not = 2,3 or 4	is not = 2,3
60	NVD =2 and/or NVE =2 and/or FVP = 2,3,4 and/or laser = 2						is not = 5		is not = 2,3 or 4	
70	<u>either</u> NVD =3; <u>or</u> PRHVH =2,3,4 and (NVD =2 and/or						is not = 5			

	NVE =3)									
71	FVP = 5 and (NVD = 2,3 and/or NVE= 2,3 and/or PRHVH = 2,3,4)									
72	PRHVH = 3,4 and all other attributes are = 90									
90	all attributes = 90									

Appendix 2 – Noise reduction

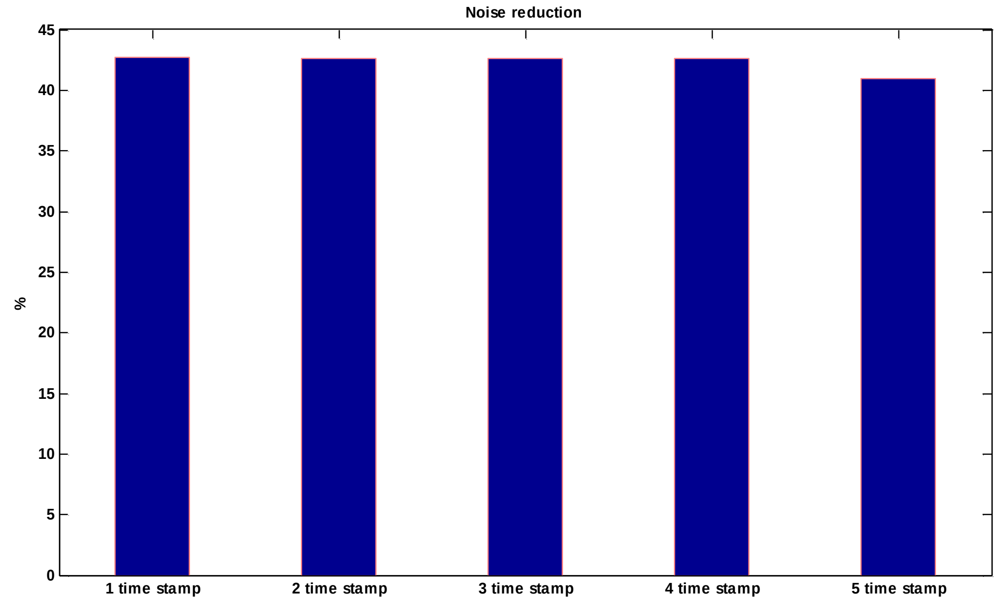


Figure A2.1: Noise reduction – 5 time stamps – Series 1

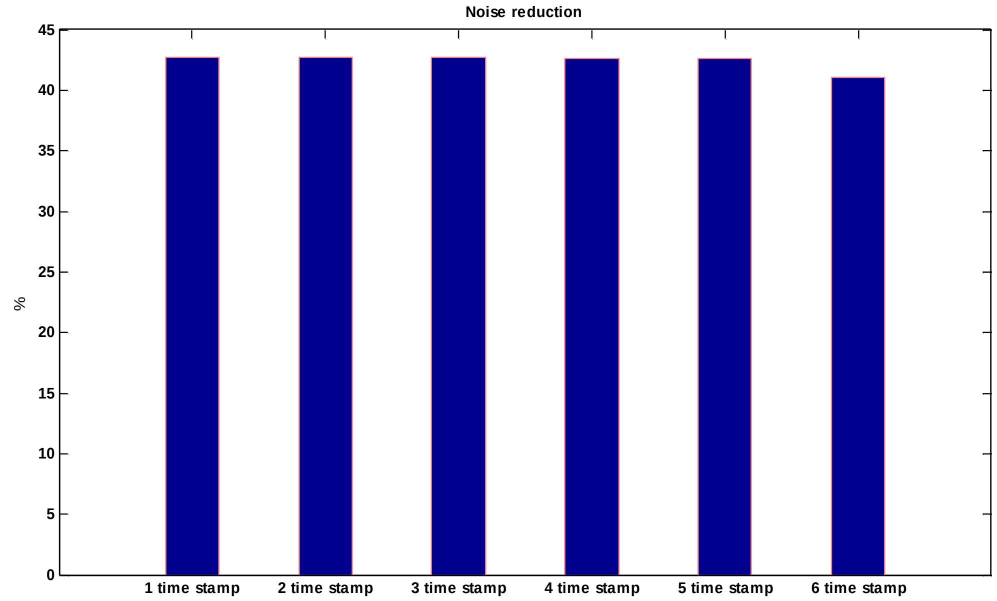


Figure A2.2: Noise reduction – 6 time stamps – Series 1

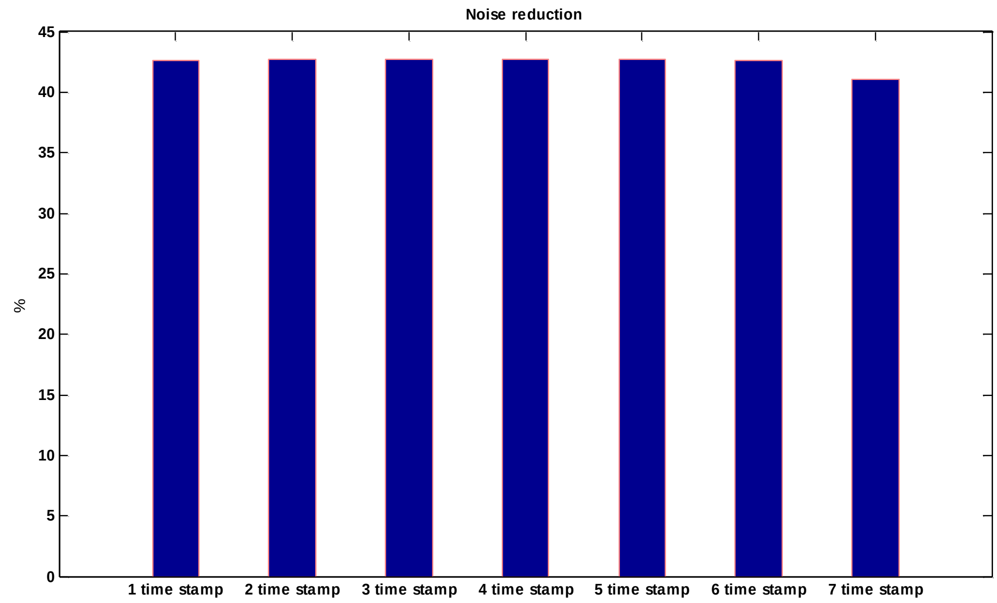


Figure A2.3: Noise reduction– 7 time stamps – Series 1

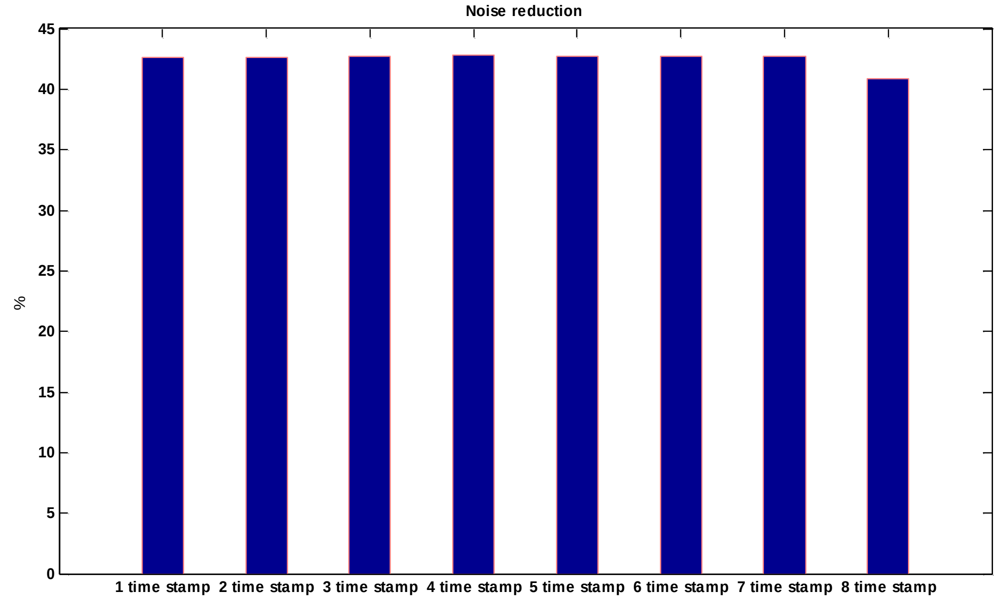


Figure A2.4: Noise reduction– 8 time stamps – Series 1

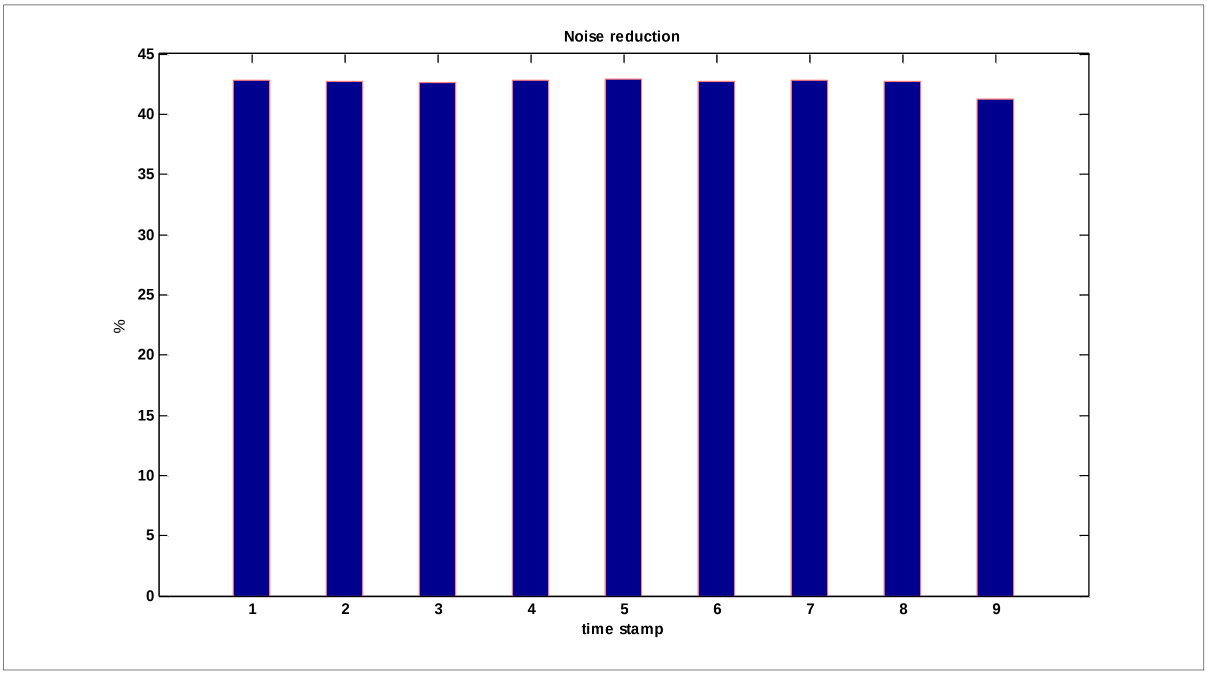


Figure A2.5: Noise reduction– 9 time stamps – Series 1

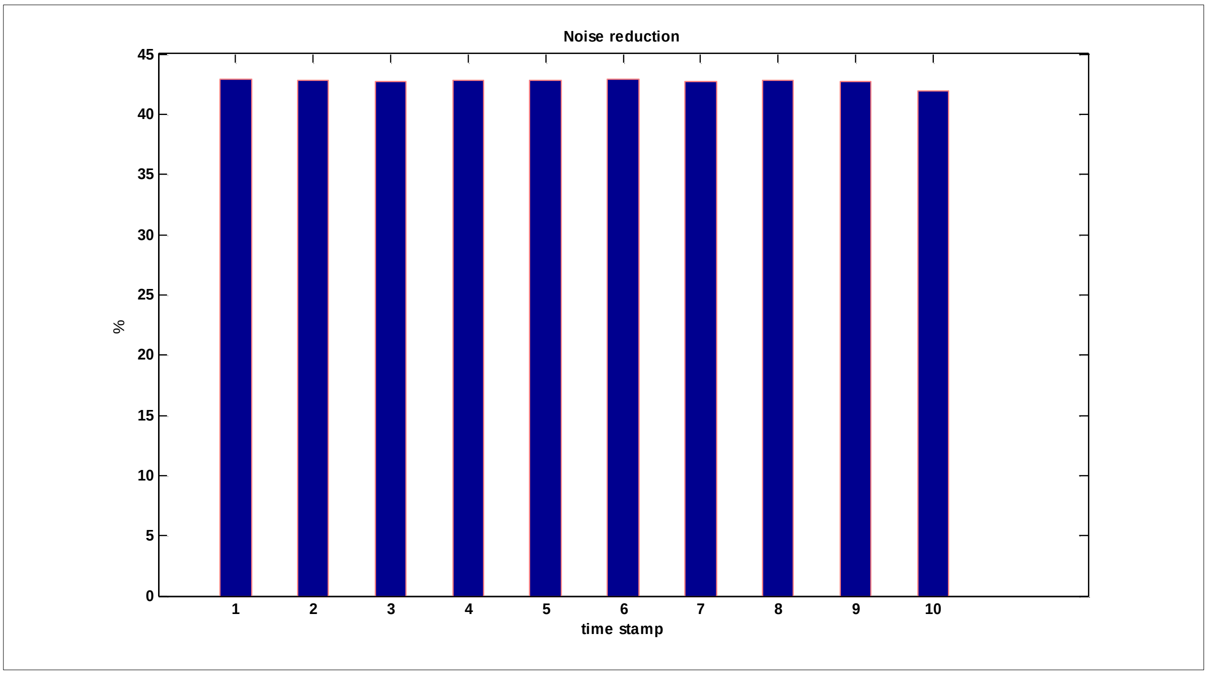


Figure A2.6: Noise reduction– 10 time stamps – Series 1

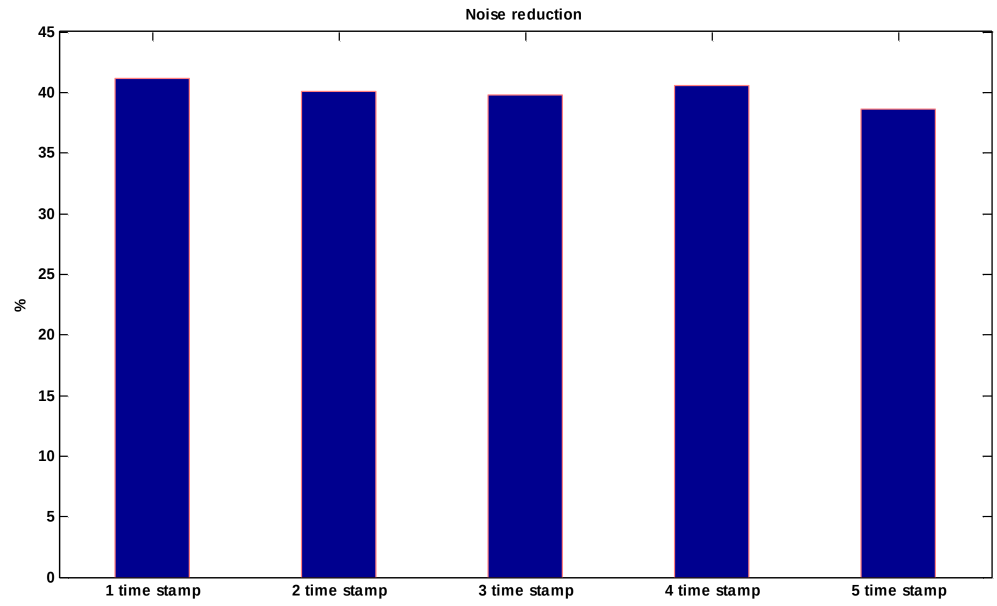


Figure A2.7: Noise reduction– 5 time stamps – Series 2

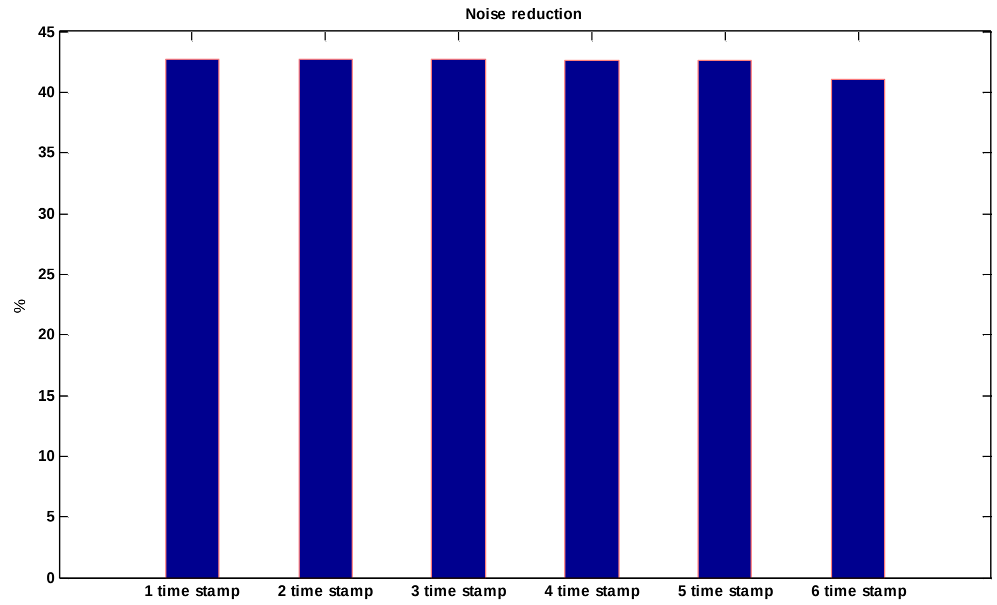


