DATA MINING FOR PRECISION MEDICINE IN CLINICAL DECISION SUPPORT SYSTEMS

A Thesis Submitted
By

Wee Pheng Goh

For the award of
Doctor of Philosophy

2020

Dedication

Dedicated in loving memory of my late mother-in-law Lee Kay Chee
(13 Jan 1933 - 10 Aug 2020); her friendliness and faith will always be remembered.

Precious in the sight of the Lord is the death of his saints. Psalm 116:15

Abstract

With the addition of new drugs in the market each year, the number of drugs in drug databases is constantly expanding, posing a problem when prescribing medications for patients, especially elderly patients with multiple chronic diseases who often take a large variety of medications. Besides the issue of polypharmacy, the need to handle the rapid increase in the volume and variety of drugs and the associated information exert further pressure on the healthcare professional to make the right decision at point-of-care. Hence, a robust decision support system will enable users of such systems to make decisions on drug prescription quickly and accurately.

Although there are many systems which predict drug interactions, they are not customised to the medical profile of the patient. The work in this study considers the drugs that the patient is taking and the drugs that the patient is allergic to before deciding if a specific drug is safe to be prescribed. To exploit the vast amount of biomedical corpus available, the system uses data mining methods to evaluate the likelihood of a drug interaction of a drug pair based on the textual description that describes the drug pair. These methods lie within the prediction layer of the conceptual three-layer framework proposed in the thesis. This framework enables drug information to be used in a decision support system which associates with the medical profile of the patient. The other two layers are the knowledge layer and the presentation layer. The knowledge layer comprises information on drug properties from drug databases such as DrugBank. The presentation layer presents the results via a user-friendly interface. This layer also obtains information from the user the drug to be prescribed and the medical profile of patients. Models used in these data mining methods include the network approach and the word embedding approach.

Empirical experiments with these models support the hypothesis that drug interactions are associated with similarities derived from their feature vectors, resulting in the deployment of a decision support system for use in dental clinics. A survey conducted on dentists found positive response in the use of such a system in helping them in drug prescription which result in a better treatment outcome. They found the system useful and easy to use. The novel approach of using information on drug interaction through data mining for use in a personalised decision support system has provided a platform for further research on optimising of drug prescription, transforming the clinical workflow at point-of-care within the healthcare domain.

基于数据挖掘的精密医学决策支持系统

摘要

随着每年市场上新药的不断增加，药品数据库中不断扩大的的药品数量增大了医生给患者开药时带来的难度，尤其是给经常服用多种药物的患有多种慢性疾病的老年患者。除了多种药物的问题外，处理药物数量和种类的快速增长以及相关信息的需求，进一步增大了医疗专业人员在开药时做出正确决定的难度。因此，一个有效的决策支持系统能够帮助医疗专业人员在开药时做出快速又准确地决策。

尽管有许多系统可以预测药物之间的相互作用，但都不是依据患者的个人药物资料进行设计的。本文研究的工作要充分的考虑患者正在服用的药物和患者对药物过敏的情况下，从而决定否可以安全地开一种特定的药物。为了充分利用现有的大量生物医学语料，该系统依据描述药物对的文本描述，采用数据挖掘方法的评估药物对之间的相互作用。这些方法属于本文提出的概念三层框架的预测层。该框架使用的药物信息能够用于与患者的医疗简介相关的决策支持系统中。本框架另外两层是知识层和表示层，其中知识层包括来自药物数据库（如DrugBank）的药物属性信息，表示层通过用户友好的界面显示结果。表示层可以从用户那里获得要开的药方和病人的医疗资料。采用的数据挖掘方法有网络方法和单词嵌入方法。

通过实验验证发现，药物的相互作用与从其特征向量有密切关系，并已部署牙科诊所的一个决策支持系统中。对牙医的调查结果显明，该系统可以帮助他们更有效地开药，从而获得更好的治疗效果。他们发现这个系统既实用又易于使用。因此，本文提出的用于个性化决策支持系统挖掘药物相互作用信息的新颖方法，为进一步研究优化药物处方提供了平台、更改了医疗领域内的临床工作流程。

Certification of Thesis

This Thesis is entirely the work of Wee Pheng Goh except where otherwise acknowledged. The work is original and has not previously been submitted for any other award, except where acknowledged.

Principal Supervisor:
Xiaohui Tao

Associate Supervisors:
Ji Zhang
Jianming Yong

Student and supervisors signatures of endorsement are held at the University.

Acknowledgments

This research which results in the contribution of new approaches in effective drug prescription for healthcare establishments would not be possible without the support of the staff at the University of Southern Queensland, colleagues at Glory Dental Surgery and family members.

It has been a privilege to work closely with my knowledgeable supervisors Associate Professor Xiaohui Tao, Assoiate Professor Ji Zhang and Professor Jianming Yong who contributed their expertise, time and effort especially in the initial process of starting the research and getting my candidature confirmed. I am especially grateful to my principal supervisor Xiaohui Tao for his academic support during each phase of the research, especially the uphill task of changing my “development” mindset to a “research” mindset. Working with him has equipped me with the skills of not just implementing software programming codes but also investigating and inventing approaches to technological ideas and concepts. The research seminars that he organised with the team of researchers also provided a platform for the exchange of ideas and mutual encouragement throughout the research journey. I am deeply grateful for all of them. Appreciation also goes to Dr Elizabeth Goh and Dr Barbara Harmes who provided me with grammatical tools and support in writing styles. In addition, special thanks goes to Nurul Binte Norhaidi and Ying Xin at Glory Dental Surgery for the logistic support with the research survey and assistance with the bibliographic references.

Sincere gratitude also goes to Dr Winston Lee Piak Nam for providing assistance and motivation in the research enrolment. This research is supported by Glory Dental Surgery Pte Ltd and I would like to thank the dentists at the clinic who provided insight and expertise regarding drug interactions - in particular, Dr Voo Hwee Choo, Dr Elizabeth Goh, Dr Chris Do, Dr Ong Jinn, Dr Oh Xue Ling, Dr Noah Teo and Dr Low Peng Koon.

Table of Contents

Abstract iii

Certification of Thesis vi

Acknowledgements vii

List of Figures xii

List of Tables xiv

Notation, Terminology, and Abbreviation xvi

1 Introduction 1

1.1 Background 1

1.2 Hypotheses and Research Questions 4

1.3 Research Problem 4

1.3.1 Limitations and Assumptions 6

1.4 Research Methodology 7

1.5 List of Publications 9

1.5.1 Refereed Journal Articles 9

1.5.2 Refereed International Conferences 10

1.6 Awards 11

2 Literature Review 12

2.1 Clinical Decision Support Systems 12

2.1.1 Approaches in Understanding Clinical Decision Support Systems 14

Technologies behind Clinical Decision Support Systems 15

Classification of Clinical Decision Support Systems 17

Benefits of Decision Support Systems 19

Summary 20

2.1.2 Challenges to Adoption of Decision Support Systems 21
Lack of Perceived Usefulness 21
Complex Sociocultural and Economic Factors 22
Difficulties in Interpretability 23
Summary 25
2.1.3 How Decision Support Systems Overcome Challenges 26
Efficient Design of Knowledge Base 26
Ontology 28
Human Computer Interface 29
Summary 30
2.1.4 Trends for Clinical Decision Support Systems 31
Big Data 31
Personalised Systems 33
Standards and Privacy Issues 34
2.1.5 Support for Drug Prescription 35
2.2 Drug Interactions 35
2.3 Detection of Drug Properties 37
2.3.1 Graph-based Similarity 37
2.3.2 Word Embedding 38
2.4 Summary 41

3 Knowledge Base and Recognition of Drug Features 43
3.1 Three-layer Conceptual Framework 44
3.1.1 Knowledge Layer 45
3.1.2 Prediction Layer 46
3.1.3 Presentation Layer 49
3.2 Constructing the Knowledge Base 50
3.3 Mining Data from Text Corpus 54
3.4 Recognition of Drugs through Feature Extraction 58
3.4.1 Bag of Words 59
3.4.2 Term Frequency Inverse Document Frequency (tf*idf) 61
3.4.3 Co-occurrence 62
3.4.4 Word Embedding 63
3.4.5 Common Paths in Adverse Network 64
3.5 Summary 64

4 Discovering Drug Interactions through Data Mining 65

4.1 Cosine similarity for Vector Space Models 66

4.2 Discovering Drug Interactions Through Word Frequencies 67

4.3 Discovering Drug Interactions Through Side-effects 70

4.4 Discovering Drug Interactions from Common Paths of Adverse Network 71

4.5 Discovering Drug Interactions Through Word Embeddings 73

4.6 Summary 75

5 Empirical Experiments 76

5.1 Experimental Design 76

5.1.1 Implementation of Statistical Model 78

5.1.2 Implementation of Side-effect Model 79

5.1.3 Implementation of Adverse Network Model 79

5.1.4 Implementation of Word Embedding Model 81

5.2 Baseline Models 82

5.3 Data Preparation 85

5.3.1 Preparing Data for Statistical Model 85

5.3.2 Preparing Data for Side-effect Model 88

5.3.3 Preparing Data for Adverse Network Model 89

5.3.4 Preparing Data for Word Embedding Model 89

5.3.5 TAC2018 Dataset 90

5.3.6 Summary 91

5.4 Performance Measuring Schemes 91

5.5 Summary 93

6 Results and Discussions 95

6.1 Experimental Results 95

6.1.1 Experiment Results from Statistical Model 96

6.1.2 Experiment Results from Side-effect Model 100

6.1.3 Experiment Results from Adverse Network Model 103

6.1.4 Experiment Results from Word Embedding Model 105

Experiment using TAC2018 dataset 106

6.2 Discussions 108

6.2.1 Comparing Adverse Network Model with Word Embedding model 108

6.2.2 Comparing Statistical Model with Word Embedding model 110

6.2.3 Comparing with Baseline Models 111

6.3 Sensitivity Study 113
6.4 Summary 114

7 Clinical Deployment 116
7.1 Design of Prototype 116
7.2 Common Drugs in Dental Clinics 120
7.2.1 Antimicrobials 121
7.2.2 Analgesics 123
7.2.3 Corticosteroids 125
7.2.4 Anxiolytics 126
7.3 Point-of-Care Scenario 126
7.3.1 Patient Registration 127
7.3.2 Drug Prescription 128
7.4 Dentists' Adoption 128
7.5 Prototype Design - Mobile Learning 133
7.6 Mobile Learning Scenario 135
7.7 Summary 138

8 Conclusions 139
8.1 Summary of the Research 139
8.2 Contributions and Significance 141
8.3 Future Work 142
8.4 Overall Conclusion 143

References 144

Appendix A Survey Information 162

Appendix B Survey Questions 164

List of Figures

1.1 Scientific methodology 8

2.1 Design of a typical decision support system 13

2.2 Ways of understanding decision support systems 20

2.3 Structurational model of technology [98] 23

2.4 Role of HCI in an intelligent information system 24

2.5 CDSS model [30] 28

2.6 Dental caries detection algorithm framework [111] 30

2.7 Big data heat map [121] 32

2.8 Objects and relations in an information network [125] 38

2.9 Rating matrix between users $u_{1,2,3}$ and movies $m_{1,2}$ [125] 39

2.10 Skip-gram model of Word2Vec 40

3.1 Three-layer framework 44

3.2 Grouping of similar words 47

3.3 Grouping of similar drugs 48

3.4 User interface 51

3.5 Typical knowledge discovery process 52

3.6 Subset of Drug Taxonomy 55

3.7 Transforming data into structured text 57

4.1 Cosine similarity of documents 66

4.2 Graph of interactive drugs with drug d_1 71

5.1 Experimental design to for measuring performance 77

5.2 Experimental design of side-effect model 80

5.3 Drug interactions 83

5.4 Drug-Entity model [155] 83

5.5 Sample extract of properties of *warfarin* 86

5.6 Organisation of the properties of dataset 87
5.7 Sample XML extract of dataset from TAC2018 92

6.1 Performance comparison against different drug attributes 99
6.2 Side-effect model: precision-recall curve 101
6.3 Side-effect model: ROC curve 102
6.4 Adverse network model: ROC curve 104
6.5 Adverse network model: F-score 105
6.6 Word embedding model with TAC2018 dataset: ROC Curve 108
6.7 AUC for adverse network and word embedding models 110
6.8 Performance comparison of statistical and word embedding models 111
6.9 Effect of Word2Vec parameters on performance 114

7.1 Mapping of prototype to the three-layer framework 117
7.2 Registering patient into the system 127
7.3 Prescribing drug based on patient's profile 129
7.4 Candidate list of drugs for *ibuprofen* 130
7.5 Candidate list of drugs for *penciclovoir* 131
7.6 Summary of response 133
7.7 System flow 134
7.8 User mobile interface 135
7.9 Flowchart of learning vignette 136

List of Tables

2.1 Recent decision support systems 15

2.2 Major functionalities of CDSS [28] 15

2.3 Classification of a CDSS 17

2.4 Challenges to adopting decision support systems 26

2.5 Current and expected features of CDSS 27

2.6 Overcoming challenges 31

2.7 Training sample at window size of three 39

3.1 Methods used for drug prediction within prediction layer 46

3.2 Features of conceptual framework 49

3.3 Database of drug interaction 54

3.4 Structure of side-effect table 57

3.5 Sample data for table of side-effects of Ibuprofen 58

3.6 Structure of adverse interaction table 58

3.7 Sample data for table of adverse interactions with ibuprofen 58

3.8 Sample documents 59

3.9 Matrix of count vector 60

3.10 Co-occurrence matrix 63

5.1 Baseline models 84

5.2 Structure of dataset in experimental evaluation 87

5.3 Sample side-effects of *warfarin* 88

5.4 Levels of side-effects 88

5.5 Sample drugs that interact with *warfarin* 89

5.6 Sample tokens for *warfarin* 89

5.7 Size of dataset used for building Word2Vec model 90

5.8 Sample word vectors from Word2Vec 90

5.9 Size of data for each model 91

6.1 Sample results from positive sample 97
6.2 Sample results from negative sample 98
6.3 Results based on different attributes of the drug pairs 99
6.4 F-score at different threshold values of θ 100
6.5 Weights to normalise feature vectors 100
6.6 Side-effect model: results from positive drug pairs 101
6.7 Side-effect model: results from negative drug pairs 102
6.8 Adverse network model: sample results from similar drug pairs 103
6.9 Adverse network model: sample results from dissimilar drug pairs 104
6.10 Word embedding model: sample results from similar drug pairs 105
6.11 Word embedding model: sample results from dissimilar drug pairs 106
6.12 Performance of word embedding model at layer size=16 106
6.13 Sample results from similar drug pairs 107
6.14 Sample results from dissimilar drug pairs 107
6.15 F-score distribution 109
6.16 Top ten DDI's from [114] 112
6.17 Effect of proximity and nodes properties on performance of the adverse network model 113
6.18 Influence on performance by training parameters 113

