DATA ANALYTICS FOR PERSONALISED HEALTH CARE

A DISSERTATION SUBMITTED TO THE UNIVERSITY OF MANCHESTER
FOR THE DEGREE OF MASTER OF SCIENCE
IN THE FACULTY OF SCIENCE AND ENGINEERING

2018

By
Samuel Mugisha (10263308)
School of Computer Science

Contents

Word Count: 15798

List of Tables

List of Figures

Glossary

CSV Comma Separated Values. 35

DTW Dynamic Time Warping. 61

DTW Dynamic Time Warping. 3, 6, 7, 28, 56–58, 60, 62, 68

ESPRC Engineering and Physical Sciences Research Council. 30

IDE Integrated Development Environment. 41

IV IntraVenous. 3, 22, 36, 46, 50

KDD Knowledge Discovery in Databases. 17

MEMS Medication Event Monitoring System. 20, 24, 64, 67, 68

SQL Structured Query Language. 17

USB Universal Serial Bus. 42

VCS Version Control System. 41

VERITAS Pro Validated Haemophilia Regimen Treatment Adherence Scale-Prophylaxis. 21, 68

Abstract

Background: Poor adherence to treatment of chronic diseases is a global problem of striking magnitude. There is no gold standard for measuring treatment adherence. Traditional medication adherence measures, for instance, pill counts, self-patient reports, etc. do not account for the pharmacokinetic properties of drugs in the body, hence they misrepresent the true therapeutic exposure. Improving the effectiveness of treatment adherence measures saves lives, time and money.

Methods: We have implemented a system to model the pharmacokinetics of drugs taken by patients (with particular relevance to haemophilia) and ranked patients according to adherence from a defined therapeutic threshold. Data were obtained from Haemtrack; a patient diary system used by patients in the UK. We have implemented and compared ranking algorithms based on manhattan and euclidean distance, and Dynamic Time Warping.

Results: A list of patients, ranked by their adherence according to their euclidean and manhattan distance was obtained. The same patient listing was obtained using the Dynamic Time warping algorithm; this consistency of order acts as informal validation of the ranking. Health professionals could be prompted by email at predefined intervals informing them of non-adherent patients.

Conclusion: The proposed adherence measure captured pharmacokinetic properties of the drug and the patient drug-taking behavior. Patients were ranked according to their adherence and health professionals could be prompted by email notifying them of non-adherent patients, improving monitoring of patient adherence especially as regards to chronic diseases and potentially saving time, money and lives.

Keywords: adherence, pharmacokinetics, drug therapy, manhattan distance, euclidean distance, time series, dynamic time warping

Declaration

No portion of the work referred to in this dissertation has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright

Acknowledgements

I would like to thank the Almighty God for giving me the good health and the resources to complete this dissertation. I would like to express my sincere gratitude to my supervisor, Prof John Keane, for the motivation, guidance and valuable feedback throughout this entire project. I would also like to thank my co-supervisor, Dr. Adrian Stetcho, for all the guidance and help to do this project. Finally, I would like to thank my parents and my friends. This study would not be possible without your endless prayers, encouragement and support.

Chapter 1

Introduction

Medication adherence is defined by the World Health Organization as ”the degree to which the person’s behavior corresponds with the agreed recommendations from a health care provider” [WHO03]. Patient adherence to prescribed treatments, especially in chronic diseases is a global challenge of striking magnitude [Sim06].

Several studies show that adherence rates are higher in patients with acute conditions than those with chronic diseases [JMT02] [Hay02]. In diseases, as with haemophilia where treatment is essential to health, non-adherence is a critical concern. According to the World Federation of Haemophilia (Oliveira and Duarte , 2010, haemophilia is an X-linked congenital bleeding disorder caused by a deficiency of coagulation factor VIII (FVIII) (in haemophilia A) or factor IX (FIX) (in haemophilia B). The deficiency is the result of mutations of the respective clotting factor genes. People with Haemophilia experience bleeding disorders, a condition where the blood can not clot properly. Hemophiliac patients experience bleeding for longer and may at times experience spontaneous bleeding into muscles, joints and soft tissues. There are an estimated 400,000 people with hemophilia in the world [SBMM$^+$12].

There is currently no cure for haemophilia, though scientific research promises a cure for the condition soon [Gan17]. Hemophilia is treated by replacement of the coagulation factor administered either prophylactically (regular) to prevent bleeding episodes or episodically (on-demand) when bleeding occurs. Usually, an injection to replace the clotting factor is administered to the patient. Unfortunately, the replacement clotting factor remains active for only 12 hours. Repeated injections are necessary if the bleeding is serious [WFH18]. Most patients with haemophilia inject themselves at home.

Reports of adherence to prophylaxis in severe haemophilia show that adherence

varies from as low as 47% to 87% [LML+03] [DUVR08]. A meta-analysis conducted on chronic diseases including heart failure, human immunodeficiency virus and diabetes showed that high adherence(75 -90) %, was associated with lower mortality levels [Sim06]. Adherence to prophylaxis is associated with better outcomes in severe and moderate hemophilia [KVFD15]. Improving treatment adherence through compliance monitoring improves medication outcomes consequently saving lives, time and money [CE10]. "Increasing the effectiveness of adherence interventions may have a far greater impact on the health of the population that any other improvement in specific medical treatments" [WHO03].

There are various methods used to monitor adherence to treatments such as; patient questionnaires, monitoring rates of refilling prescriptions, pill counts, electronic medication monitors, etc. Whereas these methods are used in practice, they do not account for the absorption, distribution and elimination of treatments taken, hence misrepresenting the true therapeutic exposure [TNL+17]. The other methods that are most effective involve direct measurement of the level of the drug in blood or urine or measurement of a biological maker added to a drug formulation in the blood, but these are expensive and costly to carry out [JJ11].

This study utilizes pharmacokinetics to model the treatments taken by patients as a function of time. Pharmacokinetics is the study of the time course of drug absorption, distribution, metabolism and excretion [Ahm15]. The study uses well known distance measures such as the Manhattan and Euclidean, and dynamic time warping algorithms to rank patients according to their adherence.

Data for this study were obtained from Haemtrack; a secure therapy recording system developed by the providers of the National Haemophilia database [DHH+15]. Haemtrack is used by many haemophilia centers across the UK and is endorsed by the Haemophilia society. The system is compulsory for all patients who receive home treatments and highly recommended for all patients. Haemtrack works like a diary; patients access a calendar where they record all therapies taken from their homes as they occur. Health professionals at the Haemophilia center can then see up-to-date therapy information to help monitor, optimize and improve patient care. Haemtrack has over 1.2 million patient entries recorded [MDS17].

NHS England [Gra16] defines personalized health care as being tailored to a person's specific characteristics. According to NHS, it is "a move away from a 'one size fits all' approach to the treatment and care of patients with a particular condition, to one which uses new approaches to better manage patients' health and targets therapies

to achieve the best outcomes in the management of a patient's disease or predisposition to disease".

Recently, there has been growing interest in personalized health care to transform the quality and reduce the cost of health and care services [NHS17]. A key aspect of personalized health care is the ability to monitor patient's adherence to treatments. Using technology tools can help to tailor health care to an individual's specific characteristics [aidPcg17]. Better use of data and technology through techniques such as data analytics is seen as the driver towards accomplishment of personalized health care [KC14].

1.1 Aim

The project aim was to develop and incorporate pharmacokinetics properties of drugs and advanced, innovative decision support techniques into Haemtrack in order to monitor adherence to patient treatments and provide insightful knowledge to health professionals.

1.2 Objectives

- To assess current existing methods to measure adherence to treatments and monitor home treatments
- To develop models for the pharmacokinetics of drugs taken by Haemophilia patients over a period of time.
- To develop a system to display the developed models, sift through the Haemtrack data at predefined intervals, rank patients according to adherence and inform health professionals of non-adherent patients.
- To evaluate the effectiveness of the system developed in improving treatment adherence monitoring. The system was evaluated against approaches being used such as; patient self-reports, rates of prescription refills, pill counts, patient diaries, etc.

The focus of this work is treatment adherence monitoring; this dissertation is a companion piece to "Data analytics for personalised health care" by Denis Bahati [Bah18]. The contributions of this work are as follows:

1. development of a system to model the pharmacokinetics of medicines (with particular relevance to haemophilia);

2. development of a method to rank patients according to adherence from a defined threshold;

3. implementation and comparison of ranking algorithms based on manhattan and euclidean distance, and Dynamic Time Warping.

   This work makes specific use of the work in [Bah18] by using:

   (a) the email features developed in [Bah18] to send emails to the health professionals,

   (b) the web interactive user interfaces and dashboard that health professionals will use to interact with the system developed.

1.3 Dissertation structure

The remainder of the dissertation is structured as follows: Chapter 2 provides the theoretical background on which the project is developed; Chapter 3 is concerned with the methodology used to develop the system and its design; Chapter 4 presents and discusses the findings of the research which are based on the experiments done with the data provided by Haemtrack; Chapter 5 discusses the evaluation of the system, both individually and in comparison to others; finally, Chapter 6 concludes and gives recommendations for future research.

Chapter 2

Background

This chapter contains background information about the area of treatment adherence measurement and the theories, approaches and technologies that this project is built upon.

2.1 Data life cycle in Haemtrack

Data in Haemtrack follows the standard data management life cycle which describes the stages through which data flows in an organization. Data is collected from patients, stored, processed and then, disseminated to health professionals.

Patients record their treatment and details such as date and time, product used and the reason for treatment in a mobile phone application or a web interface. The Data collected from patients is stored in a Microsoft SQL server database. Microsoft SQL server is a relational database management system. Data is stored in tables with each table representing a particular entity. The database management system utilizes the SQL to store and retrieve records as requested by the user [Mic18].

The data is then processed. The Knowledge Discovery in Databases (KDD) [Fay96] process is used to find and interpret patterns in the data. The KDD process is shown in Figure 2.1.

Data pre-processing involves the use of automated or semi-automated methods to prepare the data for dissemination. Some data pre-processing techniques used in Haemtrack include:

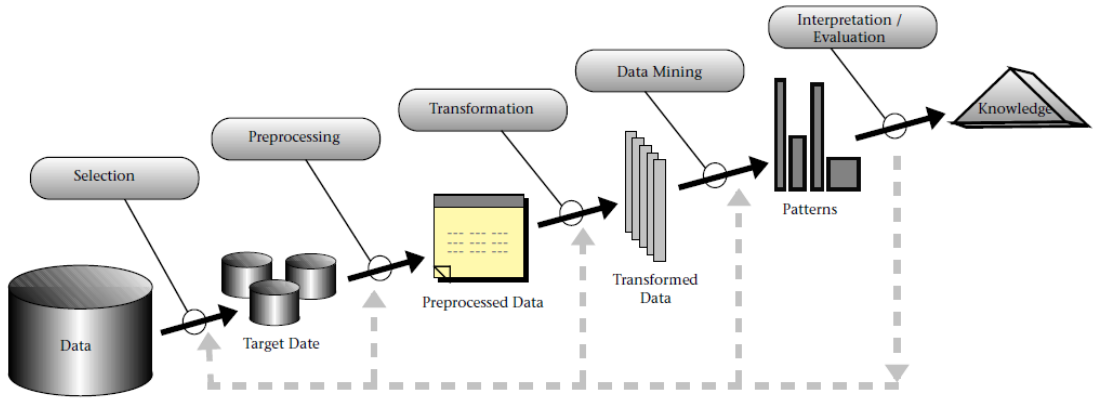


Figure 2.1: Steps involved in the KDD process [Fay96]

i. Data cleaning: which involves the removal of noise and the correction of inconsistencies in the data.

ii. Data integration, which involves merging of data from multiple sources into a coherent data store

iii. Data transformation, which involves the transformation of data into a format which improves accuracy and efficiency of dissemination methods

iv. Data reduction, which involves reduction of data size by using techniques such as aggregating, eliminating redundant features, clustering, etc.

The processed data then undergoes a process of data mining which involves searching patterns in the data that may be of interest for representing the data. Some of the approaches that could be used include: classification, regression, clustering, etc. The results and patterns are then interpreted or evaluated to get insight.

The processed data is then displayed on a web interface at the haemophilia center where health professionals monitor patient's treatment. The data is used by health professionals to determine the best treatment for the patients. Several reports are generated, for example, treatments taken by patients over specified dates. Health professionals can view reports for individuals and come up with personalized programs to help the patients

achieve health care that is particular to their needs.

2.1.1 The importance of adherence in haemophilia

The terms *adherence* and *compliance* are sometimes used as synonyms but they differ. Compliance is defined as ”the extent to which the patient’s behavior matches the prescriber’s recommendations” [HWBE05]; compliance does not account for the patient’s involvement. Adherence adds to the definition of compliance by adding that the patient has to agree to the prescriber’s recommendation and that failure to do so should not be reason to blame the patient.

In diseases such as haemophilia where treatment is essential to health, non-adherence is a critical concern. Reports of adherence to prophylaxis in severe haemophilia show that adherence varies from as low as 47% to 87% [LML$^+$03] [DUVR08]. One large bleed can lead to irreversible damage especially if the bleed occurs in joints or the central nervous system. Such bleeds can cause limitations in daily functioning and expose patients to the risk of arthropathy [FVMB$^+$02] [RNS$^+$14]. As such, adherence is a critical factor to prevent bleeding.

Non-adherence involves never taking the medication, taking reduced amounts or excessive amounts, not taking doses as per the prescriptions, or not matching medication to food-activity requirements [JJ11].

2.2 Methods to measure adherence to medication

There is no gold standard for measurement of adherence to medication [WHO03]. It is important to measure actual patient treatment to prescribed treatment to improve adherence. There are several methods to measure adherence to treatment; these include direct and indirect methods.

Direct methods are the most accurate method. They include: direct observed therapy, measurement of the drug level or its metabolite in blood or urine and detection or measurement of a biological marker added to the drug formulation in the blood. Despite being accurate, direct methods are

expensive to perform. In addition, direct methods do not consider the pharmacokinetic factors of the drug and the patient [Ost05].

Indirect methods of measuring adherence to treatments are discussed in the next sections.

2.2.1 Patient self-reports

In this method, health professionals use questionnaires and interviews to gather information from patients about their treatment behavior. Self-reports are the most used measures of adherence because they are low-cost, non-invasive, easy to administer and are flexible [GCE+11] [GSW+13]. However, several reports have found out that patients tend to overestimate their adherence and underestimate their non-adherence. As such, self-report methods are not regarded as a reliable and accurate way of measuring patient adherence [SDJC+15].

2.2.2 Pill counts

In this method, the number of pills that remain in the patient's medication bottles or vials is counted and is used as a measure of adherence. Though this method is common and simple, it also has some drawbacks, for instance, patients may discard some pills or switch medicines between containers [Pul89]. Furthermore, this method does not provide information on dose timing and drug holidays, where the medication has to be omitted on 3 or more sequential days, both of which are important in determining clinical outcomes [Mar88].

2.2.3 Electronic medication monitors

In this method, electronic monitors/devices or Medication Event Monitoring System (MEMS) powered with microprocessor technology are used to record and log the time of opening bottles, dispensing medications, opening injection containers, etc. Some of the MEMS approaches that have been used include; smart pill boxes [Tam06], wearable sensors [HFG16], and Radio Frequency Identification (RFID)devices[HYT+13], etc.

The MEMS approach captures both the frequency and the time of taking medications when patients open the bottles which other methods, for instance, pill counts do not account for [GHW+11].

The disadvantage with this method is that patients may, for instance, open the container and not take the medications. In addition, they may take the wrong amount of medication or dosage out of the container. The failure to track precisely what the patient does with the medication after opening the container is a major drawback to this method which affects the accuracy and reliability of observations made using this method [VHVD02].

2.2.4 Rates of refilling prescriptions

In this method, the rate at which patients refill their prescriptions is used as a measure of adherence to treatment. This method has proved to be an accurate measure especially in a closed pharmacy system where there is universal drug coverage as refills are measured over a period of time [SP97].

2.2.5 VERITAS Pro

VERITAS-Pro (Validated Haemophilia Regimen Treatment Adherence Scale-Prophylaxis) is a widely used questionnaire for measuring adherence to prophylactic treatment in Haemophilia patients [DKRS10]. VERITAS Pro consists of six sub-scales: These are:

- Timing: This examines the extent to which participants take their injections at the recommended time;
- Dosing: This examines how patients use the recommended dose;
- Planning: This examines the ability of the patient to plan ahead to ensure they have enough supplies;
- Remembering: This examines whether patients remember to take their injections;
- Skipping: This examines when patients skip injections;
- Communicating: This examines the extent to which patients communicate with the Haemophilia center appropriately;

The VERITAS Pro captures diverse dimensions of adherence and hence provides a comprehensive view. It provides insightful knowledge that can be used to monitor and develop targeted interventions to enhance treatment adherence to prophylactic treatments. The VERITAS Pro questionnaire has been used in several countries in several studies and has been found to be valid and reliable for use in clinical and research settings [FLD18] [LRD+14] [CBTOGP+17].

2.2.6 Pharmacokinetics-based adherence measures

Traditional medication adherence measures misrepresent the true therapeutic exposure [VHVD02]. Pharmacokinetic modeling can help understand how drugs taken by patients are absorbed, distributed and eliminated from the body .