Figure A2.8: Noise reduction– 6 time stamps – Series 2

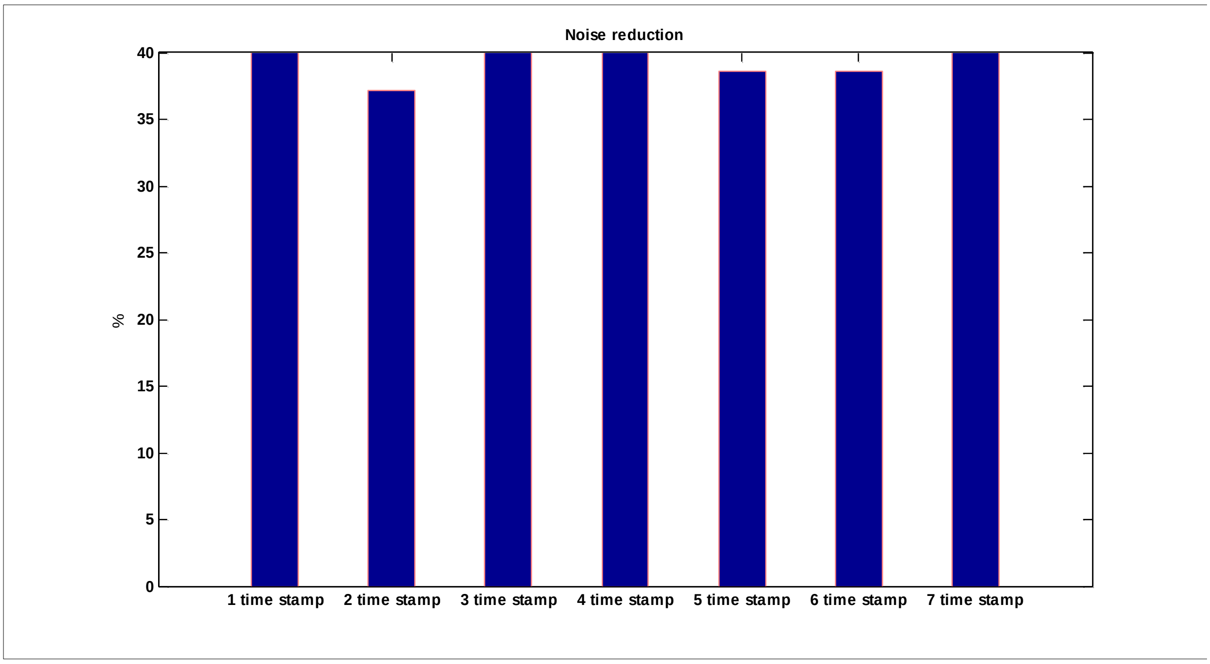


Figure A2.9:Noise reduction– 7 time stamps – Series 2

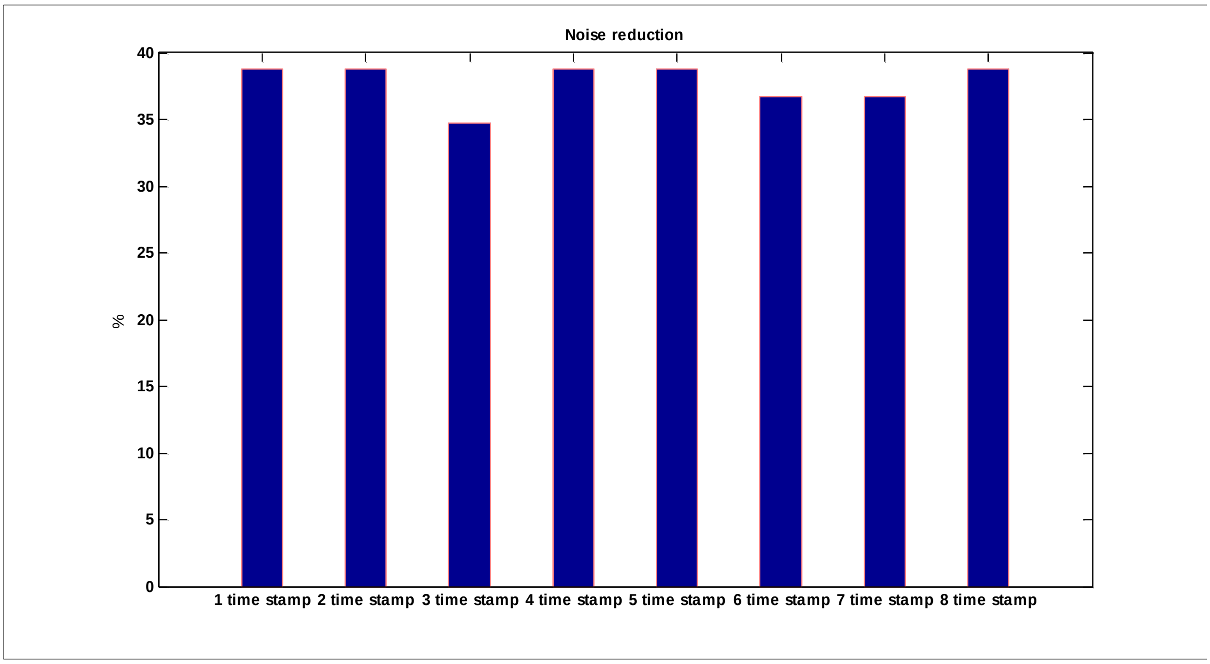


Figure A2.10:Noise reduction– 8 time stamps – Series 2

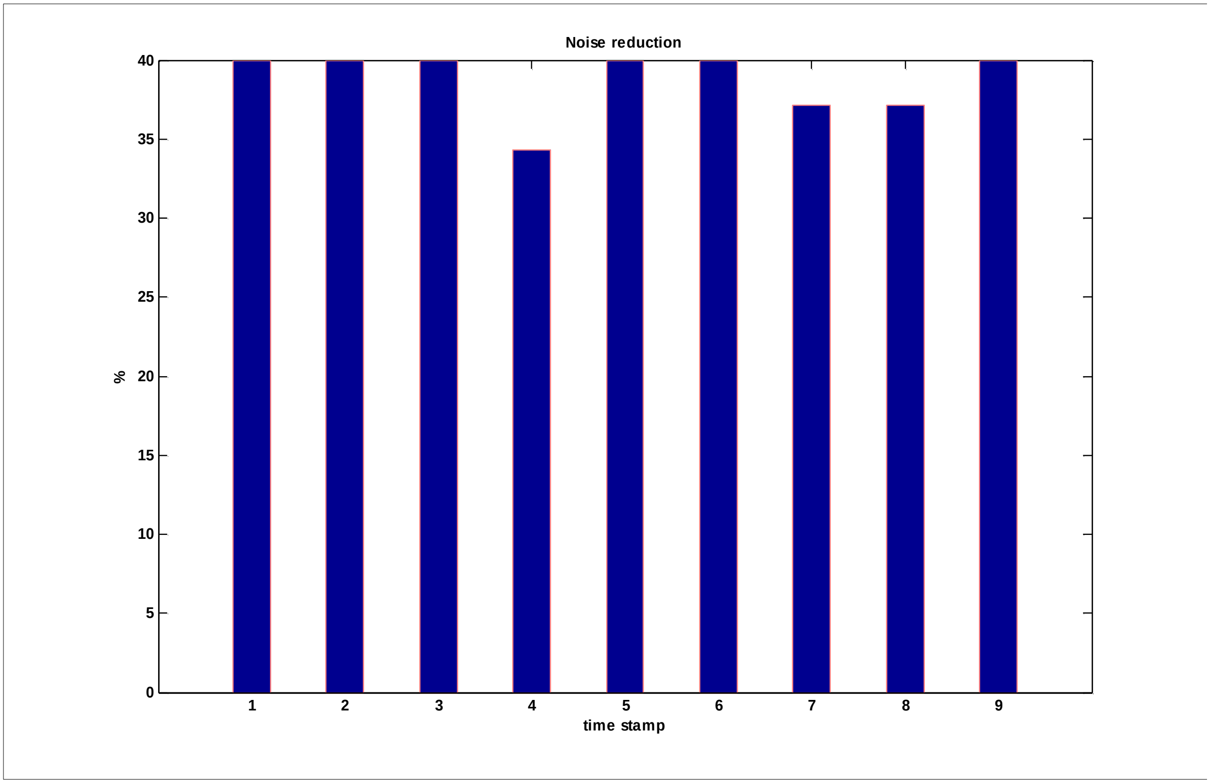


Figure A2.11: Noise reduction– 9 time stamps – Series 2

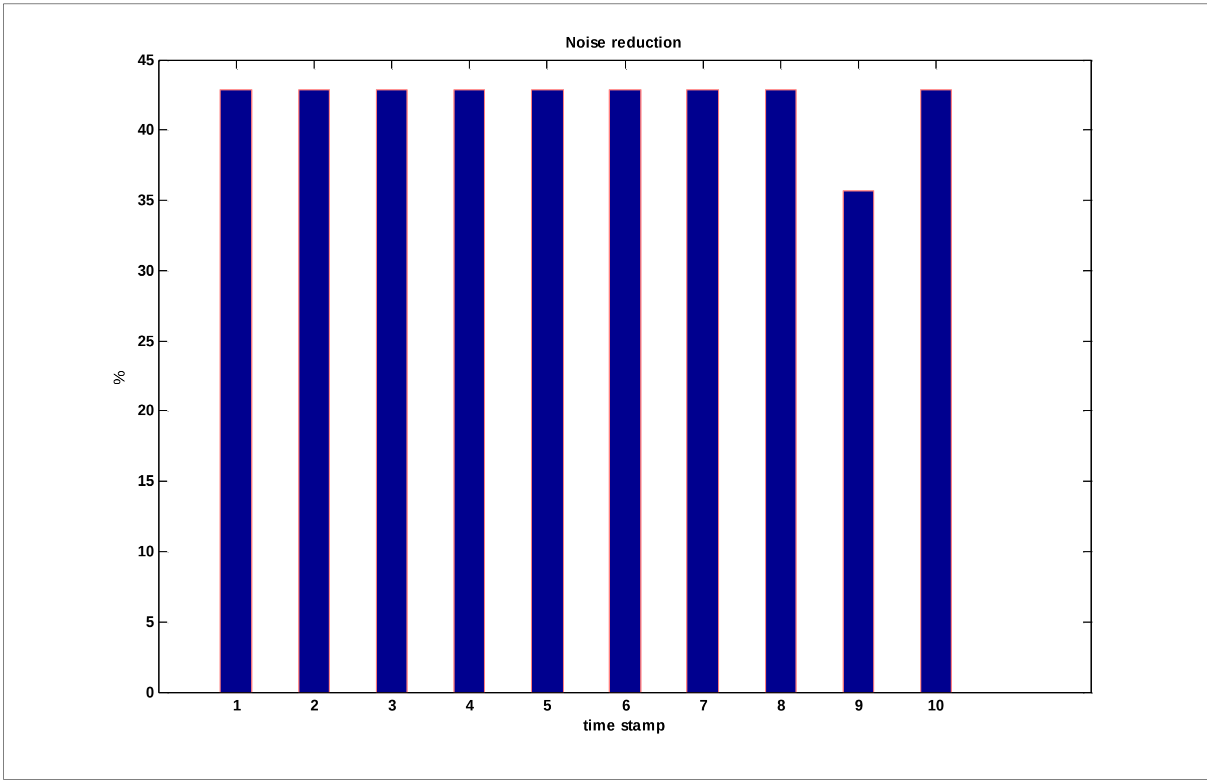


Figure A2.12 :Noise reduction– 10 time stamps – Series 2

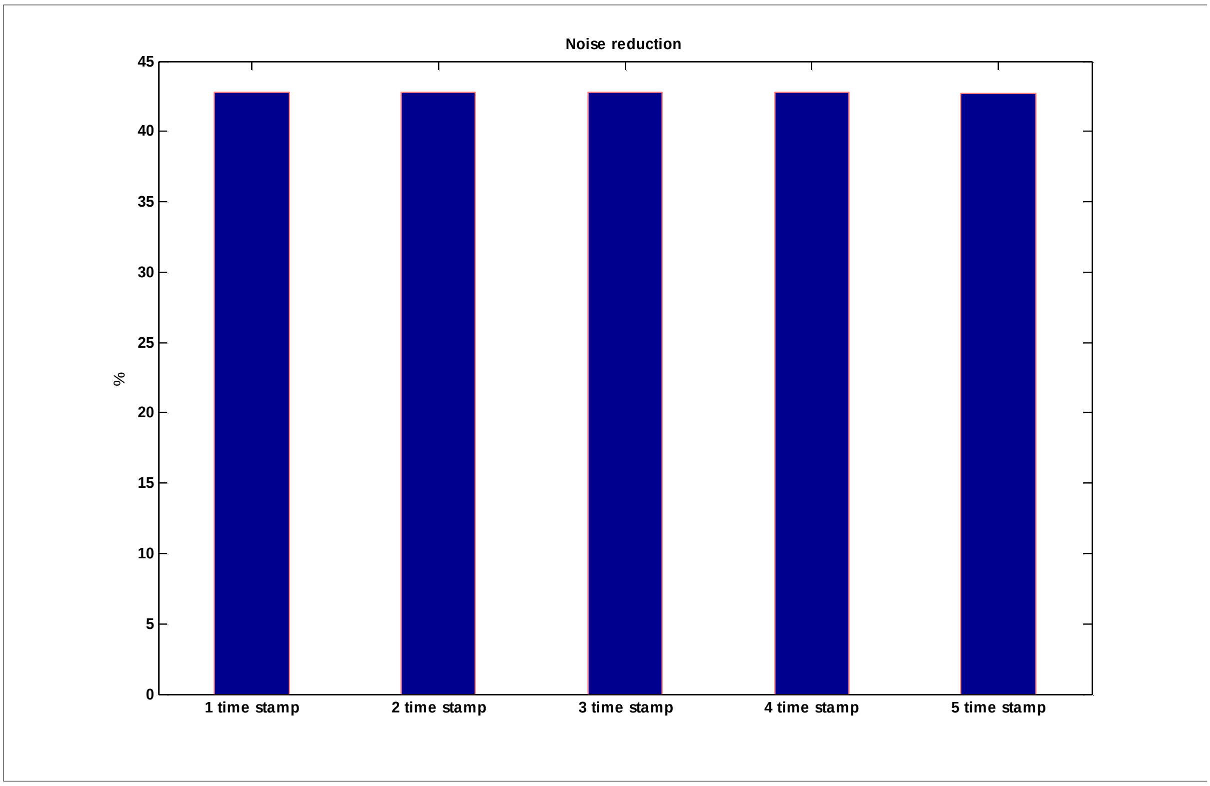


Figure A2.13: Noise reduction– 5 time stamps – Series 3

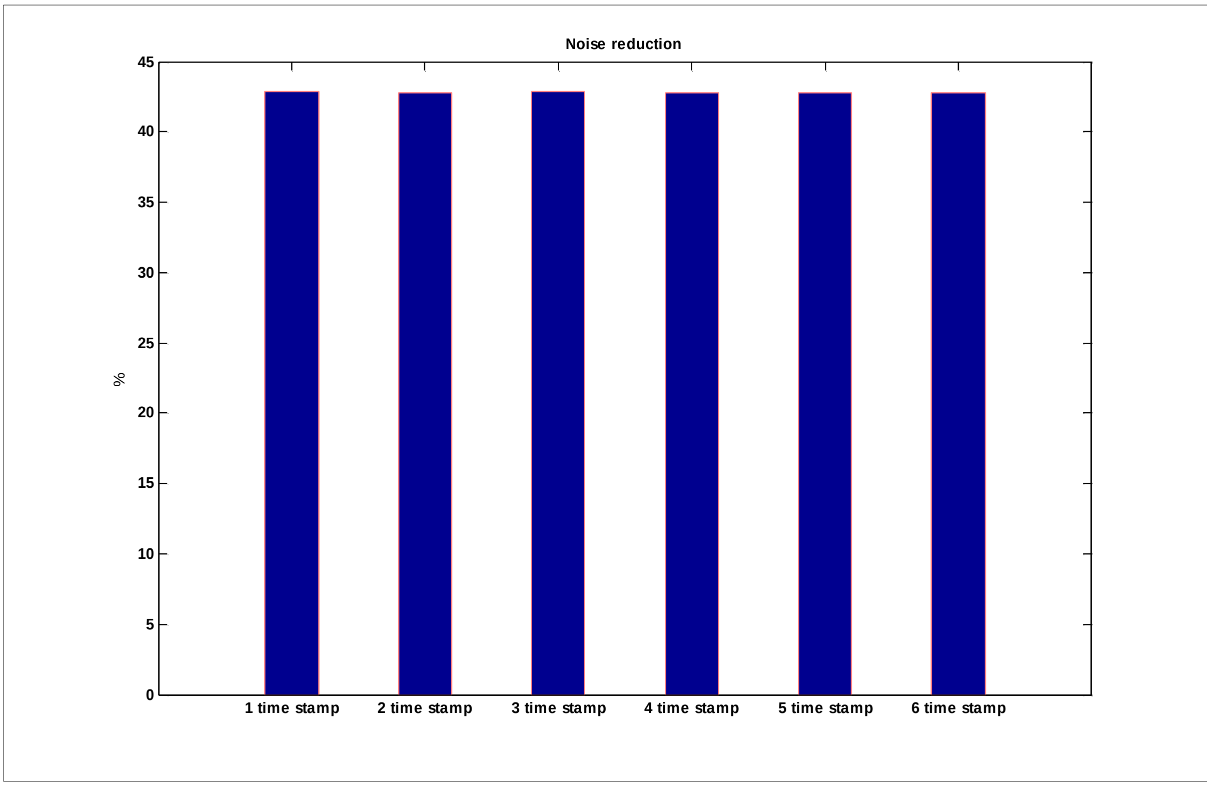


Figure A2.14: Noise reduction– 6 time stamps – Series 3

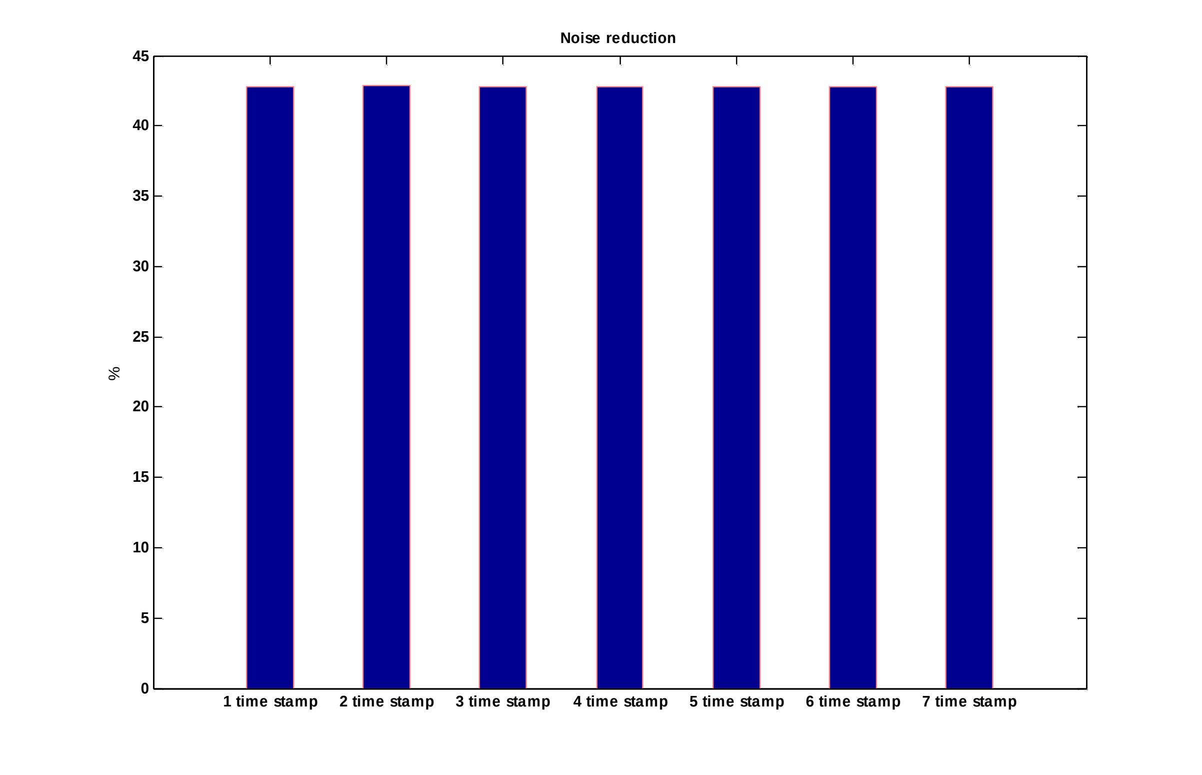


Figure A2.15:Noise reduction– 7 time stamps – Series 3

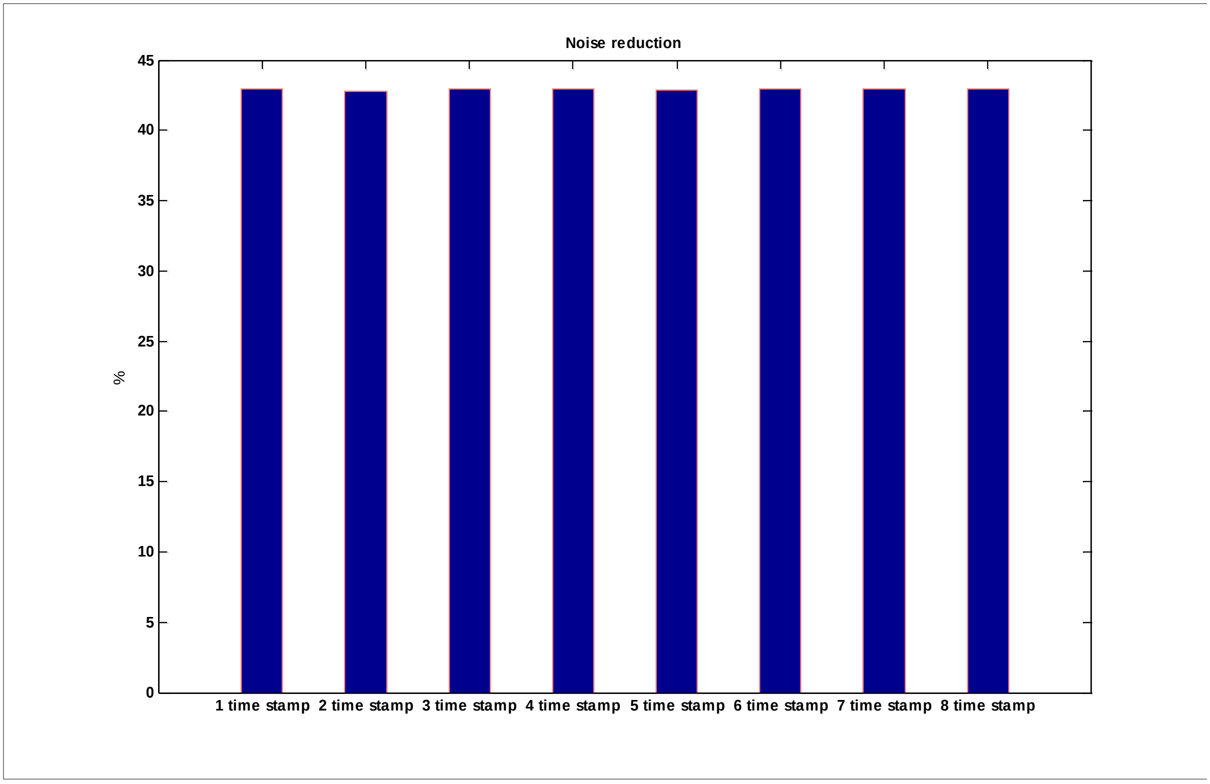


Figure A2.16: Noise reduction– 8 time stamps – Series 3

Appendix 3 – Time stamp interval distribution

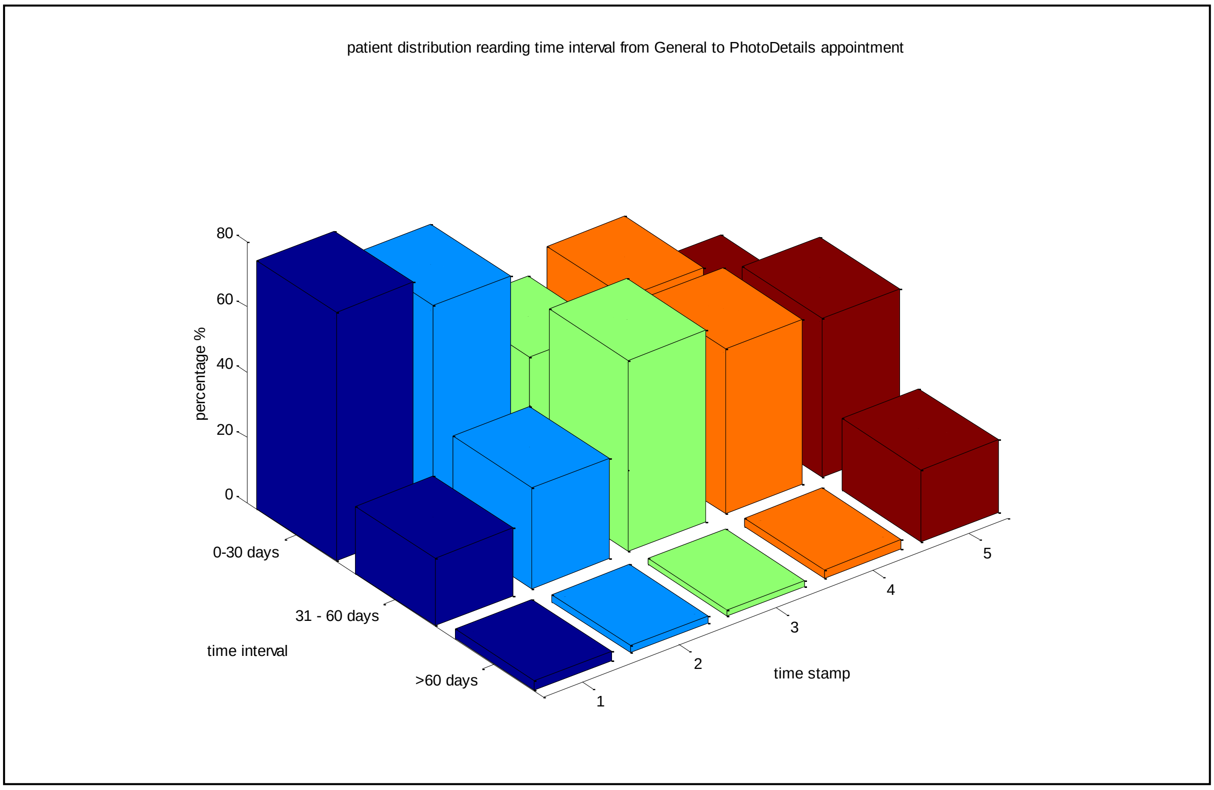


Figure A3.1: Patient distribution in intervals for every time stamp, from General to Photodetails – Series 1

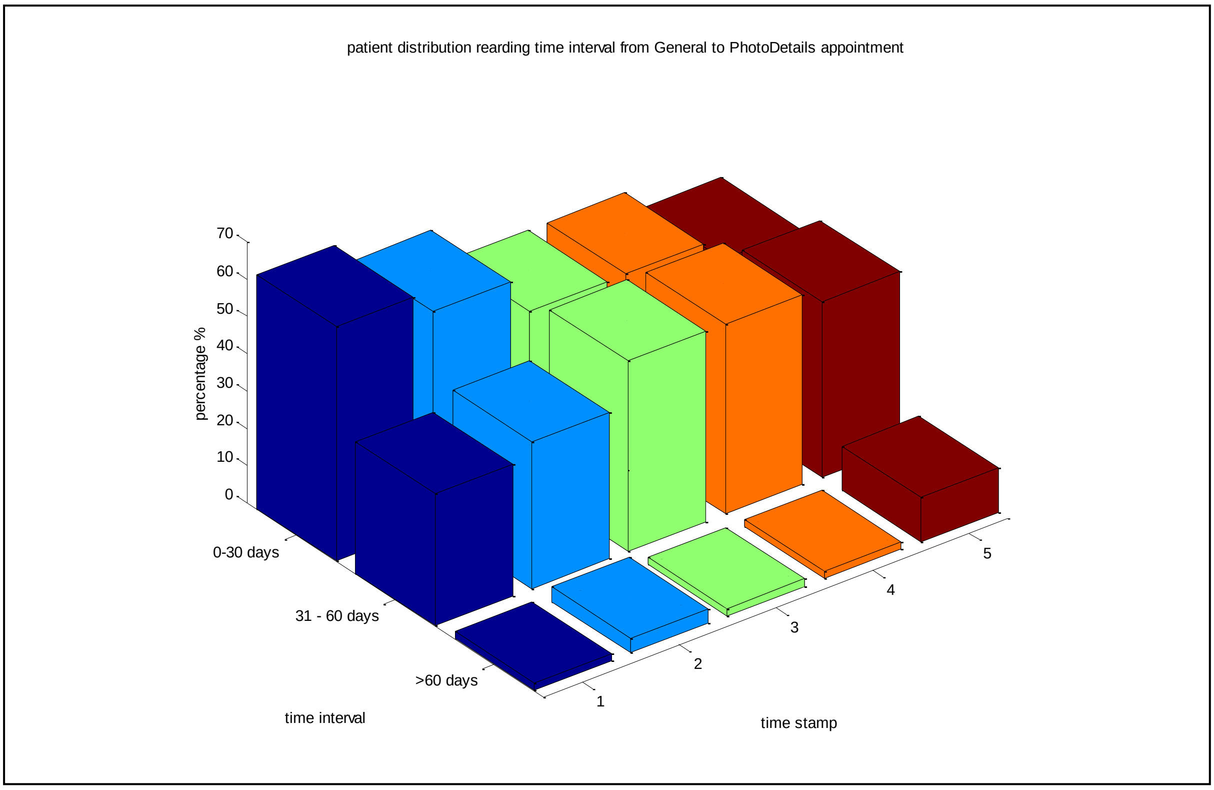


Figure A3. 2: Patient distribution in intervals for every time stamp, from General to Photodetails – Series 2