7.1 Common drugs prescribed in a dental clinic 118
7.2 Common antimicrobial drugs 121
7.3 Antimicrobial creed 122
7.4 Adverse effect of opioids [7, 102] 124
7.5 Common analgesic drugs 124
7.6 Properties of topical corticosteriods ointment used on the oral mucosa [130] . . 125
7.7 Common anxiolytics 126
7.8 Responses to CDSS Survey 132

Notation, Terminology, and Abbreviation

Notation

$\{t_i^v\}$	Set of a bag of words of drug d_i within the attribute “overview”.$\{t_i^p\}$ for “professional” and $\{t_i^s\}$ for “side-effects”
$\rightarrow$	An edge in a graph, eg $d_1 \rightarrow d_2$ shows the path between drug d_1 and drug d_2
$\mathbb{D}$	Domain set of drugs
F_1	F_1 score, the weighted average of precision and recall
$\overrightarrow{f_{d_i}}$	Feature vector of drug d_i
$\mathcal{J}$	Objective function, used in Word2Vec to predict probability of a target word within a set of context words
$N_p^+(d)$	The pth out-going node directed away from vertex node d
$\mathcal{O}$	An ontology
$\mathcal{P}$	Set of patients
$\mathbb{E}$	Domain set of side-effects
$Sim(d_i, d_j)$	Similarity ratio between drug d_i and drug d_j
$tf * idf$	Term frequency Inverse Document Frequency
$\mathcal{T}$	Drug taxonomy

Terminology

data mining — A process of discovering interesting patterns and knowledge from large amounts of data

drug interactions — Refers to the action (usually unwanted) of a drug when administrated concurrently with another drug

query — The data structure given by a user to information gathering systems for the expression of an information need

side-effects — Side effects are unwanted symptoms caused by medical treatment. They are also known as *adverse effects* or *adverse reactions*

skip gram — A method of mapping words into vectors to feed into a neural network in order to predict similar words from a target word

word embedding — Vector representation of a word

Abbreviations

AI	Artificial Intelligence
ANN	Artificial Neural Network
AUC	Area Under Curve
CDSS	Clinical Decision Support System
DSS	Decision Support System
DAG	Directed Acyclic Graph
DDI	Drug-drug interaction
FP	False Positive
FN	False Negative
FDA	Food and Drug Administration
HIN	Heterogeneous Information Networks
IS	Information Systems
P	Precision
R	Recall

ROC	Receiver Operating Characteristics
SVM	Support Vector Machine
TN	True Negative
TP	True Positive

Chapter 1

Introduction

Avoiding prescription error is crucial for the healthcare professional. The growing amount of data on drugs coupled with the prevalence of polypharmacy raise the importance of having suitable tools to harness such data to enhance the treatment outcome. With this motivation in mind, this chapter introduces the research problem and the methodology adopted to evaluate the hypothesis related to the use of drug information to help the health professional in their drug prescription. Details and awards related to the publication and presentation of the research are also included in the chapter.

1.1 Background

Dentists are trained to deal quickly and accurately with the diagnosis of oral health disease with the aid of data gathered from images, observations and interviews with patients to ascertain their medical and social history.

These clinical tasks have become a challenge due to the phenomenal growth in information technology accompanied by the prodigious amounts of data being produced. According to IBM, 2.5 quintillion bytes of data (2,500,000,000 gigabytes) are generated every day[1]. To put

[1] https://www.storagenewsletter.com/2011/10/21/ibm-cmo-study/

this into perspective, using the 250G capacity of a blu-ray disk as an example, 10 million pieces will be required to store such data, and if these disks were stacked, they would reach 1.3 times the height of Mount Everest. These unprecedented growth in data are radically transforming personalised medicine and changing the approach in biomedical research [20].

Higher volume, faster speed and greater variety of data and the way these data and their presentation have led to a paradoxical twist in the way data is managed and handled. Though the digitalisation of information offers better productivity in terms of searching and storing, it has created a burden on the user because of the need to ensure that relevant results are being applied in the right context, as well as issues of security, privacy and accuracy.

It can be seen that the way these data are presented has altered the clinical workflow and cultural climate of the entire healthcare industry. Many healthcare professionals relying on pen and paper already face the risk of going out of business as more and more data are digitalised. The skill to click, copy and paste is becoming more crucial than the ability to hold a pen, flip and clip papers. Analogue X-ray films are replaced by digital X-rays. Treatment notes written on cards are replaced by digital notes. Lead used in pencil is progressively replaced by silicon used in electronic devices for recording notes. Wooden furniture for storage is being replaced by digital media.

Technoscapes and *mediascapes* which are part of the five "scapes" (the other three are *ethnoscapes*, *financescapes* and *mediascapes*) described by Appudurai [3], offers us a lens to better appreciate and understand the phenomenal change which has lead to such an impact on the cultural and clinical flow for the healthcare professional. In *mediascapes*, images are distributed through digital and cloud technologies in increasingly complex ways. What this means is that the user throughout the world will need to experience and handle data in a complicated and interconnected repertoire of films, disks as well as local and remote storage. *Technoscapes* explain how the nature of work in the workplace has been brought about by technology like big data and machleine learning. The effect of technology combined with media can be phenomenal. Sales volume at Alibaba.com on Single's Day (11 Nov) has seen an increase of 194% over the past 5 years, rising from 91.2 billion yuan in 2014 to 268.4 billion yuan in 2019[2].

Technology and the media will surely change the way transactions are performed, not just in e-commerce but also in other areas including healthcare. Guided by relevant clinical questions, powerful Artificial Intelligence (AI) techniques can unlock clinically relevant information hidden

[2] https://cnb.double11/statistics

in the massive amount of data, which in turn can assist clinical decision making. Hence healthcare professionals cannot escape from the use of technology and should instead embrace and adopt technology as a tool to help them diagnose, record and retrieve information for their patients. Such decision support tools will help reduce the cognitive load in daily clinical practice, especially the adverse interactions of the ever increasing number of drugs being introduced by the pharmaceutical industry. As more and more patients are taking multiple drugs, it is crucial for the healthcare practitioner to be able to obtain information on drug prescription at point-of-care.

A common and crucial question in medicine is to inquire about the effect of a drug. Such a feature, if incorporated within a CDSS will enable the medical professional to quickly decide if a drug is safe for prescription. A wrong prescription can lead to many undesirable side-effects and may also be fatal. A study by Dechanont *et al.* discovered that hospital admission could have been avoided if prescription is properly administered [24]. Among hospitalised patients, nearly two-thirds are exposed to DDI while 40%-70% are discharged with potential DDI [48].

Hence it is important to make use of CDSS to reduce medication errors [135]. In fact, the ability to deliver personalised decision support is critical to the clinical success of precision medicine [60]. Therefore, a CDSS with information on drug interaction will enhance treatment efficiency of the practitioner and reduce prescription errors [88].

Although there are many such systems, there is a lacking in a personalised system where information on drug interaction is integrated with the drug profile of the patient. Imagine all the information of the patient is already stored digitally, yet the user has to enter again relevant information of the patient like age and current drug that patient is taking, it will be a waste of time and also subject to risk of errors related to data entry. Even with many decision support systems that may provide alert on adverse drug events, there is still a lack of evidence on the relevance of CDSS alerts to detect actual adverse drug events [47]. Many systems also do not cover drug related problems on an individual patient level.

With this in mind a personalised clinical decision support system for drug prescription which enables users to know if the drug to be prescribed is safe for the patient would help healthcare professionals function more efficiently a point-of-care. Although many studies have been done to ascertain drug interactions [10, 145, 156], this study uses such information as an evidence-based approach to ensure that drugs prescribed by healthcare professionals are safe based on the individual patient's profile.

1.2 Hypotheses and Research Questions

Given the crucial need for correct prescription and the way information is stored and retrieved, there is a pressing need to be able to predict if a drug pair is in an adverse relationship. Such information, with consideration of the patient's medical profile, will help the healthcare professional to prevent prescription error.

With the plethora of drug information within the bio-medical domain, the appropriate approach is to employ such information to discover the relationship within a drug pair. Intuitively, if both drugs are similar, their textual description should also be similar. This leads to the hypothesis that similar drug pairs have a higher similarity ratio than dissimilar pairs. Specifically, the thesis aims to answer these research questions:

1. What properties should be extracted from the bio-medical corpus for finding the similarity of a drug pair?
2. How can such information be extracted?
3. How can the similarity ratio of a drug pair be predict based on the features extracted?
4. How can the similarity ratio be used by the healthcare professional?

The scientific methodology is adopted (See Section 1.4) to validate the hypotheses and to answer the research questions. Various models will be proposed and tested to enable the user to make an informed choice on the drugs to be prescribed.

1.3 Research Problem

Assuming the user has knowledge of the patient's medical conditions, the medications that the patient is currently taking and the patient's drug allergies, the goal of this research is to ensure the drug that the healthcare professional is prescribing does not belong to the group of drugs that the patient is allergic to, with the system monitoring and updating the set of drug allergies within the patient profile defined in Section 3.1.3. In order to ensure that drugs can be prescribed safely, the task is then to find the relationship between the drug to be prescribed and the drug that the patient is currently taking. If their relationship is friendly, then it is safe for the user to prescribe it to the patient. Otherwise an alternative list of drugs will be needed for the user to make an informed choice.

There are many existing approaches to check if a drug pair has an adverse relationship. Since the aim is to allow clinical translation of theoretical findings, the medical profile of the patient is taken into consideration when deciding on the similarity of a drug pair. Data mining methods are used to retrieve relevant information to predict the similarity ratio of a drug pair in order to exploit the huge number of biomedical databases available.

Thus, the research problem is to evaluate the likelihood of the drug interaction of a drug pair based on semantic relationship revealed from the textual description of the drug pair. The 3-tuple definition of the patient (refer to Section 3.1.3) ensures the system captures and considers the patient's individual profile by taking into account the drugs the patient is taking as well as the drug the patient is allergic to. By achieving the research goal, the research outcome of proposed thesis will provide decision support to health professionals in practice in terms of drug prescription at point-of-care. Specifically, the research outcome will help users of such a decision support system reduce potential risk of issuing allergic drugs to a patient and as a result, improve the quality of treatment in clinics. The development of a decision support system which helps healthcare professionals reduce the chance of prescribing inappropriate drugs to a patient, thus improving the quality of treatment in clinics. In addition, the findings in this research provides the potential to contribute to the wider medical domain.

The efficient approach in the design of the clinical decision support system (CDSS) with consideration of the medical profile of the patients result in the following significant contributions:

- advancement in the design of clinical decision support systems by using similarity ratio of a drug-pair;
- attributes like adverse interactions and side effect of a drug can be used to construct feature vectors for computing similarity ratios;
- by hierarchically representing the drug-pairs within the context of a CDSS, paths linking the common drugs within the set of interacting drugs can be used to arrive at a similarity ratio;
- results support the hypothesis that similar drug-pairs have a higher similarity ratio compared to that of dissimilar pairs;
- provide a platform for further research on data mining and machine learning methods

within the medical domain which will transform the clinical work flow of the health-care industry.

1.3.1 Limitations and Assumptions

In this section, the limitations and assumptions are specified, aiming to define the scope of the research in the thesis.

While formulating the hypothesis, it is assumed that description of drugs in bio-medical text reflects the behavior and effect of the drugs. Since the decision support system requires the medical profile of the patient, it is assumed that in actual deployment of the system, information regarding drug allergy of the patient and drugs that the patient is currently taking is accurate and up-to-date. Due to the complexity of the prototype to consider fully the patient's profile, only the first two tuples of the patient's profile is taken into consideration. Moreover, medical condition is not as common and crucial compared to the drugs that the patient is taking and drugs that the patient has allergy. Hence only the first two tuples are taken into consideration. These assumptions were made as per the advices obtained while consulting with the panel of dentists at Glory Dental Surgery, Singapore. Although this project involves only dental health professionals, the same approach can be used within the medical domain as both domains share the same objective of safe prescription of drugs for their patients at point-of-care.

In the design of the model, it is assumed that it does not claim to treat the patient's medical condition - it only attempts to check for possible side-effects of the drug to be prescribed with the condition. For example, if the patient has a cardiovascular condition, the drug to be prescribed, although not considering its healing effect, should consider the adverse effects it may have on the patient as it is crucial that certain drugs be avoided for certain medical condition of the patient.

The model also assumes that the patient does not have a cross-allergy to the drugs they are currently taking. This is a safe and valid assumption since the fact that patient can attend for dental treatment shows that the patient can function normally and is not impaired by the adverse effects of the drugs. Furthermore, the drugs that the patient is currently taking is assumed to be prescribed by a medical doctor who should already have considered the patient's medical condition and known drug allergies.

Since the deployment based on the result of the study is a decision support system, the user has the liberty to overwrite the system's suggestions as the function of the system is limited to assisting the user in checking for possible adverse reactions between the drug to be prescribed and the drug that patient is currently taking.

1.4 Research Methodology

Due to the nature of this research, which involves the vigorous testing of a hypothesis relating to the way drugs interact with each other in terms of effect on the patients, the scientific methodology is used in the study. The scientific methodology is an empirical method involving making conjectures and hypothesis, deciding the predictions for testing and carrying out experiments to determine how accurate are the hypotheses (Figure 1.1)

The conjecture attempts to explain a behaviour while formulating questions. The drugs prescribed by the healthcare professional can adversely affect patients in terms of allergic reactions, as well as adverse reactions with the drugs that the patient is currently taking. Hence, the thesis aims to assist the healthcare professional in deciding if a drug is safe for prescription by investigating if that drug is similar to or adversely interacts with each drug relevant to the patient's medical profile, with the two drugs under investigation being termed a drug pair. Instead of the traditional way of checking drug interactions with their chemical properties, the thesis attempts to use data mining methods to help answer the question of how similar the drugs of a drug pair are. If the drug that the user of the system is going to prescribe is similar to the drugs that the patient is allergic to, then it will be helpful for the user to be aware of that information before making the prescription. While formulating the questions in search of a solution for the user to make a decision on the suitability of a drug for prescription, the methodology enters from the conjecture stage into the next stage where the hypothesis is formulated.