Three approaches are used for pharmacokinetic modeling: compartmental, physiological and model-independent [Bou01]. The compartmental model is an empirical approach that models the body as compartments to represent a body volume or a chemical state. In each compartment, the drug is considered to be rapidly and distributed. The compartmental model can be categorized as the one or compartmental model.

The physiological model identifies the compartments with actual body spaces basing on physiology and anatomy of humans and other animals. Though the physiological model is more complex than the compartmental models, it provides an actual representation of pharmacokinetics of drugs in the body [Gri13].

The model independent approach is a purely mathematical method that does not consider kinetic parameters which may not be valid. However, it is less complex than the physiological model.

The one-compartment model represents the simplest way to describe the process of drug distribution as well as elimination in the body. In this model, the body acts like a single, uniform unit in which the drug can enter or leave the body easily. The one-compartmental model was used because

it is simple and yet sufficient to explain the pharmacokinetics. The one-compartment model assumes that the drug in the blood is in rapid equilibrium with the drug in the extravascular system. It is not an exact representation; however, it is useful for a number of drugs to a reasonable approximation [Bou01].

2.2.7 Distribution and elimination of Intravenous (IV) bolus administration

Intravenous IV bolus administered drugs enter the bloodstream directly. The drug is distributed through the circulatory system to all tissues in the body. Assuming that the drug concentration is uniform in the body compartment at all times, the drug is eliminated by a first order process described by a constant , K_e [Ahm15]. K_e is known as the elimination rate constant which is the rate at which a drug is eliminated from the body. The equation (2.1) illustrates the one-compartment open model for intravenous bolus administration.

$$C = C_0.e^{-K_e.t} \tag{2.1}$$

From Equation 2.1, C is the Plasma Concentration, $Co = D_o/V_D$, where D_o is the Dosage and V_d is the Volume of distribution. The volume of distribution is: ”the volume of plasma in which the total amount of drug in the body would be required to be dissolved to reflect the drug concentration attained in plasma” [Gri13]. The Elimination constant is represented by Equation 2.2.

$$K_e = \frac{math.log(2)}{halfLife} \tag{2.2}$$

The half-life of the medicine is ”the time required to reduce the plasma concentration to one half its initial value” [Ahm15].

2.2.8 Absorption of oral medicines

Unlike the process of IV administration, when a drug is introduced into the body by an extra-vascular route such as oral, the drug is absorbed in the

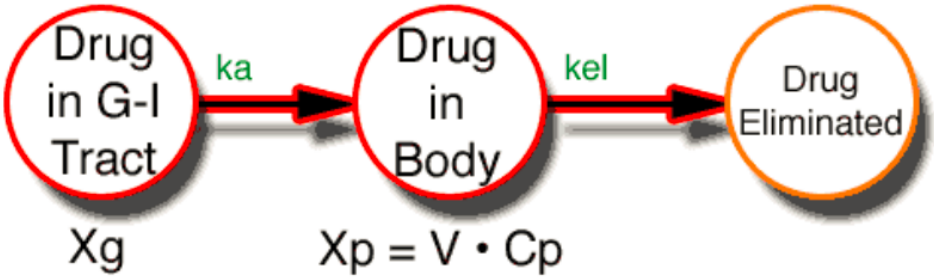


Figure 2.2: Oral Administration for one compartment Pharmacokinetic model [Bou01]

body through a process that transfers it from the administration site into the vascular system.

In Figure 2.2, X_g is the amount of drug to be absorbed, X_p is the amount of drug in the body, and K_a is the first order absorption rate constant.

The equation below shows the absorption of oral medicines by the one-compartment model.

$$Cp = [F.K_a.\frac{D_0}{V_D}.(K_a - K_e)].(e^{-K_e.t} - e^{-K_a.t})] \tag{2.3}$$

Form Equation (2.3), Cp = Plasma concentration, D_o = Dosage, K_e = Elimination rate constant, K_a = Absorption rate constant, V_D = Volume of distribution, t = Duration of infusion, F = Bio viability. Bio viability(F) refers to "the degree and rate at which an administered drug is absorbed by the body's circulatory system" [Ahm15].

IV medicines have 100% bioviability. However, when medicines are administered through other routes, for instance oral, their bioviability is affected by a number of factors including metabolism rate, age, weight, etc. As such, bioviability decreases from the point of administration and varies from patient to patient with oral administration [Gri13].

2.2.9 Using Pharmacokinetic-based adherence measures

Several studies have measured adherence to treatments using pharmacokinetics. In Kenya, a study was carried out to measure adherence to antiretroviral therapy in HIV-infected children [TNL+17]. The one-compartment model was used to model pharmacokinetics of dosing times recorded in a Medication Event Monitoring System (MEMS). The Mean plasma concentration (Cp) of individual patients during 1 month of follow-up was obtained. The intended plasma concentration (Cp') where the patient perfectly follows the dosing regimen and frequency was also calculated. The ratio between the two ($R = Cp/Cp'$) was used to determine the exposure of patients to the drug as compared to the intended level. Smaller values of R values indicated poorer adherence which was true when compared with the adherence shown by the MEMS.

Patients can be classified as adherent or non-adherent by pharmacokinetic modelling of their serum drug concentration in comparison to a therapeutic threshold. According to a study that was carried out on 7 psoriasis patients who were taking treatments for 40mg subcutaneous adalimumab at an interval of 14 days for one year, patients were classified as adherent if their serum drug concentrations remained within the therapeutic range and non-adherent if the concentrations fell below the threshold [MSK15].

Together, these studies indicate that the use of pharmacokinetic measures is a usable and effective way of monitoring treatment adherence.

2.2.10 Summary

In summary, several measurement strategies can used to measure adherence. No single method has shown to be the gold standard. Some approaches are costly, for example, direct methods. Most indirect methods, for instance, pill counts, counting rates of refilling, etc., do not account for the true pharmacokinetic exposure of the patient to the drug and thus, misrepresent the true therapeutic exposure. Using pharmacokinetic-based measures is a useful way to consider the patient therapeutic exposure to the drug.

2.3 Data analytics

The Institute for Operations Research and the Management Sciences defines data analytics as the scientific process of transforming data into insights for making better decisions [DGK+15]. Data analytics has enormous potential and application in health care in areas such as research and development, public health, genome analytics, device/remote monitoring of patients, patient profile analysis and treatment compliance monitoring [RR14].

Computer systems to aid home treatments, where users log their times, dosages and medicines taken through smartphones, laptops, wearable devices, etc. are becoming popular especially for chronic diseases [AJM18]. Haemtrack is one such system where hemophiliac patients log their treatments on a smart phone and the generated data is stored on an online database system for viewing and analysis by health professionals.

Incorporating advanced decision making analytics tools (data analytics) in such systems can provide enormous insights to health professionals and patients that can improve medication outcomes, save money and improve decision making. Applying data analytics techniques to Haemtrack can reveal valuable information that could be used to improve the quality of health care of haemophilia patients [OBK05].

2.3.1 Time series

A time series is an ordered sequence of values of a variable at equally spaced time intervals [Agr13]. Time series are a measure of unit change over time for any variable under observation. Time series are useful for predicting and forecasting from historical data. Time series can be categorized as constant (when there is no change in the variable over time), trended (when there is net increase or decrease in the time series over time), un trended (when the variable is increasing and decreasing according to the seasons of the year but there is no change in the average value of the time series) and trended (when the variable decreases or increases with the season but there is a net increase or decrease over time) [Ham94]. Figure 2.3 shows the different categorizations of time series.

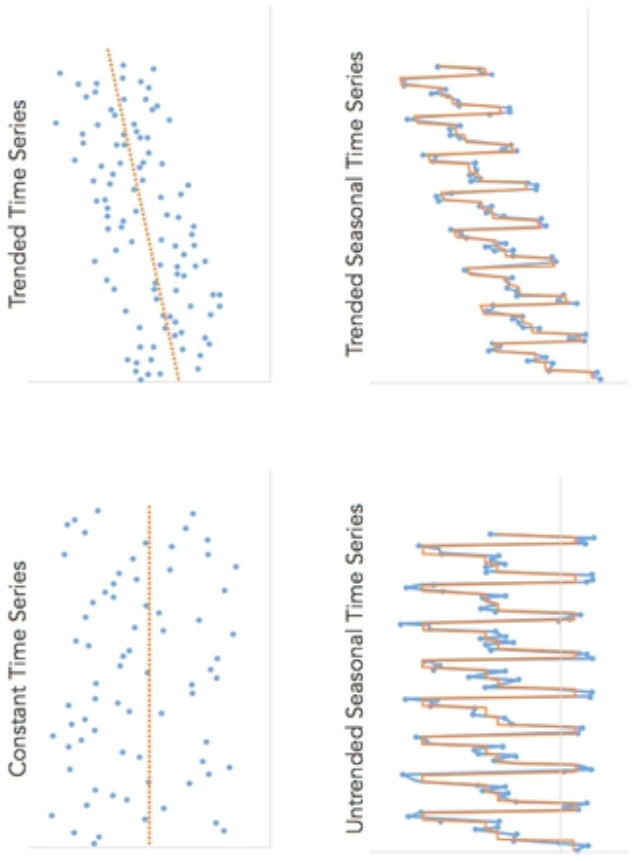


Figure 2.3: Spotting trend and seasonality in time series

2.3.2 Summary

Using pharmacokinetics Equations (2.1) and (2.3), explained in sections 2.2.7 and 2.2.8 respectively, patient treatments can be modeled as a time series. The time series would represent the drug absorption, distribution and elimination over a period of time. In practice, the euclidean, manhattan and Dynamic Time Warping Algorithms are often used to compare time series. These are discussed in the next section.

2.3.3 Euclidean distance

The Euclidean distance between two points in space measures the length of a segment connecting the two points. The Euclidean distance is obtained by taking the square root of the sum of the squares of the differences of the coordinates of the coordinates as shown in Equation (2.4).

$$\delta(p,q) = \delta(q,p) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \tag{2.4}$$

For example, if x=(a,b) and y=(c,d), the Euclidean distance between x and y is:

$$\sqrt{(a-c)^2 + (b-d)^2} \tag{2.5}$$

Euclidean distance can be used to measure the distance between two time series.

2.3.4 Manhattan distance

The Manhattan distance is obtained by taking the sum of the absolute values of the differences of the coordinates. For example, if $x = (a,b)$ and $y = (c,d)$, the Manhattan distance between x and y is $|a-c| + |b-d|$

The manhattan distance between two points or vectors on N-dimensional space is given by Equation (2.6)

$$\sum_{i=1}^{n-1} |(x[i] - y[i])|) \tag{2.6}$$

The manhattan distance can also be used to measure the distance between two time series.

2.3.5 Minkowski distance

Euclidean and Manhattan distances can be calculated using the Minkowski distance [Wik18]. The Minkowski distance between two points:
$X = (x_1, x_2, ...x_n)$ and $Y = (y_1, y_2, ...y_n) \in R^n$ is defined by as:

$$D(X,Y) = (\sum_{i=1}^{n} |x_i - y_i|^p)^{\frac{1}{p}} \tag{2.7}$$

From equation (2.7),

When p = 1, we obtain the Manhattan distance

When p = 2, we obtain the Euclidean distance

2.3.6 Dynamic Time warping

Dynamic Time Warping DTW is an algorithm used to compare similarity between two time series. DTW is often used in speech recognition to determine if two wave forms represent the same spoken phrase [Fur08]. In addition to speech recognition, DTW has been useful in; gesture recognition [GD95], robotics [Schmil [SOC98], manufacturing [GP96] and medicine [FLR$^+$12]. DTW is commonly used in data mining as a distance measure between time series [KP01].

The DTW algorithm calculates both the warping path and the distance between time series [SC78]. DTW uses dynamic programming to minimize the difference between two time series. DTW accounts for different time scales and speeds. Figure 2.4 shows an example of how one time series warps into another.

The basics of DTW is founded on the computation of distance matrix between two time series. Consider an n*m matrix D, where $D_i, j = d(x_i, y_i) = |x_i - y_i|$ or $d(x_i, y_i) = \sqrt{(x_i - y_i)^2}$

A warping path w is a contiguous set of matrix elements which defines a mapping between x and y that satisfies the following conditions:

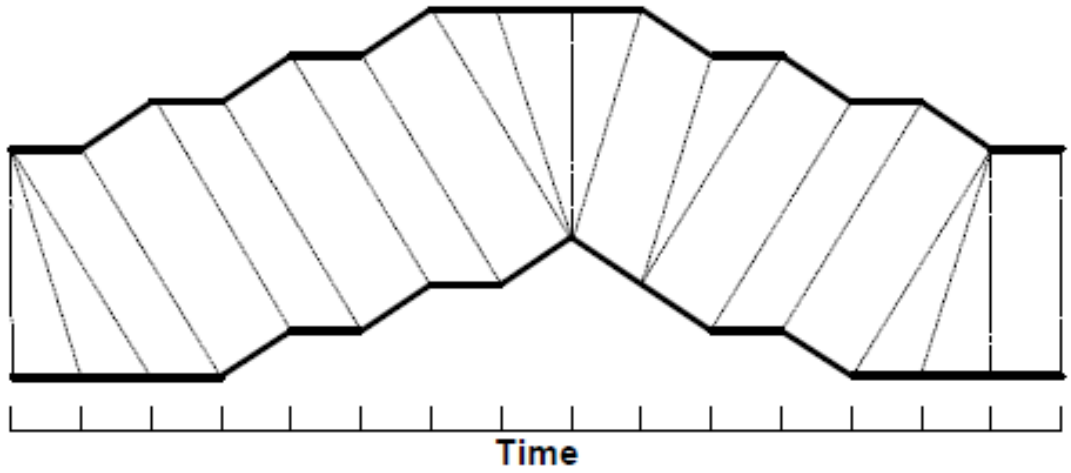


Figure 2.4: Warping between two time series

- Boundary conditions: $w_1 = (1,1)$ and $w_k = (m,n)$ where k is the length of the warping path
- Continuity: if $w_i = (a,b)$ then $w_(i-1) = (a'-b')$ where $a-a' \leq 1$ and $b-b' \leq 1$
- Monotonicity: if $w_i = (a,b)$ then $w_(i-1) = (a'-b')$ where $a-a' \geq 0$ and $b-b' \geq 0$

Using dynamic programming to prevent construction of the whole matrix, The cumulative distance between $\gamma(i,j)$ is given by Equation (2.8) below:

$$\gamma(i,j) = d(x_i,y_j) + min\gamma(i-1,j-1),\gamma(i-1,j),\gamma(i,j-1) \qquad (2.8)$$

The optimal principle used in DTW is applied using the ”backward” technique. As a result, finding the warp path uses a dynamic structure known as ”Stack”. DTW, as with all dynamic programming algorithms, has polynomial complexity. This causes challenges when memorizing large matrices of numbers and performing large numbers of calculations and such thus, DTW is limited to small time series data sets. However, there is an improvement to the standard DTW algorithm, known as fast DTW that uses; ”a multilevel approach to recursively project a warp path from a coarser resolution to the current resolution and refines it.”[SC07]

2.3.7 Summary

In summary, incorporating data analytics techniques in patient diaries such as haemtrack can help to analyze treatment trends and generate useful insights. Using time series, one could model treatments behavior as a function of time. Using the manhattan and euclidean distances, one can measure the distance between time series. In addition, the DTW algorithm can be used to compare time series. These data analytics techniques can be applied to Haemtrack to measure patient adherence to treatments.

Chapter 3

Methodology and design

This chapter discusses the methodology used to develop the system. In addition, the project design, risk management, schedule and ethical considerations are discussed.

3.1 Methodology used

After considering various software development models [Bob09], the **waterfall model** was chosen for this project. This is because the project requirements were well understood from the start being identified in an earlier ESPRC-funded project.

Figure 3.1 shows the **waterfall model** methodology of software development. Source: Wikipedia [Wik18]

The **waterfall model** model comprises several phases including requirements gathering and analysis, design, implementation, verification and maintenance; table 3.1 summarizes these activities.

3.2 Requirements gathering and analysis

A meeting was held with Medical Data Solutions and Services (MDSAS) company, the developer of Haemtrack, to understand the current system and discuss the requirements articulated in the previous project.