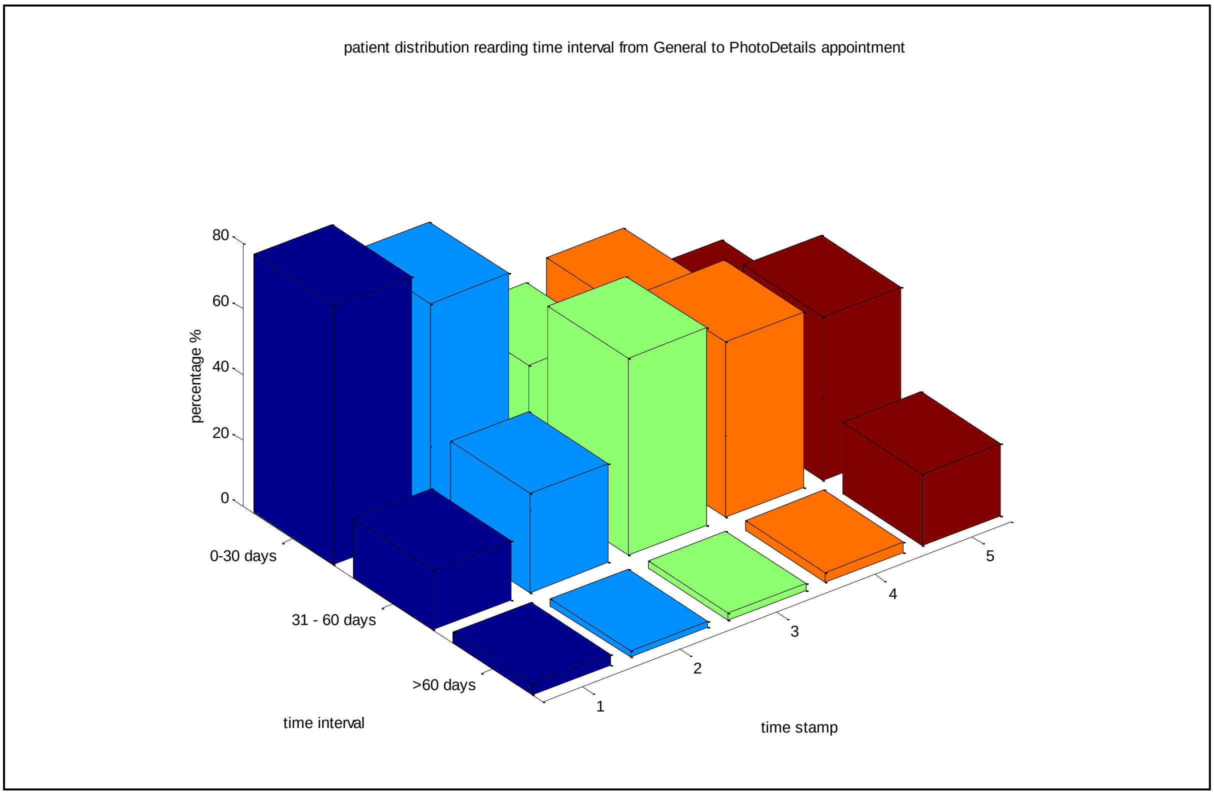


Figure A3. 3: Patient distribution in intervals for every time stamp, from General to Photodetails – Series 3

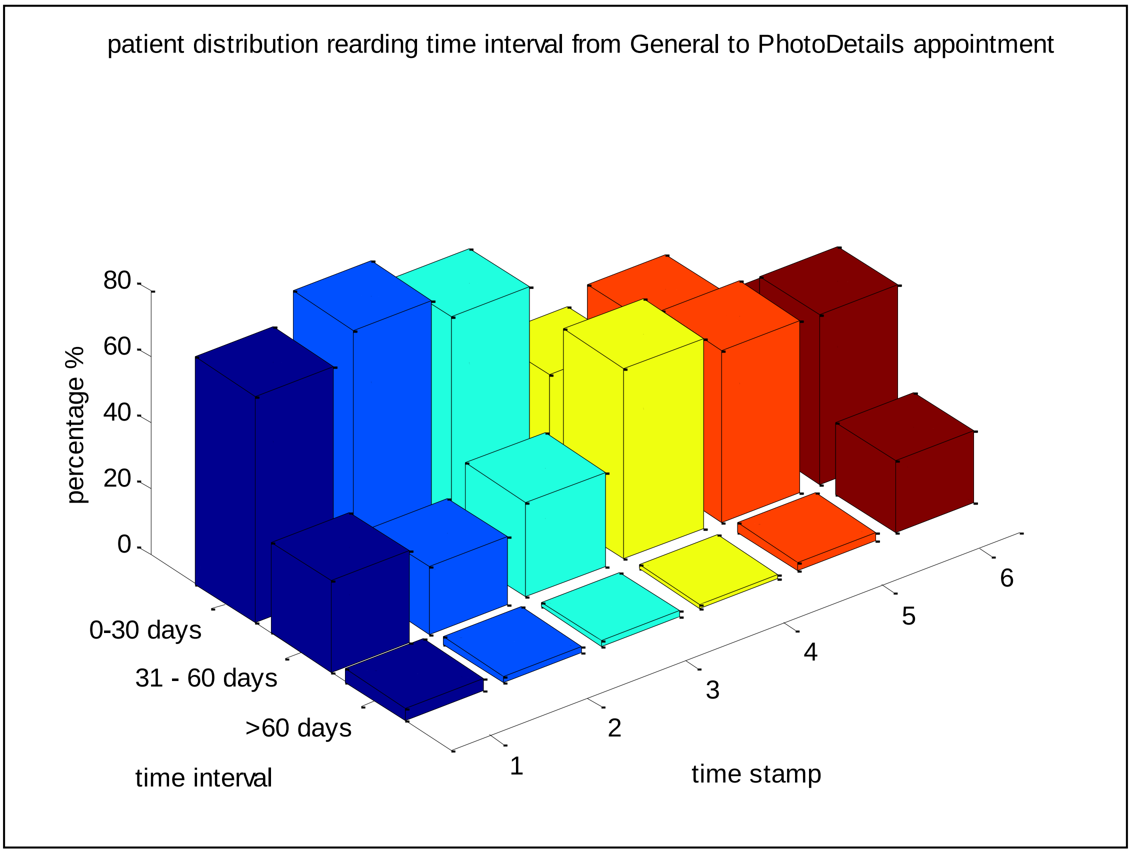


Figure A3. 4: Patient distribution in intervals for every time stamp, from General to Photodetails – Series 1

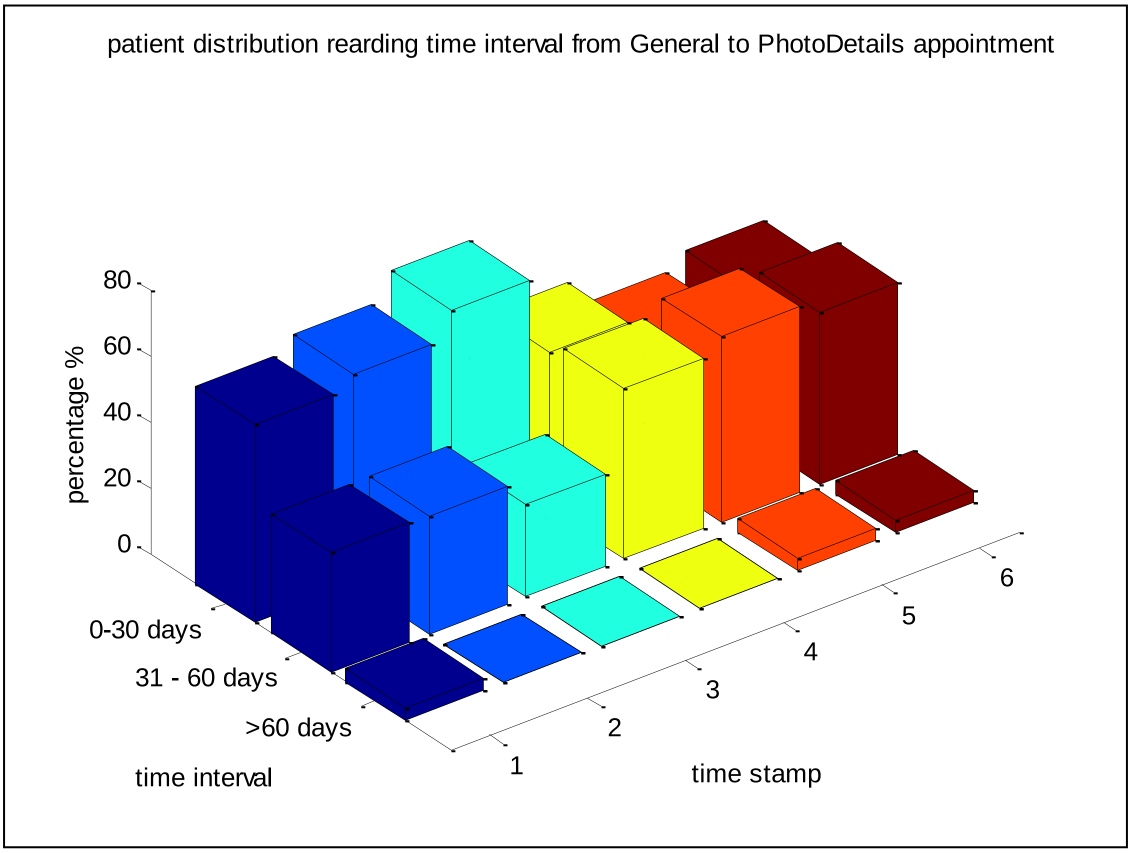


Figure A3. 5: Patient distribution in intervals for every time stamp, from General to Photodetails – Series 2

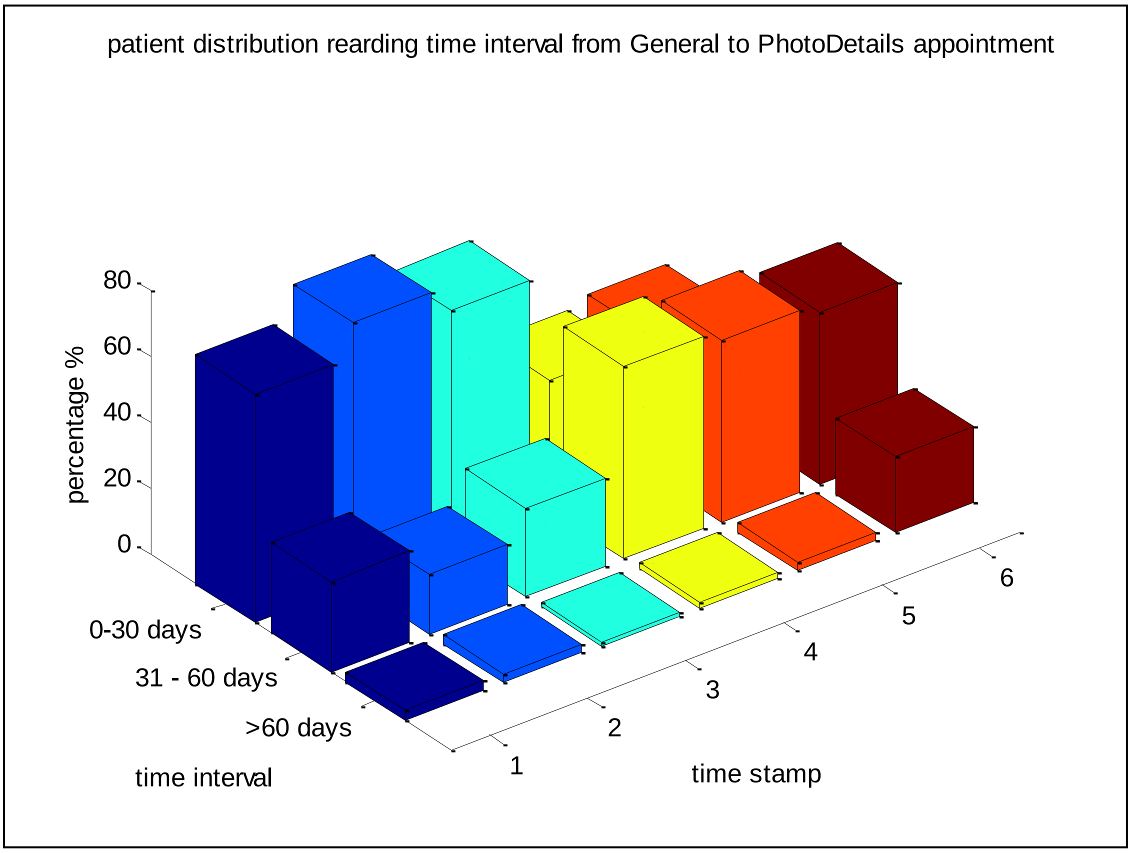


Figure A3. 6 : Patient distribution in intervals for every time stamp, from General to Photodetails – Series 3

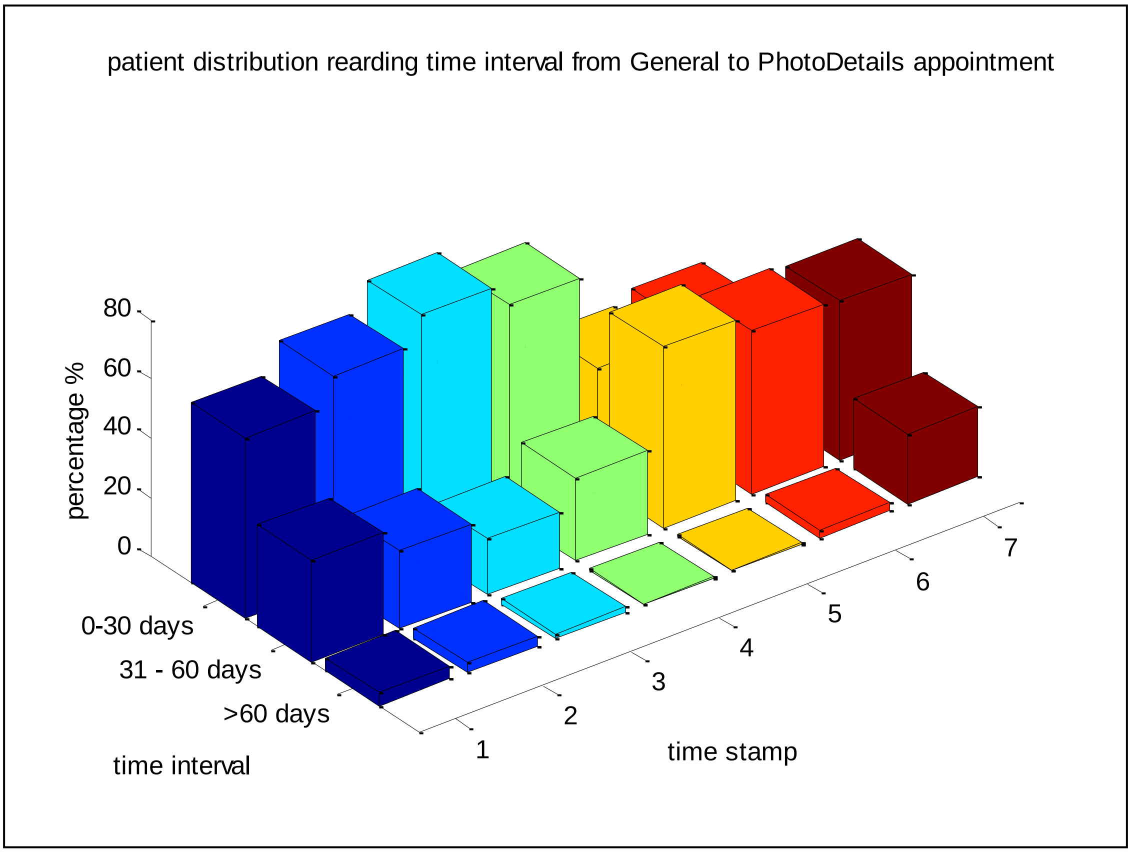


Figure A3. 7: Patient distribution in intervals for every time stamp, from General to Photodetails – Series 1

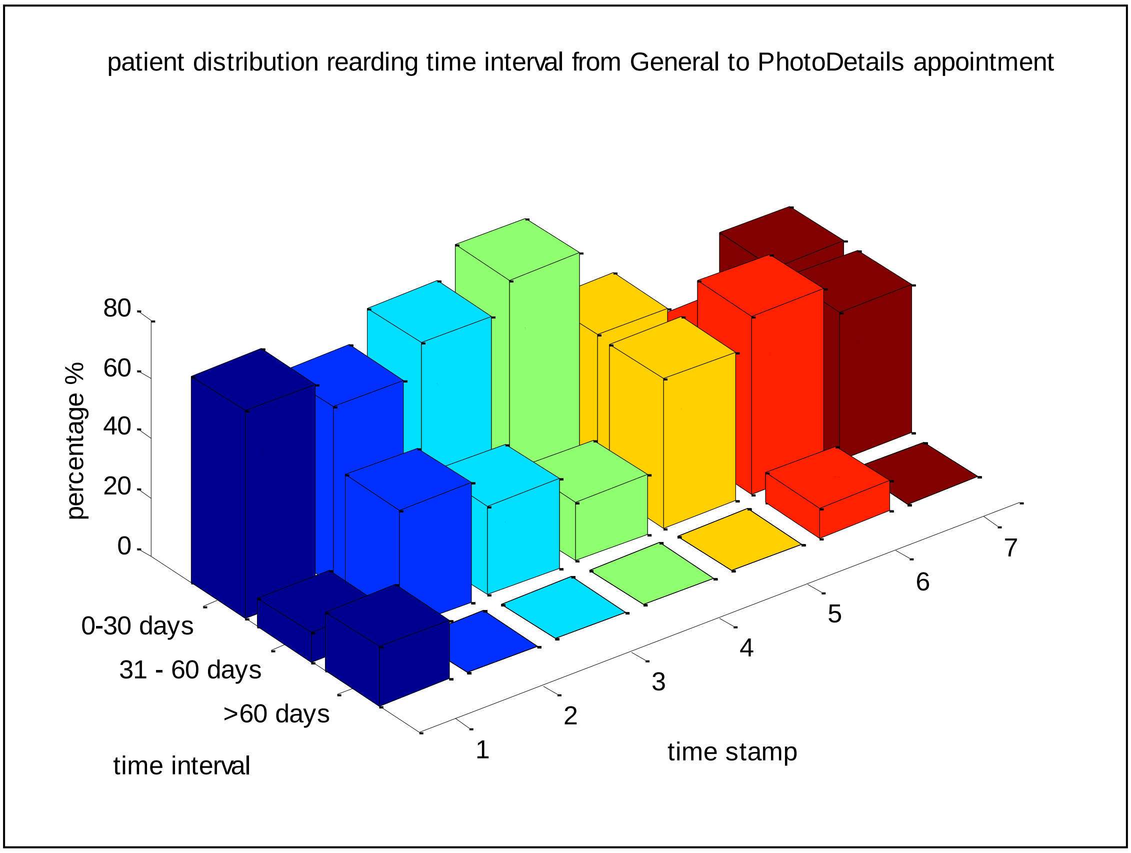


Figure A3. 8: Patient distribution in intervals for every time stamp, from General to Photodetails – Series 2

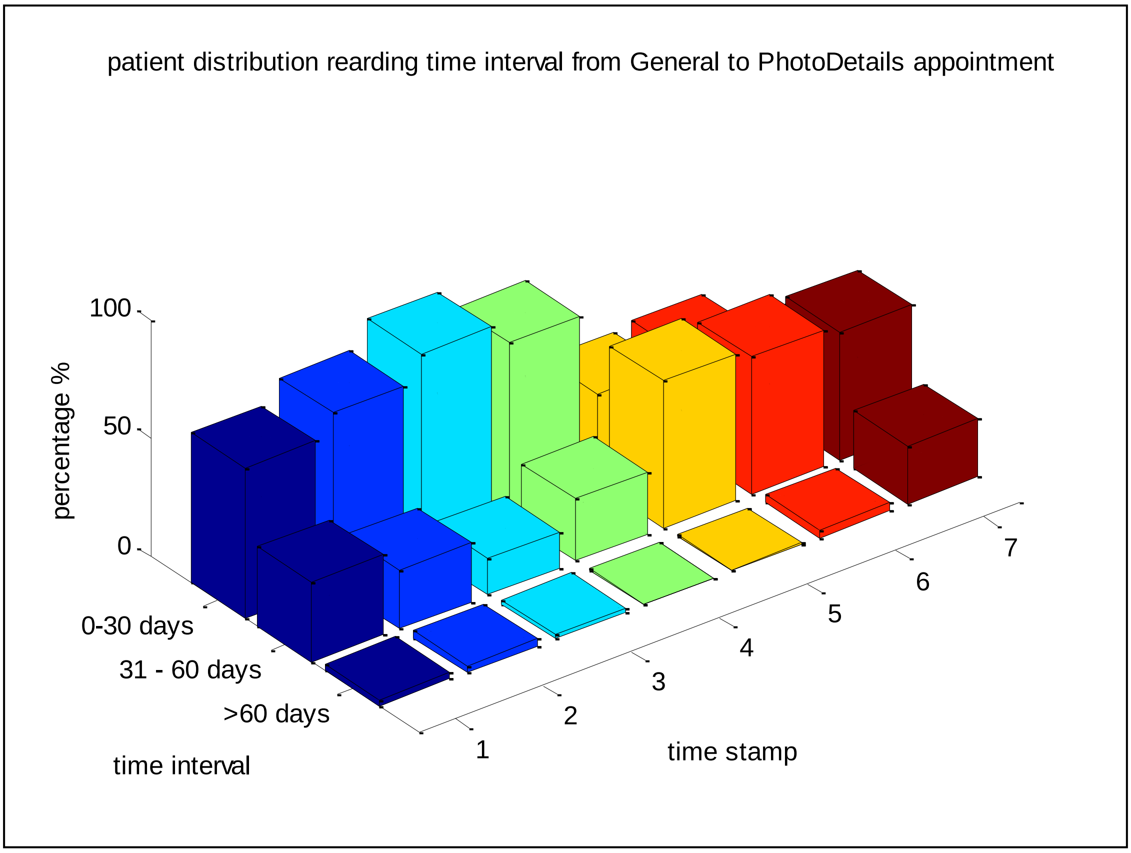


Figure A3. 9 : Patient distribution in intervals for every time stamp, from General to Photodetails – Series 3