Consequently, this leads to the prediction stage where we set a threshold to predict if a group of drug pairs are indeed similar. The more unlikely that a prediction would be correct simply by coincidence, the more convincing the prediction will be. At the Hypothesis and Prediction stage, current theories, concepts obtained from the Literature Review stage are taken into consideration. This will ensure the subsequent stages of model and experiment design can effectively test if the prediction is accurate enough to support the hypothesis.

The Testing stage gathers evidence by conducting experiments to test the accuracy of the predictions, quantified by performance measures. The hypothesis may have to be amended if the results do not support the hypothesis. Otherwise, the methodology progresses to the Evaluation and Analysis stage to determine what the experimental results show, and what the underlying reasons are for any outliers. Any possible extension of the experiment or alternative ways of conducting the experiment can be highlighted in the Conclusion stage of the methodology. Very often, the findings are communicated by presentation in conferences and reporting in relevant scientific journals. Accordingly, the findings of this research have been presented in scientific conferences and published in many journals. They have also been deployed in a clinical environment as a decision support system for drug prescription by the healthcare practitioner.

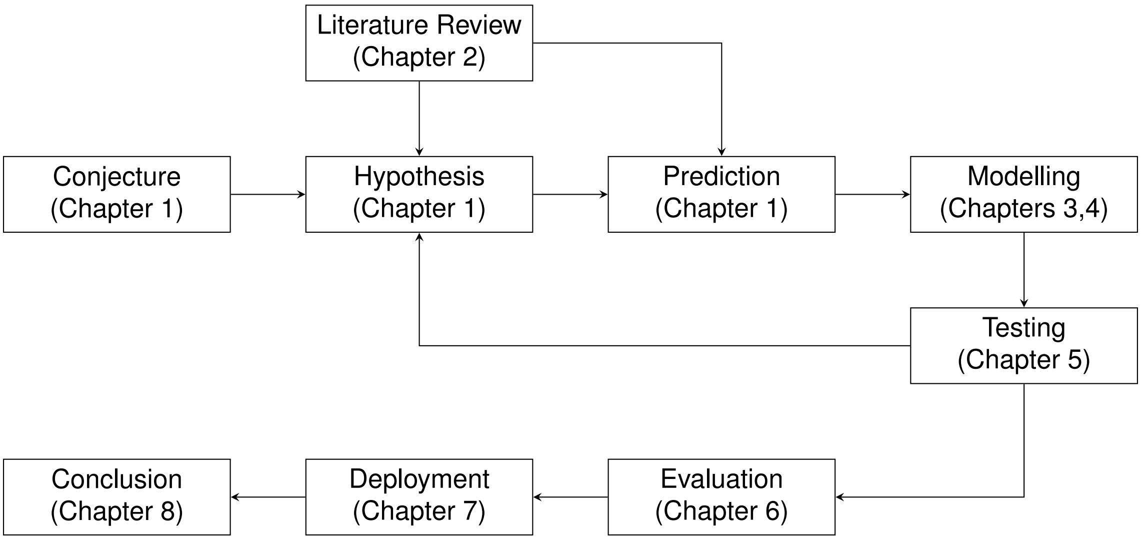


Figure 1.1: Scientific methodology

Although Figure 1.1 shows the different stages of the scientific methodology as a sequence of steps, these steps may not always be in that fixed sequence. As noted in the account of William Whewell's (1794-1866) epistemology of science, elements of "invention, sagacity and genius" are needed at each step of discovery of scientific knowledge [56].

The thesis being part of the reporting and communication process within the scientific methodology, notwithstanding, comprehensively presents the entire process including the way the hypothesis and the models are formulated as well as how the experiment is designed, conducted and analysed. The thesis is organised as follows:

Chapter 1 introduces the background of the research and how the hypothesis is formulated. The research problem is formulated and defined along with assumptions and limitations of the study.

Chapter 2 is the literature review which surveys existing work and examines their limitations and how the gaps in the area of decision support systems are identified and addressed in the study. Various technical approaches in predicting the similarity of a drug pair are also described in this chapter.

Chapter 3 explains how relevant information is extracted from text corpus to be used in the decision support system. The conceptual framework, which consists of the knowledge layer, prediction layer and presentation layer, is explained.

Chapter 4 continues to describe in more detail the way drug similarity is obtained for the different models adopted within the prediction layer.

Chapter 5 gives details in the experimental design of each model as well as how data are collected and measured.

Chapter 6 discusses and analyses the results.

Chapter 7 illustrates the relevance of the research findings in terms of deployment within a clinical setting. The concept of a mobile learning application for drug prescription is also demonstrated in this chapter.

Chapter 8 concludes the thesis by discussing the contributions and possible future work.

1.5 List of Publications

Key results and contributions in this research together with the conceptual framework have been accepted by reviewers, presented in major conferences and published in peer-reviewed journals. Some of these publications include:

1.5.1 Refereed Journal Articles

[1] W. P. Goh, X. Tao, J. Zhang, and J. Yong. Decision support systems for adoption in dental clinics: A survey. *Knowledge-Based Systems*, 104:195-206, 2016, doi: 10.1016/j.knosys.2016.04.022

[2] W. P. Goh, X. Tao, J. Zhang, J. Yong, W. Zhang, and H. Xie. Drug prescription support in dental clinics through drug corpus mining. *International Journal of Data Science and Analytics*, 6(4):341-349, 2018, doi: 10.1007/s41060-018-0149-3

[3] X. Tao, W. Goh, J. Zhang, E. Goh, and X. Oh. Mobile-based Learning of Drug Prescription for Medical Education using Artificial Intelligence Techniques. *International Journal on Mobile Learning and Organisation*, 2020 (in press). https://rebrand.ly/x0o4and

[4] X. Tao, T. Pham, J. Zhang, J. Yong, W. Goh, and N. Zhong. Mining Health Knowledge Graph for Health Risk Prediction. *World Wide Web Journal*, 23(4):2341-2362, 2020, doi: 10.1007/s11280-020-00810-1

[5] W. Goh, X. Tao, J. Zhang, and J. Yong. Feature-based Learning Drug Prescription System for Dental Clinics. *Neural Processing Letters*. Springer International Publishing, 2020, doi: 10.1007/s11063-020-10296-7

[6] W. Goh. Data Mining for Personalised Clinical Decision Support Systems. *The PhD Thesis Abstract Track, IEEE Intelligent Informatics Bulletin*, 2020, (in press). https://rebrand.ly/47agmj7

1.5.2 Refereed International Conferences

[7] W. Goh, X. Tao, J. Zhang, and J. Yong. A study of drug interaction for personalised decision support in dental clinics. In *2015 IEEE/WIC/ACM Workshop Proceedings on International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)* , volume 1, pages 88-91, 2015, doi: 10.1109/WI-IAT.2015.28

[8] W. Goh, X. Tao, J. Zhang, and J. Yong. Mining drug properties for decision support in dental clinics. In J. Kim, K. Shim, and L. Cao, editors, *PAKDD 2017:Advances in Knowledge Discovery and Data Mining*, pages 375-387. Springer International Publishing, 2017, doi: 10.1007/978-3-319-57529-2_30

[9] W. Goh, X. Tao, J. Zhang, J. Yong, Y. Qin, E. Z. Goh, and A. Hu. Exploring the use of a network model in drug prescription support for dental clinics. In *2018 5th International Conference on Behavioral, Economic, and Socio-Cultural Computing (BESC)*, pages 168-172, 2018, doi: 10.1109/BESC.2018.8697814

[10] W. Goh, X. Tao, J. Zhang, J. Yong, X. Oh, and E. Z. Goh. Personalised drug prescription for dental clinics using word embedding. In L. U, J. Yang, Y. Cai, K. Karlapalem, A. Liu, and X. Huang, editors, *Web Information Systems Engineering 2020: Communications in Computer and Information Science,* volume 1155, pages 45-54. Springer Singapore, 2020, doi: 10.1007/978-981-15-3281-8_5

1.6 Awards

1. Best Student Paper Award for achieving the highest SNIP (Source Normalized Impact per Paper) values at University of Southern Queensland, 2016:

- W. Goh, X. Tao, J. Zhang, and J. Yong. Decision support systems for adoption in dental clinics: A survey. *Knowledge-Based Systems*, 104:195-206, 2016, doi: 10.1016/j.knosys.2016.04.022

2. First prize in Publication Excellence Awards awarded by Division of Research and Innovation, University of Southern Queensland, 2016:

- W. Goh, X. Tao, J. Zhang, and J. Yong. Decision support systems for adoption in dental clinics: A survey. *Knowledge-Based Systems*, 104:195-206, 2016, doi: 10.1016/j.knosys.2016.04.022

Chapter 2

Literature Review

Many studies have been done on drug interactions and decision support systems. This chapter reviews relevant studies in these areas in order to identify the gaps this research aims to address.

2.1 Clinical Decision Support Systems

The Clinical Decision Support System (CDSS) in this research belongs to the larger group of decision support systems (DSS), where the purpose is to provide decision makers a means to make decisions. Such an understanding from the perspective of general DSS will help support awareness of the functions and features expected of clinical DSS. Figure 2.1 illustrates a design structure of a typical decision support system.

When there is more than one decision maker, the process can be complicated, all the more when the information available can be subjective, objective, a combination of both, or even fuzzy. The problem and solution in a dental clinic refers to the treatment that best suits the patient. As shown in Figure 2.1, the decision made by the decision maker will depend on the problem itself, which would influence the criteria adopted by the decision maker as well as relevant information pertaining to the problem. Such an approach is reflected in the popular

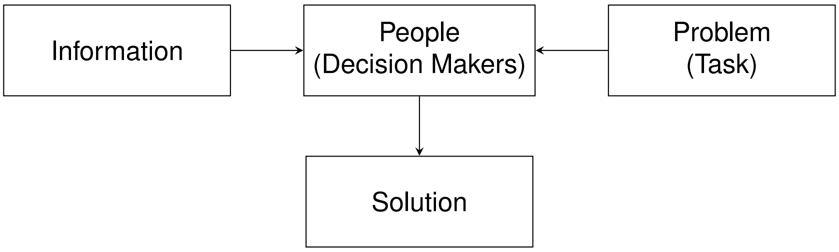


Figure 2.1: Design of a typical decision support system

PICO model [117] used by doctors in clinical assessment within evidence-based practice. This framework guides the practitioner in gathering information by asking questions related to information on the Patient (P), Intervention process (I), Comparison with other alternatives (C) and the Outcome to be achieved (O) which is the clinical problem the practitioner is trying to solve or diagnose. In most clinical situations, the patient can also act as the decision maker where information in terms of financial cost and aesthetic demands can influence the final outcome.

Therefore, a decision making process would normally be influenced by the individual's role as the decision maker, their preferences and the criteria used to make the decision [74, 76].

Generally, such complex decision making structure is determined by classical decision theories such as classical formal and empirical-cognitive decision theory, the theory of multi-criteria and/or multi-objectives decision making and the theory of group decision making. Interested readers may refer to [75] for more in-depth discussions.

Applications exist to put such theories into practice especially in the growing area of multi-criteria group DSS. For example, the “Decider" system, a fuzzy multi-criteria group DSS [78], takes into account the nature of information which in reality is usually expressed in linguistic terms, and the hierarchic structure of the problem and the decision makers.

Other areas of application where group decision making is based on multiple criteria include new product development such as for clothes [77] and digital scales [57] where preferences regarding the product have to be considered, and the car manufacturing industry where budget and time constraints are critical [70]. Another interesting application of DSS is to support a group of users in the choice of vacation packages [83]. Besides commercial applications, DSS are also found in areas that require long-term planning for sustainable development, for example, energy policy planning [115] and forest management [97]

Similar to the system established by [76] where fuzzy numbers are used to handle the uncertainties in the role of decision makers and the criteria used to arrive at the solution, a recent system by [108] uses fuzzy logic to construct a clinical DSS based on information input in the form of probability distributions. The unique aspect of this system is that the outcome given is not the single most desired solution, but rather, a set of solutions. By expressing the conclusion in this way, patients are more likely to accept the diagnostic decision from the health practitioner [43].

In terms of DSS for dentists, the information needed before the final treatment is decided will include the preferences of the patient in terms of cost and quality. For example, a fuzzy cognitive map is used to help the dentist decide on a suitable implant abutment for patients [68], combining expert knowledge from dentists and suppliers in the decision making process. Similarly, fuzzy logic is used by the system proposed by [79] to identify symptoms from patients, which are usually vague, making it difficult for the dentist to reach a detailed and definitive diagnosis.

Another DSS described by [105] represents an attempt to personalise the treatment plan by considering patient preferences to reach a mutual agreement between the patient and the doctor. Although this is a positive move to attract more users to adopt decision support, it only focuses on treatment planning for a single treatment.

As dentists are limited by and differ in their cognitive functions, such as in the recall and application of possible risk factors, there can be potential differences in the decisions made by different dentists, or even by one dentist at different times. In order to minimise such divergence, the system proposed by [8] considers expert domain knowledge and risk factors in decision support for caries management.

Though these systems utilise expert knowledge from dentists, there is no integration with a drug database which is essential within the clinical workflow.

The next section looks at some means of understanding CDSS in terms of the technology underlying their designs, classifications and benefits they can bring to the dentist.

2.1.1 Approaches in Understanding Clinical Decision Support Systems

A CDSS is an information system (IS) which has the ability to provide knowledge and personalised information to users, intelligently filtered to enhance health and healthcare

System	Comments
Selection of implant components [68]	Uses expert knowledge from dentists and abutment suppliers
Treatment of tooth fracture [79]	Uses fuzzy logic to help identify complaints from patients
Treatment of decay [105]	Considers the patient's preference
Management of tooth decay [8]	Considers the patient's oral history and health risk factors

Table 2.1: Recent decision support systems

outcomes [99]. They are not intended to replace the dentist's judgment and responsibility for decision making, but to provide assistance in diagnosis and treatment planning [146]. Table 2.2 presents some system capabilities with examples. They can be general or targeted at specific situations such as implant placement, and the output can be delivered to the user before, during or after the clinical decision is made [25]. The functionalities of DSS should follow the "five rights concept" [128] as a framework for planning and implementation:

the right information (treatment planning, drug interactions);

to the right people (dentists, patients);

through the right channels (mobile devices, workstations);

in the right intervention formats (alerts, graphics, info-buttons);

at the right time within the clinical workflow (before drug prescription, at point-of-care).