The requirements were categorized into two: functional and non-functional

Phase	Activities	Deliverable
Requirement Analysis	I. Collected all the requirements II. Profiling of the data collected III. Did requirements feasibility testing to ensure that the requirements are testable	Requirements document Use case diagrams
System Design	I. Created the design from the collected requirements II. Captured hardware and software requirements III. Drafted the architectural designs of the system	System architecture Database architecture List of hardware and software requirements
Implementation	I. Implemented the programs as per the design	Programs Unit tested code
Verification	I. Implemented unit tests for each of the modules developed II. System integration testing of unit tested modules	Test cases Test reports
System maintenance	I. Deploy the system in its expected working environment II. Test the system functionality in the deployed environment III. Continously update the system IV. Fix any issues such as bugs and errors that may arise	System yet to be deployed and maintained

Table 3.1: Activities carried out during each phase of the water fall model

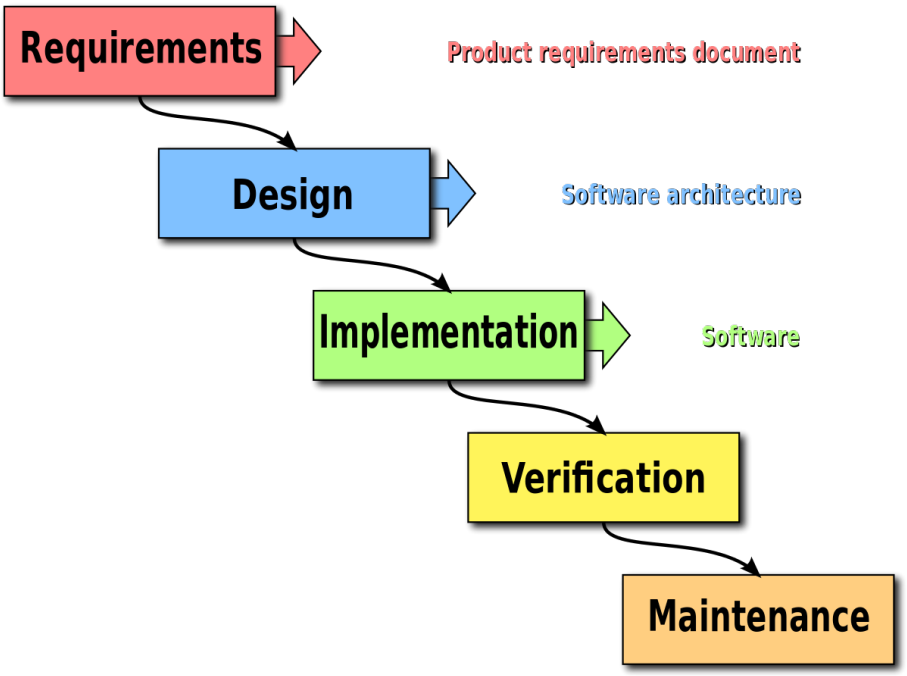


Figure 3.1: Waterfall model of software development [Wik18]

requirements. Functional requirements are those which relate to the technical functionality; non-functional requirements are quality attributes. Every requirement selected was classified according to the degree of risk, complexity and priority with rating as low, medium or high. Table 3.2 summarizes the system requirements.

Based on this, a use case diagram that shows the functions that different actors in the system perform was developed (see Figure 3.2).

The use case diagram shows the main actor as the health professional. The health professional generates several reports from the system by inputting search details, for example, details about drugs to view pharmacokinetics of the drugs, specific dates to view patient rankings in the date period, etc. The system outputs a number of reports which the health professional can view. The system periodically sends emails to the **health professional** at specific times notifying them of non-adherent patients.

As mentioned in Section 2.1, the data for this project was provided by Haemtrack Haemtrack – this data is real data gathered from the operational system. The data was stored in a Microsoft SQL Server database management system.

ID	Description of requirement	Risk	Complexity	Priority
FR1	Health professionals should be able to input drug details and view pharmacokinetic visualizations of the drug	Medium	Low	High
FR2	Health professionals should be able to compare different parameters that affect the pharmacokinetics of different drugs and visualize the effect of varying the parameters	Low	Low	High
FR3	Health professionals should be able to input patient and drug details and view pharmacokinetic visualizations of the patient's treatments	Low	High	High
FR4	Health professionals should be able to enter date ranges and get a list of patients, in the date range, with their adherence rankings	Low	Medium	High
FR5	System should generate communication messages/ notifications to informing of uncompliant patients	Low	Low	High
NFR1	The system should be secure to keep patient data safe	High	Medium	Medium
NFR2	The system should be efficient and perform desired tasks in minimum time	High	High	High
NFR3	The system should be free of errors	High	High	High
NFR4	The system should produce accurate results	High	High	High

Table 3.2: List of system requirements

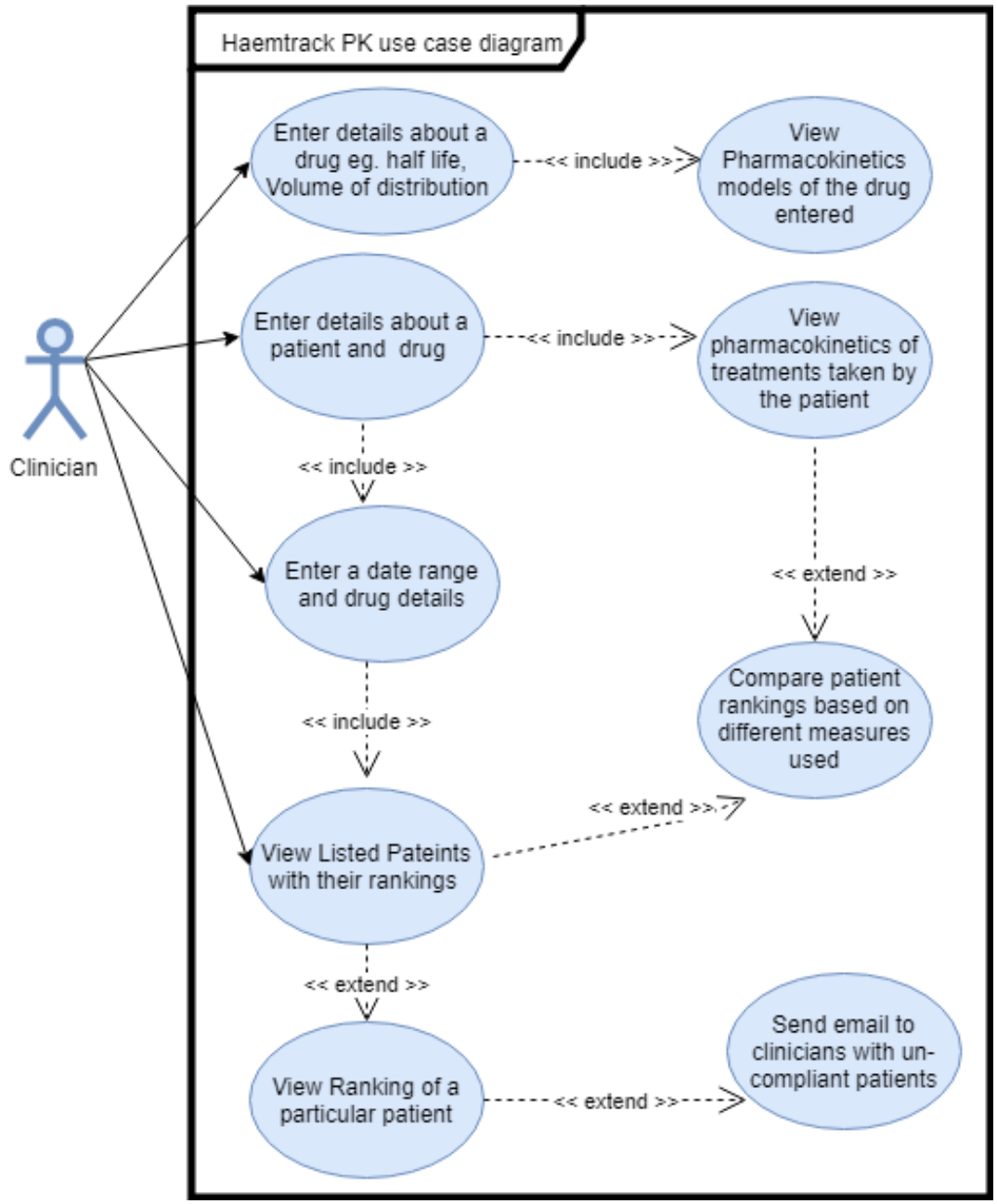


Figure 3.2: Use case diagram

It was necessary to analyze this database and profile the data before continuing with other phases of development. According to Felix Naumann [Nau14], data profiling is: "the process of examining the data available in an existing data source and collecting statistics and information about that data." Section 3.2.1 explains the steps taken in profiling the data in Haemtrack.

3.2.1 Steps taken in profiling the data

The Python Pandas library and Microsoft Structured Query Language (SQL) were used to profile the data. The **python pandas** library offers data structures and operations for manipulating numerical tables and time series. **Python pandas** is particularly powerful as it makes it easy to convert files in CSV, SQL and creates data frames with rows and columns which look like tables. The **python pandas**library is also very robust and can handle large volumes of data [sF18].

The following were the steps taken in profiling the data:

i. Exploratory analysis: This involved visually inspecting the data to learn about it. By inspecting the data, we obtained an overview of the Haemtrack data, as well as identify any inconsistent column names, missing data, untidy data and column types that have unexpected data values, etc.
   It was important to understand how many tables are held in the Haemtrack database. Figure 3.4 shows an Entity relationship diagram of Haemtrack database highlighting the key tables and their relationships. There are 104 tables in the database.
   Some tables were identified as key to this research project including:
   - Patients table, which contains information about patient demographics
   - Treatments table, which contains information about treatments taken by the patients including drugs, date and time taken, etc.
   - Bleed details table, which contains information about any bleed that happened to the patient including its nature, cause, severity, etc.

- Combo Products, which contains information about drugs including; units of measure, dosage form and half-life.
- Centres, which contains information about the haemophilia center that the patients are registered at. This has information including; the center number(CentreNo), name, address, post code, phone number, email, etc.

There were 2809 patients in the database. The Patients table had 16 columns; all columns had descriptive names with valid data types.

Some of the drugs registered in the database include Advate, Alphanate, Factor 13, Hemofil, etc. There were 61 drugs in the database. There were some inconsistencies in the data, for example, the Dosage form had not been registered for each drug and there were some inconsistent dosage form names in the table Combo Products. This data was cleaned.

The treatments table had 838,001 patient treatment records. Important data captured about treatments in the database includes:

- Patient ID, the patient identifier referenced from the Patients table
- Product, the drug taken by the patient, referenced from the "Combo Products" table
- Treated Date and Treated Time, the time the patients took the drug. This is critical when constructing time series
- Total units, the dosage amount of the drug that the patient took

ii. Combining the data: The process of combining the data was important to understand the data better since the data was stored in separate tables, each representing a particular entity. This involved combining rows of data, concatenating and merging data. The data is held in a relational database and foreign keys were used to reference the data across the tables, for instance, Treatments and Bleed Details table were combined to when constructing time series and examining bleeds that occurred to the patient.

iii. Cleaning data: This involved removing duplicate records, missing data, dropping duplicate data and filling missing data etc. to ensure that the data does not have noise.

iv. Statistical profiling : This involved qualitative and quantitative description of the data. Some of the common figures looked out for

include: counts, median, mean, standard deviation, variance, quartiles etc. This helps to understand the data more in detail, detect errors and outliers. We ,for instance, did analysis to find out the most frequently taken medicines from the treatments table. The most frequently taken medicines were used for pharmacokinetic modelling.

Once the requirements were collected and analyzed, the system design was undertaken (see Section 3.3).

3.3 System design

In this phase, the database and system architecture of the system were developed. In addition, the software and hardware requirements of the system to be developed were identified. Figure 3.3 shows the major components and how they interact.

From Figure 3.3, the system consists of 4 main modules: These are:

- Database
- Processing: This modules involves the implementation codes that were implemented to process and present the data
- Data visualizations and analysis: This includes visualization of analysis results, such as patent rankings, via graphs etc.
- Communication: Sends notifications for instance, email notifications to health care professionals notifying them of non-adherent patients.

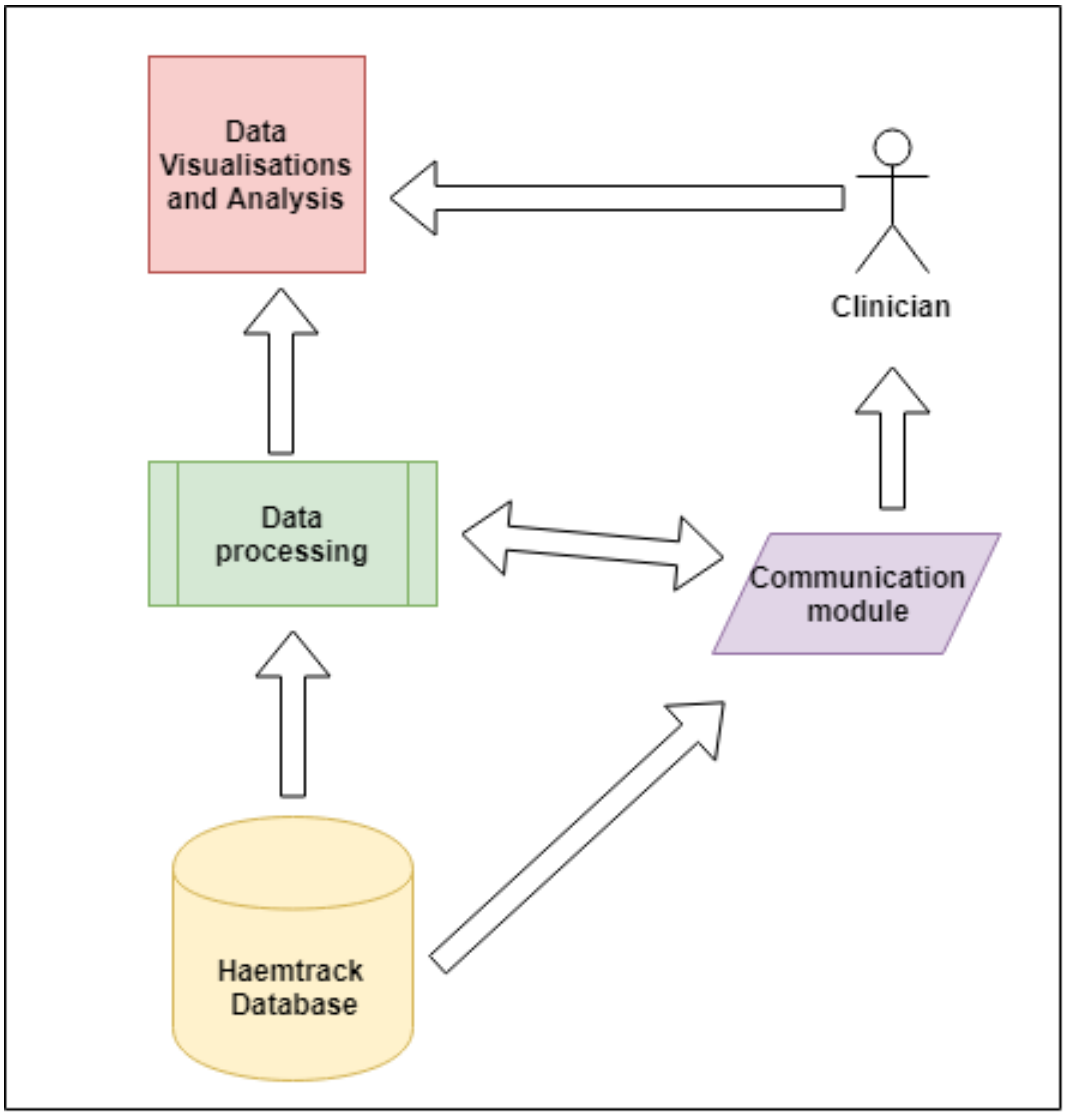
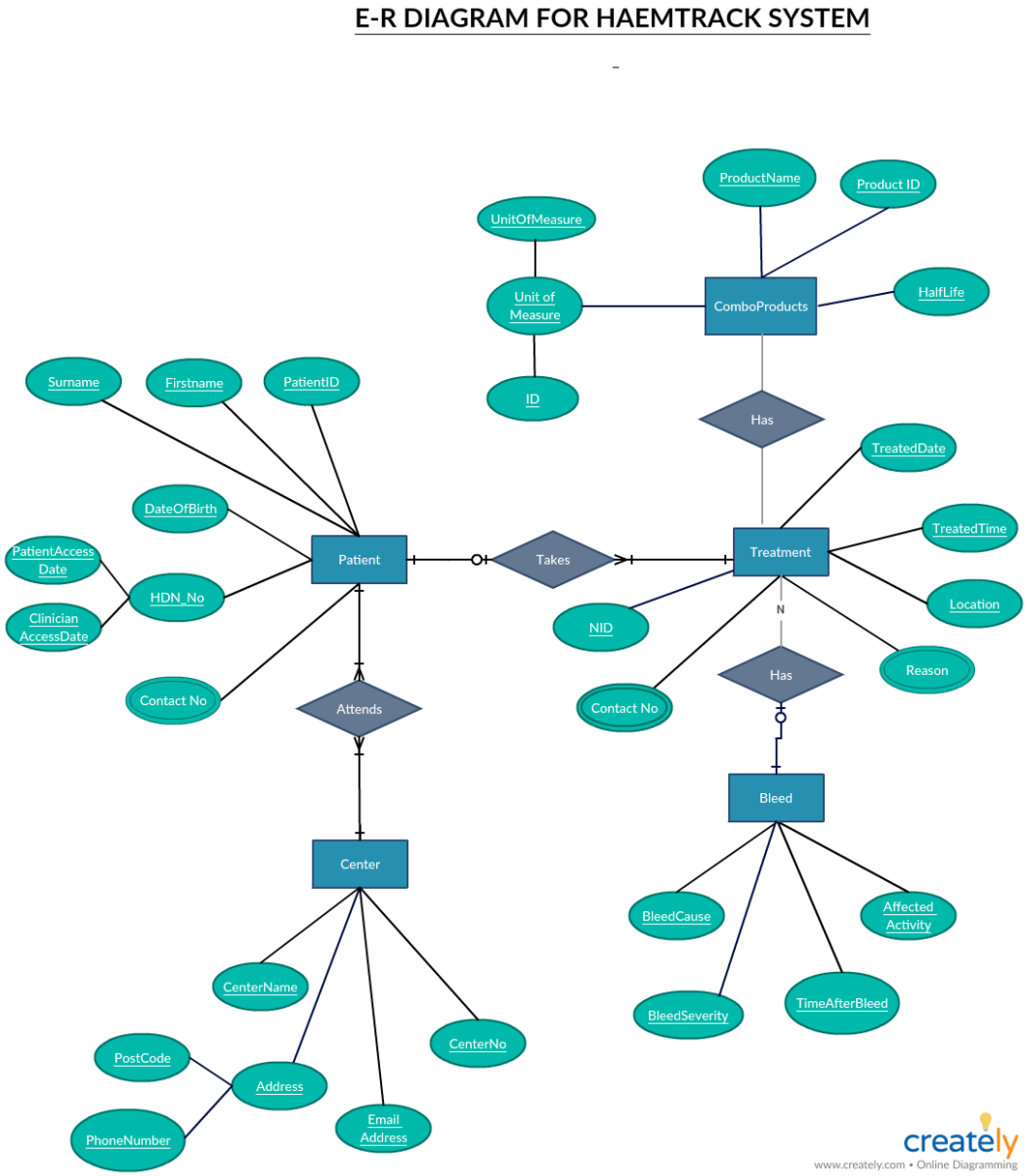


Figure 3.3: System block diagram


E-R DIAGRAM FOR HAEMTRACK SYSTEM
ProductName
Product ID
UnitOfMeasure
Unit of Measure
ComboProducts
HalfLife
ID
Has
Surname
Firstname
PatientID
TreatedDate
DateOfBirth
TreatedTime
PatientAccess Date
Patient
Takes
Treatment
HDN_No
Location
Clinician AccessDate
NID
N
Reason
Has
Contact No
Contact No
Attends
Bleed
Center
BleedCause
Affected Activity
CenterName
BleedSeverity
TimeAfterBleed
CenterNo
PostCode
Address
Email Address
PhoneNumber
creately
www.creately.com • Online Diagramming

3.3.1 Languages and tools

This section describes the tools used to develop, test, debug and implement the system.