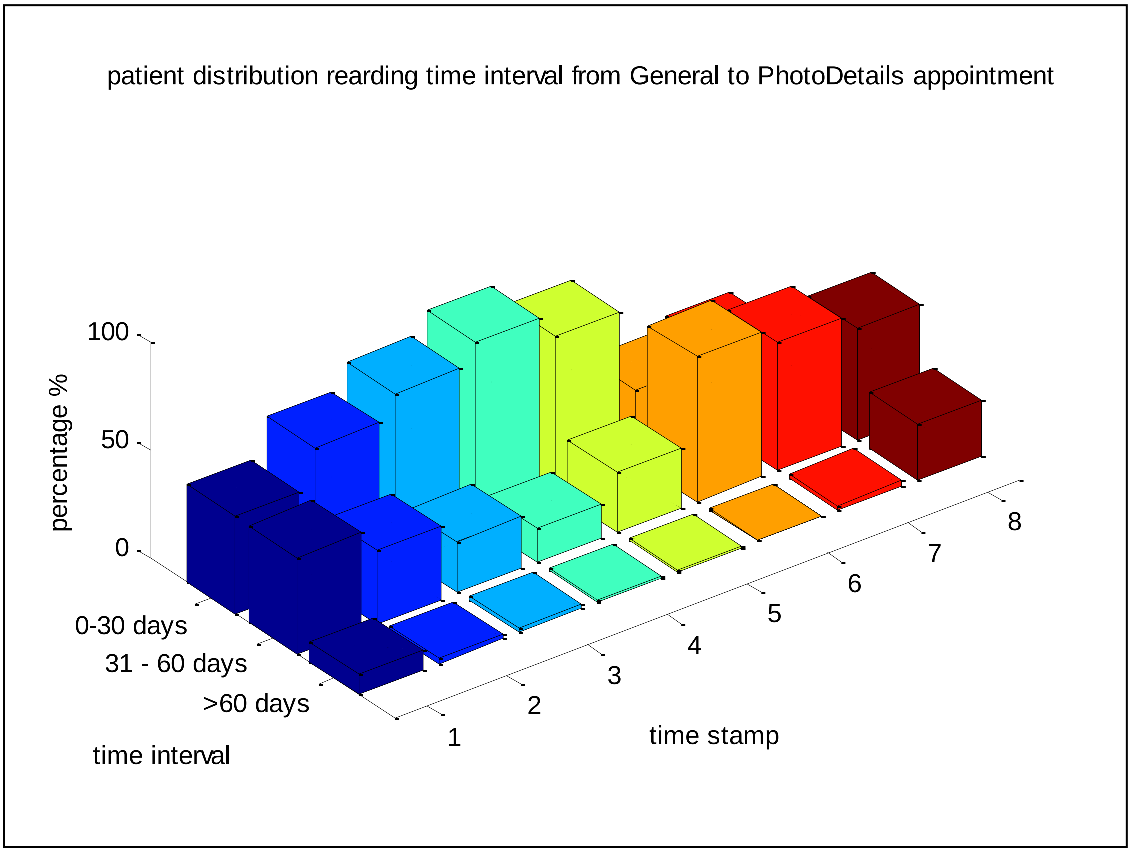


Figure A3.10: Patient distribution in intervals for every time stamp, from General to Photodetails – Series 1

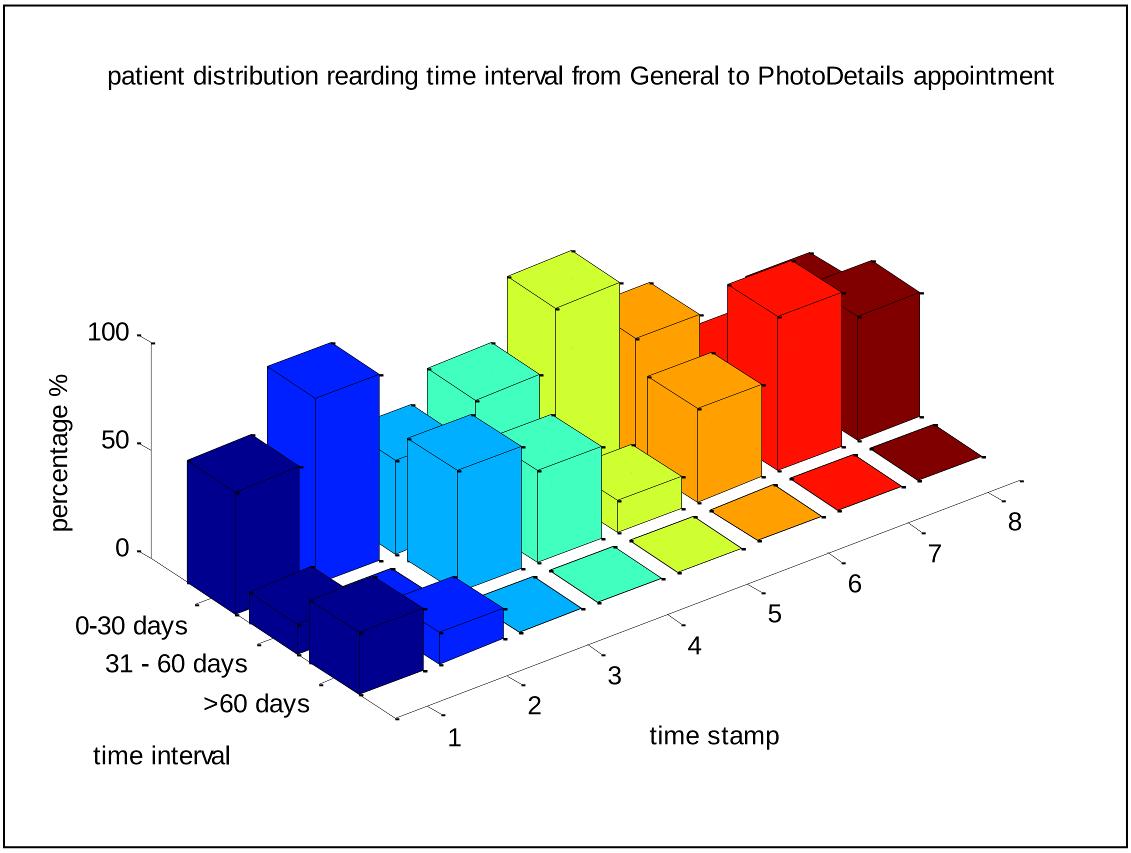


Figure A3. 11 : Patient distribution in intervals for every time stamp, from General to Photodetails – Series 2

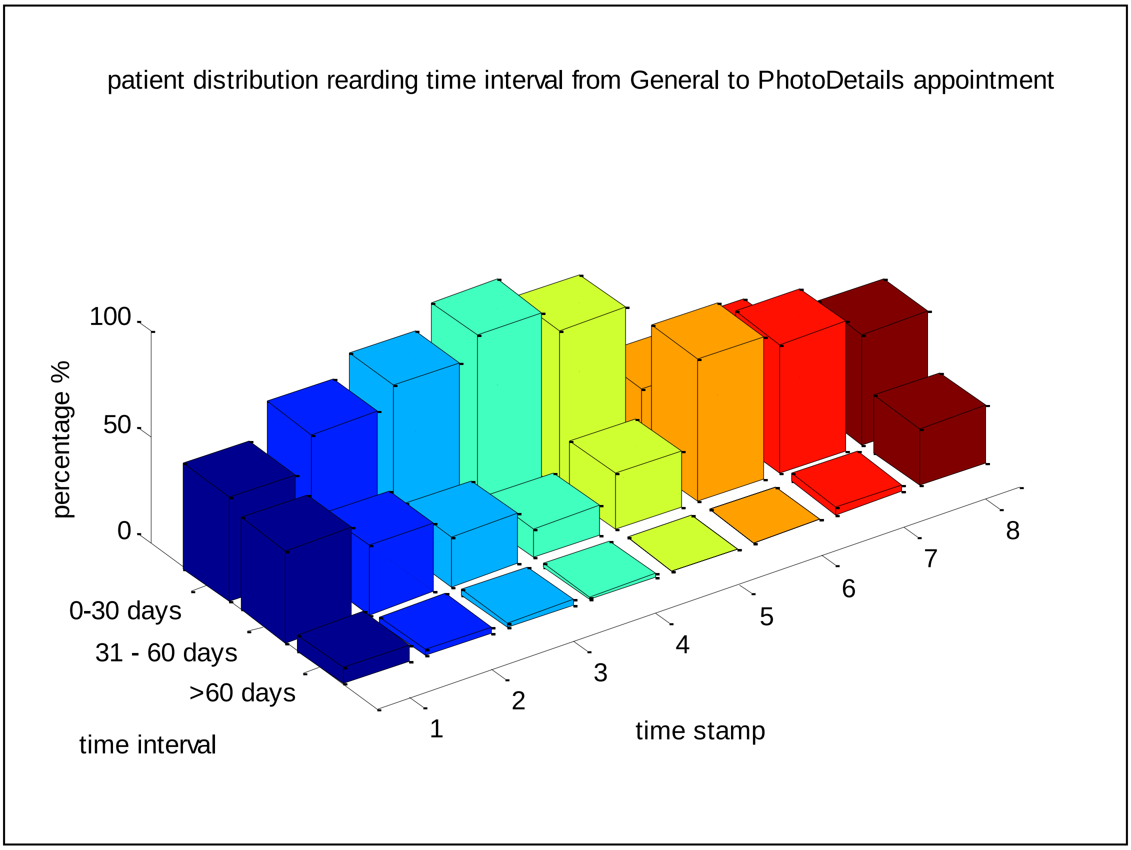


Figure A3. 12: Patient distribution in intervals for every time stamp, from General to Photodetails – Series 3

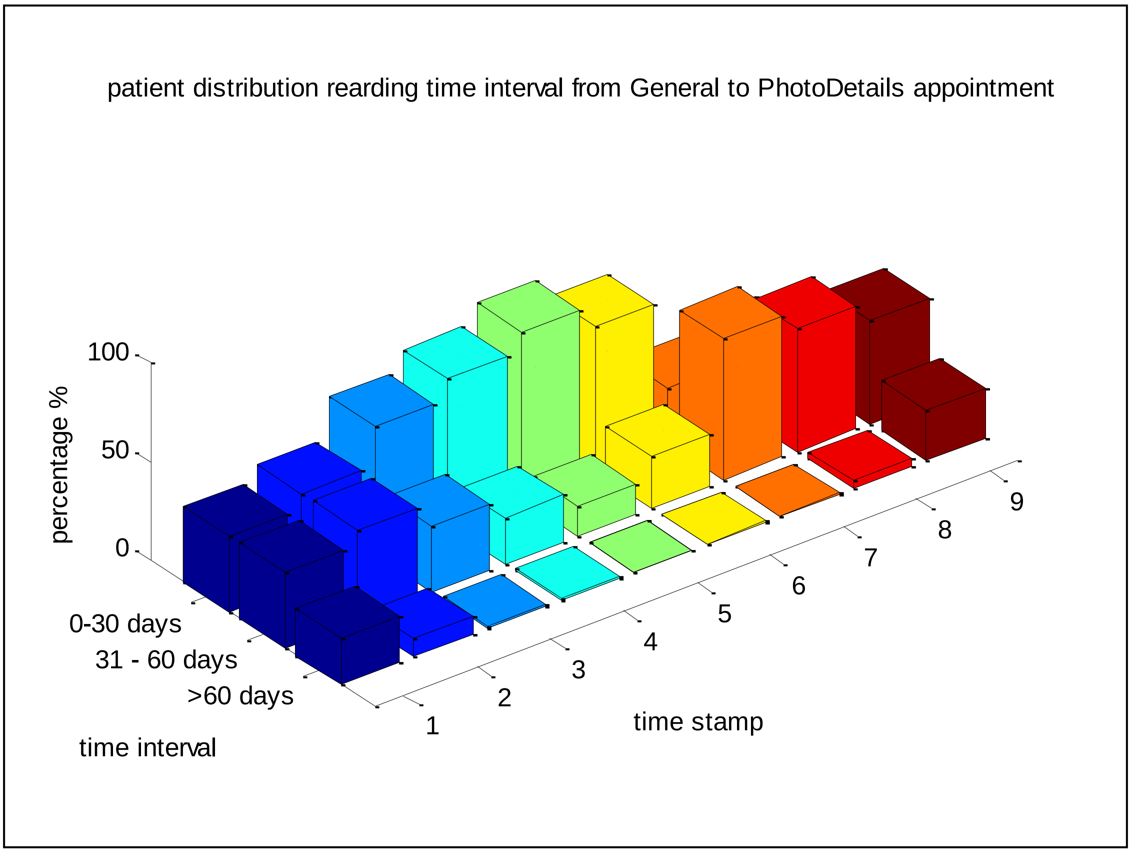


Figure A3. 13 : Patient distribution in intervals for every time stamp, from General to Photodetails – Series 1

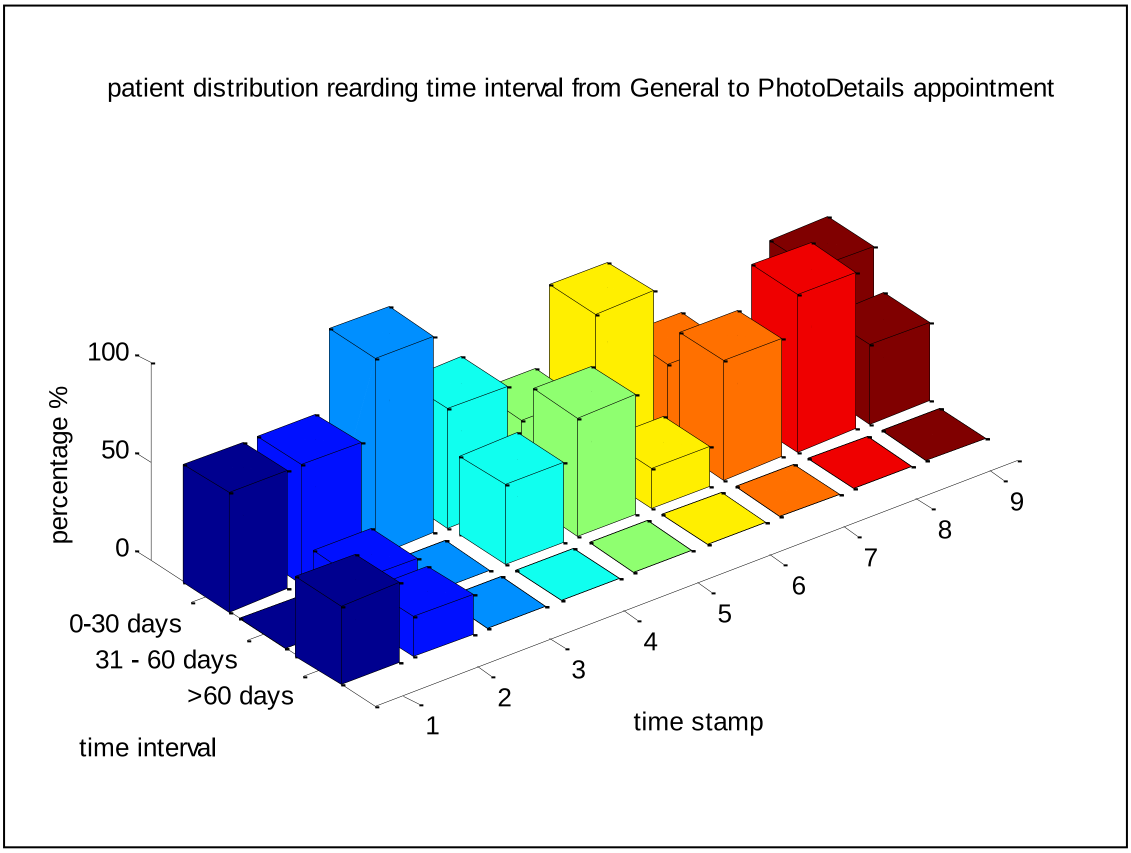


Figure A3. 14 : Patient distribution in intervals for every time stamp, from General to Photodetails – Series 2

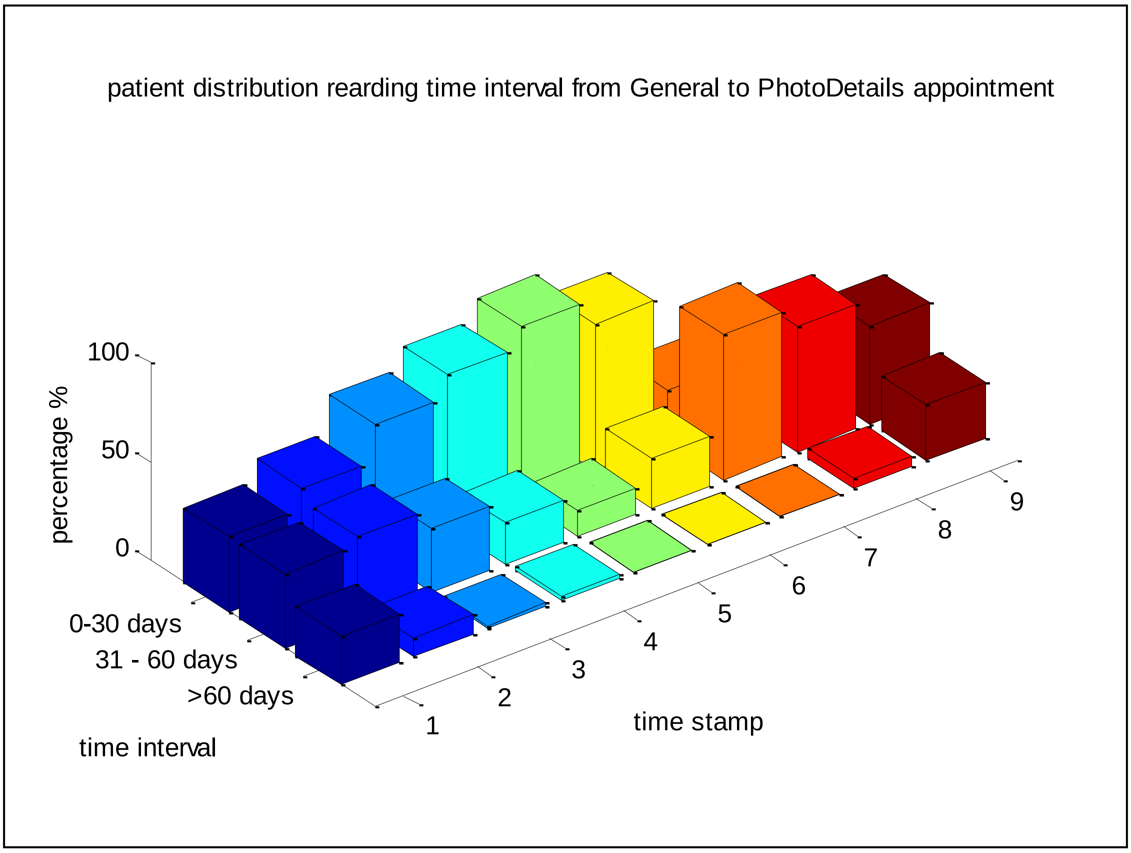


Figure A3. 55 : Patient distribution in intervals for every time stamp, from General to Photodetails – Series 3

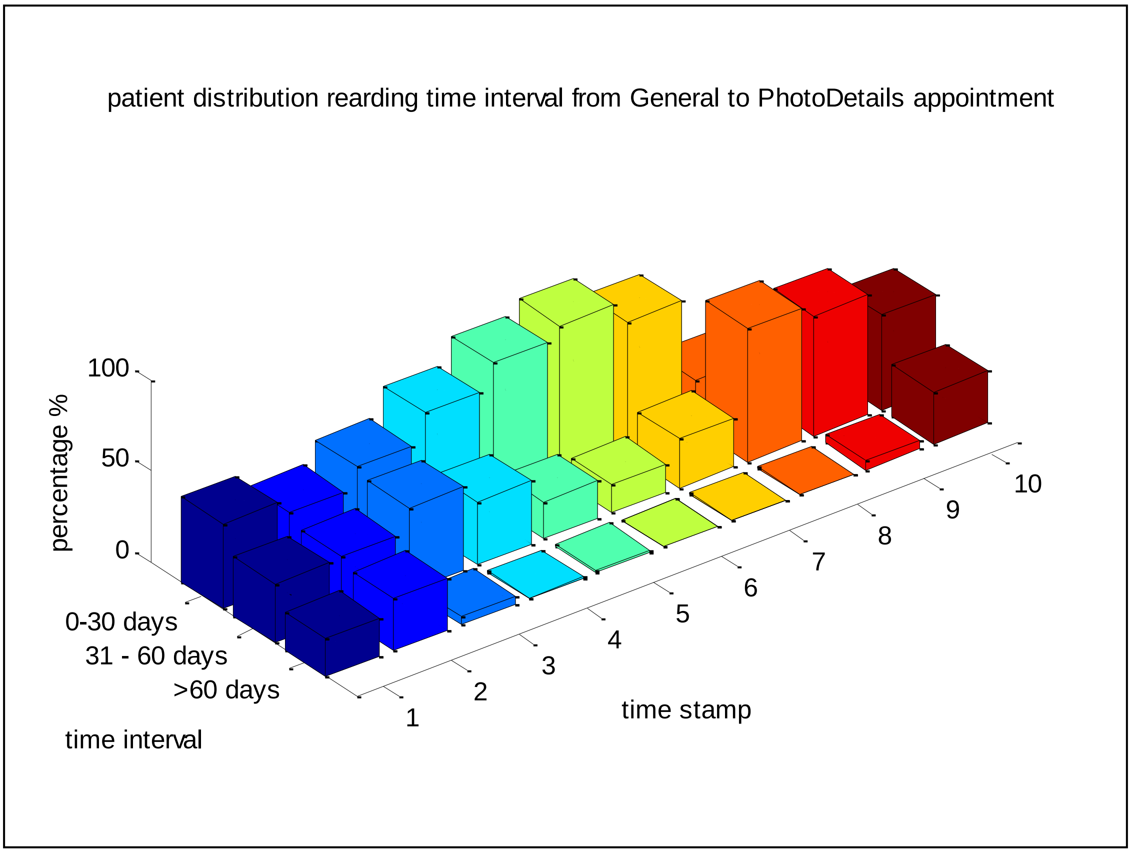


Figure A3. 16 : Patient distribution in intervals for every time stamp, from General to Photodetails – Series 1

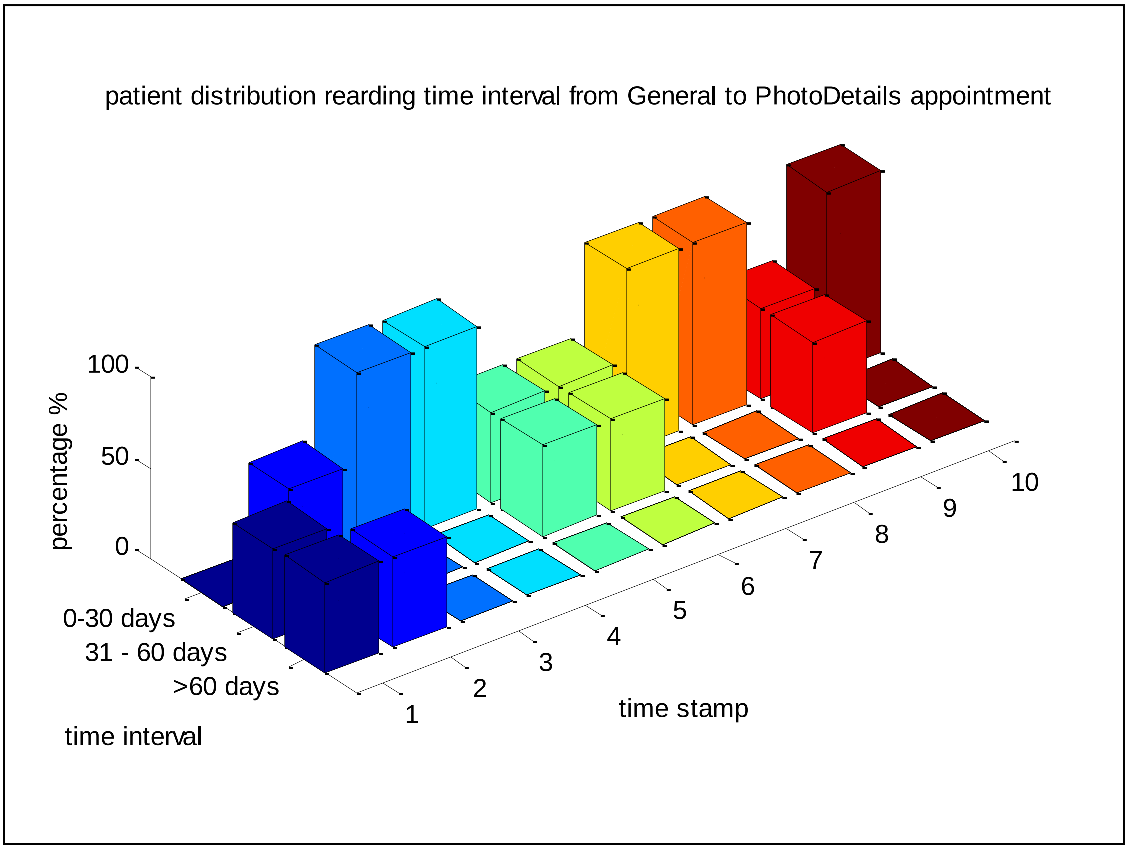


Figure A3. 17 :Patient distribution in intervals for every time stamp, from General to Photodetails – Series 2