CDSS Capabilities	Examples
Preventive care	Screening, immunisation and disease management suggestions
Diagnosis	Lists of ranked differential diagnoses
Treatment plans	Treatment guidelines and drug dosage recommendations

Table 2.2: Major functionalities of CDSS [28]

Technologies behind Clinical Decision Support Systems

A CDSS can be implemented as a passive system, a semi-active system or an active system according to how it is being triggered [35]. Depending on the clinical tasks to be achieved, typical technologies used to develop such a system include machine learning, knowledge representation and data mining.

Machine Learning Machine learning is an appealing technique for its predictive ability based on existing representative data for diagnosis. Common machine learning techniques include Artificial Neural Network (ANN), logistic regression and support vector machines (SVM). ANN attempts to simulate the non-linear processing pattern of the human brain and is a very powerful tool for generalising acquired knowledge and data analysis by interweaving artificial neurons across input, hidden and output layers. For example, Georgios *et al.* used ANN for periodontal disease diagnosis and to classify patients according to their immune responses [103]. ANN was also applied to support decisions on implant placements, where the system mimicked choices made by implant experts [118]. Though data learning and training in the hidden layer is not transparent to the user, ANN is simple to implement as it requires minimal statistical training. The logistic regression method utilises a simpler linear model, and unlike ANN which can handle arbitrary relationships between input and output variables, it can only be used if such relationships can be explicitly identified [4]. Thus, the logistic regression method is not as robust as ANN. To classify non-linear datasets for an effective diagnosis, SVM can be used, which separates complex datasets with a linear hyperplane. Due to the complex nature of the datasets, training time can be high, especially when the volume of the datasets is large. However, this can be reduced by excluding outlier data points. For example, Kang *et al.* were able to obtain highly reliable drug failure prediction results with SVM when superfluous data points were excluded from the SVM ensemble construction [63].

Knowledge Representation Instead of learning from clinical knowledge as in machine learning, knowledge representation focuses on creating a knowledge description language which, when combined with a reasoner, is able to make diagnostic inferences. One approach in knowledge representation is the use of fuzzy logic, which is important in DSS as many applications deal with imprecise data and expect the results to have a dispositional rather than categorical validity. Unlike binary logic methods such as the above described ANN or SVM where the output is either true or false, fuzzy logic allows for different degrees of truth. In [79]'s design of a DSS for dental treatment, fuzzy logic was used to accept inaccurate and vague values of dental signs and symptoms associated with fractured teeth to produce possible treatment plans. Under rigorous testing conditions, the system was found to be similar to the dentist's professional predictions with respect to treatment for such situations.

Besides fuzzy logic, ontology-based systems can also be used to represent expert domain knowledge. Park *et al.* developed a shared DSS for dental fillings [105]. An ontology was built

Dympna *et al.* [29]	**Khalilfa [64]**	**Proposed Method**
Simple: interactive query	Basic: checking on drug interactions	Static: EHR, appointment reminders, drug allergy alerts
Complex: prediction of diseases using ANN	Advanced: individualised dosing support	Dynamic: knowledge base integration, self-learning

Table 2.3: Classification of a CDSS

based on tooth anatomy, diseases and treatment options. This enables ontology-generated evidence-based alternatives to be made available for dentists and patients to reach a shared decision on the most effective treatment plan.

Since the use of radiographs feature prominently in oral disease diagnosis [127], information from the images should also be stored in the knowledge base. This focus on the problem rather than the technology corresponds to an improvement from the conventional method of diagnosis and meaningful use of DSS [72].

Data Mining For unstructured data, text mining techniques can be used to discover context-specific knowledge based on patient-specific profile in supporting dentists in their decision-making process for a specific oral health situation.

Semantic meanings can be extracted from textual data through data mining methods based on rules created from concepts and relationships within the appropriate ontology. [153] used a data mining method to identify relationships between medications for diabetes patients. By identifying patterns within the drug database, the system was able to predict, with significant accuracy, the subsequent medication to be prescribed.

Classification of Clinical Decision Support Systems

In the literature, there are many ways to classify CDSS, according to their features and functions. For example, as shown in Table 2.3, Dympna *et al.* classifies them according to complexity of the system's functions [29]. A simple system is one that accepts a command from the user and produces a response to the user. As an illustration, the user may use the system to check for drug reactions to a particular drug by entering the drug name, with the system then displaying the results to the user. Complex systems use a “black-box" approach, including artificial intelligence, logistic regression and data mining, to produce advice or diagnostic predictions to the user. Examples include systems for identification of prostate cancer, sleep apnoea and psychiatric problems. Unfortunately, there are no examples in the area of dental pathology. This is expected as even in the medical domain, complex systems are difficult to customise to local clinical workflow, not to mention being difficult to develop as it requires both design expertise from the researchers and relevant knowledge from the users within their clinical domain [29]. Similarly, Khalilfa refers to systems that perform checking on drug-drug interactions as basic systems [64]. Those with more elaborate features such as checking on contra-indications and dosage support are referred to as advanced systems.

In the context of DSS for dental clinics, it is recommended that such systems be classified as static and dynamic to reflect the approach taken in the design and implementation of the system within the clinical workflow. Static systems are those which do not possess the learning ability that dynamic systems can provide to the dentist. With machine learning features incorporated into the design, dynamic systems are able to provide real-time personalised support to the dentist where the medical profile of each individual patient is taken into consideration. In the prototype designed as a result of this study, the drug that is prescribed is stored in the database against the particular patient's record. Subsequent checking will refer to the updated drugs that were being prescribed.

According to these definitions, the systems that correspond to the simple or basic groups of DSS referred to earlier will be known as static systems since such systems are not personalised to the individual patient. Systems that provide logistic and administration support also come under this category. Examples are programs that allow storing, searching and retrieval of information on the clinic's inventory, accounting and patient information.

On the other hand, dynamic systems are designed to incorporate reasoning and self-learning capabilities so as to provide personalised support at point-of-care to the dentist within the clinical workflow. One critical feature in personalised support is in the area of drug prescription, where the system should be able to support the dentist in determining if the drug to be prescribed is safe for the patient by considering the individual's relevant medical history - the drugs the patient is currently taking, the drugs the patient is allergic to, and the medical conditions of the patient [40].

Hence, dynamic systems typically incorporate a drug knowledge base to store decisions made by the dentist and information on side-effects and interactions of drugs. It is crucial in a dynamic system to ensure that the drug knowledge base is updated regularly, not only with the latest information on drugs, but also with the decisions made by the dentist. This will allow the system to capture the ground truths from the dentist and in turn, become more efficient in providing relevant information.

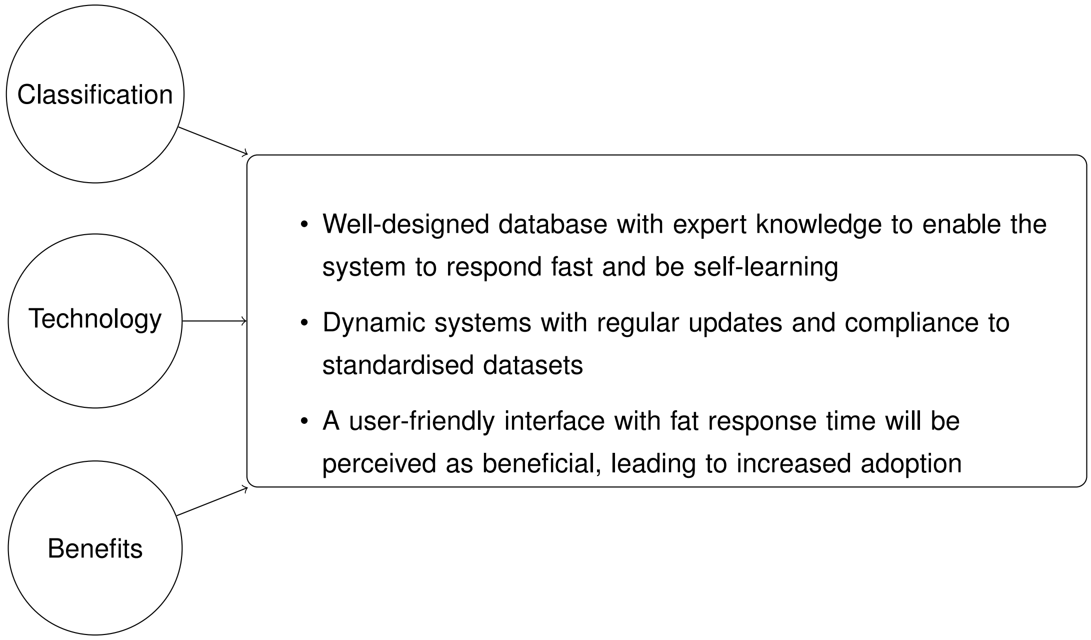
Following the suggested approach in the classification of DSS, if the system is not self-learning, it will be grouped as static even if it provides advanced features such as the dosing support mentioned by [64]. On the other hand, a system that provides answers to simple queries [29] or basic functions [64] on drug interactions can be considered a dynamic system if such queries take into account the relevant medical history of the individual patient and is able to learn from previous decisions of the dentist.

Benefits of Decision Support Systems

Besides assisting dentists to make timely and informed treatment decisions, a DSS is also useful in the following areas [94, 150]:

1. keeping electronic health records (EHR);
2. drug prescription, medication dosing support;
3. clinical reference count;
4. point-of-care alerts and reminders.

In addition, a well-designed system which integrates patients' EHR will complement the dentist's evidence-based decision making with benefits including less paperwork, better tracking of data, accounting and reporting functionality [18]. Storing the daily clinical decisions and treatment outcomes will enable the system to "learn" and possess more knowledge to solve subsequent clinical problems. With datasets stored in ontology and made available using the techniques and technologies of the Semantic Web, the data will become accessible for further data analysis and knowledge discovery. This produces a platform that supports a "range of scientific research activities intended to advance our understanding of dental conditions and the relative success of different treatment interventions" [129]. Consistent and reliable information will also avoid misdiagnosis and malpractice, which can lead to expensive legal suits. With comprehensive drug information and diagnostic support provided in real-time at point-of-care within the clinical workflow, there will be improved clinical efficiency, oral health outcomes for patients and job satisfaction for the whole dental team.



- Well-designed database with expert knowledge to enable the system to respond fast and be self-learning
- Dynamic systems with regular updates and compliance to standardised datasets
- A user-friendly interface with fat response time will be perceived as beneficial, leading to increased adoption

Figure 2.2: Ways of understanding decision support systems

Summary

This section outlined approaches in understanding CDSS in terms of the technology underlying their designs, classifications and benefits to the user. As summarised in Figure 2.2, the right technology behind the design of CDSS will ensure the system is self-learning and has the most relevant information on the patient's medical conditions and drug allergies. Besides providing static information, it should be dynamic where the knowledge base is updated regularly and able to give alternative suggestions based on the personalised medical status of the patient. A dynamic DSS will be perceived as beneficial which results in increased adoption by users within the clinical workflow. Hence, it is important for clinical DSS to be able to progressively learn from the user's decisions and make diagnostic personalised inferences in a user-friendly manner.

Despite the benefits that a DSS can potentially bring to the user as described in this section, many dentists are still not adopting it as a diagnostic tool in their daily practice. The next section looks at the common challenges that hinder such an adoption.

2.1.2 Challenges to Adoption of Decision Support Systems

Although DSS have existed since the 1990s, adoption in the clinical workflow is still poor. This section looks at some of the major challenges in the adoption and implementation of DSS as a treatment planning tool for dentists.

Lack of Perceived Usefulness

As mentioned in Section 3.1.2, the focus of diagnosis should be on the problem and not on the technology [72]. Many dentists feel that they can diagnose the problem better than the DSS, perceiving that such systems are not useful within their clinical workflow. Poor usability is often cited as a reason for slow adoption of IS as it "makes it difficult for providers to navigate through the information and obtain an integrated view of patient data" [140]. Besides, most systems only support a particular kind of treatment, such as treatment planning for tooth decay [79,104] or the selection of implant components [68].

Such limited scope also contributes to their slow adoption rate [129]. A qualitative case study with thirty-seven doctors found that usefulness in relation to consultation issues is one of the driving factors for adoption of DSS in diagnosing clinical problems [126]. Though it investigated medical doctors, the findings can be applied to dentists as well, with other studies also supporting this conclusion. For example, Venkatest *et al*'s findings [144] are in agreement with the Technology Acceptance Model [23] which posits perceived usefulness as a determinant in usage intention of technology. In another study which examines challenges in adopting CDSS by using the Unified Theory of Acceptance and Use of Technology (UTAUT) model, performance expectancy (which includes perceived usefulness), defined as "the degree to which an individual believes that using the system will help him or her to attain gains in job performance" [144], is again the strongest predictor of usage intention. A literature review conducted by Devaraj has identified challenges to performance expectancy of DSS [26], with the top five being:

- time constraints;
- obscure workflow issues;
- authenticity/reliability of information;
- disagreement with the system;

- interoperability/standards.

A study on how clinicians diagnose and treatment plan also revealed that sources of information used by dentists come as separate blocks which distract the users and have adverse effects on efficiency [140].

Complex Sociocultural and Economic Factors

Dentists envisage that a DSS is not very useful in aiding diagnosis, and they are used to depending on their own clinical skills or at most a quick discussion with colleagues before arriving at a treatment plan. Medical practitioners are used to the culture of autonomy, and using such a system will disrupt that autonomy leading to resistance to their adoption within the clinical workflow [143]. A study on the challenges perceived by a group of rheumatologists discovered a sense of ambivalence relating to concerns that using technology could impair doctor-patient communication [165]. Many studies have also noted that practitioners are reluctant to use the system in front of patients [26]. This is expected since practitioners do not wish to be perceived as lacking in diagnostic skills or appear to be inefficient in navigating the system. Resistance to new technology is not just confined to DSS, as can be seen from the introduction of the blood pressure monitor into the clinical workflow during the early 20th century. At that time, physicians deemed that their unique skill in taking blood pressure by palpation was being challenged and thus felt uneasy about using such technology [22]. However, it is so common nowadays, to the point that it has become a do-it-yourself gadget and can be used by anyone at home.

Majid, in a study to understand the lag in IS adoption, discovered that financial gain and time savings are crucial factors in influencing technology adoption, suggesting that for a clinical DSS to be used at point-of-care, a fast response time is required [80]. Research by Mamatela with African doctors identified environmental factors to contribute to the practitioner's propensity to adopt the use of electronic health technology [81]. Zande *et al.* also discovered that the diffusion rate of a new technology depends on social influence from peers and the perceived advantages that the system will bring to their workflow [142].