The system was developed using Python. Python is a high level, interpreted general purpose dynamic language that focuses on code readability. Python supports many libraries for data science, such as Pandas, Matplotlib and seaborn which analyze data. In addition, python is open source and community driven with a lot of online support.

Python libraries; Matplotlib, seaborn and Plotly were used to visualise the data. Matplotlib is a graph plotting library for Python, Plotly makes it easy to create, edit and share interactive data visualizations on the web [Plo17]. The seaborn library is a high-level interface built on matplotlib for drawing attractive and informative statistical graphics [Was18].

An integrated Development Environment (IDE) provides a source code editor, build automation tools, and a debugger Wikipedia [Wik18]. Atom IDE was chosen as it is a free and open-source text and source code editor developed by Github. Atom can be used across multiple platforms including Windows, Linux and macOS platforms [Git18].

Jupyter notebook was used to supplement ATOM as an IDE. Jupyter Notebook is ”an open-source web application that allows sharing of documents (including live code, equations, visualizations and narrative text)” [Jup18].

Microsoft SQL server Management studio was used for querying the database. To connect to the database, we used the Python Open Database Connectivity (ODBC) driver [ODB17].

Cmd/Bash scripts were used to generate cron jobs that sift through the data and notify doctors.

3.3.2 Source code management

Git VCS was chosen because of its distributed functionality and popularity [Git18]. Git enables rolling back to previous changes when some edits are made in the source code. In addition, it offers an automatic safe backup for the code at all times. If there is more than one person working on a

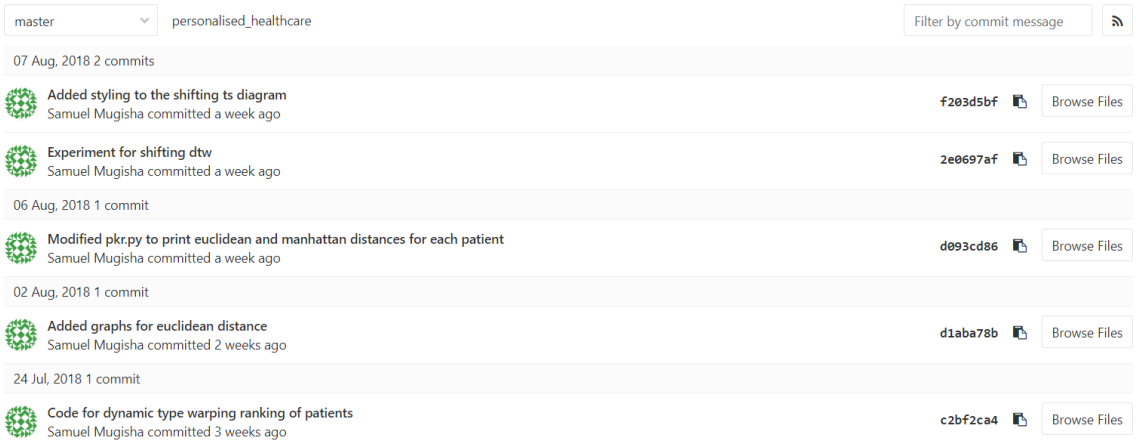


Figure 3.5: Graph showing usage of git

project, Git enables collaboration and merging of files automatically, syncing changes made by any member of the team. The School of Computer Science has an installation of a remote hosting software known as Gitlab which was used.

Figure 3.5 shows usage of the Git with the GitLab repository at the University of Manchester.

3.3.3 Project management

Git software uses issues as the beginning point of project management. GitHub describes issues as ”a great way to keep track of tasks, enhancements and bugs for your products”, adding that ”they’re kind of like email, except they can be shared and discussed with the rest of your team ”[Git18]. Issues are opened up with their details eg. start and due dates. The issues are grouped by milestone. Issues can also be tagged with labels that signal their priority, importance, etc. Git also has a board which is a good visual tool to look at the project as a whole. The figure below 3.6 shows some project issues categorized as: To do, Ongoing, BackLog and Closed. The board is a great way to keep track of the project status.

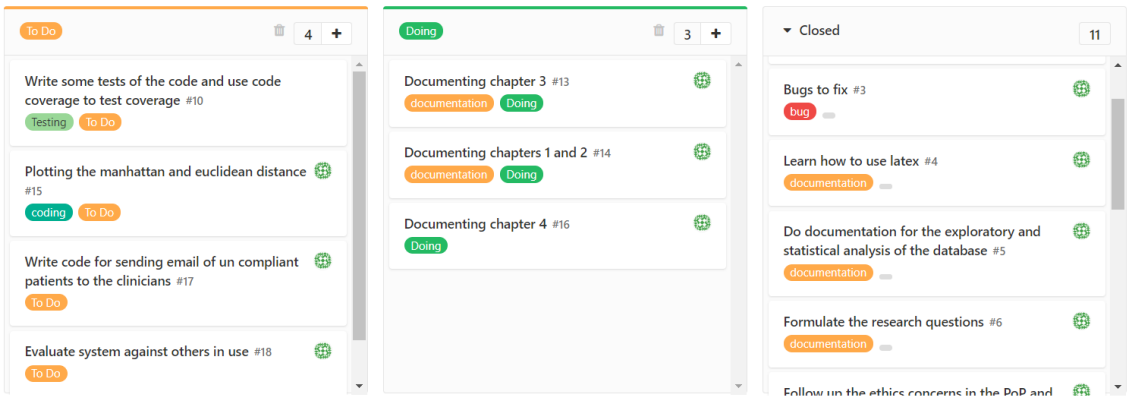


Figure 3.6: Board showing the status of the project

3.4 Ethics and professional considerations

When working with real-world data, ethical and professional issues arise. It is important to ensure security, privacy and autonomy of data.

The data provided by Haemtrack was held on an encrypted USB hard drive. Precaution was taken to ensure that the drive is safely protected.

In addition, we ensured autonomy. Data were encrypted so that it would be very difficult for someone to associate treatments and other data to the respective patients in case the USB drive lands in wrong hands.

3.5 Risk management

Table 3.3 shows the risk management plan that was put in place to ensure project completion.

3.6 Project plan

Figure 3.7 is a gantt chart showing time lines and milestones at each stage.

Risk statement	Risk impact	Mitigation plan
No process for measuring and managing the project progress to meet the objectives	Failure to achieve the project objectives	Use GitLab software to monitor progress Weekly meetings with supervisor
Failure to backup project work and dissertation	Loss of project work in case of any hardware or software problems	Backup all the project work automatically on Google drive using the Backup and Sync tool from Google
Failure to meet the project's technical requirements	Failure to achieve the project objectives	I took extra python classes courses to polish up on programming skills
Failure to produce good documentation of the dissertation	Poorly written dissertation	I attended the English in-sessional writing classes offered by the university to improve writing skills

Table 3.3: Risk management

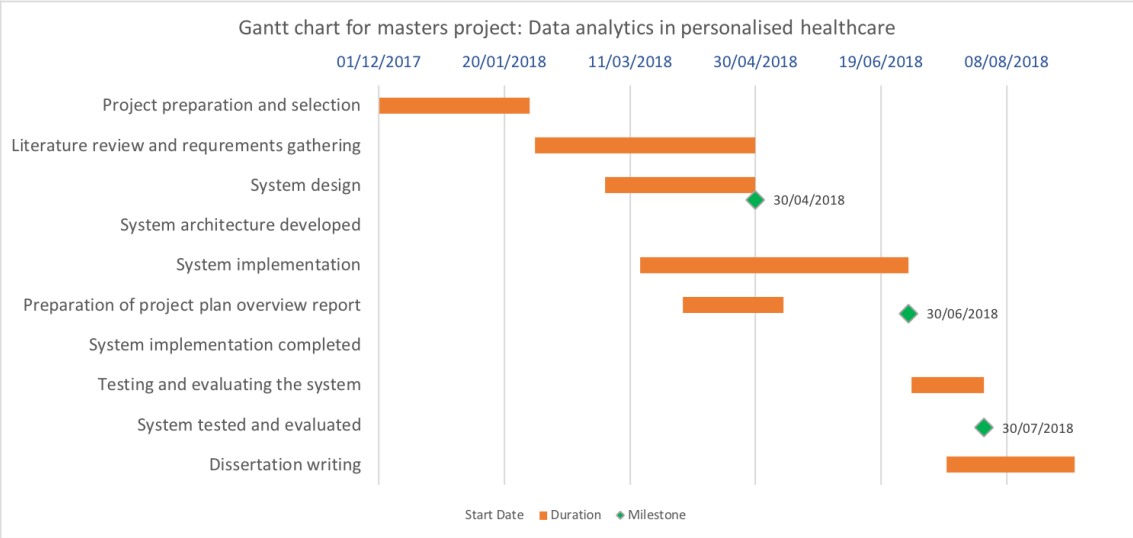


Figure 3.7: Gantt chart for project showing key activities, milestones and their time lines

3.7 Summary

This chapter discusses the methodology used to develop the system. The water fall model was used to develop the system. The system design that includes: the Entity Relationship diagram, the database structure, the languages and tools have also been discussed in this chapter. In addition, the project management tools, ethics and professional issues, risk management and project plan been discussed as well. The next chapter presents and discusses the results obtained after implementation of the system.

Chapter 4

Results and discussion

This chapter discusses the results obtained basing on the background and methodology presented in Chapter 2 and 3.

4.1 Modeling Pharmacokinetics for IV medicines

Objective 2 of the project was to develop models for the pharmacokinetics of medicines taken by Haemophilia patients over a period of time.

Equation (2.1) was used to model the first order pharmacokinetics of intravenous drugs. Figure 4.1 shows the elimination of IV drugs from the body.

Health professionals input details of a drug, that is, half-life, dosage and the Volume of distribution into the system and generate a plot of the drug pharmacokinetics into the body. For example, in Figure 4.1, the half-life used is 6 minutes, the volume of distribution is 3.2L and the dosage is 10mg.

Figure 4.1, shows a steady decrease in concentration as the drug is eliminated from the body. The slope of this curve at various times gives the actual measure of the rate of change of concentration at each time point. The rate of change of Plasma concentration (Cp) versus time = $\frac{\Delta.Cp}{\Delta.t} = Ke.Cp.$ where Ke is the elimination rate constant described in Section 2.2.7. Understanding the elimination rate constant for each drug helps health professionals to understand the rate at which a drug taken by the patient will be eliminated from the body. This is useful when determining dosages during

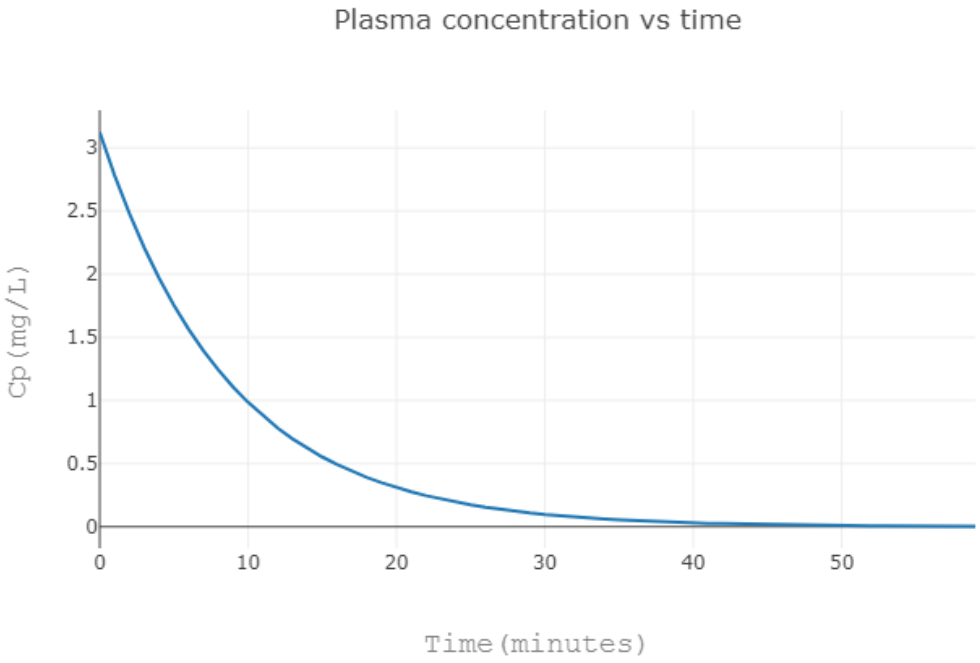


Figure 4.1: Graph showing absorption of IV drugs

prescription [D.M15].

4.1.1 Multiple IV bolus dosage pharmacokinetics

A parameterized model was implemented which takes as input a discrete set of temporal treatments from a database and produces a time series of pharmacokinetic curves representing drug concentration in the body based on patient specific parameters such as half-life of a medicine, volume of distribution, etc. Figure 4.2 shows an example of a Pharmacokinetic modelling for patient 105 for a month worth of treatments from 15th March 2016 to 15th April 2016.

Each injection is indicated by spikes on the Y-axis (Factor VIII IU/dL)) as well as the exponentially decaying characteristic of the medicine in body as it gets excreted. As well as drug concentrations, Figure 4.2 shows major events in red (Joint Ankle Trauma) which resulted in a bleed in this patient. Health professionals can specify a threshold of acceptable drug blood concentrations for each patient. The legend (top right in the Figure) shows the

amount of time the blood concentration was above and below this threshold. The Volume of Distribution is 32; the half life used is 10 hours; the blue line in the graph shows the threshold defined which is 40.