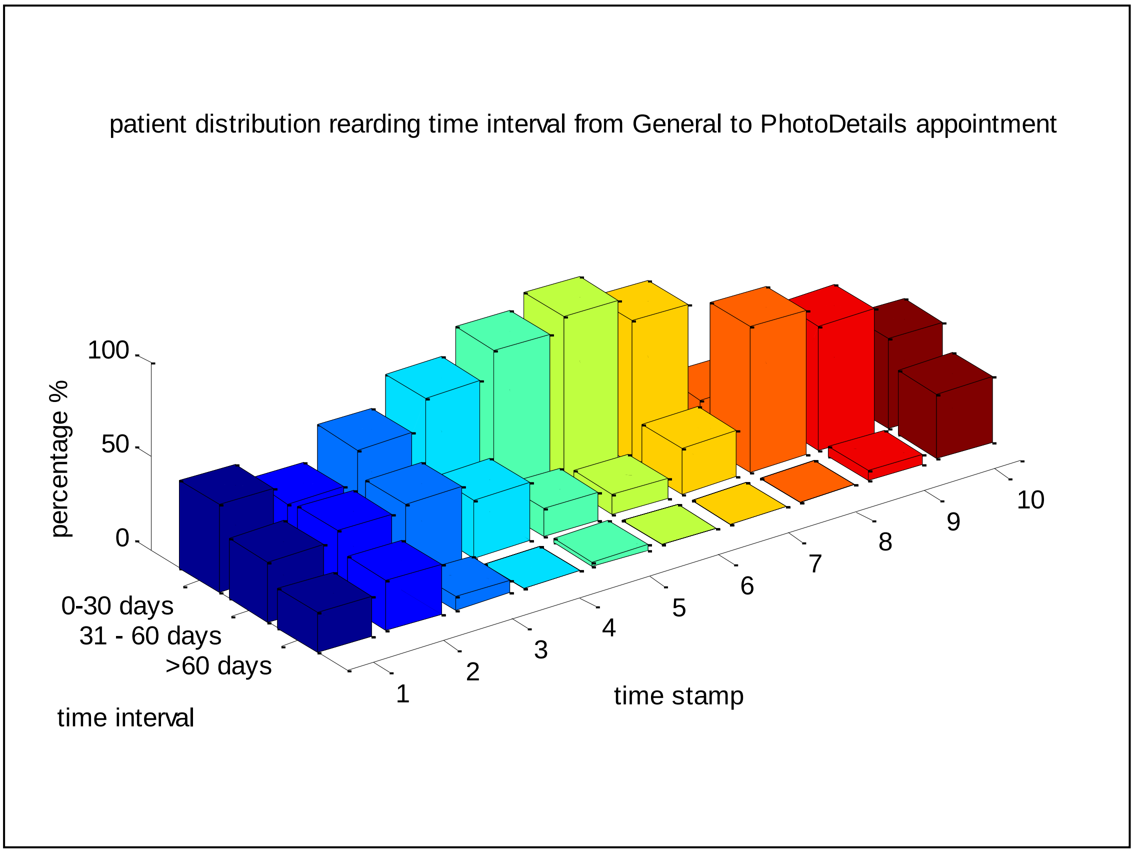


Figure A3. 18 :Patient distribution in intervals for every time stamp, from General to Photodetails – Series 3

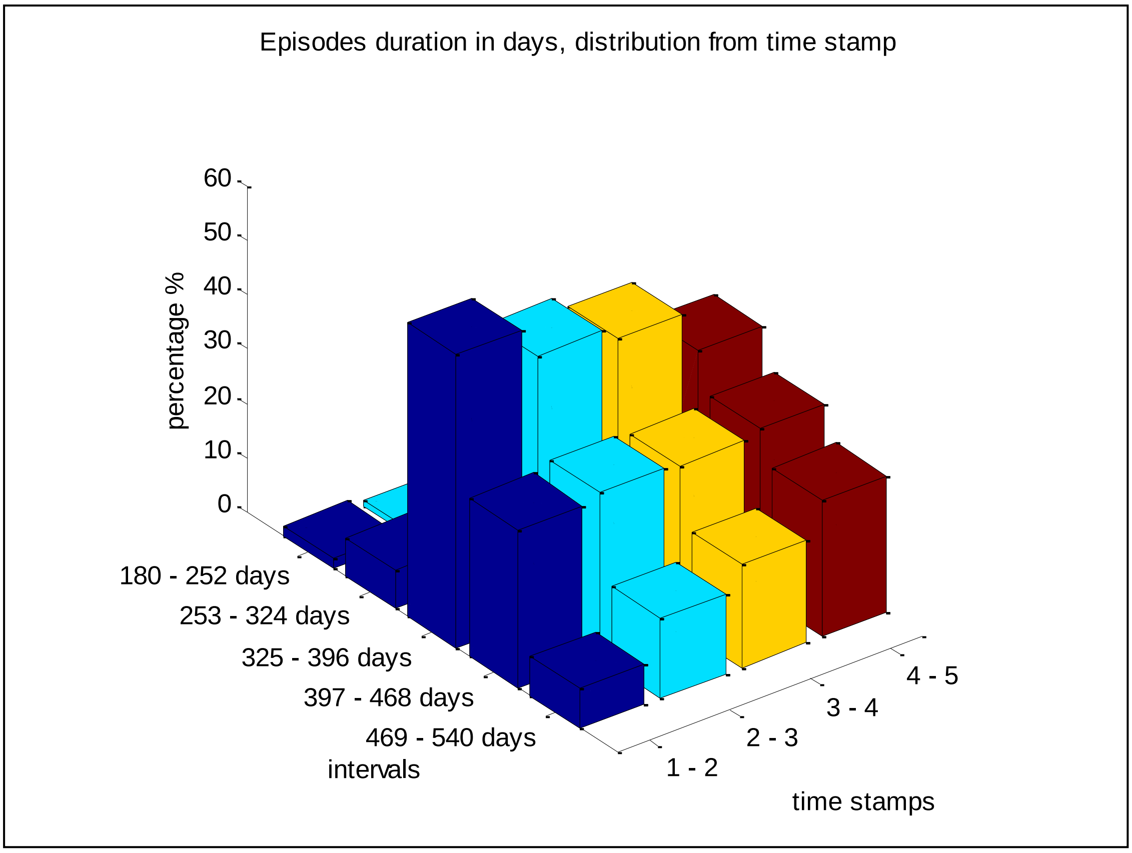


Figure A3. 19 :Patient distribution in intervals from previous to next time stamp – Series 1

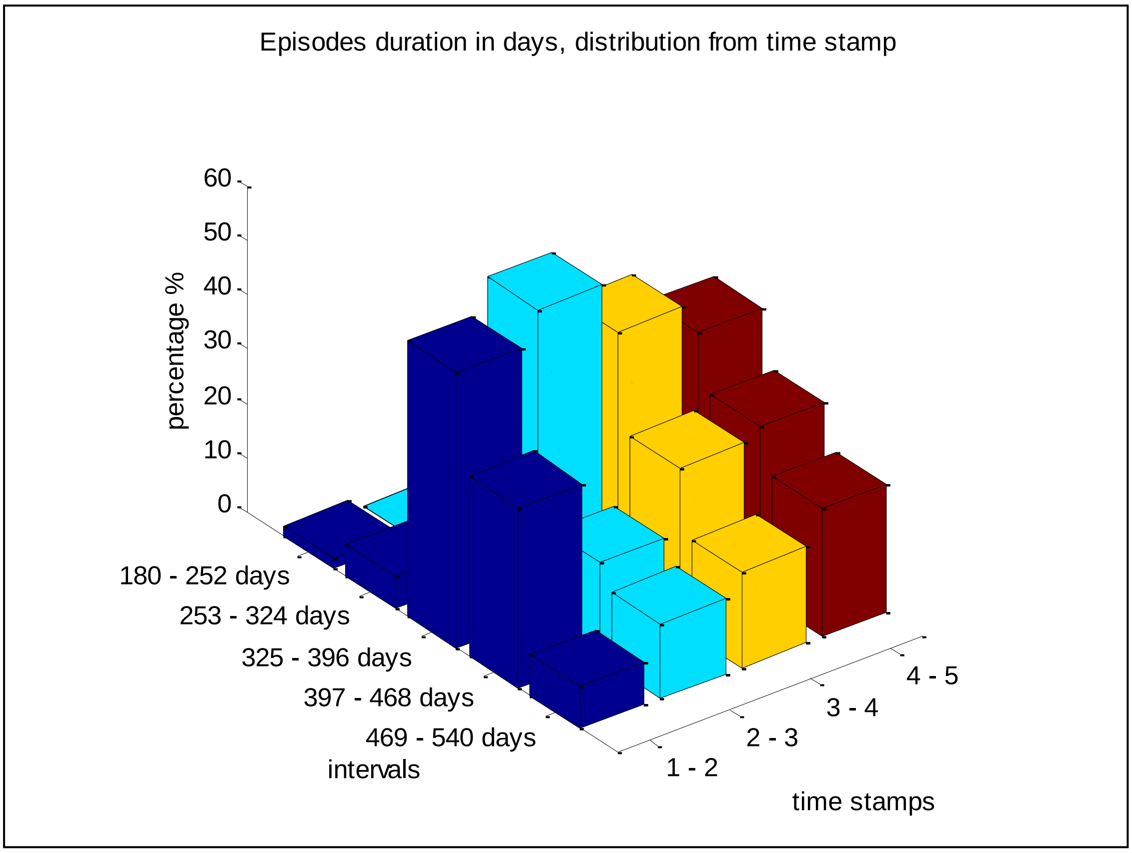


Figure A3. 20 :Patient distribution in intervals from previous to next time stamp – Series 2

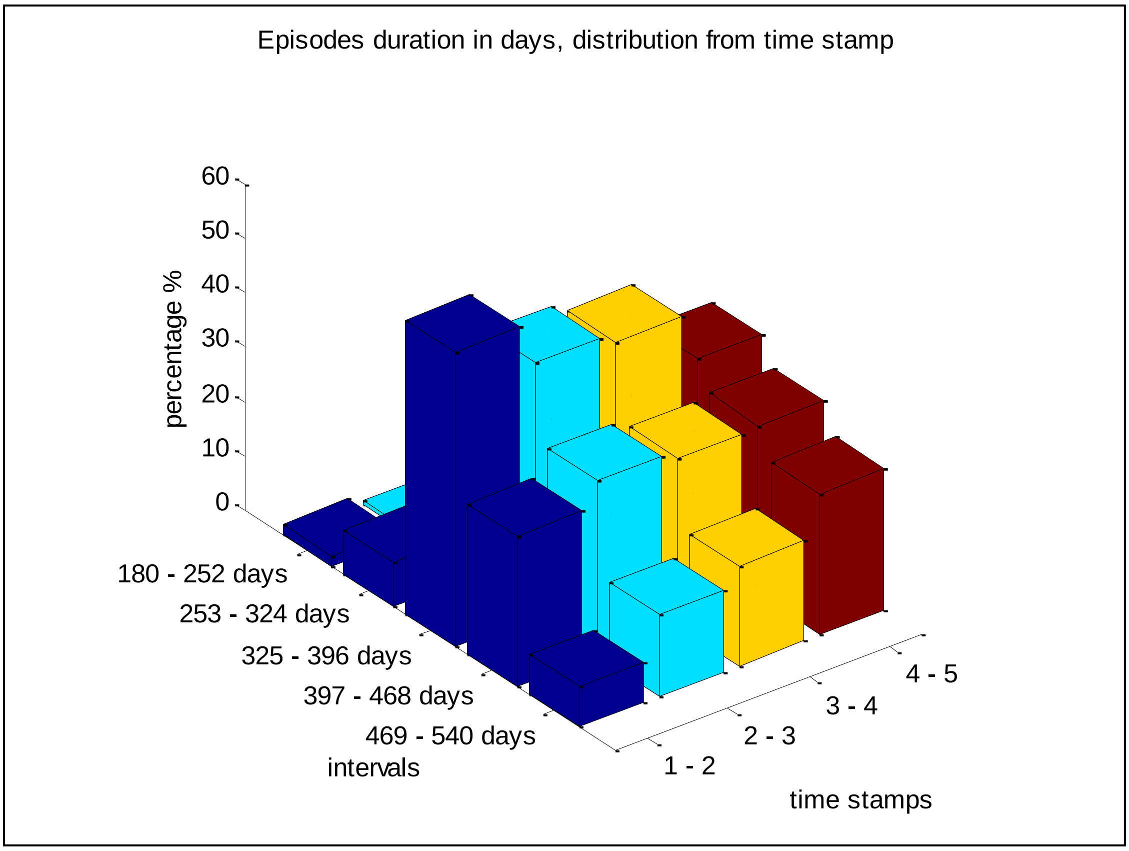


Figure A3. 61 :Patient distribution in intervals from previous to next time stamp – Series 3

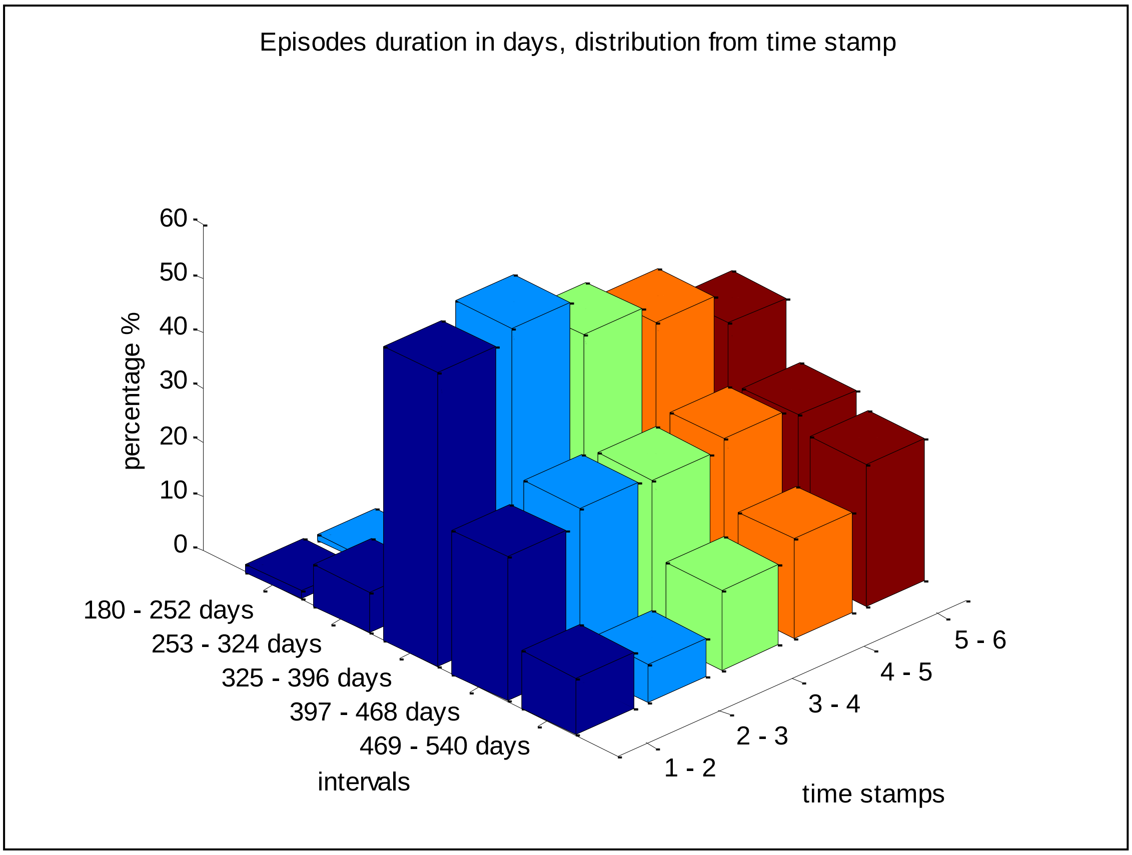


Figure A3. 22 :Patient distribution in intervals from previous to next time stamp – Series 1

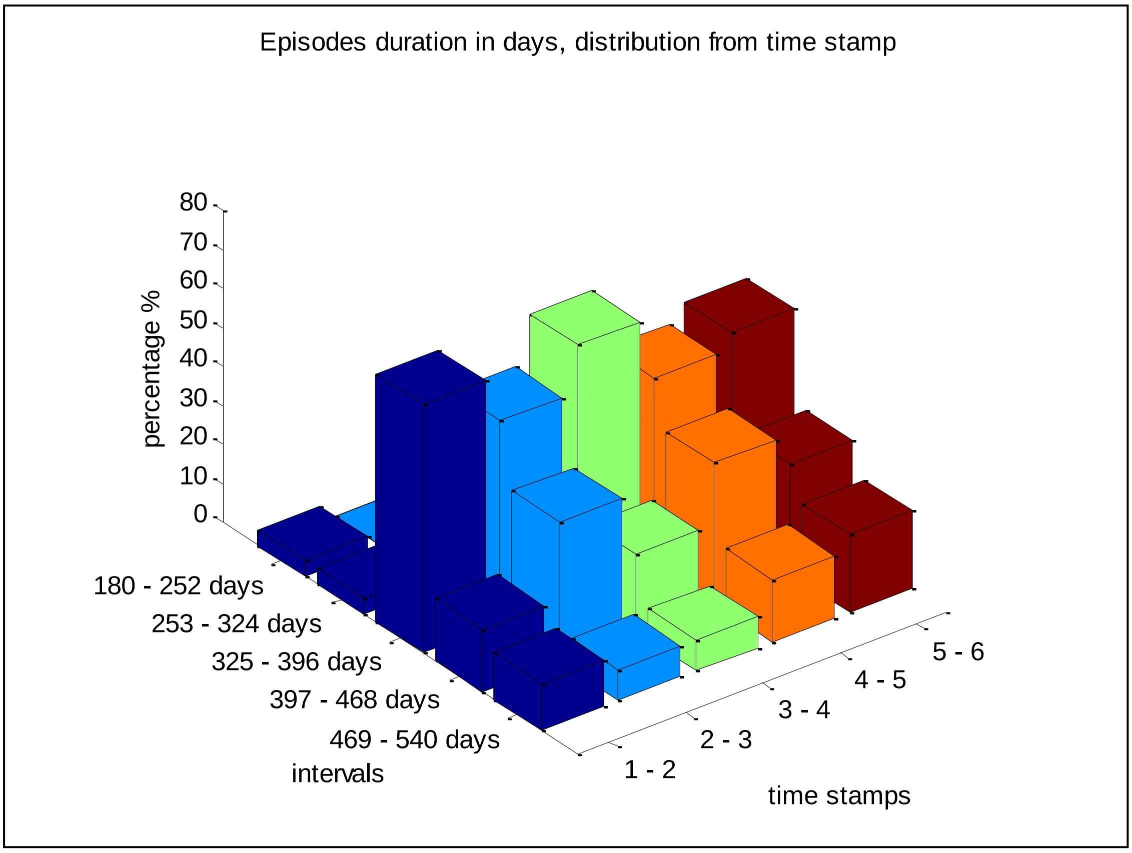


Figure A3. 23 :Patient distribution in intervals from previous to next time stamp – Series 2

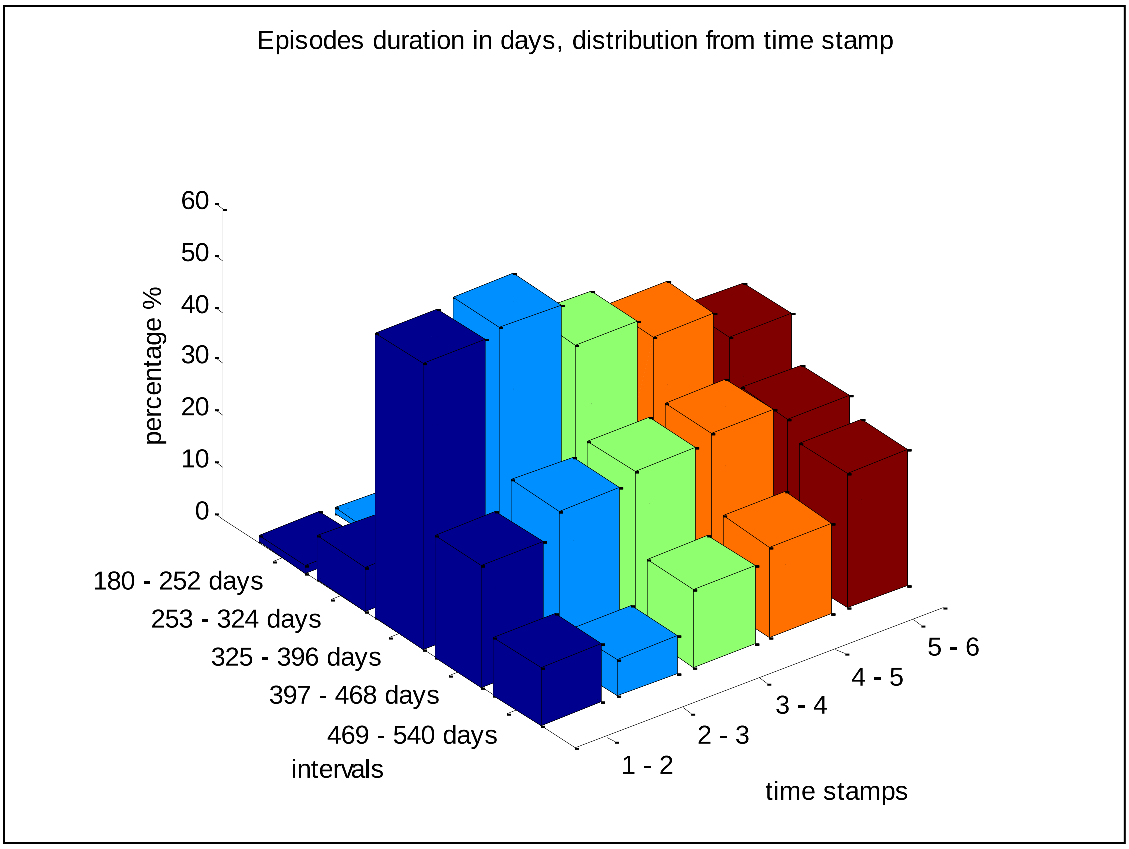


Figure A3. 24 :Patient distribution in intervals from previous to next time stamp – Series 3