Horgan performed a comprehensive survey of why personalised medicine is not being accepted by many clinical establishments, and found that important factors included differences in company cultures and the practitioner's ability to use and interpret results from the IS [51].

In an effort to explore the challenges that come with the implementation of an IS at Stockholm's Karolinska University, Ovretveit *et al.* discovered that consultation before implementation is a prominent factor [100], and that the perceived usefulness of the new system aligns with Roger's Theory on the diffusion of innovation which seeks to examine factors that influence the adoption of new technology [113]. While most studies focus on the economic aspect of technology [92], this study looks at challenges to adoption from the sociocultural aspect.

The findings also appear to support Orlikowski's theoretical model [98] which explains how user interaction with technology is influenced by the corporate culture within the clinic. The model is an attempt to explain that technology is a product of human design and yet used by humans to accomplish the designed task. As illustrated by the model in Figure 2.3, such actions are often confounded by the social environment of the work place.

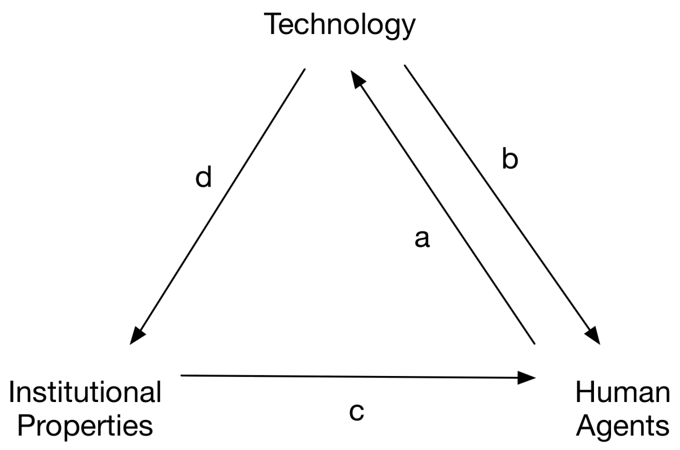


Figure 2.3: Structurational model of technology [98]

In an evaluation of the factors that impact on information systems, it was also strongly suggested that socio-technical connectives between users and technology should be considered when developing electronic health systems [95].

Difficulties in Interpretability

Interpretability, in terms of interfacing and standards, is another issue that can influence CDSS adoption.

Human Computer Interface

Usability and human factors are the first recommended domains within the research agenda tasked by the American Medical Informatics Association [84], which highlights the significance of the user interface in CDSS. Without a well-designed interface, the personalised and smart learning features of the IS will not be fully utilised and its usefulness will not be perceived by the user. In fact, user-friendliness is important in increasing the "usability" of the system as it will make it easier for the dentist to navigate and obtain an integrated view of the patient's data [140].

As shown in Figure 2.4, the human computer interface (HCI) plays an important role within the cyclic path of the local expert knowledge base and diagnostic result from the DSS. An effective system will have a user-friendly interface to enable the dentist to understand the given result from the system. Based on the result, the dentist will be able to further update the local knowledge base. With the updated knowledge base and data mining techniques, the system will be able to continue to produce useful and relevant information for the dentist to make subsequent decisions. A smooth and efficient human-computer integration such that knowledge can be obtained with ease will result in more clinicians accepting and using the technology [36].

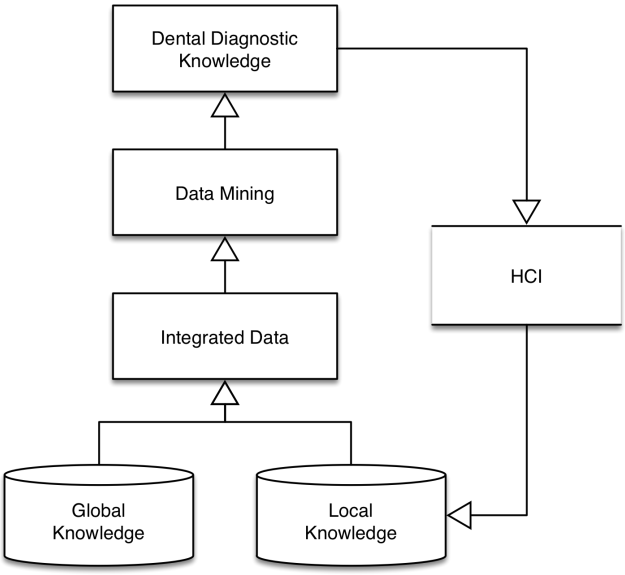


Figure 2.4: Role of HCI in an intelligent information system

As an efficient and effective IS involves communication between the system and the user, a comprehensive interface design is crucial for the successful construction and flow of an appropriate knowledge base. Thyvalikakath *et al.* observed that there is little research on the application of cognitive engineering methods to support system design [141]. More studies are required to observe how dentists interact with patients and computers as the results will contribute to the design of an IS that can enhance cognitive support for dentists [140].

Lack of Standards

Lack of standards and lack of time act as challenges to the adoption of CDSS [26]. While a DSS needs to simulate the decision-making process of the dentist, the result of the process may appear difficult to interpret for the dentist due to emerging standards of healthcare information technology [29], and may cause dentists to spend too much time on the system at point-of-care. Since good design of a system requires the efficient collaboration of knowledge from patient profiles and other knowledge bases, standardisation of data is important to ensure the system performs efficiently.

Summary

This section has identified perceived usefulness as one of the main challenges against CDSS adoption by dentists. Other challenges (Table 2.4) include various complex sociocultural factors, system interface and the issue of standards.

As perceived usefulness also implies a system with an acceptable response time and a user-friendly interface, many users are reluctant to use CDSS as current systems have limited functions and features, are perceived to be difficult to use, and require unwarranted effort to interpret the results produced by the system.

A lack of concern for the user's needs and expectations contributes further to the lack of propensity to adopt the system within their clinical workflow. Perceived advantages that the system will bring about, such as possible time savings within the user's workflow (thus leading to cost savings), are also crucial factors in influencing technology adoption by the dentist.

The next section surveys some of the current DSS that attempt to overcome these challenges.

Challenges	**Remarks**
Perceived usefulness	Limited functions
Sociocultural and economic factors	Resistance towards technology Social and corporate influences
User interface and standards	Lack of standards for datasets Difficulty in interpreting results

Table 2.4: Challenges to adopting decision support systems

2.1.3 How Decision Support Systems Overcome Challenges

The need for a robust and intelligent self-learning system has been identified by IBM as one of the challenges in effective healthcare delivery [1]. Such a system should have appropriate tools and techniques to provide decision support to users [132]. Most current systems consist of only simple alerts and reminders with no sophisticated advisory functions [29]. Table 2.5 presents some features available in the design of current DSS, in comparison with the features expected to appear in future systems as suggested by some researchers.

The following sections highlight some of the important features in these systems that help dentists overcome the challenges in the adoption of CDSS within their clinical workflow.

Efficient Design of Knowledge Base

It appears that many designs contain a knowledge base of rules pertaining to the expert knowledge of the application. For example, for an application that targets implants, [158] described a dental expert system, which stores facts on symptoms and diseases with static general information of patient profiles to assist the dentist in disease diagnosis. It stresses the importance of an evidence-based diagnostic approach instead of an experimental one and provides a modular design framework containing a knowledge acquisition database, a general database, an inference engine and the user interface. The knowledge acquisition database is important for any DSS to be useful for the users, and is critical in assisting the dentist to make an intelligent treatment plan [90].

Similarly, Lee *et al.* have researched the optimal selection of dental implant abutments [68]. A fuzzy cognitive map is used to contain rules and expert domain knowledge from both dentists

Current Clinical DSS Design Features	Expected Design Features
Separate display of information sources [140]	Integration of medical and dental history [116, 121]
Simple, static and non-learning [29]	Intelligent and personalised [51]
Perceived as not useful and time consuming	Efficient searching, retrieval algorithm and user-friendly HCI
Limited scope [129]	Interoperability and accessibility [37]

Table 2.5: Current and expected features of CDSS

and domain experts from implant manufacturers. To enhance patient satisfaction and effective treatment, the clinical DSS not only stores expert knowledge but also generates treatment options using ontology that contains the patient's profile and their preference of options [105]. Mago *et al.* also developed a system to reduce inconsistencies in treatment planning for a fractured tooth [79]. Fuzzy logic, first introduced by Takagi and Sugeno [136], was used for its strength in dealing with imprecision pertaining to dental disease and symptoms.

In another CDSS, anatomy and diseases are stored in a database according to standards from FMA and ICD-10 respectively [104]. By linking treatment with information from the database, the system was able to aid the dentist in treatment planning and reduce the need to primarily rely on memory of similar cases, or on trial and error.

As seen from the design of current CDSS, the knowledge base plays an important role in providing treatment options to the users. Naturally, such domain knowledge needs to be regularly updated to maintain options that are relevant. With the help of the Delphi technique, a six-month study at King Faisal Specialist Hospital and Research Center collecting experiences and suggestions on strategies for successful implementation of decision support reported that updated knowledge bases is one of the success factors for a CDSS to be useful and acceptable to users [64]. To allow such knowledge bases to be reviewed, updated and managed effectively, an important feature in current systems is to the separation of these clinical rules and knowledge from the main IS application. This leads to cheaper service integration of DSS into existing IS [66] and also enables such systems to utilise information from local knowledge base with those from other ontology. Figure 2.5 is an illustration of this model.

Though the current CDSS utilise knowledge bases in their design, they are of a limited nature, restricted to a particular kind of treatment plan. Even if it is focused on diagnosis of a common disease such as dental caries, the knowledge base is not self-learning. For example, the system developed by Park *et al.* [105] for dental fillings needs to be expanded to include clinical guidelines from global dental ontology in a real-time manner and integrated

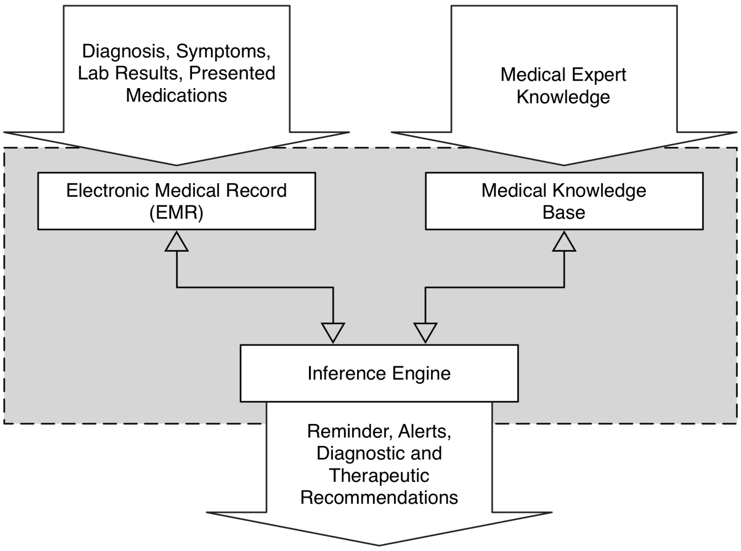

Diagnosis, Symptoms, Lab Results, Presented Medications
Medical Expert Knowledge
Electronic Medical Record (EMR)
Medical Knowledge Base
Inference Engine
Reminder, Alerts, Diagnostic and Therapeutic Recommendations

Figure 2.5: CDSS model [30]

with local knowledge, in order for the system to be self-learning and to allow practise of evidence-based dentistry. This involves semantic annotation that requires complex machine learning techniques [129]. Since dental ontology can enable decision support system to automatically update their knowledge base with expensive expert medical and dental knowledge, it will be easier and cheaper to maintain the system with the current expertise of dentists and the latest existing knowledge in scientific and clinical evidence [129]. Additionally, the efforts of researchers and dentists can be harnessed easily through a Semantic Web interface provided by dental ontologies which act as a consensual representation of knowledge in the dental domain [129]. Good design and fast response time will increase the appeal of such a system.

Ontology

We expect a CDSS to not only be efficient enough to appear helpful to dentists, but also to fit the clinical workflow at point-of-care, which commonly requires it to handle multiple diseases and drug allergy information. Bhatia and Singh designed a CDSS to produce a treatment plan for tooth decay [9]. Based on the different degree of oral symptoms, the system suggests possible treatment plans based on the Bayesian Network. Another system proposed by Bessani *et al.* also used the Bayesian Network as an inference engine to produce treatment options based on the individual's oral health history and risk factors [8]. Though these systems help the dentist to treat patients more confidently, they are only restricted to situations involving tooth decay. Furthermore, there is no interfacing with ontology knowledge based on dental disease and drug information.

The inclusion of drug ontology is important as drug information is commonly needed within the clinical workflow and is a basic point-of-care activity in oral health therapy. In a study to encourage health professionals to use CDSS by identifying the potential challenges that they are facing, over half the literature short-listed for review utilised patient disease in their CDSS [26]. This reflects that patient disease/condition management is the area where healthcare practitioners require the most assistance in decision-making. Therefore, a CDSS which integrates with drug knowledge bases to advise on drug suitability before prescription will appear helpful to dentists and overcome the performance expectancy challenge .

Ontology should be updated in real time without the need for manual intervention. Using this technology also requires the standardisation of datasets, with the need to only be familiar with one set of terminology increasing the attractiveness of usage.

Human Computer Interface

As described in Section 2.1.2, a poorly designed user interface downgrades the performance and reduces the benefits to clinicians [52], posing a challenge to system adoption. A well designed interface enhances usability and cognitive support for the user to make better and faster decisions. The system proposed by Park *et al.* also integrates expert knowledge from the patient and existing ontology though it is unclear if the ontology is updated in real-time [105]. Overall, it is a good system except for the lack of a drug checking function, which is essential within the clinical workflow.

Many existing systems lack the usability and friendliness that users expect from an IS. In a survey on factors influencing implementation and outcomes of a dental recording system, less than a third of the respondents (n=130) thought that the system improved productivity when asked: *“What do you like about the Electronic Patient Record System?”* The majority favoured its increase of legibility and improved access to patient charts [147]. The results suggest a need to enhance the usability of IS. In order to transform patient profiles and data in a knowledge base into useable and useful knowledge, the design of the HCI must consider who are the potential users. It should combine the cognitive and reasoning ability of the expert user with the fast and accurate data mining processing power of the IS [50]. A good interface is also crucial in the technology diffusion process to enable high acceptance and absorption rates.