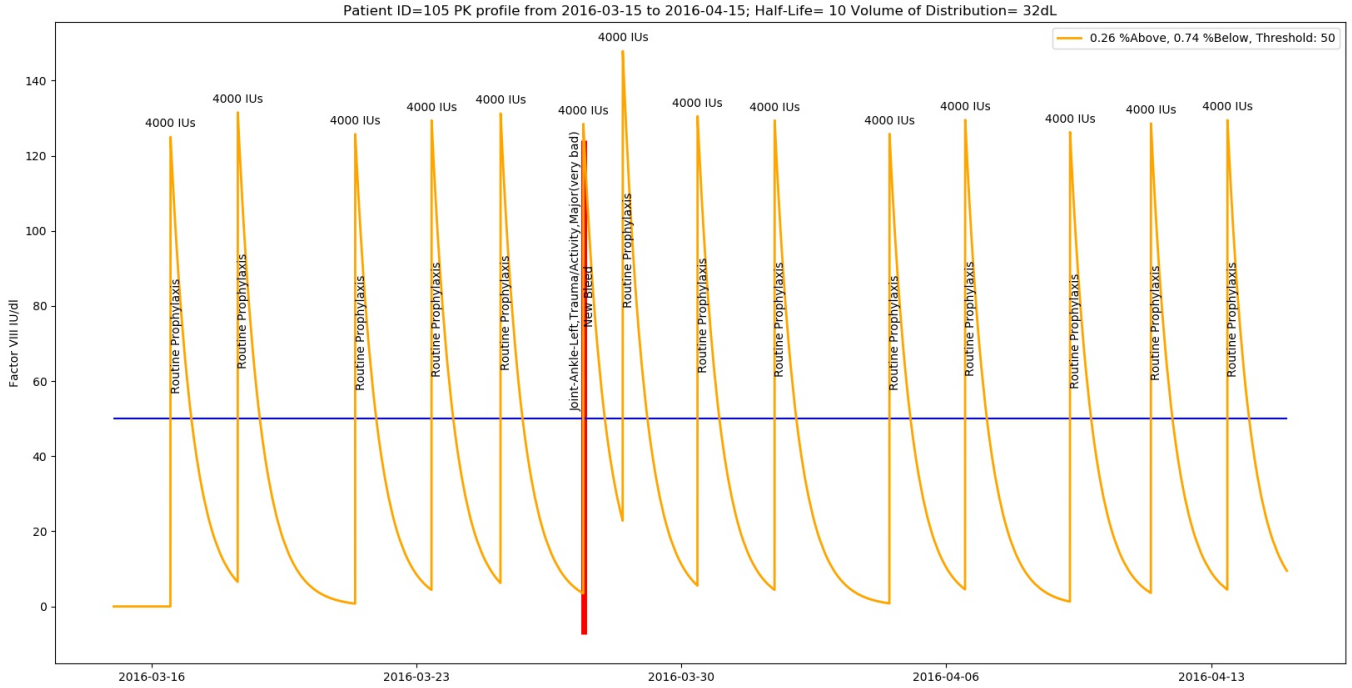


Figure 4.2: Time series for patient 105

Figure 4.2 shows an un-trended seasonal time series, defined in Section 2.3.1. The plasma concentration increases and decreases according to the times of the month but there is no change in the average value of the time series. The difference in the spikes is because of the times at which the next injection is administered.

If the next injection is given early enough, not all of the previous dose is eliminated from the body. The patient will still have some drug in their body and the drug will start to accumulate and we will get higher concentrations with the proceeding injections.

However, if the injections are given far enough apart (not prescribed frequency), the plasma concentration will have fallen to approximately zero before the next dose. There will be no accumulation of drug in the body [Bou01].

Figure 4.3 shows the time series for patient 79 for the period from 1st October 2008 to 28th October 2008; the volume of Distribution is 3 and the half Life is 8 hours. The red lines in the graph represent the various bleeds that the patient had in the time period.

From Figure 4.3, the doses were given at time intervals there were far a part. The drug concentration fell to zero. The patient had no drug in their body for some time period before the next drug was administered. This possibly explains why the patient gets multiple bleeds. People with severe hemophilia usually experience bleeding in their muscles or joints. They may bleed once to twice per week. Bleeding is often spontaneous with no obvious cause.[WFH18] [RNS$^+$14]. It is important for haemophilia patients to adhere to taking their medications so as to avoid such bleeds which can have far reaching effects such as arthropathy or even death [FVMB$^+$02].

Patient 79 is a typical example of a non-adherent patient. They do no adhere to the prescribed dosage and therefore, they have no drug in their body to help in blood clotting. They experience a number of bleeds in their elbow, knee and muscle.

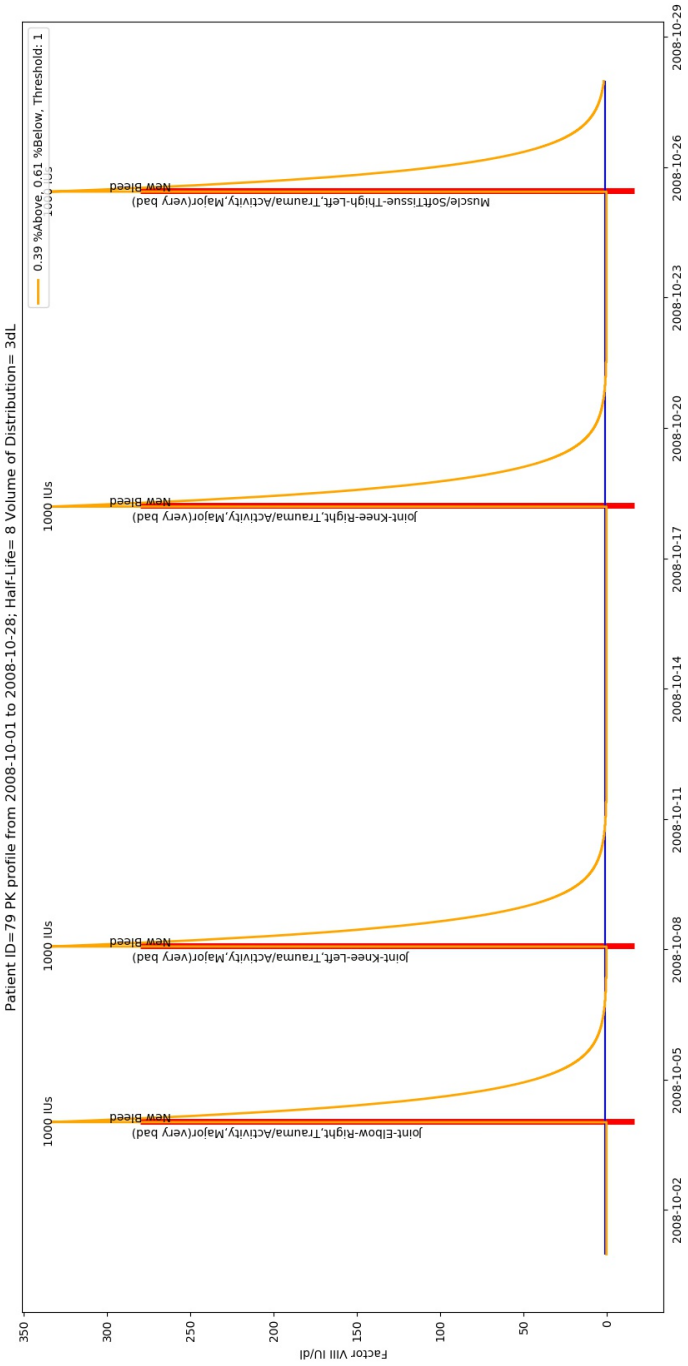


Figure 4.3: Time series for patient 79 with multiple bleeds

4.2 Modeling Pharmacokinetics for oral medicines

Using Equation 2.3 for the pharmacokinetics of oral medicines, we plot a graph of plasma concentration against time. Figure 4.4 shows pharmacokinetics of oral medicines.

Health professionals can input details about an oral medicine; eg. the half-life, volume of distribution, dosage, elimination constant (K_e), absorption constant (K_a), and system outputs a graph showing the medicine's absorption and elimination from the body.

In Figure 4.4, the half life is 6 minutes, the dosage is 300mg, the elimination constant (K_e) is 0.0025, the absorption constant (K_a) is 0.025, Bio viability constant(F) is 1 and the volume of distribution is 25.

Unlike with IV medicines, oral medicines are absorbed from the point of administration into the vascular system through a process defined by the absorption rate constant K_a and later eliminated by a process defined by the elimination constant, K_e as discussed in Section 2.2.8.

4.2.1 Effect of Ka

Health professionals can input different values for different parameters in the system such as K_a and F to study their effect on the pharmacokinetics of the drug. Figure 4.5 shows the output from the system on varying the absorption rate constant, K_a.

Figure 4.5 shows that the higher the values of the absorption constant (K_a), the higher and earlier the peak plasma concentrations (tpeak) are. When absorption rate constant K_a is 3hr-1, the peak plasma concentration tpeak is 1 hour. With $K_a = 0.6$, tpeak is 2.75 hours and with $K_a = 0.125hr - 1$, tpeak is 6.25 hours.

This means that a drug with a higher K_a is absorbed in the body faster than that with a lower K_a. The peak plasma concentration helps health professionals to estimate the rate of absorption of drugs. This is important in determining which drugs to administer for different situations, for instance, emmergencies may require drugs with higher absorption rates (K_a)[D.M15].

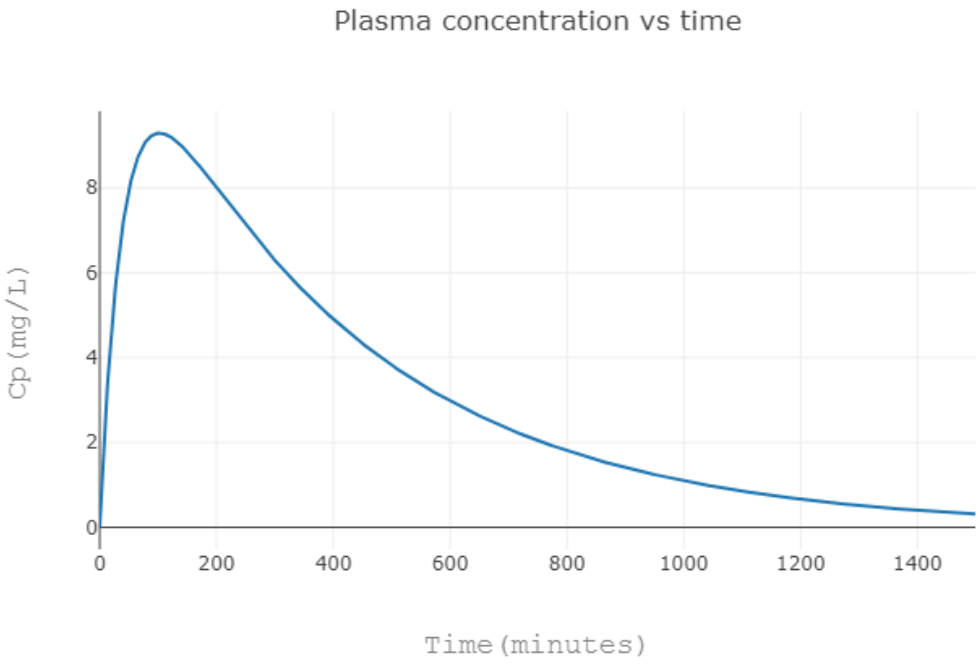


Figure 4.4: Graph showing absorption of oral medicines

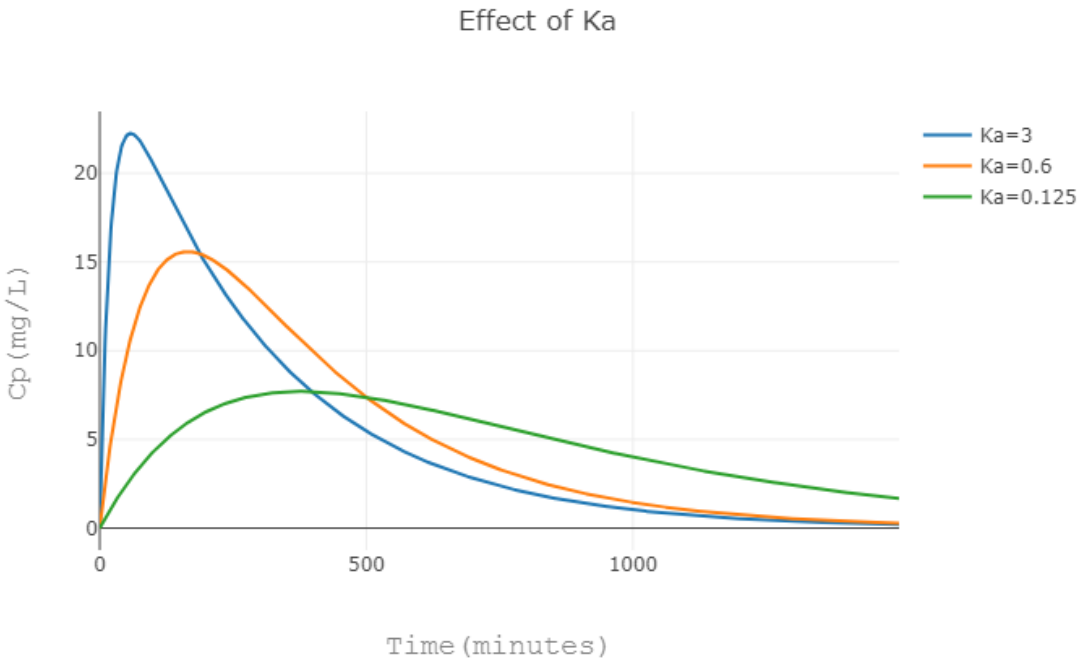


Figure 4.5: Effect of varying the absorption rate constant,Ka

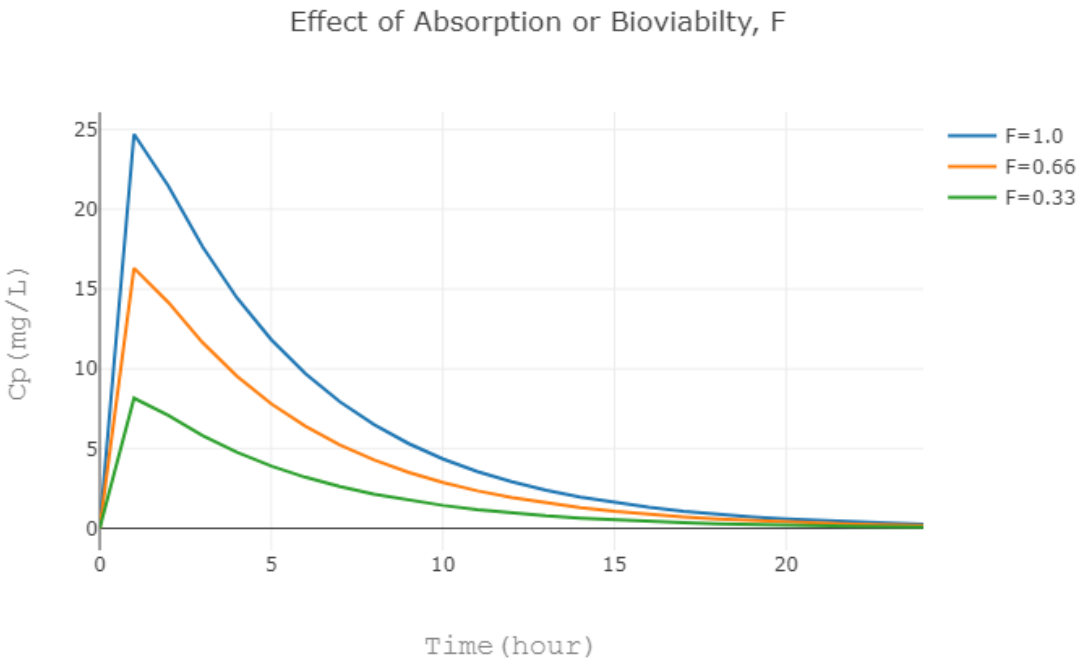


Figure 4.6: Effect of varying bioviability

4.2.2 Effect of varying bioviability

As discussed in Section 2.2.8, when oral drugs are administered to the body, their bioviability reduces depending on various factors such as age, weight of the patient, etc.

Figure 4.6 shows the effect of varying the bio viability. Health professionals can input values into the system including; half life, volume of distribution, dosage and bioviability and then, study the effect of varying the bioviability values.

Figure 4.6 shows that the peak plasma concentration (tpeak) is the same that is, 1 hour in all the cases with different values of the Bioviability (1, 0.66 and 0.33). As the values of Ke and K_a are unchanged, the time of peak plasma concentration is unchanged.

Bioviability is important to health professionals when calculating dosages for non-intravenous routes of administration to determine the fraction of an administered dose that reaches the systemic vascular system [Gri13].

4.3 Ranking of patients

Objective 3 of the project was to implement a program to sift through the Haemtrack data at predefined intervals, to rank patients according to compliance and inform health professionals of non-adherent patients.

4.3.1 Using the manhattan and euclidean distance

We use Equation 2.4 defined in Section 2.3.3 to compute the euclidean distance between the threshold and the patient time series. Health professionals input the threshold Plasma concentration that is recommended for a particular drug.

We also computed the Manhattan distance between the threshold and the patient time series using Equation 2.6.

We repeated the calculation of the manhattan and euclidean distances using the minkowski distance in Equation 2.7. As discussed in section 2.3.5, when p = 1, we obtain the Manhattan distance and when p = 2, we obtain the Euclidean distance. This helped us to verify the correctness of the distance values that were obtained.

Health professionals input a date range, the product (drug), the half-life, volume of distribution and the Center Number(CentreNo) for the patients they would like to rank for adherence.

Figure 4.7 shows the manhattan and euclidean distances of 24 patients when a threshold of 50 is defined for the period of 15th March 2016 to 15th April 2016. The volume of distribution used is 32, the half life is 10, the product used is 3(Advate) and the CentreNo is 999.

Figure 4.7 shows a list of patients in a random order.

To rank the patients, we sort the list by manhattan or euclidean distance. Results show that sorting the patient list in Figure 4.7 by either the manhattan or the euclidean distance, gave the same patient listing. Figure 4.8 shows the results obtained by sorting the patient data by manhattan or euclidean distance in ascending order.

Manhattan distance may be preferable to Euclidean distance for the case of high dimensional data [AHK01].

Patients with higher values of the manhattan and euclidean distance are considered to be non-adherent.