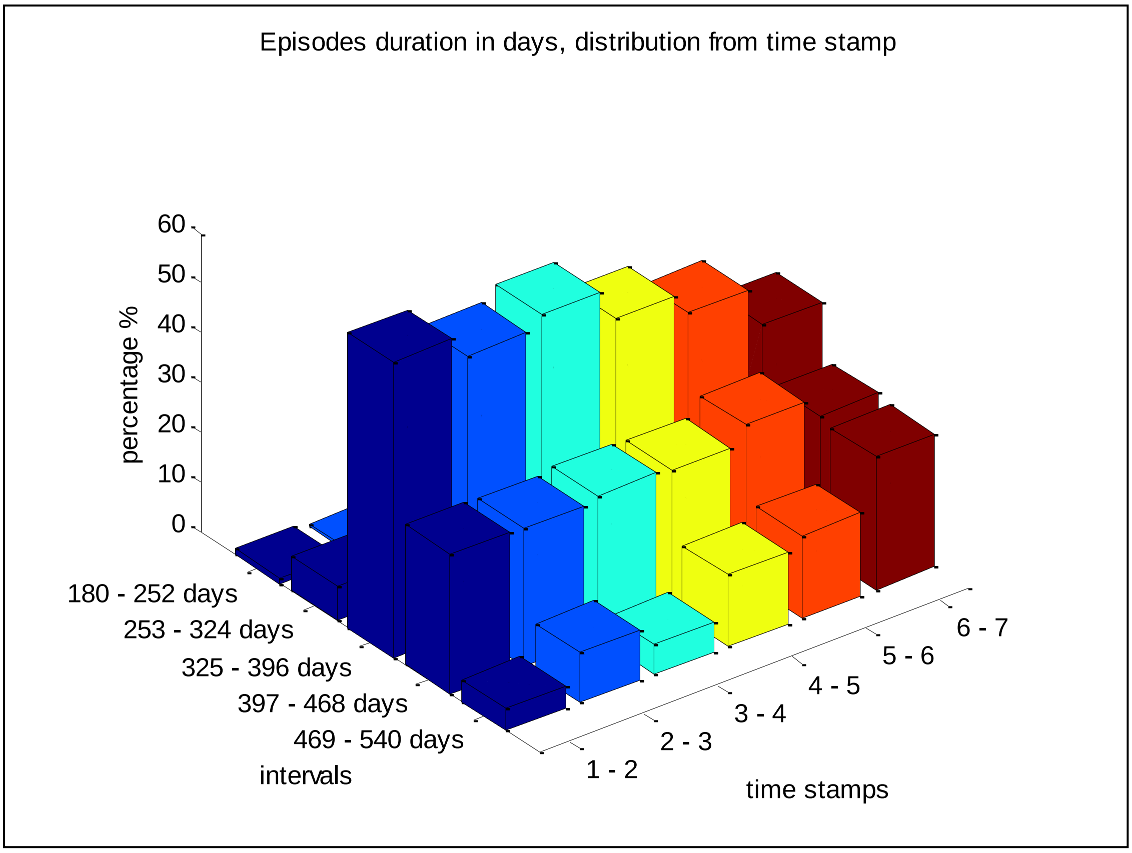


Figure A3. 25 :Patient distribution in intervals from previous to next time stamp – Series 1

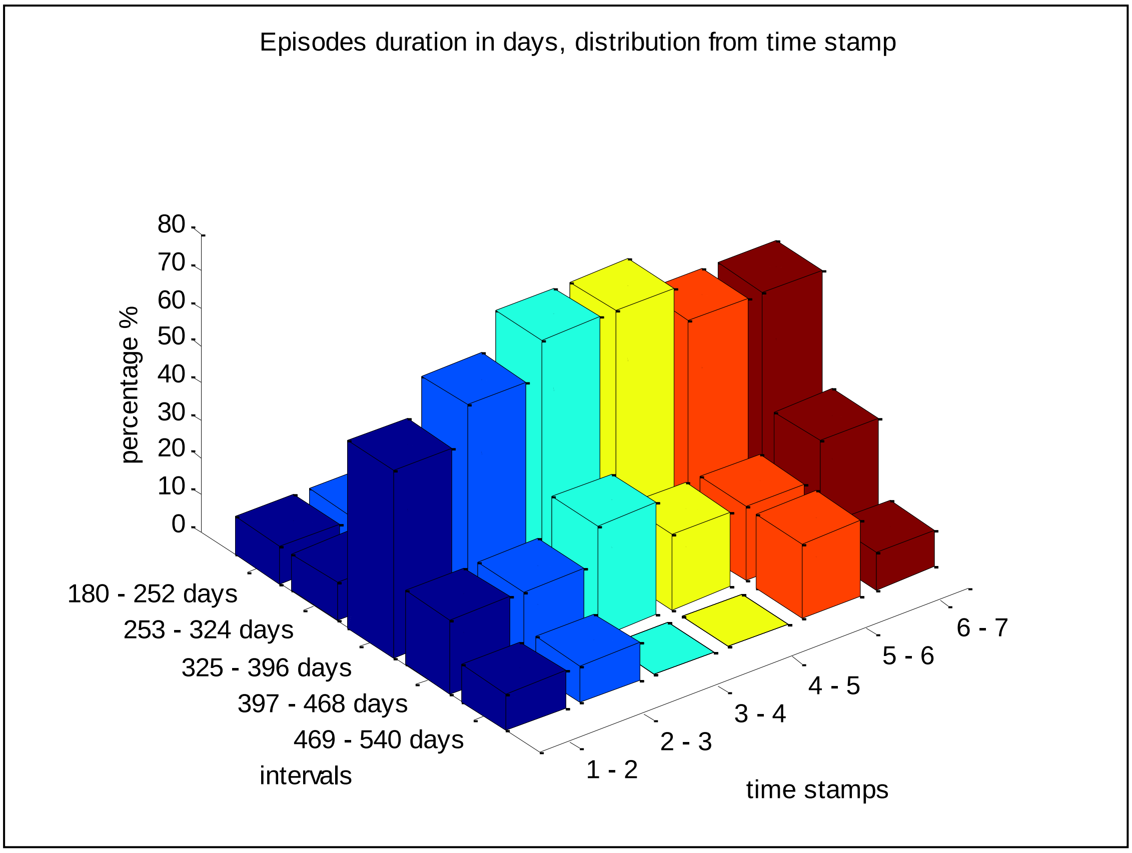


Figure A3. 26: Patient distribution in intervals from previous to next time stamp – Series 2

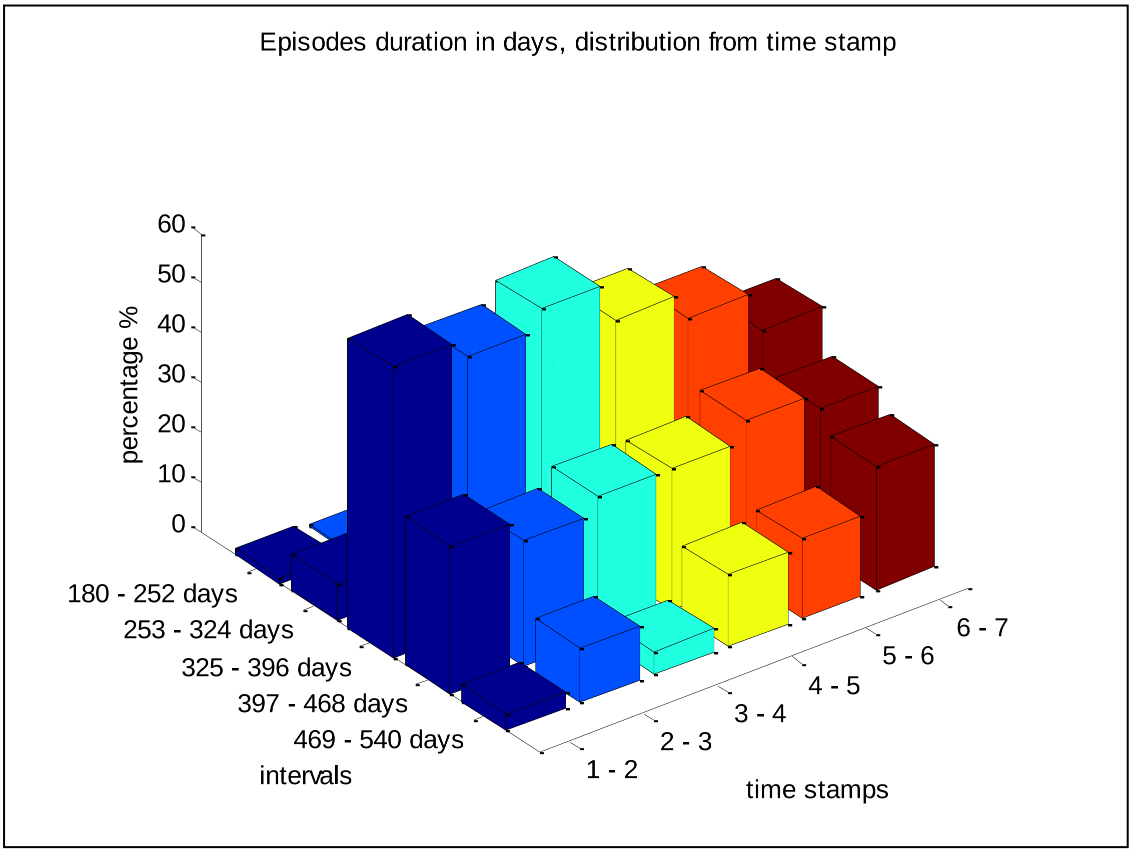


Figure A3. 27 : Patient distribution in intervals from previous to next time stamp – Series 3

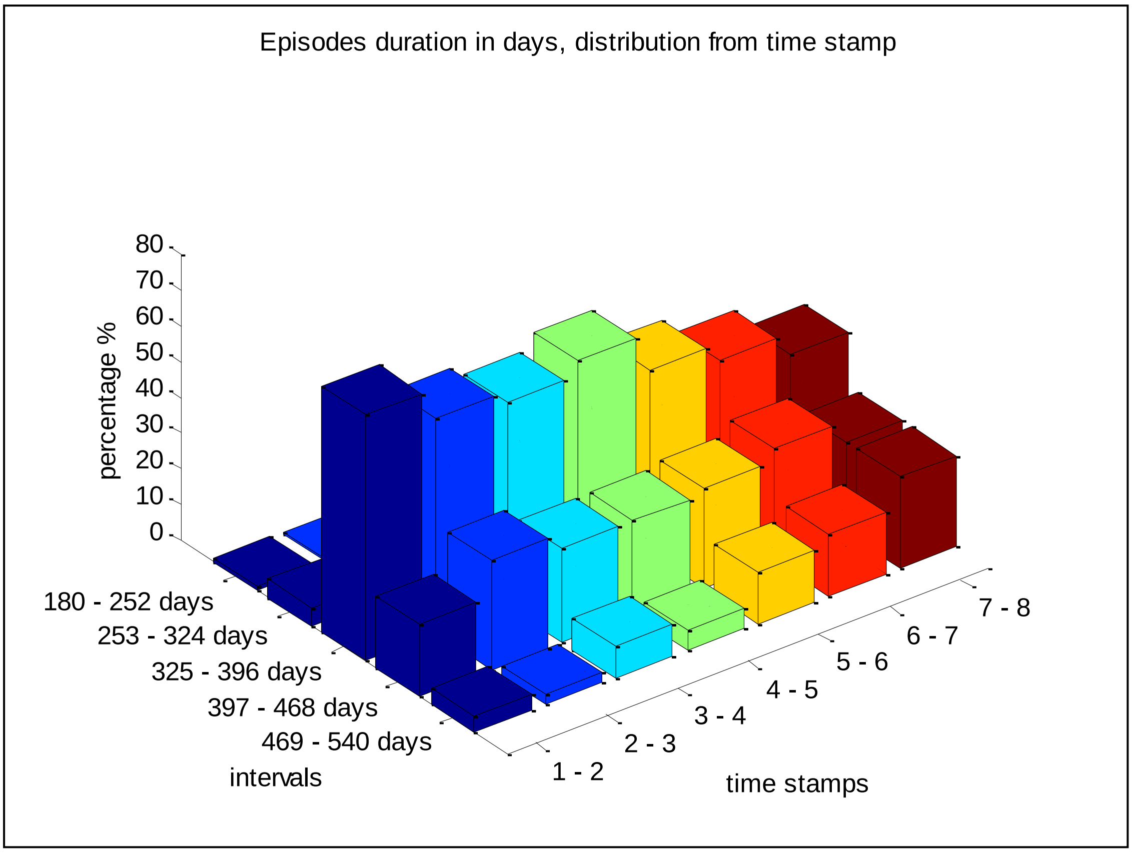


Figure A3. 28 : Patient distribution in intervals from previous to next time stamp – Series 1

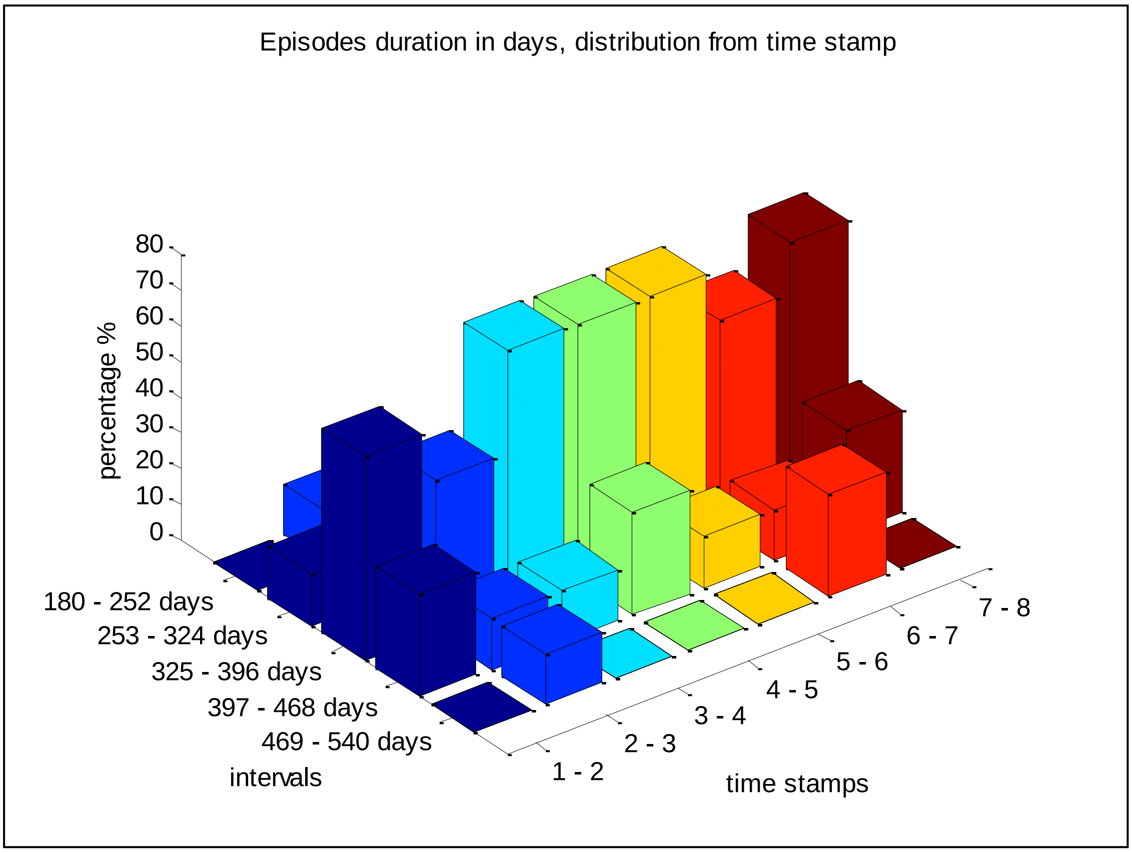


Figure A3. 29 : Patient distribution in intervals from previous to next time stamp – Series 2

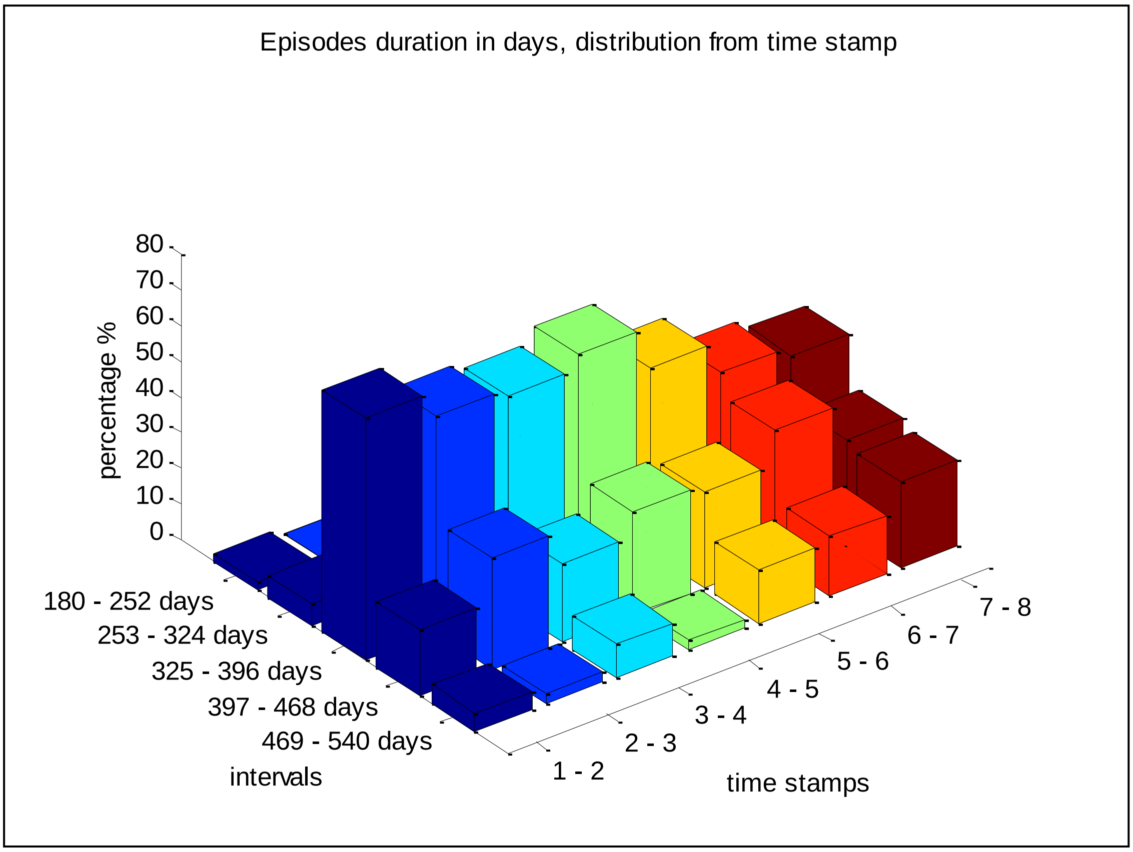


Figure A3. 30: Patient distribution in intervals from previous to next time stamp – Series 3

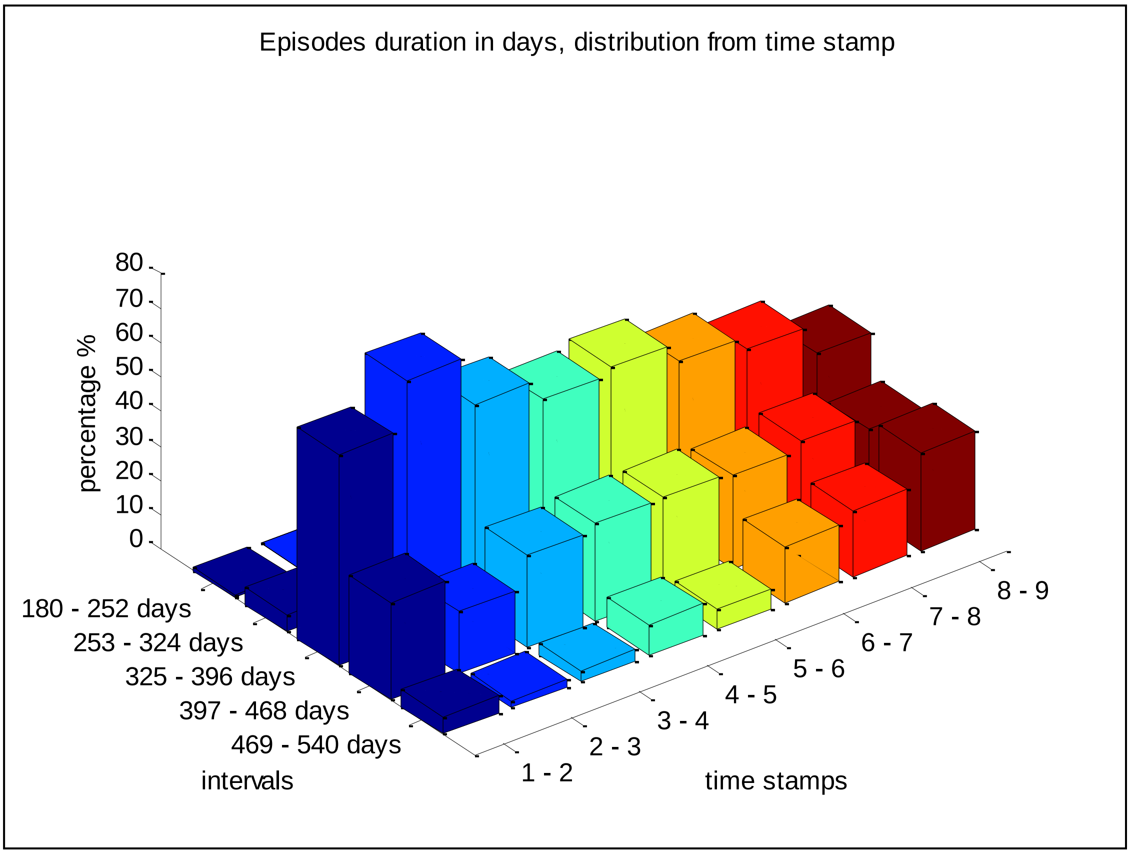


Figure A3. 31: Patient distribution in intervals from previous to next time stamp – Series 1