As mentioned previously, radiographs are useful diagnostic tools for the dentist to identify oral diseases. A system has been developed to analyse x-ray images for patients with tooth decay in order to assist the dentist in making accurate and timely decisions on diagnosis and treatment planning [111]. As illustrated in Figure 2.6, the original image is enhanced to enable the user to more accurately identify the exact location of the tooth decay. The image is then segmented to eliminate misjudgment, and feature extraction performed to enable the algorithm to identify the location of the lesion for the dentist to make further judgment on the treatment plan.

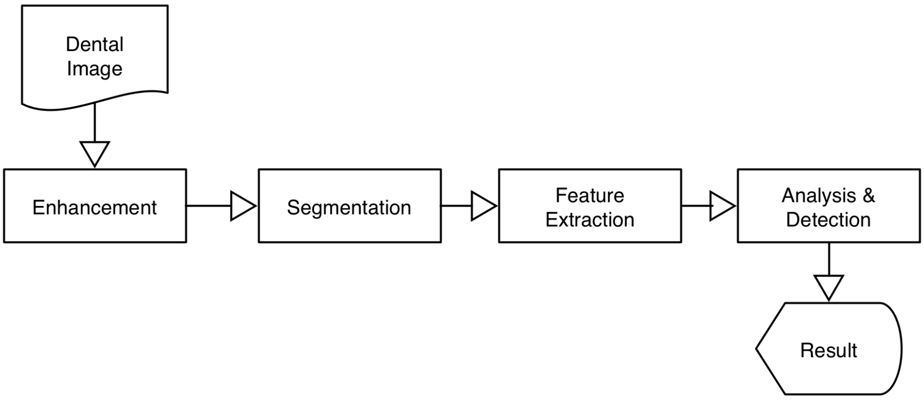


Figure 2.6: Dental caries detection algorithm framework [111]

ORAD (Oral Radiographic Differential Diagnosis) is a system developed for identifying intra-bony lesions from radiographs to produce a list of possible diseases [151]. It was found that the system is useful as an adjunct for the dentist in diagnosing oral diseases [127]. As can be seen from current systems, there is yet an ideal design to cater for real-time updating of ontology and treatment planning for multiple oral therapies as well as drug information checking before prescription at point-of-care.

Summary

This survey explored key features that are crucial for the adoption of CDSS by dentists, such as effective design and a user-friendly interface. Systems should be well designed to enable the user to make effective and efficient treatment plans without having to depend on memory of past cases. As indicated in Table 2.6, systems which incorporate visual representations in identifying oral disease with user-friendly interface will help the dentist overcome the performance expectancy challenge.

A survey of existing systems with features that support treatment planning for the dentist found that such systems offer treatment options only for a single aspect such as selection of implant components or the identification of tooth decay.

Systems that are personalised to the patient's oral health profile with a user-friendly interface will be perceived by dentist as more useful. This will help them to overcome challenges in their decision to adopt a DSS within their clinical workflow. Even within such a personalised system, there is still a lack in real-time interfacing with drug and disease knowledge bases to enable treatment planning for multiple oral therapies and recommendations in drug prescription.

Features	Remarks
Effective design of database	Insightful use of expert knowledge
Ontology	Important to link to drug and disease knowledge bases
User-friendly interface	Overcome the performance expectancy challenge

Table 2.6: Overcoming challenges

2.1.4 Trends for Clinical Decision Support Systems

Research and development on CDSS should keep pace with technology changes so that the system can fit the diagnostic requirements of users and be adopted into the clinical workflow. This section highlights some of the emerging trends such as the use of big data in personalised systems - recently mentioned as a top contribution in a survey of 1,254 papers published in 2014 in the field of clinical decision support [13] - as well as the issue of privacy.

Big Data

With an ever-increasing volume and different types of knowledge to be stored in an IS (for example, structured, semi-structured and/or unstructured), it remains a challenge for the system to allow processing and searching techniques to interact efficiently with human intelligence. Compared to other fields such as education and finance, the velocity and variety

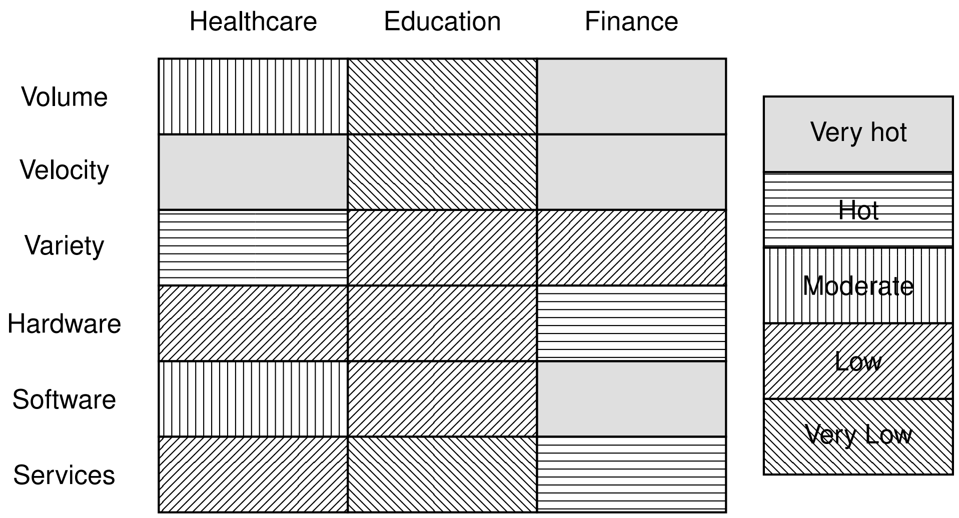


Figure 2.7: Big data heat map [121]

of data generated in healthcare is much more significant, with Figure 2.7 illustrating a big data heat map covering these domains.

From the heat map, it is evident that the quantity and expected speed of processing, analysing and distributing of information in healthcare will “bring the potential to discover new knowledge that can improve work practices and produce better outcomes” [121]. This is particularly true in the dental clinic where the dentist needs to consider information from intra-oral images, 3D images, unstructured clinical notes and the patient’s profile in real-time at point-of-care before deciding on a personalised treatment plan.

Big data, which integrates knowledge through analytic tools such as Semantic Web, offers advisory functions such as personalised treatment options, in addition to the typical administrative functions. Furthermore, the indexing of clinical and non-clinical datasets of big data will help researchers discover new knowledge and relationships among multiple variables, which is impossible with unconnected and disparate datasets.

Hence, design and implementation of CDSS should exploit the notable potential of big data. The system should effectively and efficiently analyse, integrate and interpret knowledge to be used by the user in enhancing treatment outcomes and patient health [50]. Due to information silos, which fragment the medical and dental domains [122], it is important that data from both domains is seamlessly integrated for efficient processing and distribution to clinicians. Besides early medical prognosis (as many medical conditions are manifested first in oral cavity), other benefits of medical and dental record integration are [116]:

- improved decision-making;
- improved patient outcomes through prevention, early detection, and proper intervention;
- transparent information across medical and dental providers;
- reduced cost to providers.

In addition to the challenge of information silos in knowledge bases is the need for DSS to be able to reference and reason from these databases to produce an effective personalised treatment plan. For example, by using OWL 2 (an ontology language for the web), Park *et al.* [104] designed a system to generate dental treatment options by querying knowledge bases that represent the type of disease and tooth location. Datasets containing drug information will also be very useful for the dentist when prescribing drugs at point-of-care. This is to allow dentists to ensure that the patient will not suffer from an adverse effect from a cross-allergy to the prescribed drug (usually due to similarities to a drug that the patient is known to be allergic to) or an interaction between the prescribed drug and the drugs that the patient is currently taking.

Personalised Systems

Among the many knowledge domains to be stored in a typical CDSS, there is a growing interest in the field of genomics to cater to genetic variations among patients. Focusing on such personalised information will result in greater quality of care and reduced healthcare cost, so it is not surprising that pharmocogenomics, the use of patient genotypes to explain individual differences in drug responses [59], is one of the most common examples of personalised medicine [51]. In fact, it has been predicted that personalised medicine will replace the traditional trial and error approach in healthcare [39]. More specifically in oral healthcare, personalised medicine based on the individual's unique genetic, molecular and clinical profile should be the aim for researchers and dental practitioners in providing quality, customised and effective healthcare [37]. It has been anticipated that applying genomic information to oral disease diagnosis will allow a better understanding of disease aetiology, leading to preventive measures being implemented prior to disease onset [31].

A proposal for a framework to support the sequencing of genomes predicted that a decision support system provides the greatest opportunity to enable the use of genetically-guided personalised medicine [150]. Hopefully, the collaboration between eMERGE [41] and the clinical decision support consortium [21] will lead to a standard for genome-informed IS and fulfill the vision for personalised medicine in the near future.

Hence, it is important that potential systems are personalised to the patient's profile to align with the trend towards personalised medicine.

Standards and Privacy Issues

As discussed in Section 2.1.2, interoperability and standards are one of the top challenges in adopting CDSS. The difference in data formats from different vendors and countries not only reduces interoperability but also makes the merging of complex datasets complicated [51]. Thus, the challenge is to standardise the knowledge base format to enable the system integration less painful.In fact, focusing on a standard approach to knowledge sharing is one of the most active areas in current research in translating support from campus research to clinical point-of-care [89]. To ease practical development of CDSS, design should endeavour to conform to standards such as those produced by Health Level Seven International [49] and incorporating clinical terminologies [1] which adhere to interoperability specifications like those owned and distributed by the International Health Terminology Standards Development Organisation [55]. This will remove another challenge to the adoption of a CDSS. While it is important to unite and standardise different data and coding standards, there may be potential issues of privacy with regards to patient information. For dentists to adopt and integrate DSS, there is a need to convey both to patients and practitioners that secure protection of information is in place within the system. Privacy regulations are required to balance against the need for exposure of data between researchers and developers [1].

In summary, CDSS should be capable of integrating the ever-increasing volume of data appearing in different genres and formats, and to make inferences to effectively process and produce clinically relevant knowledge to support decision-making by dentists. The challenge of information silos requires systems to work on standardised datasets stored in an ontology which can be inferred and retrieved through the latest Semantic Web technology. While research efforts are focusing on maintaining a uniform knowledge base format for effective sharing and reasoning, the delicate issue of privacy needs to be addressed carefully so that personalised features of a CDSS can be fully utilised by dentists without the risk of compromising patient confidentiality.

[1] https://www.snomed.org/

2.1.5 Support for Drug Prescription

In order for the design of CDSS to fit the workflow for the health professional, critical features within the workflow should be included. As mentioned in Section 1, a system to reduce medication errors is important to enhance treatment efficiency. Such features will also enhance the adoption of the system as management of diseases is the area where support is most needed [26]. As reviewed earlier, many works do not include the medical profile of the patient [8, 27, 34, 68, 105] . The lack of such features clearly results in the inability to provide a personalised system. Although there are many studies that examine drug-drug interaction [12, 14, 155], they do not associate them with the patients medical profile to facilitate individual drug prescription. The unique approach adopted in the thesis uses the medical information of the patient to support the decision-making process for doctors at point-of-care within the clinical work-flow.

2.2 Drug Interactions

In the study of pharmacology, drug interactions can refer to pharmacokinetics and pharmacodynamics [45]. Pharmacokinetics refers to the movement of drugs in the body, which is what the body does to the drugs in terms of absorption, distribution, metabolism and excretion. Such interactions occur when the perpetrator drug alters the concentration of another drug, the object drug with clinical consequences, which can be positive or negative. On the other hand, pharmacodynamics is what the drug does to the body. A wrong prescription of a drug may result in unwanted side-effects like rash or dizziness. The textual description of side-effects can also be used in data mining procedures to study the similarity of a drug pair.

Many systems are using data mining techniques to explore drug-drug interaction (DDI). In fact, such techniques are evolving quickly to improve the accuracy of the experiments, though in most situations results may not be sufficient to derive DDI [154]. A recent work by Bokharaeian *et al.* attempts to determine DDI by identifying neutral candidates, negation cues and scopes from bio-medical text [12]. Features extracted from these articles include linguistic definitions of negation, the position of the drugs discussed in the sentence and the linguistic-based confidence level of an interaction. By using datasets from DrugBank, it is reported that the results achieved an F-score of 68.4%. Text mining techniques have also recently used to predict protein interactions from bio-medical text [69]. To increase the prediction rate by ensuring information from tables and figures are also extracted, Milosevic *et al.* suggested a 7-step methodology [87]. These 7 steps are table detection, functional processing, structural processing, semantic tagging, pragmatic processing, cell selection and syntactic processing and extraction. Such an approach will ensure both figures and text are considered when mining the clinical literature. The approach achieved a F-score of at least 82% depending on the complexity of the tasks.

Another common way of examining DDI is to extract relevant information from text. For example, Tari *et al.* developed a method that combines text mining and automated reasoning to predict enzyme-specific DDI [138]. Yan *et al.* also used text mining techniques to create features based on relevant information such as genes and disease names extracted from drug databases to augment limited domain knowledge [155]. These features were then used to build a logistic regression model to predict DDI. Another method to extract information on DDI from bio-medical text was proposed by Bui *et al.* [14]. DDI pairs are mapped according to their syntactic structure, and subsequently generated feature vectors are used to produce a predictive model which classify the drug pair as interacting or not interacting. Drug similarity has also been shown to be associated with literature-based similarity from a recent work by Zeng *et al.*. They attempted to measure drug similarity, which can be used to predict drug interactions [27, 34], from electronic medical records [157].

Though these studies use data mining methods to extract relevant information to predict DDI, these works are limited to two tiers, the knowledge layer and prediction layer, unlike the three-layer framework in this paper.

The crucial need to use the knowledge obtained from data mining motivated us to develop this three-layer conceptual framework. Although our system is similar to that proposed by Casillas *et al.* in terms of using information from the patient [17], the unique approach adopted in the thesis goes one step further in using such information to support the decision-making process for the dentist at point-of-care within the clinical workflow. In this model, an additional presentation layer is introduced, providing an important interface between the user and the knowledge mined from bio-medical sources. Besides the use of the proposed presentation layer, the use of features like $tf * idf$ and word vectors in getting the drug similarity to decide if the drug pair is in adverse relationship distinguish the approach against existing DDI methods. Moreover, innovative approaches adopted in the prediction layer allow the efficient extraction of features that relate the similarity of a drug pair in terms of the shared difference in their term frequencies. Experimental results for this approach was favorable compared to existing models.