	RankID	Patient	DateFrom	DateTo	Threshold	ManhatanDistance	EuclideanDistance	Product	Centre
1	1	1921	2016-03-15 00:00:00	2016-04-15 00:00:00	50	446450	4724.669	3	999
2	2	2585	2016-03-15 00:00:00	2016-04-15 00:00:00	50	446450	4724.669	3	999
3	3	1589	2016-03-15 00:00:00	2016-04-15 00:00:00	50	367629.469	3973.326	3	999
4	4	687	2016-03-15 00:00:00	2016-04-15 00:00:00	50	446450	4724.669	3	999
5	5	1566	2016-03-15 00:00:00	2016-04-15 00:00:00	50	446450	4724.669	3	999
6	6	2562	2016-03-15 00:00:00	2016-04-15 00:00:00	50	336168.094	3816.461	3	999
7	7	2894	2016-03-15 00:00:00	2016-04-15 00:00:00	50	435801.156	4620.226	3	999
8	8	1875	2016-03-15 00:00:00	2016-04-15 00:00:00	50	446450	4724.669	3	999
9	9	1543	2016-03-15 00:00:00	2016-04-15 00:00:00	50	446450	4724.669	3	999
10	10	1374	2016-03-15 00:00:00	2016-04-15 00:00:00	50	446450	4724.669	3	999
11	11	2539	2016-03-15 00:00:00	2016-04-15 00:00:00	50	446450	4724.669	3	999
12	12	1188	2016-03-15 00:00:00	2016-04-15 00:00:00	50	366242.156	3902.99	3	999
13	13	192	2016-03-15 00:00:00	2016-04-15 00:00:00	50	304264.906	3505.478	3	999
14	14	856	2016-03-15 00:00:00	2016-04-15 00:00:00	50	446450	4724.669	3	999
15	15	1065	2016-03-15 00:00:00	2016-04-15 00:00:00	50	446450	4724.669	3	999
16	16	1520	2016-03-15 00:00:00	2016-04-15 00:00:00	50	446450	4724.669	3	999
17	17	401	2016-03-15 00:00:00	2016-04-15 00:00:00	50	446450	4724.669	3	999
18	18	524	2016-03-15 00:00:00	2016-04-15 00:00:00	50	446450	4724.669	3	999
19	19	501	2016-03-15 00:00:00	2016-04-15 00:00:00	50	294091.188	3402.426	3	999
20	20	2748	2016-03-15 00:00:00	2016-04-15 00:00:00	50	317253.312	3635.749	3	999
21	21	1443	2016-03-15 00:00:00	2016-04-15 00:00:00	50	446450	4724.669	3	999
22	22	2794	2016-03-15 00:00:00	2016-04-15 00:00:00	50	446450	4724.669	3	999
23	23	1998	2016-03-15 00:00:00	2016-04-15 00:00:00	50	311347.531	3525.451	3	999
24	24	919	2016-03-15 00:00:00	2016-04-15 00:00:00	50	446450	4724.669	3	999

Figure 4.7: List of patients with their manhattan,euclidean distances

	RankID	Patient	DateFrom	DateTo	Threshold	ManhatanDistance	EuclideanDista	Product	Centre
1	19	501	2016-03-15 00:00:00	2016-04-15 00:00:00	50	294091.188	3402.426	3	999
2	13	192	2016-03-15 00:00:00	2016-04-15 00:00:00	50	304264.906	3505.478	3	999
3	23	1998	2016-03-15 00:00:00	2016-04-15 00:00:00	50	311347.531	3525.451	3	999
4	20	2748	2016-03-15 00:00:00	2016-04-15 00:00:00	50	317253.312	3635.749	3	999
5	6	2562	2016-03-15 00:00:00	2016-04-15 00:00:00	50	336168.094	3816.461	3	999
6	12	1188	2016-03-15 00:00:00	2016-04-15 00:00:00	50	366242.156	3902.99	3	999
7	3	1589	2016-03-15 00:00:00	2016-04-15 00:00:00	50	367629.469	3973.326	3	999
8	7	2894	2016-03-15 00:00:00	2016-04-15 00:00:00	50	435801.156	4620.226	3	999
9	8	1875	2016-03-15 00:00:00	2016-04-15 00:00:00	50	446450	4724.669	3	999
10	9	1543	2016-03-15 00:00:00	2016-04-15 00:00:00	50	446450	4724.669	3	999
11	10	1374	2016-03-15 00:00:00	2016-04-15 00:00:00	50	446450	4724.669	3	999
12	11	2539	2016-03-15 00:00:00	2016-04-15 00:00:00	50	446450	4724.669	3	999
13	4	687	2016-03-15 00:00:00	2016-04-15 00:00:00	50	446450	4724.669	3	999
14	5	1566	2016-03-15 00:00:00	2016-04-15 00:00:00	50	446450	4724.669	3	999
15	1	1921	2016-03-15 00:00:00	2016-04-15 00:00:00	50	446450	4724.669	3	999
16	2	2585	2016-03-15 00:00:00	2016-04-15 00:00:00	50	446450	4724.669	3	999
17	21	1443	2016-03-15 00:00:00	2016-04-15 00:00:00	50	446450	4724.669	3	999
18	22	2794	2016-03-15 00:00:00	2016-04-15 00:00:00	50	446450	4724.669	3	999
19	14	856	2016-03-15 00:00:00	2016-04-15 00:00:00	50	446450	4724.669	3	999
20	15	1065	2016-03-15 00:00:00	2016-04-15 00:00:00	50	446450	4724.669	3	999
21	16	1520	2016-03-15 00:00:00	2016-04-15 00:00:00	50	446450	4724.669	3	999
22	17	401	2016-03-15 00:00:00	2016-04-15 00:00:00	50	446450	4724.669	3	999
23	18	524	2016-03-15 00:00:00	2016-04-15 00:00:00	50	446450	4724.669	3	999
24	24	919	2016-03-15 00:00:00	2016-04-15 00:00:00	50	446450	4724.669	3	999

Figure 4.8: Sorted list of patients by manhattan and euclidean distance

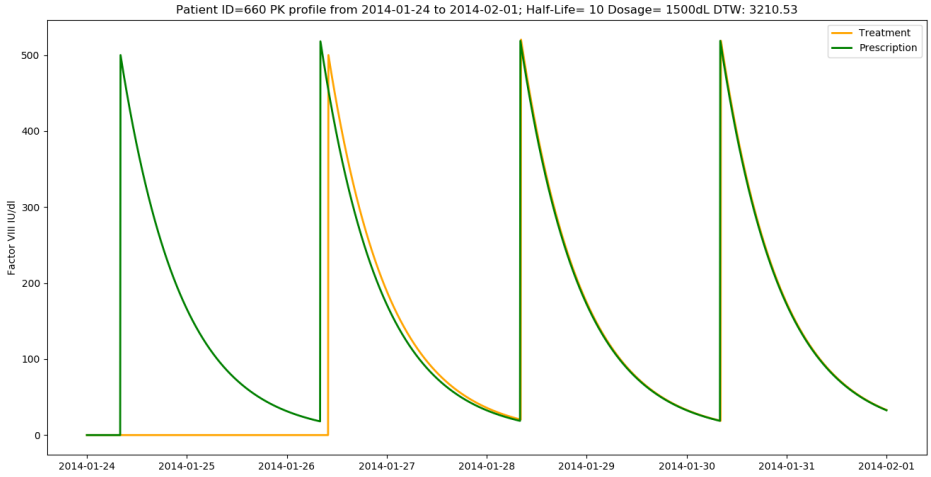


Figure 4.9: Dynamic Time Warping between the patient and treatment time series where patient adheres

4.3.2 Using the Dynamic Time Warping algorithm (DTW)

Following the discussion in Section 2.3.6, we used the DTW algorithm to measure the similarity between two time series. We compare a prescription time series and a patient treatment time series. The prescription time series consists of the dosage, frequency, number of doses and the dosing interval that the patient is supposed to take in a period of time. The patient treatment time series has the actual time, frequency and half life of the drug that the patient takes as modelled by the pharmacokinetic equation 2.1

Figure 4.9 shows a Prescription time series in green and a treatment time series in orange. The patient adheres to the prescription and that is why the curves are uniform at some points.

Figure 4.10 shows the wrapping path between the two time series in Figure 4.9. A straight line is expected if the time series are the same even when they are time shifted.

Figure 4.11 shows a treatment and prescription time series of a patient whose treatment deviates from the prescription (non-adherent). Figure 4.12 shows the wrapping path between these two time series.

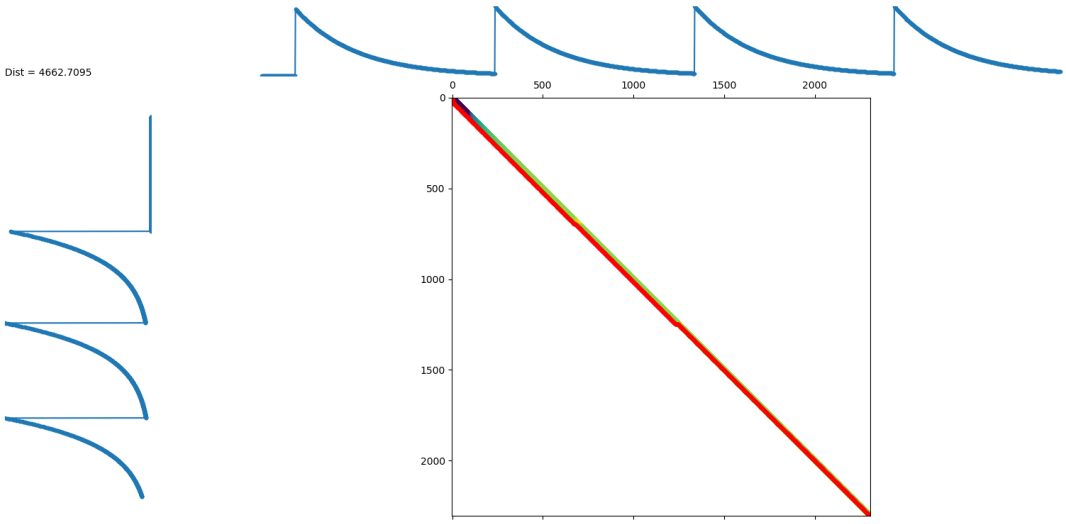


Figure 4.10: Warping path

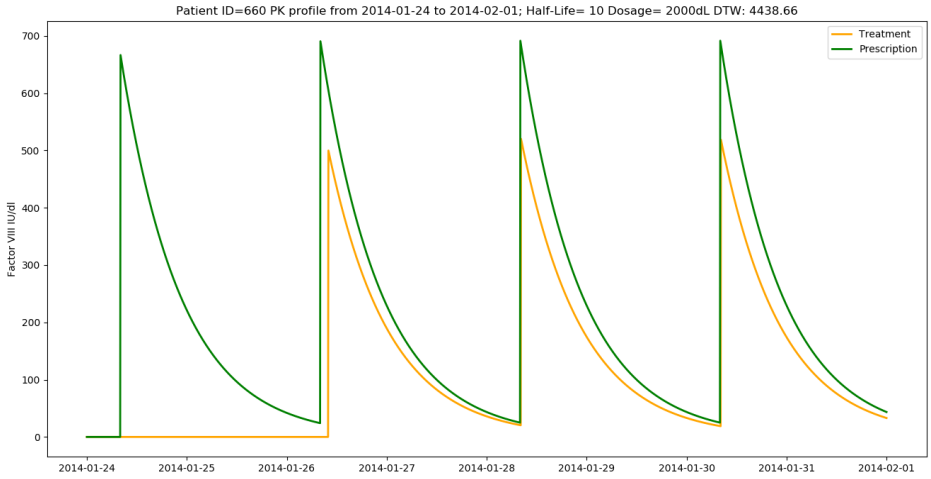


Figure 4.11: DTW between the patient and treatment time series where patient does not adhere

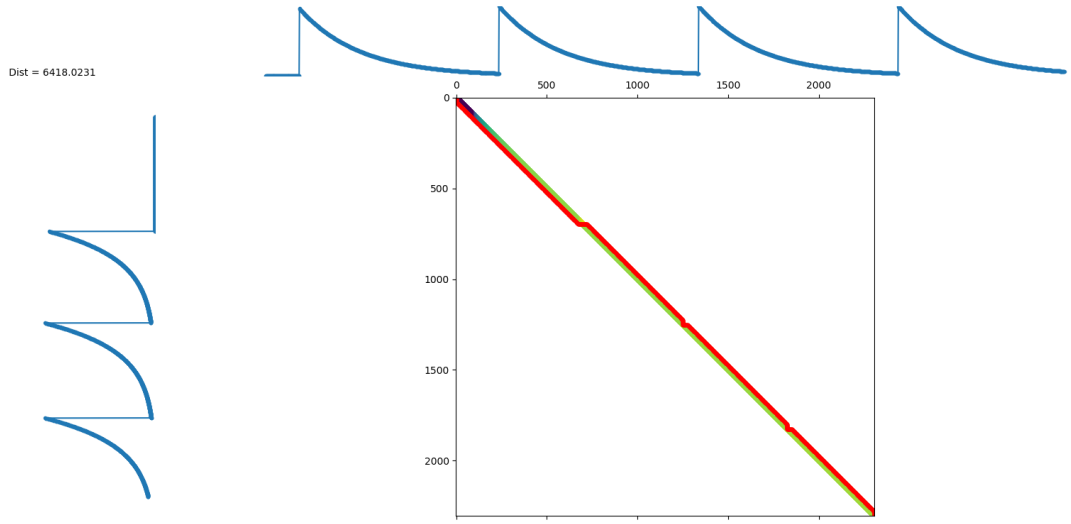


Figure 4.12: DTW path

	RankID	Patient	DateFrom	DateTo	Dosage	Frequency	Number_Doses	Dosing_Interval	DTW	Product
1	1	1921	2016-03-15 00:00:00	2016-04-15 00:00:00	1500	5	4	48	898.756243108167	3
2	2	2585	2016-03-15 00:00:00	2016-04-15 00:00:00	1500	5	4	48	898.756243108167	3
3	3	1589	2016-03-15 00:00:00	2016-04-15 00:00:00	1500	5	4	48	712.395462770971	3
4	4	687	2016-03-15 00:00:00	2016-04-15 00:00:00	1500	5	4	48	898.756243108167	3
5	5	1566	2016-03-15 00:00:00	2016-04-15 00:00:00	1500	5	4	48	898.756243108167	3
6	6	2562	2016-03-15 00:00:00	2016-04-15 00:00:00	1500	5	4	48	1090.06331623324	3
7	7	2894	2016-03-15 00:00:00	2016-04-15 00:00:00	1500	5	4	48	458.243489011205	3
8	8	1875	2016-03-15 00:00:00	2016-04-15 00:00:00	1500	5	4	48	898.756243108167	3
9	9	1543	2016-03-15 00:00:00	2016-04-15 00:00:00	1500	5	4	48	898.756243108167	3
10	10	1374	2016-03-15 00:00:00	2016-04-15 00:00:00	1500	5	4	48	898.756243108167	3
11	11	2539	2016-03-15 00:00:00	2016-04-15 00:00:00	1500	5	4	48	898.756243108167	3
12	12	1188	2016-03-15 00:00:00	2016-04-15 00:00:00	1500	5	4	48	550.974503864727	3
13	13	192	2016-03-15 00:00:00	2016-04-15 00:00:00	1500	5	4	48	1273.35179172393	3
14	14	856	2016-03-15 00:00:00	2016-04-15 00:00:00	1500	5	4	48	898.756243108167	3
15	15	1065	2016-03-15 00:00:00	2016-04-15 00:00:00	1500	5	4	48	898.756243108167	3
16	16	1520	2016-03-15 00:00:00	2016-04-15 00:00:00	1500	5	4	48	898.756243108167	3
17	17	401	2016-03-15 00:00:00	2016-04-15 00:00:00	1500	5	4	48	898.756243108167	3
18	18	524	2016-03-15 00:00:00	2016-04-15 00:00:00	1500	5	4	48	898.756243108167	3
19	19	501	2016-03-15 00:00:00	2016-04-15 00:00:00	1500	5	4	48	1338.59471707851	3
20	20	2748	2016-03-15 00:00:00	2016-04-15 00:00:00	1500	5	4	48	1292.69315788125	3
21	21	1443	2016-03-15 00:00:00	2016-04-15 00:00:00	1500	5	4	48	898.756243108167	3
22	22	2794	2016-03-15 00:00:00	2016-04-15 00:00:00	1500	5	4	48	898.756243108167	3
23	23	1998	2016-03-15 00:00:00	2016-04-15 00:00:00	1500	5	4	48	1055.79020448499	3
24	24	919	2016-03-15 00:00:00	2016-04-15 00:00:00	1500	5	4	48	898.756243108167	3

Figure 4.13: Calculating the Dynamic Time Warping for each patient

4.4 Ranking patients using DTW

We used the Dynamic Time Warping (DTW) algorithm in section 2.3.6 to rank patients. Health professionals can define a prescription of a drug which includes dosage, frequency, number of doses and dosing interval. A prescription time series is generated using this information. The prescription time series acts as the threshold or the ideal which the patient is expected to follow. We obtain the patient time series which is the actual treatments taken by the patient. We then compare the patient time series with the prescription time series and compute the DTW distance.