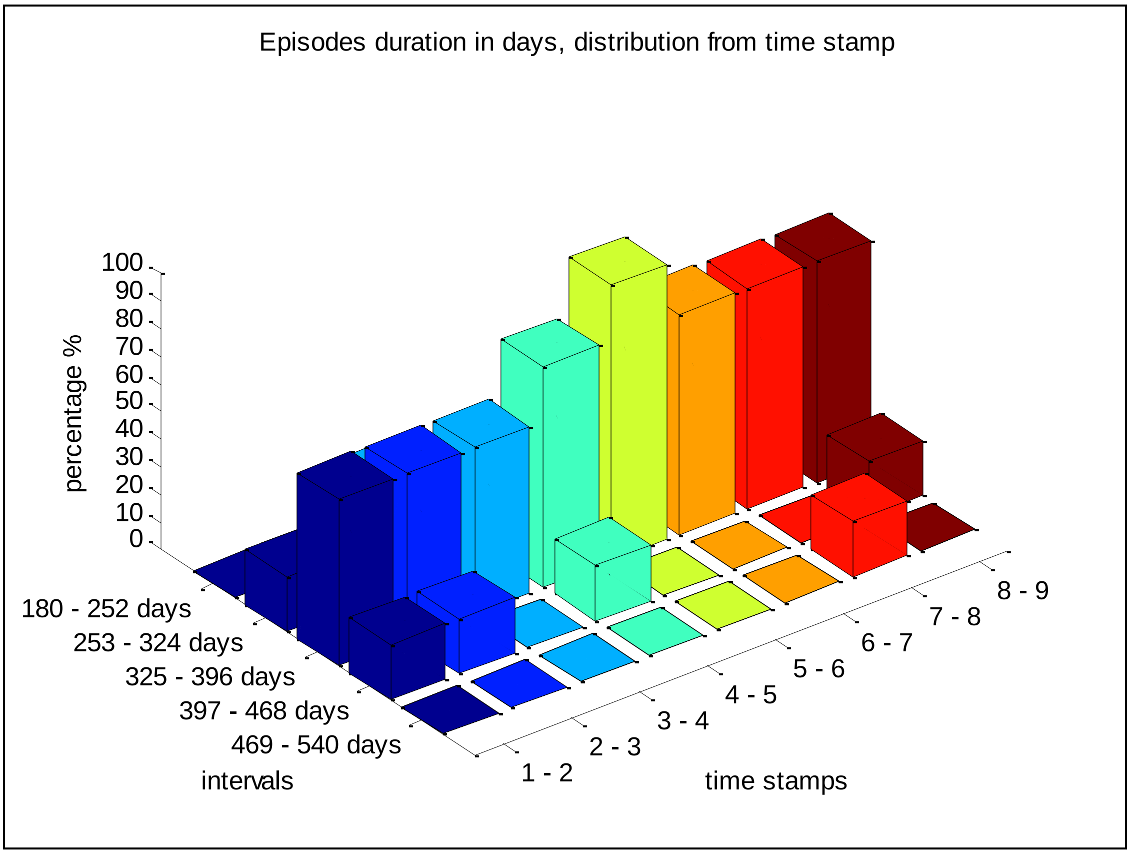


Figure A3. 32 : Patient distribution in intervals from previous to next time stamp – Series 2

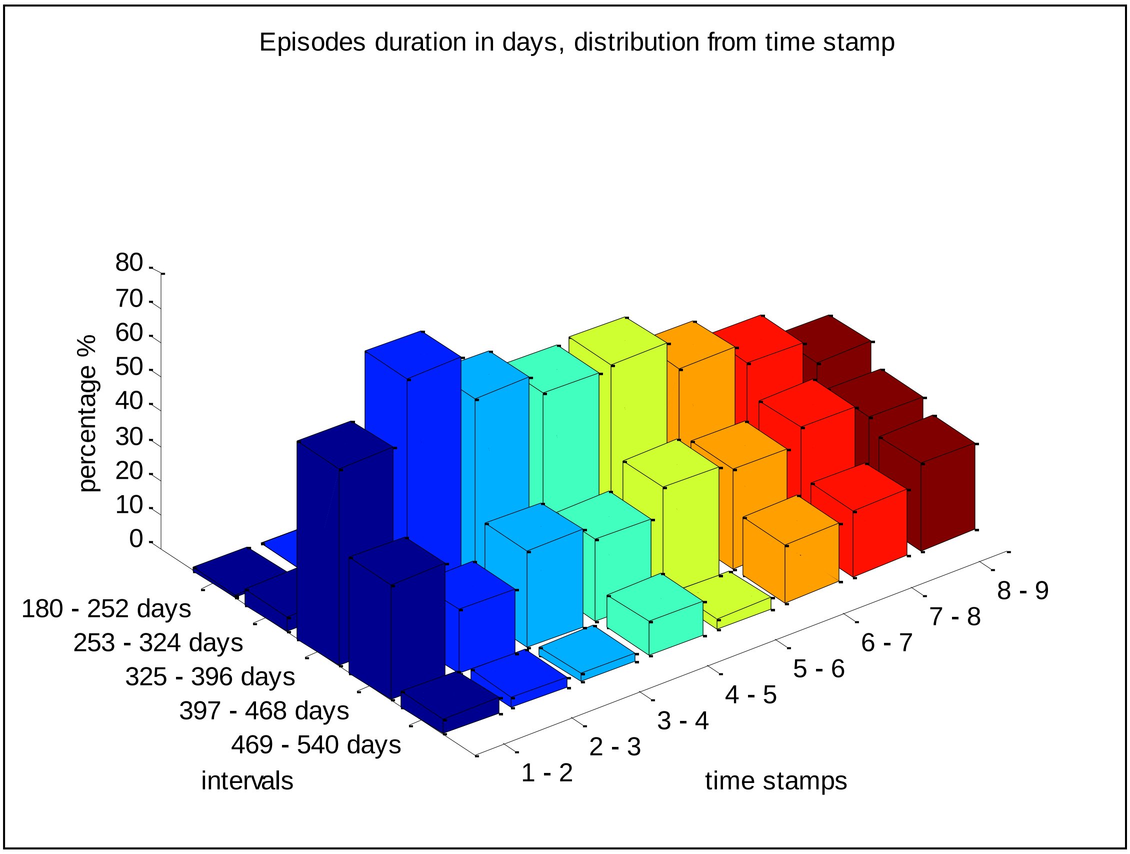


Figure A3. 33 :Patient distribution in intervals from previous to next time stamp – Series 3

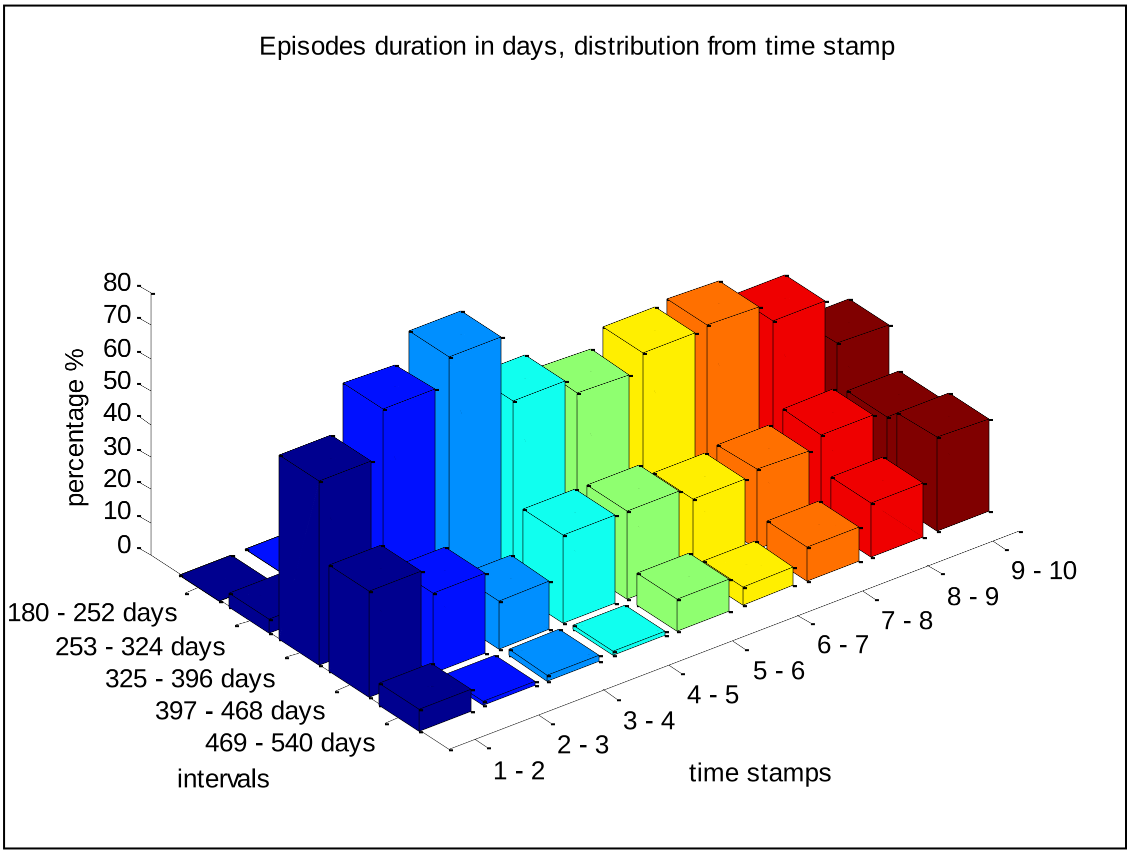


Figure A3. 34 : Patient distribution in intervals from previous to next time stamp – Series 1

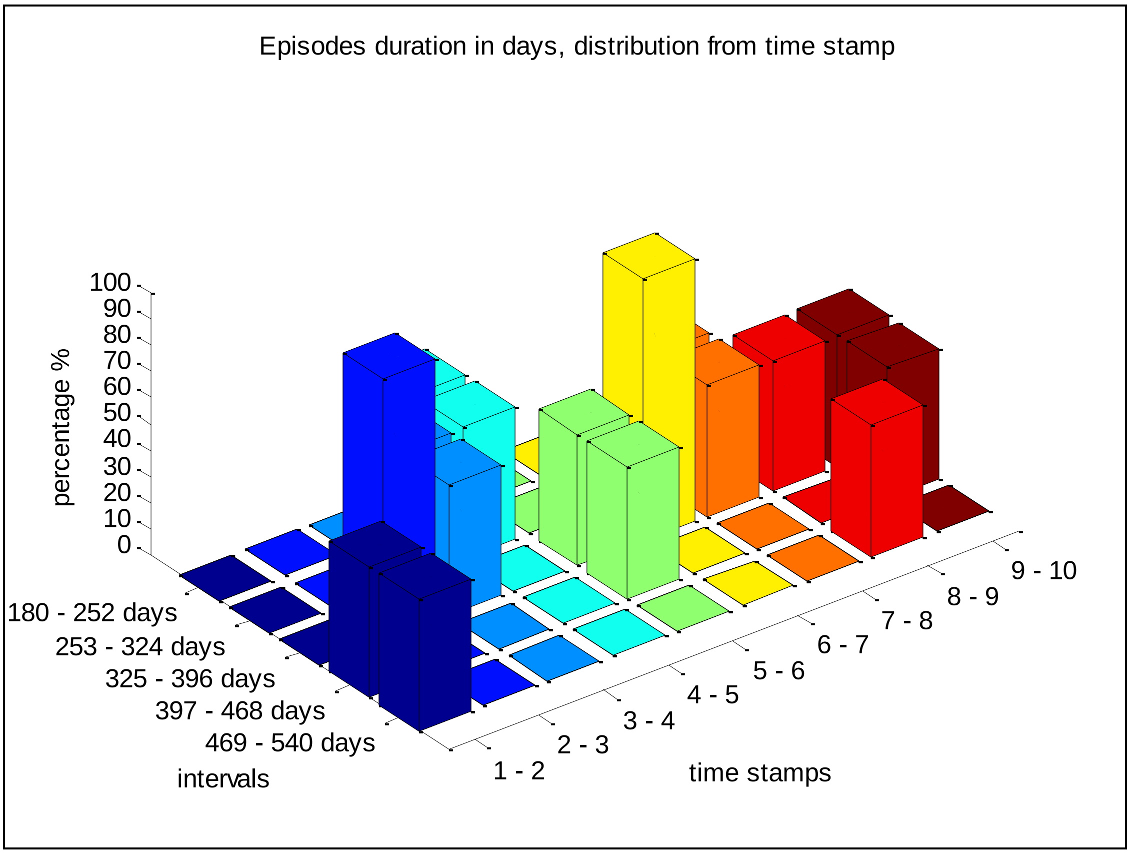


Figure A3. 35: Patient distribution in intervals from previous to next time stamp – Series 2

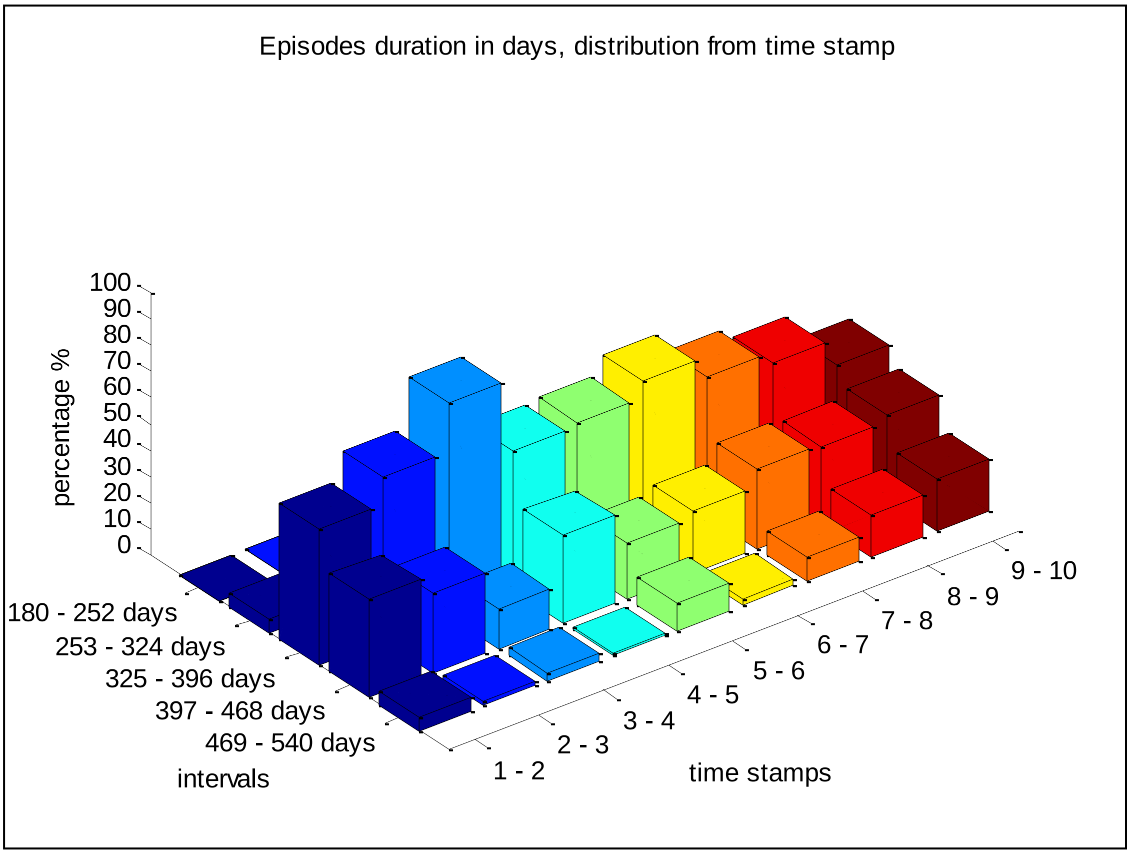


Figure A3. 36: Patient distribution in intervals from previous to next time stamp – Series 3

Appendix 4 – Trends representation

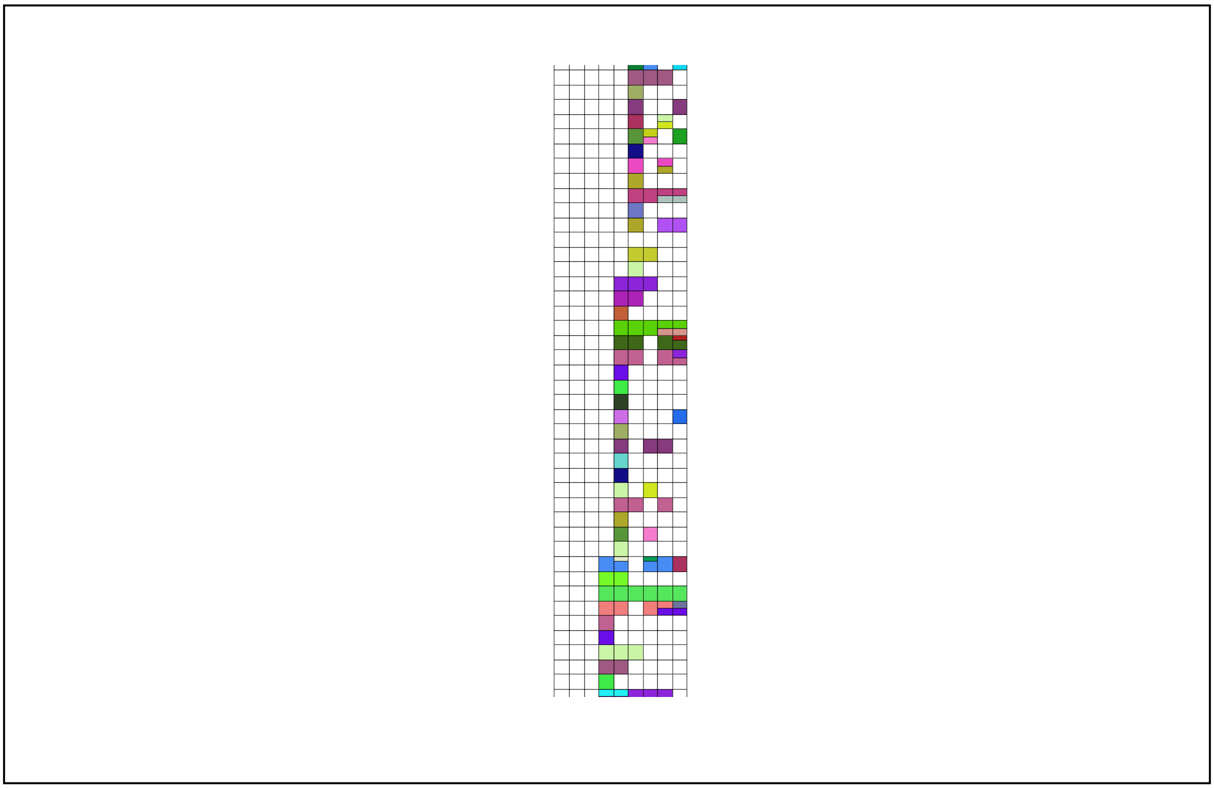


Figure A4.1: Snapshot of mosaic representation of trends from experiment with 9 time stamps

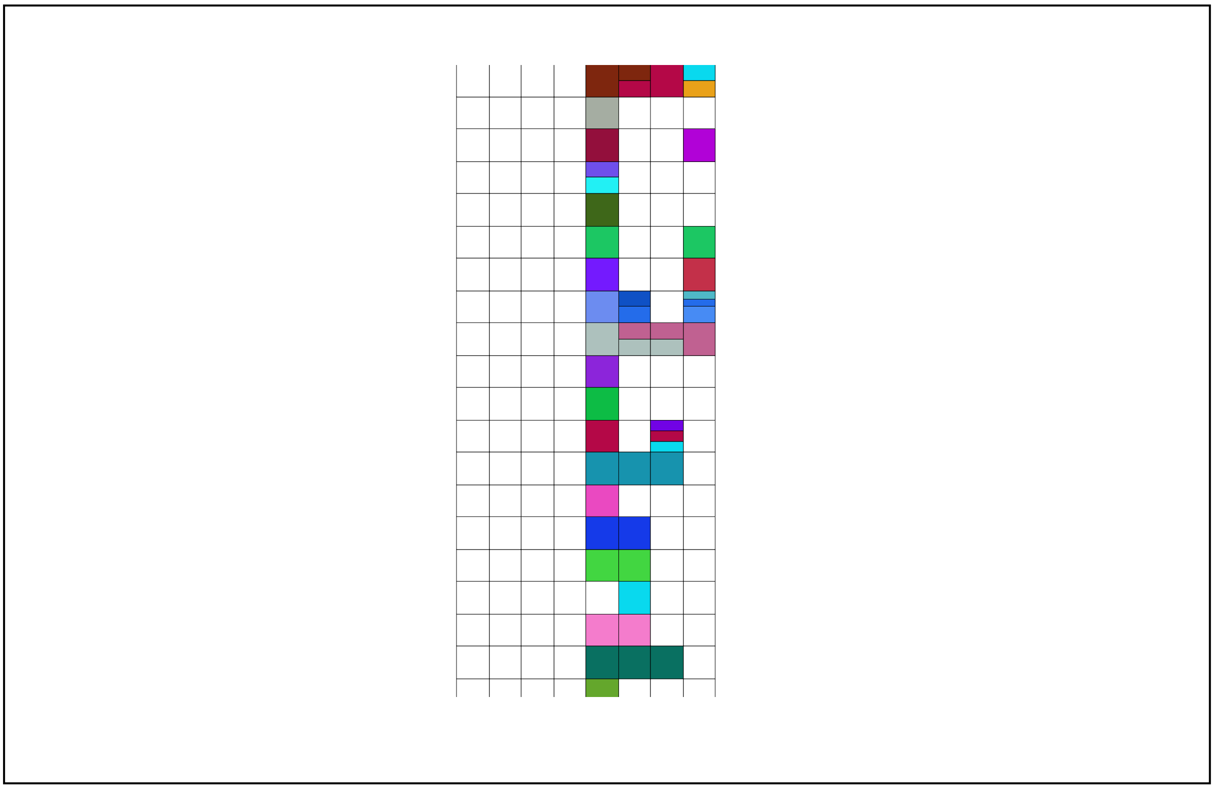


Figure A4.2: Snapshot of mosaic representation of trends from experiment with 8 time stamps

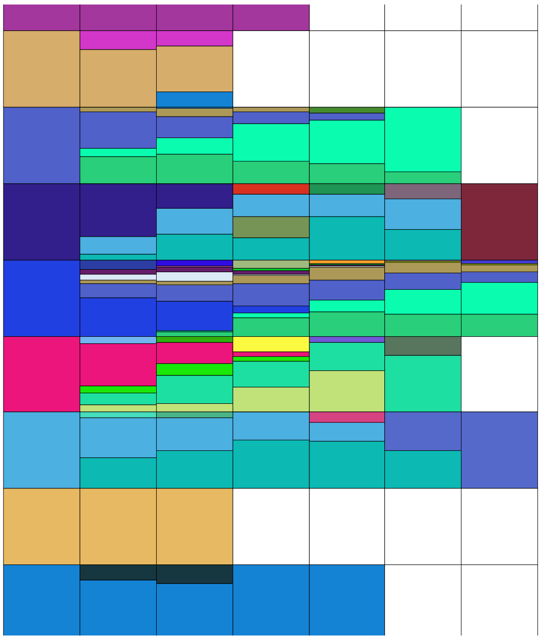


Figure A4.3 Snapshot of mosaic representation of trends from experiment with 7 time stamps

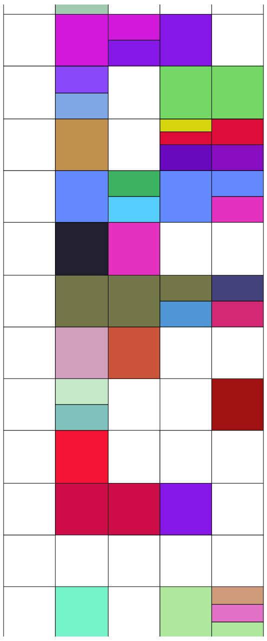


Figure A4.4 Snapshot of mosaic representation of trends from experiment with 7 time stamps

Bibliography

Agrawal R, Srikant R 1995. Mining sequential patterns. Proceedings of the 11th international conference on data engineering, 3–14.

Agrawal, R., Mannila, H.; Srikant, R., Toivonen, H., and Verkamo, I. 1996. Fast Discovery of Association Rules. In Advances in Knowledge Discovery and Data Mining, eds. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 307–328. AAAI Press.

Agrawal,R., and Srikant,R., 1994. Fast Algorithms for Mining Association Rules.20th VLDB Conference, 487-499.

Agrawal,R., Imielinski,T. and Swami,A.,1993. Database Mining: A Performance Perspective. IEEE Trans. Knowl. Data Eng., 5, 914-925

Apte,C. and Hong,S.J., 1996.Predicting equity returns from securities data with minimal rule generation.Advances in Knowledge Discovery and Data Mining,514 - 560.
Araki, A., Ito, H., Hattori,A., et al. 1993.Risk factors for development of retinopathy in elderly Japanese patients with diabetes mellitus. *Diabetes Care*.;16(8):1184–1186.