2.3 Detection of Drug Properties

2.3.1 Graph-based Similarity

As data representation is crucial to machine learning in terms of retrieval and recommendation tasks [163], this study attempts to represent the domain knowledge on drug interactions by a graph. This enables the similarity of a drug pair to be predicted based on their common paths. By representing the drugs as nodes and the interactions between them as edges, algorithms can be developed to find common paths connecting two nodes. If the number of common paths between a drug pair is large, it indicates that the pair is very similar.

In recent years, there has been a growing interest in comparing text and computing similarity between entities by representing them as a graphical model. For example, in the model experimented by Palma *et al.*, the semantic similarity between drug pairs is used to predict drug target interactions [101]. Based on the hypothesis that similar targets interact with the same drugs, and similar drugs interact with the same targets, a heterogeneous graph was constructed with edges that included the drug-target interaction as well as drug-drug and target-target similarity edges. Jeh and Widom also proposed a framework to compute the similarity between two objects by representing them and their relationship as a graph [58].

With the objects as nodes and their relationship as edges, this framework assumes that two objects are similar if the objects related to them are also similar. For example, two publications are considered similar if the papers cited by each publication are also similar. The directed graph $\mathcal{G}$ used to represent such a framework with nodes V and edges E can be formally defined as $\mathcal{G}$ = (V,E) where the nodes V represents the objects and the edges E represent the relationship between the objects. If $I_i(v)$ represents individual incoming objects and $O_i(v)$ individual outgoing objects, then the similarity score between any two nodes A and B is defined as:

$$\text{Sim(A,B)} = \frac{C1}{|O(A)||O(B)|} \sum_{i=1}^{|O(A)|} \sum_{j=1}^{|O(B)|} s(O_i(A), O_j(B)) \tag{2.1}$$

In another model with special emphasis on Heterogeneous Information Networks (HIN) [148], similarities between two entities can be found by considering the number of paths between them. Nodes and edges are defined in this HIN as $G = (V, E)$ where the nodes are the set of entities A and edges are the set of links R between the items in the entities. The entity type mapping is given by $\phi = V \rightarrow A$ and the relation type mapping given by $\psi = E \rightarrow R$. The similarity between two entities A and B is defined as:

$$\text{sim(A,B)} = \frac{\text{2.\# paths between A and B}}{\text{\#circles with entity A + \#circles with entity B}} \tag{2.2}$$

In yet another attempt to represent entities in a directed graph, Shi *et al.* focused on an approach which also assigns weights to the relations between the entities [125]. Hence, this method of representation becomes appropriate for a recommender system. Besides weightage, this method also assigned attributes to the links between the entities.

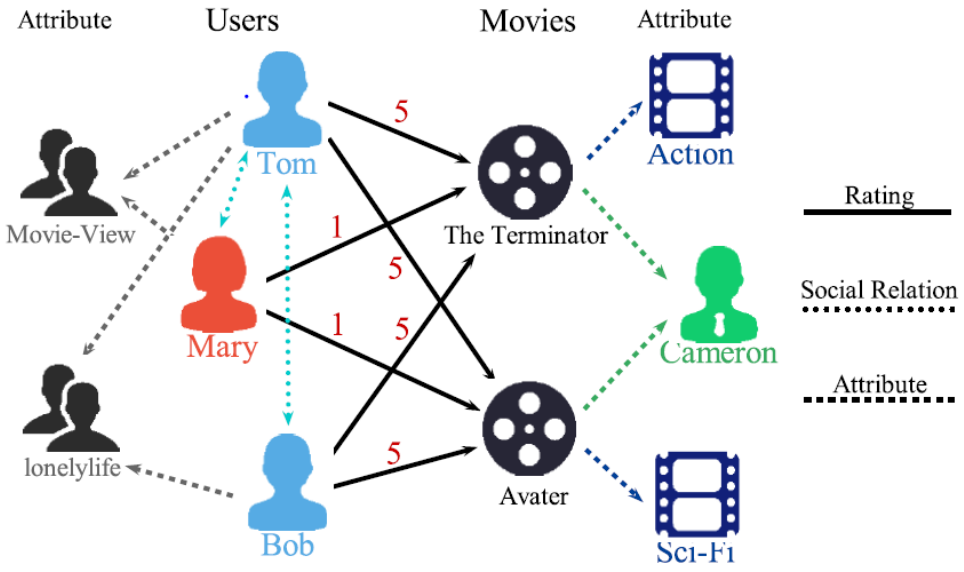


Figure 2.8: Objects and relations in an information network [125]

For example, users a and b may have a common liking for movie $m1$ (see Figure 2.8), besides other movies, so we can say that $m1$ is in a direct neighborhood of a. In addition, the model suggested here incorporates attributes to the relationship between a and $m1$, a rating matrix (Figure 2.9). In this case, the similarity between user Tom and user Bob is higher than the similarity between user Tom and user $Mary$

2.3.2 Word Embedding

Word embedding is a method inspired by deep neural network models to represent the semantic and syntactic similarities between words. It is used in many areas including sentiment

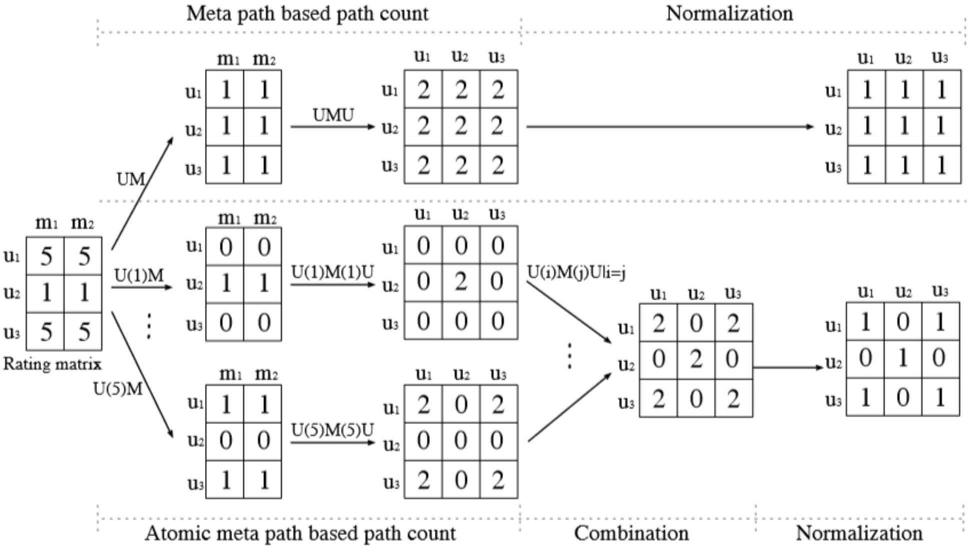


Figure 2.9: Rating matrix between users $u_{1,2,3}$ and movies $m_{1,2}$ [125]

analysis [137] and sentence classification [159]. Figure 2.10 shows one of the models employed by a popular platform Word2Vec. The figure shows two context words before and after the target word being predicted from the popular Skip-gram model. Performance of the training of models using Word2Vec depends on window size and layer size. Window size refers to the number of words before and after the word to extract for the training sample. Table 2.7 shows the training sample for the sample input word "the" and "jumps" with a window size of three [6].

Source Text	Input text	Training Sample
The quick brown fox jumps over the lazy dog	The	The,quick,brown The,quick,fox The,brown,fox
The quick brown fox jumps over the lazy dog	jumps	quick,brown,jumps brown,fox,jumps quick,fox,jumps jumps,over,the jumps,the,lazy jump,over,lazy

Table 2.7: Training sample at window size of three

The more frequent the combination of words occur in the training sample, the more likely the word will be selected. If quick, brown occur more frequently, and if quick is chosen as an

[6] Illustration adapted from http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/ (Accessed 11 Dec 2016)

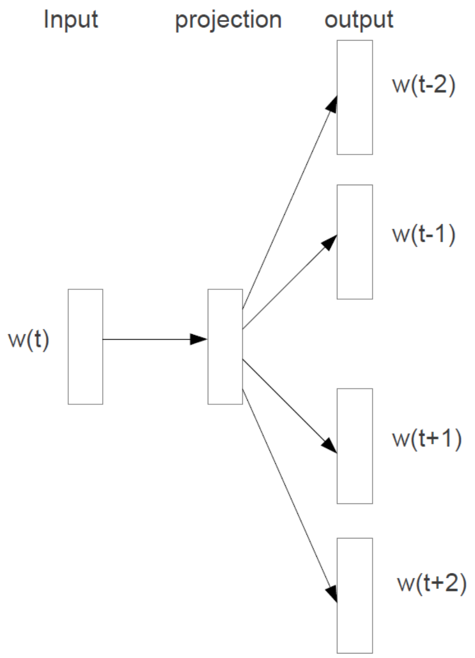


Figure 2.10: Skip-gram model of Word2Vec

input word, then brown will be selected as the nearest word.

Layer size refers to the number of desired features in the word vector. Thus, if we wish to have three features, and there are four words in the vocabulary, the size of the hidden layer matrix will be four rows by three columns:

$$\begin{bmatrix} 0 & 0 & 1 & 0 \end{bmatrix} X \begin{bmatrix} 17 & 24 & 1 \\ 23 & 5 & 7 \\ \underline{4} & \underline{6} & \underline{13} \\ 10 & 12 & 19 \end{bmatrix} = \begin{bmatrix} \underline{4} & \underline{6} & \underline{13} \end{bmatrix}$$

The input vector of a word is a single row vector with all zeros except the word itself. Take the sentence “quick fox jumps over” for example. If we sort them in alphabetical order, the vocabulary list will be $\begin{bmatrix} fox & jumps & over & quick \end{bmatrix}$. Then the input vector for the word “jumps” will be $\begin{bmatrix} 0 & 1 & 0 & 0 \end{bmatrix}$. There are 4 items as we assumed our vocabulary only has 4 words. The word vector obtained from the word2Vec model will be the product of the two matrices.

A recent work by Wang *et al.* [149] used this approach to extract information on DDI from bio-medical corpus. The method was to capture the core meaning of the sentences in the text and incorporate the syntatic contexts into the embeddings. Zhu *et al.* examined the ability of Word2vec in deriving semantic relatedness and similarity between biomedical terms in published articles [164]. It is interesting to note that models that are trained on specific text like abstracts yielded better results than those trained with the main text of the articles.

Liu *et al.* attempted to use word embeddings to capture semantic information of words for the classification of DDI [73]. Word embeddings was also used to exploit the syntactic information of a sentence to extract DDI [162]. Both systems deliver promising performance, notwithstanding that neither are customised to the drug profile of the patient.

2.4 Summary

This chapter has presented a comprehensive review of decision support systems, how drug pairs are identified as unsafe for prescription and the different data mining approaches associated with obtaining the similarity ratio of a drug pair.

Although a CDSS is helpful for the healthcare professional, the review has identified challenges to adoption of such technology, such as perceived usefulness and social factors, as well as some of the key features in current systems that attempt to overcome these challenges. In order to gain acceptance, a personalised CDSS is crucial which hence motivated the research, the results of which has already been used for deployment by dental healthcare professionals. The deployed system differentiates itself from existing systems reviewed in this chapter [8, 27, 34, 68, 105] by integrating with the medical profile of the patient. Based on algorithms created in the research, the system will then advise the user if the drug to be prescribed will interact with the drug the patient is currently taking, or belongs to the same group of drugs to which the patient is allergic.

Much work has been done to discover information on DDI using data mining techniques [14, 138, 155, 157], as it is known that drug similarity can be used to predict drug interaction [27, 34]. In designing the current framework for the research, the study introduces the concept of layers. While reviewing these current systems, the term *knowledge layer* is used to identify the information related to the drugs and all the processing algorithms have been allocated to the *prediction layer*. By adding a presentation layer, the conceptual framework in the research enables knowledge obtained from data mining to be deployed for use. This enables the user to customise such knowledge for drug prescription at point-of-care thereby enhancing the efficiency of treatment by the healthcare professional. While creating features to determine if a drug pair is similar, many methods including the graph-based approach [101] and word-embedding approach [73, 149] have been deployed.

The rest of the thesis further describes each of these models within the three-layer conceptual framework. The incidence of adverse drug interaction as well as associated hospital admission can be reduced [24] when such models are integrated with the medical profile of the patients to support the users in prescribing the drugs through a comprehensive decision support system. Such a system will also relieve the healthcare practitioner from having to rely on search engines like Google or Yahoo which suffer from low recall and precision rates [15] as the results may not be relevant to their needs. Therefore, a CDSS which integrates with drug knowledge bases to identify adverse drug events and advises on drug suitability before prescription will appear helpful to users. With timely and accurate DDI information embedded within a CDSS, more comprehensive treatment options can be made available to patients and practitioners, thus contributing to a more positive treatment experience, better oral health outcomes and job satisfaction for the healthcare professional.

Chapter 3

Knowledge Base and Recognition of Drug Features

The review of the literature in the previous chapter has identified the gap where the prediction of a drug interaction needs to be associated with the drugs that the patient is currently taking and the drugs that the patient is allergic to. This chapter explains the design idea behind the approach to achieve the ability for drug prediction, which will enable a reliable CDSS to be implemented for clinical applications. The conceptual framework is also introduced in this chapter, with description of how data associated with the drugs is selected and transformed into different kinds of features to enable discovery of potential interactions within drug pairs.

3.1 Three-layer Conceptual Framework

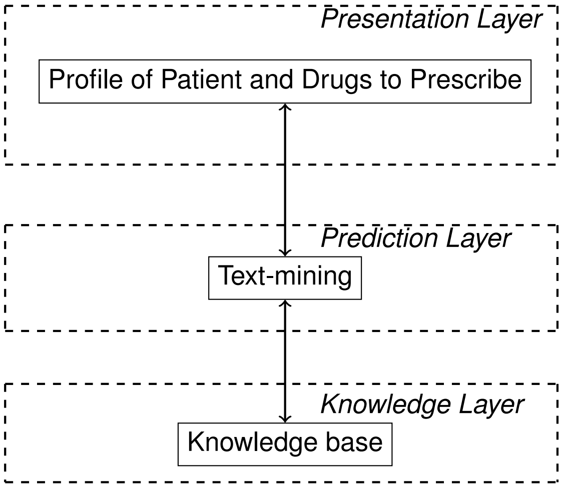


Figure 3.1: Three-layer framework

The aim of the study is to propose a unique approach in supporting the professional user in drug prescription, with consideration of the drugs that the patient is currently taking and the drugs that the patient is allergic to. In order to advise the dentist if the drug to be prescribed is suitable, a three-tier conceptual model was used: the knowledge layer, the prediction layer and the presentation layer (Figure 3.1).