Figure 4.13 shows the list of patients with their DTW distance. The prescription is shown by the columns: Dosage, Frequency, Number-Doses and Dosage-Interval.

Figure 4.14 shows the results obtained by sorting the patient data by DTW. The same patient ranking list is obtained using the euclidean, manhattan and DTW distances; this consistency of order acts as informal validation of the ranking.

	RankID	Patient	DateFrom	DateTo	Dosage	Frequency	Number_Doses	Dosing_Interval	DTW	Product
1	19	501	2016-03-15 00:00:00	2016-04-15 00:00:00	1500	5	4	48	1338.59471707851	3
2	13	192	2016-03-15 00:00:00	2016-04-15 00:00:00	1500	5	4	48	1273.35179172393	3
3	6	2562	2016-03-15 00:00:00	2016-04-15 00:00:00	1500	5	4	48	1090.06331623324	3
4	23	1998	2016-03-15 00:00:00	2016-04-15 00:00:00	1500	5	4	48	1055.79020448499	3
5	24	919	2016-03-15 00:00:00	2016-04-15 00:00:00	1500	5	4	48	898.756243108167	3
6	25	1251	2016-03-15 00:00:00	2016-04-15 00:00:00	1500	5	4	48	898.756243108167	3
7	21	1443	2016-03-15 00:00:00	2016-04-15 00:00:00	1500	5	4	48	898.756243108167	3
8	22	2794	2016-03-15 00:00:00	2016-04-15 00:00:00	1500	5	4	48	898.756243108167	3
9	14	856	2016-03-15 00:00:00	2016-04-15 00:00:00	1500	5	4	48	898.756243108167	3
10	15	1065	2016-03-15 00:00:00	2016-04-15 00:00:00	1500	5	4	48	898.756243108167	3
11	16	1520	2016-03-15 00:00:00	2016-04-15 00:00:00	1500	5	4	48	898.756243108167	3
12	17	401	2016-03-15 00:00:00	2016-04-15 00:00:00	1500	5	4	48	898.756243108167	3
13	18	524	2016-03-15 00:00:00	2016-04-15 00:00:00	1500	5	4	48	898.756243108167	3
14	1	1921	2016-03-15 00:00:00	2016-04-15 00:00:00	1500	5	4	48	898.756243108167	3
15	2	2585	2016-03-15 00:00:00	2016-04-15 00:00:00	1500	5	4	48	898.756243108167	3
16	4	687	2016-03-15 00:00:00	2016-04-15 00:00:00	1500	5	4	48	898.756243108167	3
17	5	1566	2016-03-15 00:00:00	2016-04-15 00:00:00	1500	5	4	48	898.756243108167	3
18	8	1875	2016-03-15 00:00:00	2016-04-15 00:00:00	1500	5	4	48	898.756243108167	3
19	9	1543	2016-03-15 00:00:00	2016-04-15 00:00:00	1500	5	4	48	898.756243108167	3
20	10	1374	2016-03-15 00:00:00	2016-04-15 00:00:00	1500	5	4	48	898.756243108167	3
21	11	2539	2016-03-15 00:00:00	2016-04-15 00:00:00	1500	5	4	48	898.756243108167	3
22	3	1589	2016-03-15 00:00:00	2016-04-15 00:00:00	1500	5	4	48	712.395462770971	3
23	12	1188	2016-03-15 00:00:00	2016-04-15 00:00:00	1500	5	4	48	550.974503864727	3
24	7	2894	2016-03-15 00:00:00	2016-04-15 00:00:00	1500	5	4	48	458.243489011205	3

Figure 4.14: Sorted list of patients ranked by DTW

ED	DTW
Only works with time series of the same length	Can be used on time series of any length
More resistant to noisy or spikey test data than DTW	Generally weaker when data is spikey or noisy
Fails when data is shifted or transformed with respect to time	Robust when data is shifted or transformed with respect to time
Quicker/simpler	Much slower

Table 4.1: Euclidean distance vs DTW [Bev17]

It is important to note that in order to obtain the same ranking as shown in figure 4.8, the data in figure 4.14 is sorted in descending order.

The query to output results in Figure 4.8 with manhattan and euclidean distances took 2 seconds to execute. However, to output patient listing in 4.14 with DTW algorithm took up to 25 minutes. Therefore, the DTW algorithm takes more time to output patient rankings than the Euclidean or Manhattan distance. This is consistent with the discussion in Section 2.3.6 because DTW uses dynamic programming which causes polynomial complexity and as thus, memorizing large matrices of numbers and performing large numbers of calculations is challenging.

4.4.1 Difference between Euclidean and DTW distances

Figure 4.15 shows the results obtained by shifting a time series to study the difference between the euclidean and DTW distances. A patient's treatments time series was obtained and shifted as in Figure 4.15. The results showed that euclidean distance is higher than the DTW distance.

Euclidean distance is a one-to-one comparison. In contrast, the DTW) distance is a one-to-many comparison. According to [Bev17], the difference between Euclidean distance and DTW are as shown Table 4.1

4.5 Summary

Using pharmacokinetic first-order equations, health professionals specify values of a drug and view graphs showing drug pharmacokinetics. **Health**

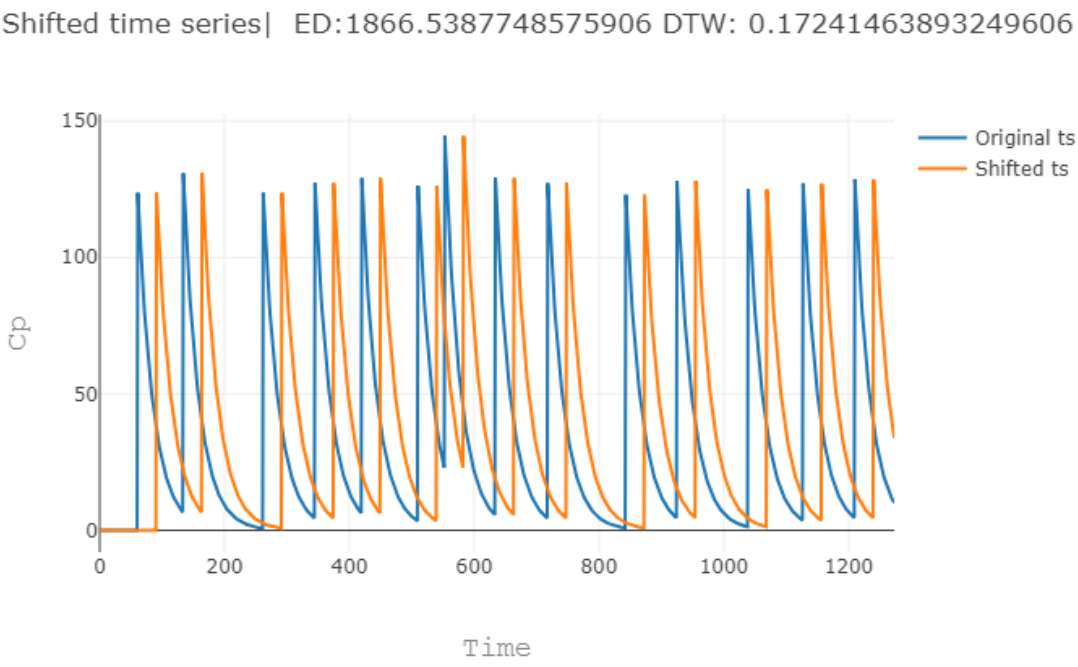


Figure 4.15: Shifting time series

professionals can vary different properties of the drug to study the effect. The results presented in this chapter demonstrate the use of time series to model patient treatments. By defining the drug threshold, **health professionals** can use Euclidean or Manhattan distance to rank patients according to their adherence rates. The dynamic time warping algorithm can also be used to rank patients but it is slower by as it compares the treatment time series with the prescription time series.

Chapter 5

Evaluation

A detailed discussion on how the system was evaluated, both individually and in comparison to others is presented in this chapter.

5.1 Evaluation of the system against traditional approaches

The system was evaluated against traditional approaches identified in the literature search such as questionnaires, patient self-reports, rates of prescription refills, direct methods and patient diaries. Table 5.1 shows how the current system addresses the current challenges in measuring adherence to treatments.

5.1.1 Comparing the developed system with VERITAS Pro

The VERITAS Pro is a widely used questionnaire that has been proved to be reliable and valid in monitoring adherence to prophylactic treatments. Both the system developed here and VERITAS-Pro enhance communication and understanding between providers and patients by allowing providers to understand patient's adherence and identify specific modifiable determinants of non-adherence associated with poor outcome. Table 5.2 compares the VERITAS Pro with the developed system.

Identified significant challenge	Developed solution
Traditional methods of adherence monitoring for instance, patient self-reports do not account for the pharmacokinetic properties of the drug and the medication-taking behavior, hence misrepresent the true therapeutic exposure. [TNL+17] [VHVD02]	The developed system considers the pharmacokinetic properties of the drug and models the treatment behavior as a function of time making it possible to understand the medication-taking behavior and the patient therapeutic exposure.
Existing methods such as patient diaries do not provide the ability to visualize the treatments taken by a patient over a specified period of time [JJ11].	The developed system models patient treatments as a time series making it easy to analyze patient medication behavior over a specified period of time.
Methods such as questionnaires require clinical knowledge to make sense out of the data and to be able to rank non-adherent patients [DKRS10] [LRD+14]	The developed system has been designed to be easy for all health professionals and administrators to quickly get a ranked list of non-adherent patients.
Inability to visualize each drug and its pharmacokinetics [JJ11] [VHVD02].	Health professionals can model the pharmacokinetics of drugs and analyze the effect of varying several parameters such as half-life, bio-viability, etc.
A significant amount of time is taken to go through patient treatments, one at a time to identify non-adherent patients when using traditional methods such as patient diaries, pill counts, monitoring drug refills, etc. [JJ11] [VHVD02]	The developed system is automated. Cron jobs sift through patient data and send e-mail lists of patients with their adherence ranking to health professionals, saving significant and expert analysis, and allowing real-time continuous monitoring to occur.
Methods that account for the pharmacokinetic properties of the drug, for instance, measurement of the drug or its metabolite in blood or urine are very expensive to perform. [JJ11] [VHVD02]	The developed system is less costly than direct measures of adherence. The system can be integrated as a plugin in existing patient diary systems, such as Haemtrack to provide extra functionality of pharmacokinetic properties of treatments taken.

Table 5.1: Comparison of traditional adherence methods to the developed system in addressing identified challenges

VERITAS Pro	**Developed System**
Only used for prophylactic treatments in hemophilia.[DKRS10]	Can be used for both prophylaxis and on-demand treatments. Could be adopted to monitor adherence to treatments in other chronic diseases.
VERITAS-Pro is comprehensive. Including diverse indicators of adherence to prophylactic regimens, including, timing, dosing, communicating, planning, skipping and remembering. [DKRS10] [FLD18]	The developed system considers some indicators that the VERITAS -Pro covers, for example, Timing and Dosing. The other parameters - Communicating, Planning, Skipping and Remembering are not covered.
VERITAS-Pro is brief, easy to use and well-suited for use during a clinic visit. [LRD$^+$14] [CBTOGP$^+$17]	The developed system can be incorporated into patient diary systems to improve home treatments.
Usefulness of the VERITAS Pro requires expert understanding of the clinical implications of specific total and sub-scale scores as defined by cutoff scores and norms. [DKRS10]	The developed system automates the process of patient ranking making it easy for health professionals to quickly view the non-adherent patients without expert knowledge of clinical scores.

Table 5.2: Comparison of the VERITAS Pro and the developed system

MEMS	Developed system
Allows sending data from patients to health professionals in real time. Data usually sent includes log time of opening bottles, dispensing medications, opening injection containers, etc. This data is used to measure adherence to treatments [AJM18].	Also allows sending data from patients to health professionals in real time. Emails are sent to health professionals at predefined times informing them of non-adherent patients
There is doubtful reliability on the accuracy of results used by MEMS. Patient may open the bottles but not take the medication [VHVD02].	The reliability of the system developed depends on the accuracy of the treatments data collected by other means for instance; patient diaries, Electronic Medical Records software, etc.
Both approaches are not accessible to everyone. They rely on the use of technology which requires internet connectivity. This may not be easy to achieve in resource poor settings. Also, patients may require some skills to operate MEMS devices which some people find challenging [ZN07].	
Both approaches need additional custom healthcare software to operate. The developed system requires treatment data collected from either patient diary systems or Electronic Medical Records Software. MEMS requires additional software to analyse the log data generated [GHW+11]. The developed system could be used with MEMS to do pharmacokinetic analysis.	

Table 5.3: Comparison of the developed system with MEMS

5.1.2 Comparing the developed system with MEMS

As discussed in Section 2.2.3, Medication Electronic Monitoring Systems use microprocessor powered technology to record and log the time of opening bottles, dispensing machines, opening injections, etc. Table 5.1.2 shows the difference between the MEMS approach and the developed system in measuring adherence to treatments.

5.2 Summary

In this chapter, we evaluated the system against other approaches being used to measure adherence to treatment such as patient self reports, questionnaires, pill counts,etc. A detailed comparison of the system with the VERITAS Pro questionnaire and with MEMS was also carried out.

Apart from direct methods, all the other methods do not account for the

pharmacokinetic properties of the drugs that patients take and hence, misrepresent the true therapeutic exposure. The developed system approach that utilizes pharmacokinetics and ranks patients using the euclidean, manhattan distances and the DTW algorithm could be used with approaches like patient diaries and MEMS to provide the pharmacokinetic analysis.

Chapter 6

Conclusion and recommendations

6.1 Conclusion

This project was undertaken to incorporate data analytics technologies in Haemtrack and evaluate the impact of such technologies in improving adherence to treatments.

Our objectives were to:

i. assess current methods used to measure adherence to treatments,
ii. develop pharmacokinetic models for treatments taken by haemophilia patients over a period of time,
iii. develop a system to sift through the patient's treatment's data and rank patients according to adherence
iv. evaluate the effectiveness of the system developed in improving treatment adherence monitoring.

We have developed a system that can model pharmacokinetics of different drugs. The system can be used to study the effect of varying different pharmacokinetic properties of a drug.

This study has shown that patient treatments can be modeled as a time series using pharmacokinetic properties of drugs taken. Using well known distance measures, namely manhattan and euclidean distance, patients can be ranked by the difference of their time series with a threshold specified

by the health professional. In addition, the Dynamic Time Warping (DTW) algorithm can be used to compare the patient time series with a prescription time series to yield the same ranking list. However, the DTW is slower by an order of magnitude.

Overall, the study strengthens work on previous studies that use pharmacokinetic based adherence measures of adherence to treatments. A study was classified patients as adherent or non-adherent by pharmacokinetic modelling of their serum drug concentration in comparison to a therapeutic threshold. Patients were classified as adherent if their serum drug concentrations remained within the therapeutic range and non-adherent if the concentrations fell below the threshold [MSK15]. Another study carried out in Kenya to measure adherence to antiretroviral therapy in HIV-infected children used the ratio between the Mean plasma concentration (Cp) of individual patients during 1 month of follow-up and the intended plasma concentration (Cp') where the patient perfectly follows the dosing regimen. The ratio was used to determine the exposure of patients to the drug as compared to the intended level. Smaller values of the ratio indicated poorer adherence. [TNL$^+$17]

The findings of this study should be of interest to the Haemophilia society and the implementer of the National Haemophilia database Haemtrack. These could integrate this system into Haemtrack to improve on monitoring of patient treatments. Further, the methods could be adopted for use in measuring adherence to treatments in other chronic diseases such as heart failure, human immunodeficiency virus, diabetes, etc.

This system could be used in combination with patient diaries, MEMS, pharmacy databases, etc. to account for the pharmacokinetic properties of the drugs. Combining several methods is a recommended way of measuring treatment adherence, since there is no gold standard [WHO03].

This study does not consider specific patient demographics, such as, age, weight, etc as the data set does not contain these parameters. The patient specific demographics will have an effect on the pharmacokinetics of the drugs taken.

This study is limited to the one-compartment model of pharmacokinetics. The one-compartment model assumes that the drug in the blood is in rapid equilibrium with the drug in the extra-vascular tissues [Bou01]. This is

not an exact representation. However, it is useful for several drugs to a reasonable approximation.

6.2 Recommendations

This project provides the following insights for future research:

This study could be repeated using the method of clustering patients based on their treatment. Using clustering algorithms such as k-means and hierarchical clustering, analytics on the existing time series data models could be implemented to find insights that will help health professionals adjust dosages given or identify risky patients. Using this method, patients could be automatically grouped into clusters based on their pharmacokinetic profiles. Clusters can be associated with different outcomes and patients grouped under the ”poor outcome” category can be recommended treatment plans from the ”good outcome” cluster.

Forecasting the treatment pattern behavior in patients would be a fruitful area for further research as well. This could be done applying forecasting algorithms on past pharmacokinetic data to predict what treatment paths the patient is likely to take.