Baralis, E., Chiusano, S. and Graza, P., 2004. On support thresholds in associative classification. Proceedings of the 2004 ACM Symposium on Applied Computing, 553–558.

Bayardo,J.R. and Rakesh Agrawal,R., 1999. Mining the Most Interesting Rules, KDD, 145-154

Bellazzi, R., Larizza, C., Magni, P. and Bellazzi, R., 2005.Temporal data mining for the quality assessment of hemodialysis services. Artif. Intell. Med., 34, 25–39.

Branstad,M.A. and Cherniavsky,J.C., 1982. Validation, Verification, and Testing of Computer Software. ACM Comput. Surv., 14, 159-192.

Breault,J., Goodall,C. and Fos,P.,2002. Data mining a diabetic data warehouse. Artificial Intelligence in Medicine,26, 37-54.

Breiman,L., Friedman,J., Stone,C.J. and Olshen, R.A., 1984.Classification and Regression Trees.

Brin, S., Motwani, R., Ullman, J. and Tsur,S., 1997. Dynamic item set counting and implementation rules for market basket data. Proceedings of the 1997 ACM SIGMOD international conference on Management of data, 255-264.

Chen, M.S., Kao ,C.S., Fu ,C.C., Chen, C.J., Tai, T.Y., 1995. Incidence and progression of diabetic retinopathy among non-insulin-dependent diabetic subjects: a 4-year follow-up. *Int J Epidemiol*.,24,4,787– 795.

Chen, X., Petrounias, I. and Heathfield, H., 1998. Discovering temporal association rules in temporal databases. Proceedings of the international workshop on issues and applications of database technology (IADT'98), 312–319

Cheung, D.W., Han, J., Ng, V., Fu, A., and Fu, Y., 1996. A fast distributed algorithm for mining association rules. In: Proceeding of the 1996 international conference on parallel and distributed information systems, Miami Beach, FL, 31–44.

Cios, K.J. and Moore, G.W.,2000. Medical data mining and knowledge discovery: an overview. In: Cios KJ, editor. Medical data mining and knowledge discovery. Heidelberg: Springer, 1–16.

Cios, K.J., Teresinska, A., Konieczna, S., Potocka, J., Sharma, S., 2000.Diagnosing Myocardial Perfusion from PECT Bull's-eye Maps - A Knowledge Discovery Approach, *IEEE Engineering in Medicine and Biology Magazine*, Special issue on Medical Data Mining and Knowledge Discovery, 19, 17-25.

Cohen ,O., Norymberg, K., Neumann, E., Dekel, H., 1998. Complication-free duration and the risk of development of retinopathy in elderly diabetic patients. *ArchIntern Med*., 158,641–644.

Cohen, P.R., 2001. Fluent learning: elucidating the structure of episodes. In Hoffmann, F., Hand, D., Adams, N., Fisher, D., Guimaraes,G.,(eds), Proceedings of the 4th International conf. on Intelligent data analysis (IDA 2001), 268–277.

Claster, W., Shanmuganathan, S., & Ghotbi, N. ,2008. Text Mining of Medical Records for Radiodiagnostic Decision-Making. Journal of Computers, 3(1), 329-333.

Delen,D., Walker,G. and Kadam,A., 2005.Predicting breast cancer survivability: a comparison of three data mining methods. Artificial Intelligence in Medicine,34,113-127

Dong, G. and Li,J., 1999.Efficient Mining of Emerging Patterns: Discovering Trends and Differences. Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, 43-52.

Dong, G., Zhang, X., Wong, L. and Li, J., 1999. CAEP: Classification by aggregating emerging patterns.Proceedings of the 2nd international conference on discovery science.

Fan, H. and Kotagiri, R., 2003. A Bayesian Approach to Use Emerging Patterns for classification. Proceedings of the 14th Australasian database conference, 17, 39-48.

Fan,H. and Ramamohanarao,K, 2006. Fast discovery and the generalization of strong jumping emerging patterns for building compact and accurate classifiers. IEEE Transactions on Knowledge and Data Engineering,18, 721-737.

Fayyad, U. M., Djorgovski, S. G. and Weir, N. 1996.From Digitized Images to On-Line Catalogs: Data Mining a Sky Survey. AI Magazine, 17, 51–66.

Fayyad, U. M., Piatetsky-Shapiro, G., and Smyth, P., 1996a. From Data Mining to Knowledge Discovery in Databases.AI Magazine, 17, 37-54.

Fayyad, U. M., Piatetsky-Shapiro, G., and Smyth, P., 1996b. From Data Mining to Knowledge Discovery:An Overview. In Advances in Knowledge Discovery and Data Mining, eds. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 1–30.

Fayyad, U.M., Piatesky-Shapiro, G., Smyth, P., and Uthurusamy, R., 1996. *Advances in Knowledge Discovery and Data Mining*, AAAi/MIT Press,

Glymour,C., Scheines,R., Spirtes,P. and Kelly,K., 1987.Discovering Causal Structure: Artificial Intelligence, Philosophy of Science, and Statistical Modeling. Academic Press.

Guimarães, G., Peter, J.H., Penzel,T. and Ultsch, A., 2001. A method for automated temporal knowledge acquisition applied to sleep-related breathing disorders. Artif. Intell. Med., 23, 211–37.

Guimarães,G. and Ultsch,A., 1999.A method for temporal knowledge conversion. Proceedings of the 3rd international symposium on advances in intelligent data analysis (IDA 1999), pp 369–382.

Han, J., Pei, J., and Yin, Y., 2000. Mining frequent patterns without candidate generation. In Proc. 2000 ACMSIGMOD Int. Conf. Management of Data (SIGMOD'00), 1–12.

Han,J., Kamber.,M and Pei,J., 2012.Data mining concepts and techniques. Elsevier

Hand, D.J., 1981.Discrimination and Classification. Wiley.

Harding, S.P., Broadbent, D.,2009- personal communication.

Harris ,E.L., Sherman, S.H., Georgopoulos ,A. 1999. Blackwhite differences in risk of developing retinopathy among individuals with type 2 diabetes. *Diabetes Care*.; 22,779–783.

Heckerman, D. 1996. Bayesian Networks for Knowledge Discovery. In Advances in Knowledge Discovery and Data Mining, eds. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 273–306. AAAI Press.

Höppner, F., 2003. Knowledge discovery from sequential data. PhD thesis, Technical University Braunschweig, Germany.

Höppner,F. and Klawonn,F., 2002. Finding informative rules in interval sequences. Intell. Data Anal.Int. J., 6,237–256.

Jain, A. K. and Dubes, R. C. 1988. Algorithms for Clustering Data. Prentice-Hall.

Kalton, G., and Kasprzyk, D., 1986. The treatment of missing survey data. Survey Methodology, 12, pp1-16.

Kalton,G., and Kasprzyk,D .,1986. The treatment of missing survey data , in Survey Methodology, 12, 1-16.

Kam, P.S. and Fu, A.W.C., 2000. Discovering temporal patterns for interval-based events. Proceedings of 2nd international conference on data warehousing and knowledge discovery (DaWaK), 317–326.

Kanski,J., 2007.Clinical Ophthalmology: A Systematic Approach, 6th Edition, Elsevier.

Khan, M.S., Coenen, F., Reid, D., Taw_k, H., Patel, R. and Lawson, A., 2010. A Sliding Windows based Dual Support Framework for Discovering Emerging Trends from Temporal Data. Journal Knowledge-Based Systems, 23, 316-322.

Kim, H.K., Kim ,C.H., Kim, S.W., et al. 1998. Development and progression of diabetic retinopathy in Koreans with NIDDM. *Diabetes Care*.,21,1, 134–138.

Kimm, S.Y.S., Glynn, N.W., Kriska, A.M., Fitzgerald, S. L., Aaron, D. J., Similo,S.L., McMahon, R.P., Barton, B.A., 2000. Longitudinal changes in physical activity in a biracial cohort during adolescence. Medicine and Science in Sports and Exercise,32, 1445-1454.

Kohavi,R., Rothleder,N. and Simoudis,E., 2002. Emerging Trends in Business Analytics. Communications of the ACM, Evolving data mining into solutions for insights, 45, 45-48.

Kurgan, L., Cios, K.J., Sontag, M., and Accurso, F.J., 2003. Mining a Cystic Fibrosis Database, In: Zurada, J., and Kantardzic, M. (Eds.), *Novel Applications in Data Mining*.

Kurgan, L., Cios, K.J., Tadeusiewicz, R., Ogiela, M. and Goodenday, L.S., 2001. Knowledge Discovery Approach to Automated Cardiac SPECT Diagnosis, *ArtificialIntelligence in Medicine*, 23:2, pp. 149-169.

Larsson, L.I, Alm, A., Bergenheim, T., Lithner, F., Bergstrom, R., 1999. Retinopathy in diabetic patients aged 15–50 years in the county of Umea, Sweden. *Acta Ophthalmol Scand*.,77,430–436.

Last, M., Klein, Y. and Kandel,A., 2001. Knowledge discovery in time series databases. IEEE Trans. Syst. Man. Cybernet, 31,160–169.

Lavrac, N., 1999.Selected techniques for data mining in medicine. Artificial Intelligence in Medicine, 16, 3-23.

Levy, M.L., Cummings, J.L., Fairbanks, L.A., Bravi, D., Calvani M. and Carta, A.1996. Longitudinal assessment of symptoms of depression, agitation, and psychosis in 181 patients with Alzheimer's disease. American Journal of Psychiatry; Num.153, pp1438-1443.

Li J, Dong G, and Ramamohanarao K 2001. Making use of the most expressive jumping emerging patterns for classification. Knowl. Inf. Syst. Int. J., 3, 131–145.

Li, J., Ramamohanarao, K. and Dong,G., 2000. The Space of Jumping Emerging Patterns and Its Incremental Maintenance Algorithms. Proceedings of the Seventeenth International Conference on Machine Learning, 551-558.

Li, Y., Ning, P., Wang, X.S. and Jajodia, S., 2003. Discovering calendar-based temporal association rules. Data Knowl. Eng., 44, 193–218.

Li,J., Zhang,X., Dong,G., Ramamohanarao,K. and Qun Sun.Q., 1999. Efficient mining of high confidence association rules without support thresholds. Proceedings of Principles of Data Mining and Knowledge Discovery (PKDD), 406-411.

Li,L., Tang,H., Wu,Z., Gong,J., Gruidl,M., Zou,J., Tockman,M. and Clark,R., 2004.Data mining techniques for cancer detection using serum proteomic profiling.Artificial Intelligence in Medicine, 32, 71-83.

Lim, T.S., Loh,W.Y. and Shih, Y.W., 2000. A comparison of prediction accuracy, complexity and training time of tirthy three old and new classification algorithms.Machine Learning, 40 , 203-228.

Lin, J.L. and Dunham, M.H., 1998. Mining association rules: anti-skew algorithms. Proceedings of 14th International Conference on Data Engineering, 486 - 493.

Lin, J.L. and Dunham, M.H., 2000. A low-cost checkpointing technique for distributed databases. Distributed And Parallel Databases, 10, 241-268.

Lin, M-Y. and Lee, S-Y., 2005. Fast discovery of sequential patterns through memory indexing and database partitioning. J. Inf. Sci. Eng., 21, 109–128.

Little, R. J., and Rubin, D.B., 2002 . Statistical Analysis with Missing Data. Second Edition. John Wiley and Sons, New York.

Liu, B., Hsu, W. and Ma, Y. 1998. Integrating classification and association rule mining.Proceedings of the International Conference on Knowledge Discovery and Data Mining, 80–86.

Liu, B., Hsu, W. and Ma, Y., 1999. Mining association rules with multiple minimum supports, Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 337 – 341.

Liu,Y.; Li,Z., Xiong,H., Gao,X., Wu,J.and Wu,S., 2010.Understanding of Internal Clustering Validation Measures. Proceeding of 10th International Conference on Data Mining,911-916.

Maberley ,D.A., King, W., Cruess, A.F., Koushik A., 2002.Risk factors for diabetic retinopathy in the Cree of James Bay. *Ophthalmic Epidemiol*. ;9(3):153–167.

Mannila, H. and Toivonen, H., 1996 .Discovering generalized episodes using minimal occurrences. Proceedings of the 2nd international conference on knowledge discovery and data mining (KDD-96), 146–151.

Mannila, H., 1998.Database Methods for Data Mining.KDD-98 tutorial.
MMWRMorbMortalWkly 1993.Public health focus: prevention of blindness associated with diabetic retinopathy. *Rep*.,42,10,191–195.

Moore, G.W., Berman ,J.J., 2000.Anatomic pathology data mining. In: Cios KJ, editor. Medical data mining and knowledge discovery. Heidelberg, Springer,. 61–108.

Moreo, G., Mariani, E., Pizzamiglio, G., Colucci, G.B., 1995. Visual evoked potentials in NIDDM: a longitudinal study. *Diabetologia*,38,573–576.

Morrish ,N.J., Wang, S.L, Stevens ,L.K., Fuller ,J.H., Keen ,H., 2001. Mortality and causes of death in the WHO Multinational Study of Vascular Disease in Diabetes. *Diabetologia* , 44 Suppl 2,14–21.

Mumoz, J.F., and Rueda,M., 2009.New imputation methods for missing data using quantiles. in Journal of Computational and Applied Mathematics, Vol. 232.

Murrell,S. and Plant,R.T., 1997.A Survey of Tools for Validation and Verification 1985-1995.Decision Support Systems, 21, 307-323.

Miwa, M., Thompson, P., McNaught, J., Kell, D. B., & Ananiadou, S. ,2012. Extracting semantically enriched events from biomedical literature. BMC bioinformatics, 13(1), 108-132.

Nathan, D.M., 1995. Prevention of long-term complications of non-insulin-dependentdiabetes mellitus. *Clin InvestMed*.;18(4):332–339.

Nohuddin, P.N.E., Coenen, F., Christley, R., Setzkorn, C., Patel, Y. and Williams, S., 2012.Finding "interesting" trends in social networks using frequent pattern mining and self organizing maps.Knowledge-Based Systems, 29, 104-113.

O'Keefe,R.M. and Preece,A.D.,1996. The Development, Validation and Implementation of Knowledge-Based Systems. European Journal of Operational Research, 92,458-473.

O'Leary,D., 1993. Expert System Verification and Validation. Artificial Intelligence Review, 7, 3-42.

Pang-Ning, T., Steinbach, M. and Vipin,K., 2005.Introduction to Data Mining. Addison-Wesley.

Papapetrou, P., Kollios, G., Sclaroff, S. and Gunopulos, D., 2005. Discovering frequent arrangements of temporal intervals. Proceedings of the 5th IEEE international conference on data mining (ICDM-05). Houston, Texas, 354–361.

Park, J.S., Chen, M.S. and Yu, P.S., 1995. An effective hash-based algorithm for mining association rules. Proceeding of the ACM-SIGMOD international conference on management of data, 175–186.

Quinlan, J. R. 1990. Decision trees and decision making. IEEE Trans. Syst. Man Cybern., 20, 339–346.

Robu and Hora 2012. Medical data mining with extended WEKA. 16th International Conference on Intelligent Engineering Systems (INES),IEEE, 347-350.

Roglic, G., Unwin ,N., Bennett ,P.H., Mathers ,C., Tuomilehto, J., Nag ,S .,et al., 2005.The burden of mortality attributable to diabetes: realistic estimates for the year 2000.*Diabetes Care*, 28,9,2130–2135.

Sacchi, L., Bellazzi ,R., Larizza ,C., Magni .P., Curk ,T., Petrovic, U., Zupan, B., 2004. Clustering gene expression data with temporal abstractions. Stud Health Technol Inform.,107,Pt 2,798-802.

Sacchi,L., Larizza,C., Combi,C. and Bellazzi,R., 2007.Data mining with Temporal Abstractions: learning rules from time series. Data Mining and Knowledge Discovery, 15, 217-247.

Sacha, J.P., Cios, K.J., and Goodenday, L.S., 2000.Issues in Automating Cardiac SPECT Diagnosis. *IEEE Engineering in Medicine and Biology Magazine*, special issue on Medical Data Mining and Knowledge Discovery, 19, 78-88.

Saul, J.M., 2000. Legal policy and security issues in the handling of medical data. In: Cios KJ, editor. Medical data mining and knowledge discovery. Heidelberg, Springer,. 17–31.

Savarese, A., Omiecinsky,E. and Navathe,S., 1995. An e±cient algorithm for mining association rules in large databases, in: Proceedings of the 21st International Conference on Very Large Databases.

Schneier, B., 1996. Applied cryptography. Protocols, algorithms, and source code in C. 2nd ed. New York: Wiley.

Silva, A., Cortez, P., Santos, M.F., Gomes, L. and Neves, J., 2008. Rating organ failure via adverse events using data mining in the intensive care unit. Artificial Intelligence in Medicine, 43, 179-193.

Silverman, B. 1986. Density Estimation for Statistics and Data Analysis.Chapman and Hall.

Singer, J.D., and Willet, J.B., 2003 .Applied longitudinal data analysis modelling change and event occurrence . Oxford University Press.

Skinner, J.D., Carruth, B.R., Wendy, B., and Ziegler, P.J., 2002. Children's Food Preferences A Longitudinal Analysis. Journal of the American Dietetic Association,Volume 102, Issue 11, pp1638-1647.

Somaraki, V., Broadbent, D., Harding,S.P. and Coenen,F., 2010. Finding Temporal Patterns in Noisy Longitudinal Data: A Study in Diabetic Retinopathy. Lecture Notes in Artificial Intelligence, 6171, 418-431.

Soulet,A., Cremilleux,B. and Francois R., 2004.Condensed representation of EPs and patterns quantified by frequency-based measures. proceedings of KDID 2004, Lecture Notes in Computer Science, 3377,173-189.

Srimani, P. K. and Koti, M. S., 2011. A comparison of different learning models used in data mining for medical data. The Smithsonian/NASA Astrophysics Data System, AIP Conference, 1414, 51-55.

Srinivas,K., Rani,B.K. and Govrdhan,A., 2010. Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks. International Journal on Computer Science and Engineering, 2, 250-255.

Stagos,A., 2009 –personal communication.

Streibel, O.,2008.Trend Mining with Semantic-Based Learning. European Semantic Web Conference ESWC2008.

Sweeney, L., 2001.Computational disclosure control: a primer on data privacy protection. PhD Thesis. Spring: Massachusetts Institute of Technology, Draft, (http://www.swiss.ai.mit.edu/classes/6.805/articles/privacy/sweeney-thesis-draft.pdf).

Thabtah, F., Cowling, P., and Peng, Y., 2005. MCAR: Multi-class classification based on association rule approach. Proceeding of the 3rd IEEE International Conference on Computer Systems and Applications , 1-7.

Thabtah, F; Hadi, W; Abu-Mansour, H; McCluskey, L., 2010. Proceeding of 7th International Multi- Conference on Systems, Signals and Devices.

Terlecki,P. and Walczak,K., 2007. Jumping emerging patterns with negation in transaction databases – Classification and discovery. Information Sciences, 177, 5675-5690.

Theodoridis, S. and Koutroumbas.K, 1999. Pattern recognition. Academic Press.

Titterington, D.M., Smith, A. F.M. and Makov,U.E., 1985. Statistical Analysis of Finite Mixture Distributions. John Wiley.

Toivonen,H. ,1996.Sampling large databases for association rules. Proceedings of the 22nd International Conference on Very Large Data Bases , 134-135.

Tung, A.K.H., Lu, H., Han, J. and Feng, L., 2003.Efficient mining of intertransaction association rules. IEEE Trans. Knowl. Data Eng., 15, 43–56.

Twisk, J.W.R., 2003 .Applied longitudinal data analysis for epidemiology: a practical guide. Cambridge University Press.

Van der Kamp, L.J.T., and Bijleveld, C.C.J.H., 1988.Methodological issues in longitudinal research In: Bijleveld, C.C.J.H., van der Kamp, L.J.T. , Mooijaart, A., van der Kloot, W., van der Leeden, R. and van Der Burg, E., Longitudinal Data Analysis Designs Models and Methods. SAGE publications, pp.1-45.

Villafane,R., Hua, K.A., Tran, D. and Maulik, B., 2000.Knowledge discovery from series of interval events. J. Intell. Inform. Syst., 15, 71–89.

Wagner, M, "and others" 1992. What Happens Next? Trends in Postschool Outcomes of Youth with Disabilities: The Second Comprehensive Report from the National Longitudinal Transition Study of Special Education Students. SRI International, 333 Ravenswood Ave., Menlo Park, CA 94025-3493.

Wang,K., Tang,L. Han,J. and Liu,J., 2002. Top-down FP-growth for association rule mining. Advances in Knowledge Discovery and Data Mining, Proceedings of the Sixth Pacific-Asia Conference (PAKDD 2002), Springer Lecture Notes in Artificial Intelligence 2336.

Weiss,S. and Kulikowski,C., 1991.Computer Systems That Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning and Expert Systems. Morgan Kaufmann.

Winarko, E. and Roddick J.F., 2005.Discovering richer temporal association rules from interval-based data. Proceedings of the international conference on data warehousing and knowledge discovery (DaWaK), 315–325.

Witten, I. and Frank,E.,2005.Data mining.Practical machine learning tools and techniques.Elsevier.

World Health Organization , 1999 .Definition, diagnosis and classification of diabetes mellitus and its complications. Part 1: Diagnosis and classification of diabetes mellitus. Geneva, (WHO/NCD/NCS/99.2).

World Health Organization , 2012.Global data on visual impairments 2010. Geneva.

World Health Organization, 2011.Global status report on non communicable diseases 2010.Geneva.

Wu, C.-H., Lee, C.-H., and Yang, H.-C. ,2007. Text mining of clinical records for cancer diagnosis.Innovative Computing, Information and Control,. ICICIC'07, 172-175.

Yamaguchi, K., Tetro, A.M., Blam, O., Evano_, B.A., Teefey S.A. and Middleton,W.D. 2001. Natural history of asymptomatic rotator cu_ tears: A longitudinal analysis of asymptomatic tears detected sonographically. Journal of Shoulder andElbow Surgery, 10, 199-203.

Yanoff ,M.D., and Duker,J.S ., 2008. Ophthalmology .Basic and Clinical Science Course, Section 12: Retina and Vitreous AAO.

Yuan, Y. and Huang,T., 2005. A matrix algorithm for mining association rules. ICIC, Part I, Lecture Notes in Computer Science,3644, 370-379.

Yin X. and Han J. ,2003. CPAR: Classification based on predictive association rule. Proceedings of the SDM , 369-376.

Zaki, M. and Gouda, K. (2003). Fast vertical mining using diffsets.Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining ,326-335.

Zaki, M.J., Parthasarathy, S., Ogihara, M. and Li, W., 1997. Parallel algorithm for discovery of association rules. Data mining knowl. Discov., 1, 343–374.

Zaki,M.J., 2000. Scalable algorithms for association mining. IEEE Trans.,Knowl., Data Eng., 12,372–390.

Zembowicz, R. & Zytkow, J.M. 1996. From Contingency Tables to Various Forms of Knowledge in Databases.In Advances in Knowledge Discovery and Data Mining, eds. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 307–328. AAAI Press.

Zhao,j. and Wang,T., 2010. A General Framework for Medical Data Mining. International conference on Future Information Technology and Management Engineering (FITME), 163 - 165.

Zhu,Y. and Shasha, D., 2002.StatStream: Statistical Monitoring of Thousands of Data Streams in Real Time. Proceedings of the 28th International Conference on Very Large Databases Conference, 358-369.