In addition, an automatic anomaly detection algorithm could be implemented to detect certain anomalies (unusual patterns that do not conform to the usual expected behavior). The algorithm would help health professionals to quickly identify changes in treatments/prophylaxis patterns and thus, to take rapid action whenever they appear. There are various algorithms that could be explored in anomaly detection. These include: Support Vector Machine-Based Anomaly Detection,

Further more, multi-dimensional time series patterns associated with good/bad outcomes could be investigated further. This feature will consider multi-dimensional time series (where each modality represents a pharmacokinetic time series for a particular medicine).

Bibliography

[Agr13] R.K; Ratnadip Adhikari Agrawal. An Introductory Study on Time Series Modeling and Forecasting. *arXiv preprint arXiv:1302.6613*, 1302.6613:1–68, 2013.

[AHK01] Charu C. Aggarwal, Alexander Hinneburg, and Daniel A. Keim. On the Surprising Behavior of Distance Metrics in High Dimensional Space. pages 420–434, 2001.

[Ahm15] Tarek A. Ahmed. *Pharmacokinetics of Drugs Following IV Bolus, IV Infusion, and Oral Administration*. 2015.

[aidPcg17] Strategy and innovation directive Personalised care group. Harnessing technology for personalised health care. Technical report, 2017.

[AJM18] Murtadha Aldeer, Mehdi Javanmard, and Richard Martin. A Review of Medication Adherence Monitoring Technologies. *Applied System Innovation*, 1(2):14, 2018.

[Bah18] Denis Bahati. *Data analytics in personalised Health care*. PhD thesis, University of Manchester, 2018.

[Bev17] Jonnie Bevan. Time Series Classification. 2017.

[Bob09] Bob Hughes and Mike Cotterell. *Software project management*. McGraw-Hill, 5th editio edition, 2009.

[Bou01] David W. A. Bourne. *Basic pharmacokinetics*. 2001.

[CBTOGP+17] Rubén Cuesta-Barriuso, Ana Torres-Ortuño, Pilar Galindo-Piñana, Joaquín Nieto-Munuera, Natalie Duncan, and José Antonio López-Pina. Validation of the VERITAS-Pro treatment adherence scale in a Spanish sample population with hemophilia. *Patient Preference and Adherence*, 11:653–660, 2017.

[CE10] David M. Cutler and Wendy Everett. Thinking Outside the Pillbox — Medication Adherence as a Priority for Health Care Reform. *New England Journal of Medicine*, 362(17):1553–1555, 2010.

[DGK+15] Iain Dunning, Vishal Gupta, Angela King, Jerry Kung, Miles Lubin, and John Silberholz. Operations Research and Analytics, 2015.

[DHH+15] National Haemophilia Database, How Haemtrack, Haemtrack Using Haemtrack, Security The, and Registration Haemtrack. Haemtrack. 2015.

[DKRS10] N. A. Duncan, W. Kronenberger, C. Roberson, and A. Shapiro. VERITAS-Pro: A new measure of adherence to prophylactic regimens in haemophilia. *Haemophilia*, 16(2):247–255, 2010.

[D.M15] S.B. Jaiswal D.M Brahmankar. *Biopharmaceutics and Pharmacokinetics*. 2015.

[DUVR08] Philippe De Moerloose, W. Urbancik, H. M. Van Den Berg, and M. Richards. A survey of adherence to haemophilia therapy in six European countries: Results and recommendations. *Haemophilia*, 14(5):931–938, 2008.

[Fay96] Smyth Fayyad, Piatetsky-Shapiro. *From data mining to Knowledge discovery: An overview*. MIT PRess, 1996.

[FLD18] Adriana Aparecida Ferreira, Isabel Cristina Gonçalves Leite, and Natalie August Duncan. Validation of the Brazilian version of the VERITAS-Pro scale to assess adherence to prophylactic regimens in hemophilia. *Hematology, Transfusion and Cell Therapy*, 40(1):18–24, 2018.

[FLR+12] Germain Forestier, Florent Lalys, Laurent Riffaud, Brivael Trelhu, and Pierre Jannin. Classification of surgical processes using dynamic time warping. *Journal of Biomedical Informatics*, 45(2):255–264, 2012.

[Fur08] Titus Felix Furtună. Dynamic Programming Algorithms in Speech Recognition. *Word Journal Of The International Linguistic Association*, 2(2):94–99, 2008.

[FVMB+02] K Fischer, J G Van Der Bom, E P Mauser-Bunschoten, G Roosendaal, R Prejs, P De Kleijn, D E Grobbee, and M Van Den Berg. The effects of postponing prophylactic treatment on long-term outcome in patients with severe hemophilia. *Blood*, 99(7):2337–2341, 2002.

[Gan17] Manisha Ganguly. Hemophilia trial finds promising cure, dec 2017.

[GCE+11] Sara Garfield, Sarah Clifford, Lina Eliasson, Nick Barber, and Alan Willson. Suitability of measures of self-reported medication adherence for routine clinical use: A systematic review. *BMC Medical Research Methodology*, 11(1):149, 2011.

[GD95] D.M. Gavrila and L.S. Davis. 3-D model-based tracking of human upper body movement: a multi-view\napproach. *Proceedings of International Symposium on Computer Vision - ISCV*, pages 3–8, 1995.

[GHW+11] David Gillespie, Kerenza Hood, Andrew Williams, Rachel Stenson, Christopher Probert, and Antony Hawthorne. The use of the Medication Event Monitoring System (MEMS) for assessing medication adherence for chronic conditions: use and results from a 12 month trial of patients in remission with ulcerative colitis (UC). *Trials*, 12(Suppl 1):A130, 2011.

[Git18] Github. Atom, 2018.

[GP96] K. Gollmer and C. Posten. Supervision of bioprocesses using a dynamic time warping algorithm. *Control Engineering Practice*, 4(9):1287–1295, 1996.

[Gra16] Ellen Graham. Improving Outcomes Through Personalised Medicine. *NHS England*, pages 6–10, 2016.

[Gri13] J. P. Griffin. *The Textbook of Pharmaceutical Medicine*. BMJ Books, New Jersey, 7th editio edition, 2013.

[GSW+13] Jeffrey S. Gonzalez, Havah E. Schneider, Deborah J. Wexler, Christina Psaros, Linda M. Delahanty, Enrico Cagliero, and Steven A. Safren. Validity of medication

adherence self-reports in adults with type 2 diabetes. *Diabetes Care*, 36(4):831–837, 2013.

[Ham94] James Douglas Hamilton. *Time series analysis*, volume 2. Princeton university press Princeton, NJ, 1994.

[Hay02] R. Brian Haynes. Helping Patients Follow Prescribed Treatment<SUBTITLE>Clinical Applications</SUBTITLE>. *Jama*, 288(22):2880, 2002.

[HFG16] Niloofar Hezarjaribi, Ramin Fallahzadeh, and Hassan Ghasemzadeh. A machine learning approach for medication adherence monitoring using body-worn sensors. In *Proceedings of the 2016 Conference on Design, Automation & Test in Europe*, DATE '16, pages 842–845, San Jose, CA, USA, 2016. EDA Consortium.

[HWBE05] Rob Horne, John Weinman, Nick Barber, and Rachel Elliott. Concordance, adherence and compliance in medicine taking. *Report for the National Co-ordinating Centre for NHS Service Delivery and Organisation R & D (NCCSDO)*, pages 1 – 331, 2005.

[HYT+13] Faiz M. Hasanuzzaman, Xiaodong Yang, YingLi Tian, Qingshan Liu, and Elizabeth Capezuti. Monitoring activity of taking medicine by incorporating RFID and video analysis. *Network Modeling Analysis in Health Informatics and Bioinformatics*, 2(2):61–70, 2013.

[JJ11] Beena Jimmy and Jimmy Jose. Patient medication adherence: Measures in daily practice. *Oman Medical Journal*, 26(3):155–159, 2011.

[JMT02] Cynthia A Jackevicius, Muhammad Mamdani, and Jack V Tu. Adherence With Statin Therapy Acute Coronary Syndromes. *Jama*, 288(4):462–467, 2002.

[Jup18] Jupyter. Jupyter, 2018.

[KC14] Tim Kelsey and Will Cavendish. Personalised health and care 2020: Using data and Technology to Transform Outcomes for Patients and Citizens. A framework for action. *National Information Board*, (November):1–66, 2014.

[KP01] Eamonn J Keogh and Michael J Pazzani. Derivative Dynamic Time Warping. *First SIAM International Conference On Data Mining (SDM'2001*, pages 1–11, 2001.

[KVFD15] S. Krishnan, J. Vietri, R. Furlan, and N. Duncan. Adherence to prophylaxis is associated with better outcomes in moderate and severe haemophilia: Results of a patient survey. *Haemophilia*, 21(1):64–70, 2015.

[LML+03] Carrie D. Llewellyn, Alexander H. Miners, Christine A. Lee, Christine Harrington, and John Weinman. The illness perceptions and treatment beliefs of individuals with severe haemophilia and their role in adherence to home treatment. *Psychology and Health*, 18(2):185–200, 2003.

[LRD+14] J. Lock, H. Raat, N. Duncan, A. Shapiro, M. Beijlevelt, M. Peters, R. Y.J. Tamminga, F. W.G. Leebeek, H. A. Moll, and M. H. Cnossen. Adherence to treatment in a Western European paediatric population with haemophilia: Reliability and validity of the VERITAS-Pro scale. *Haemophilia*, 20(5):616–623, 2014.

[Mar88] Peter Rudd Richard L. Byyny Valerie Zachary Mary E. LoVerde Wayne D. Mitchell Chris Titus Gary Marshall. Pill Count Measures of Compliance in a Drug Trial: Variability and Suitability. *American Journal of Hypertension*, 1(3):309–312, 1988.

[MDS17] MDSAS. About Haemtrack, 2017.

[Mic18] Microsoft. What you'll love about SQL Server 2017, 2018.

[MSK15] Alan Morrison, Melissa E. Stauffer, and Anna S. Kaufman. Defining medication adherence in individual patients. *Patient Preference and Adherence*, 9:893–897, 2015.

[Nau14] Felix Naumann. Data profiling revisited. *ACM SIGMOD Record*, 42(4):40–49, 2014.

[NHS17] NHS. Personalised care, 2017.

[OBK05] Lars Osterberg, Terrence Blaschke, and — C Everett

Koop. Adherence to Medication. *N Engl J Med*, 353:487–497, 2005.

[ODB17] ODBC. Python ODBC, 2017.

[Ost05] Osterberg L and Blaschke T. Adherence to medication. *The new England Journal of Medicine*, 353(5):487–497, 2005.

[Plo17] Plotly. Modern Visualization for the Data Era, 2017.

[Pul89] Feely M. Pullar T, Kumar S, Tindall H. Time to stop counting the tablets? *Clin Pharmacol Ther*, 46(2):163–168, 1989.

[RNS+14] Michael Recht, Ellis J. Neufeld, Vivek R. Sharma, Caitlyn T. Solem, A. Simon Pickard, Robert Z. Gut, and David L. Cooper. Impact of acute bleeding on daily activities of patients with congenital hemophilia with inhibitors and their caregivers and families: Observations from the dosing observational study in hemophilia (DOSE). *Value in Health*, 17(6):744–748, 2014.

[RR14] Wullianallur Raghupathi and Viju Raghupathi. Big data analytics in healthcare: promise and potential. *Health Information Science and Systems*, 2(1):3, 2014.

[SBMM+12] J. S. Stonebraker, P. H. B. Bolton-Maggs, J. Michael Soucie, I. Walker, and M. Brooker. A study of variations in the reported haemophilia B prevalence around the world: A STUDY OF REPORTED HAEMOPHILIA B PREVALENCE. *Haemophilia*, 18(3):e91–e94, 2012.

[SC78] Hiroaki Sakoe and Seibi Chiba. Dynamic Programming Algorithm Optimization for Spoken Word Recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49, 1978.

[SC07] Stan Salvador and Philip Chan. FastDTW : Toward Accurate Dynamic Time Warping in Linear Time and Space. *Intelligent Data Analysis*, 11:561–580, 2007.

[SDJC+15] Michael J. Stirratt, Jacqueline Dunbar-Jacob, Heidi M. Crane, Jane M. Simoni, Susan Czajkowski, Marisa E.

Hilliard, James E. Aikens, Christine M. Hunter, Dawn I. Velligan, Kristen Huntley, Gbenga Ogedegbe, Cynthia S. Rand, Eleanor Schron, and Wendy J. Nilsen. Self-report measures of medication adherence behavior: recommendations on optimal use. *Translational Behavioral Medicine*, 5(4):470–482, 2015.

[sF18] Python software Foundation. Pandas 0.23.4, 2018.

[Sim06] Ross J. Simpson. Challenges for Improving Medication Adherence. *JAMA*, 291(21):2614–2616., 2006.

[SOC98] Matthew D Schmill, Tim Oates, and Paul R Cohen. Learned Models for Continuous Planning. *Preliminary Papers of the Seventh International Workshop on Artificial Intelligence and Statistics*, (April):278–282, 1998.

[SP97] John F. Steiner and Allan V. Prochazka. The assessment of refill compliance using pharmacy records: Methods, validity, and applications. *Journal of Clinical Epidemiology*, 50(1):105–116, 1997.

[Tam06] John M.Hunt Tamara L.Hayes. An Electronic pillbox for continuous monitoring of medication adherence. *Proceedings of the International Conference of the IEEE Engineering in Medicine and Biology Society*, 23-26(30):6400–6403, 2006.

[TNL+17] Wanzhu Tu, Winstone M. Nyandiko, Hai Liu, James E. Slaven, Michael L. Scanlon, Samuel O. Ayaya, and Rachel C. Vreeman. Pharmacokinetics-based adherence measures for antiretroviral therapy in HIV-infected Kenyan children:. *Journal of the International AIDS Society*, 20(1):1–6, 2017.

[VHVD02] E Vermeire, H Hearnshaw, P Van Royen, and J Denekens. Patient adherence to treatment: three decades of research. A comprehensive review. *J Clin Pharm Ther*, 26, 2002.

[Was18] Michael Waskom. Seaborn, 2018.

[WFH18] WFH. World Federation of Haemophilia: About bleeding disorders, 2018.

[WHO03] WHO. Adherence to long term therapies. 2003.

[Wik18] Wikipedia. Minkowski distance, 2018.

[ZN07] Diana A. Shellmer Zelikovsky and Nataliya. The challenges of using medication event monitoring technology with pediatric transplant patients. *NIH*, 11(4):442–428, 2007.

Appendix A

Source code

A.0.1 Source code version control using Git

The code below shows the steps taken in source code management using Git as explained in section 3.3.2

```
#adding changes to a file
git add pkr.py

#commiting changes to the file
git commit -m "Added graphs for euclidean distance"

#saving changes to the remote repository
git push -u origin master
```

A.0.2 Function to generate time series

The code below shows the functions that generates time series that are used to get results in Chapter 4

```
import pyodbc
import pandas as pd
import datetime
import math
import matplotlib.pyplot as plt
import matplotlib.dates as md
```

```
import os
import time
import numpy as np
np.set_printoptions(threshold=np.nan)

# We get Treatment and Bleed data into Python Structures.
# We return a Treatment Series and a Bleeds dictionary
def fillTS(cursor, dateFrom, dateTo):
    rangeFrom = datetime.datetime.strptime(dateFrom[1:-1], '%b %d, %Y'
    rangeTo = datetime.datetime.strptime(dateTo[1:-1], '%b %d, %Y')
    rangeTS=pd.date_range(rangeFrom, rangeTo, freq="5min")
    treatments = pd.Series(data=0, index=rangeTS)
    bleeds = {}
    treatmentReasons = {}
    while 1:
        row = cursor.fetchone()
        if not row:break
        dt=datetime.datetime.combine(row.TreatedDate,
datetime.datetime.strptime(row.TreatedTime, '%H:%M %p').time())
        r=divmod(dt.minute,5)[1]
        if (r != 0 ):
            # we have a time series of 5min frequency and we
#enter this branch if the time of the treatment
            # is not divisible with 5. E.g. Treatment at time
#13:06 is approximated with 13:10.
            if (r < 2.5):
                dt= dt - datetime.timedelta(minutes=r)
            else:
                dt = dt + datetime.timedelta(minutes=(5-r))
        #print(dt)
        treatments[dt] = row.TotalUnits
        if not row.Location is None:
            bleeds[dt] = [row.Location, row.BleedCause, row.BleedSever
row.TimeAfterBleed]
        if not row.Reason is None:
```

```
            treatmentReasons[dt] = row.Reason

    return treatments, bleeds, treatmentReasons

#We transform the Treatment time series into PK series knowing the half
#life and the volume of distribution
def generatePK(ts, halfLife):
    tenhalflives = (10* 60* halfLife)
    ke =  math.log(2)/(halfLife * 60)
#constant of excretion computed using halflife
    for t in ts[ts>0].index:
        for i in range(5,tenhalflives,5):
           if( (t + datetime.timedelta(minutes=i)) <= ts.index[-1] ):
              if (ts[t +  datetime.timedelta(minutes=i) ]==0):
                  ts[t +  datetime.timedelta(minutes=i) ] = ts[t] *
  math.pow(2.718,-ke*i)
              else:
                   ts[t +  datetime.timedelta(minutes=i) ] =
   ts[t +  datetime.timedelta(minutes=i) ]+
   ts[(t +  datetime.timedelta(minutes=i)) -
   datetime.timedelta(minutes=5) ]
                  break;
    return ts
```