The knowledge layer comprises of the drug properties where the model will need in order to determine if a drug pair is similar and whether there are any adverse interactions between a pair of drugs. Such properties are easily available from a plethora of bio-medical text and knowledge bases.

These drug properties are then transformed into feature vectors in order to determine if a drug pair is similar. These tasks are performed within the prediction layer where our algorithm will predict if the drug pair is suitable to be prescribed to the patient for consumption.

The top most layer is the presentation layer where our model will present the results to the user. Based on the outcome of the prediction layer, this layer will suggest alternative drugs if the drug to be prescribed is found to be unsuitable for the patient. Additionally, the presentation layer also contains the initial drug profile of the patients, consisting of the drugs the patient is currently taking and the drugs that the patient is allergic to. The presence of the drug profile of patients also distinguishes our approach against many other decision support systems as it allows the result of the prediction layer to be customised according to the drug profile of the patients. Hence as shown in Figure 3.1, the model exists in a round trip which starts with the drug profile and the intended prescription in the presentation layer. The prediction layer then co-ordinates data from the presentation layer with information from the knowledge layer to predict the suitability of the drug that is being prescribed. The results are then stored back in the knowledge base and also transmitted back to the user in the presentation layer. Though a good interface is important for the user to adopt the system, it should be noted that the focus of the research is about a system that considers the medical profile of the patient in predicting if a drug is safe for prescription. By separating the framework into three layers, there is a flexibility in the actual implementation as it is not necessary for each tier to be physically located within the same hardware platform. Each can be developed and maintained as an independent tier. This approach also ensures the interfacing between each layer conforms to standards to enable the smooth linkage between them. Moreover, the output can be used for other tasks. For example, by using the word embedding approach in building feature vectors (See Section 3.4.4), the output from the prediction layer can be fed into other neural networks to accomplish other tasks. The following sections describe in more details the functions of each tier of the framework.

3.1.1 Knowledge Layer

The knowledge layer consists of the bio-medical text which described the properties of the drugs. It comprises the domain knowledge in raw data form. Examples of such bio-medical drug database includes DrugBank, KEGG (Kyoto Encyclopedia of Genes and Genomes) and NDF-RT (National Drug File Reference Terminology). DrugBank contains an exhaustive compilation of biochemical properties about drugs. KEGG is a database of metabolic pathways of chemical structures, drug interactions and target molecules. NDF-RT contains a repository of drug interactions sourced from the United States Veteran Administration Data from the database is used to construct a new taxonomy $\mathcal{T}$ relating to interactions and side-effects. It consists of the domain of drugs linked by their semantic relations of “advantageous" and “adverse", and is defined as a 3-tuple $\mathcal{T} := \langle \mathbb{D}, \mathbb{R}, \mathcal{H}_{\mathbb{D}}^{\mathbb{R}} \rangle$, where

- $\mathbb{D} = \{d_1, d_2, ..., d_{|\mathbb{D}|}\}$ is the domain set of drugs;
- $\mathbb{R} = \{r^+, r^-, r^0\}$ is a set of semantic relations, where $r^+(d_i, d_j)$ means that the effects of drugs d_i and d_j are advantageous; $r^-(d_i, d_j)$ means that the effects of drugs d_i and d_j are adverse; $r^0(d_i, d_j)$ means that the effects of drugs d_i and d_j are not related.

- $\mathcal{H}_{\mathbb{D}}^{\mathbb{R}}$ is the taxonomical structure constructed by all $d \in \mathbb{D}$ linked by $r \in \mathbb{R}$. □

With this knowledge base, information is retrieved to using various models so as to determine and predict if the drug is in an adverse relationship with another drug. Features relating to each drug like their potential side-effects can also be obtained.

3.1.2 Prediction Layer

From the drug taxonomy $\mathcal{T}$ text mining was conducted to extract relevant information, where the text for each drug was extracted, cleaned and stored in order to provide information on the underlying properties of a drug pair, enabling the similarity of the drug pair to be computed. The flexibility and robustness of the three-layer framework allows the calculation of drug pair similarity to proceed with various approaches. In the current work, four methods were used: the statistical model, the side-effect model, the adverse network model and the word embedding model as shown in Table 3.1 (see Chapter 4 for detailed description of the models).

Model	**Method**
Statistical model	Frequency-based embedding
Side-effect model	Frequency-based embedding
Adverse network model	Network approach
Word embedding model	Prediction based embedding

Table 3.1: Methods used for drug prediction within prediction layer

The methods are not integrated and different methods result in different models. Experiments are conducted to see how well can each model predict the similarity of a drug-pair. The network approach as well as word embedding methods like frequency-based and prediction-based methods are used to derive feature vectors. These feature vectors are an indication of the probability of an adverse interaction of a drug pair. With such information, these methods can be readily applied to a decision support system to assist dentists in their drug prescription at point-of-care. One of the tools used in the prediction-based method is the Gensim implementation of Word2Vec[1] where similar words can be efficiently obtained. As an illustration, Figure 3.2 shows that not only do related words tend to cluster together, but similar entities are also within close vicinity of each other.

Figure 3.3 shows that the tool is also able to group similar classes of drugs together

[1] https://radimrehurek.com/gensim/models/word2vec.html

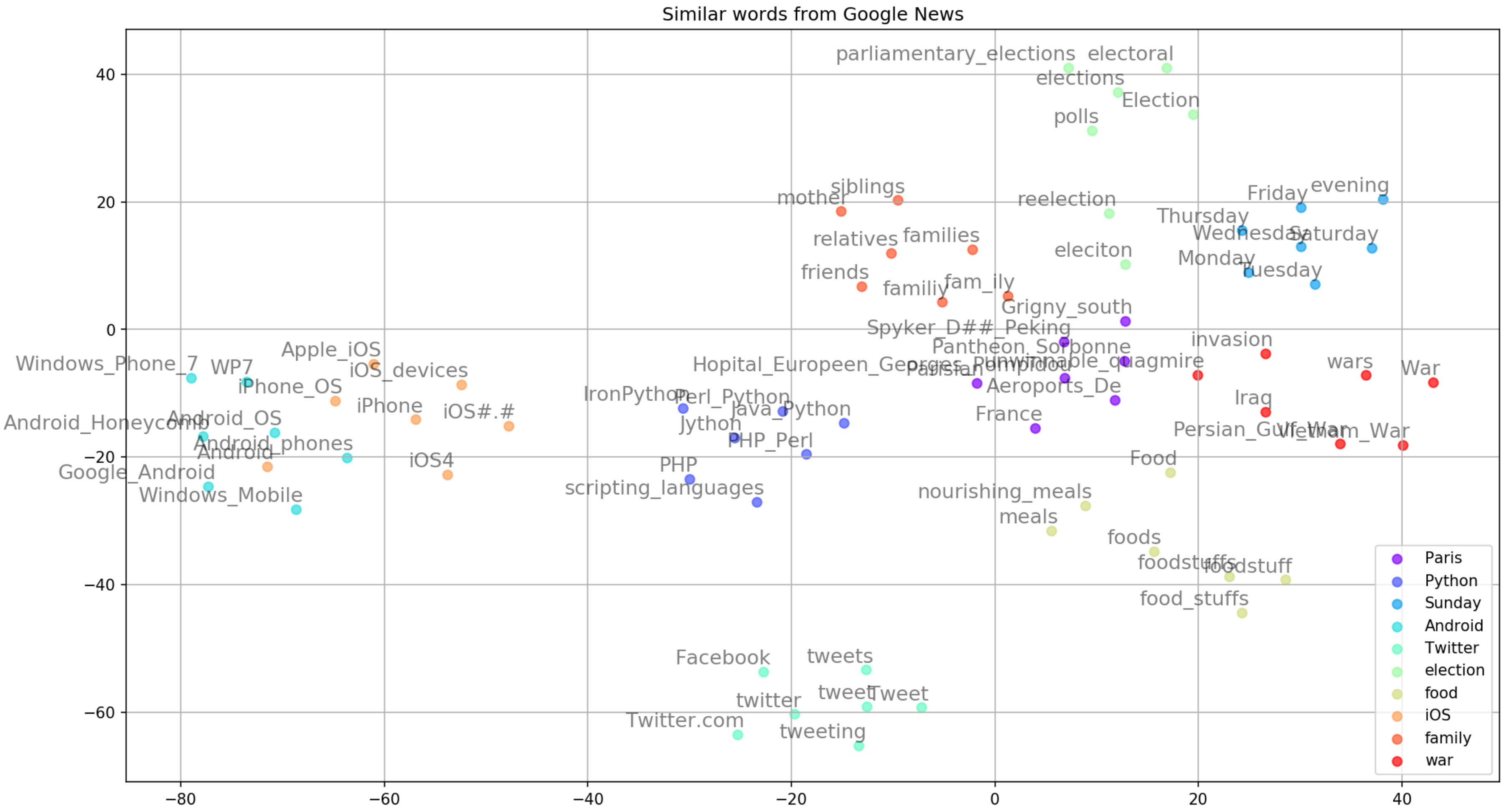


Figure 3.2: Grouping of similar words

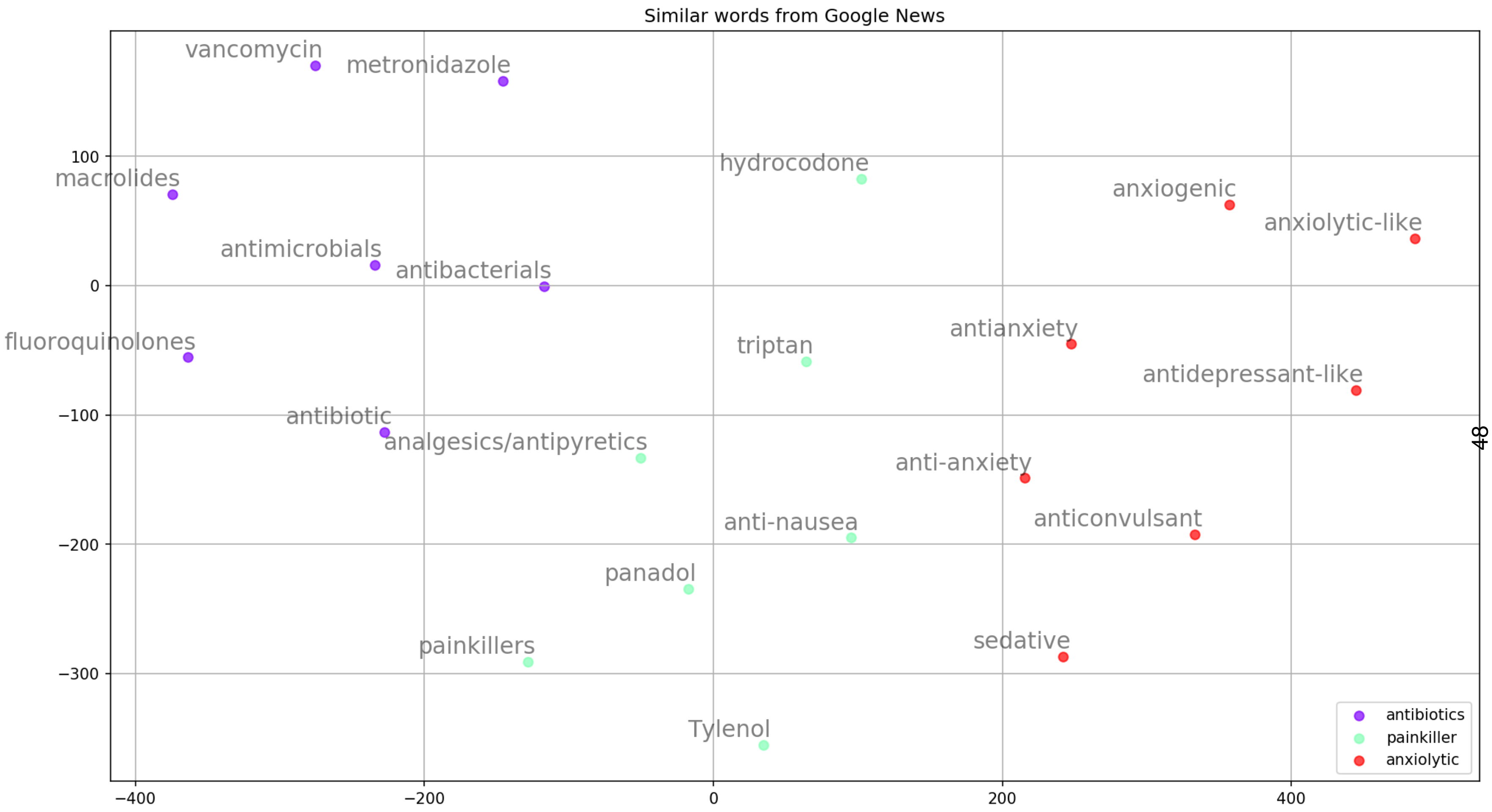


Figure 3.3: Grouping of similar drugs

even though the dataset used to build feature vectors is from Google News, which is not a bio-medical corpus.

3.1.3 Presentation Layer

The layer serves as an interface between the prediction layer and the user. A well-designed user-friendly interface will help dentists adopt such a system in their clinical workflow. As highlighted in Table 3.2, user requirements in the presentation layer need to be efficiently mapped onto the prediction layer to enable useful and relevant information to be extracted for further computing of the similarity ratio.

Presentation Layer	**Prediction Layer**	**Knowledge Layer**
• Efficient mapping of user requirements • User-friendly interface	• Efficient choice of programming approach • Implementation of data mining • Algorithm design	• Bio medical data sources, drug taxonomy • Drug properties

Table 3.2: Features of conceptual framework

The presentation layer also distinguishes the system from many other systems as it contains personalised patient information. In this system, patient $p \in \mathcal{P}$ is a 3-tuple belonging to a set of patients $\mathcal{P}$ where $p := \langle \mathcal{D}, \mathcal{D}^{-}, \mathcal{M} \rangle$ such that

- $\mathcal{D} \subset \mathbb{D}$ is the set of drugs that p is currently taking, where $|\mathcal{D}| <= \theta_{\mathcal{D}}$;
- $\mathcal{D}^{-} \subset \mathbb{D}$ is a known set of drugs that p is allergic to, where $|\mathcal{D}^{-}| <= \theta_{\mathcal{D}^{-}}$ and $\mathcal{D}^{-} \cap \mathcal{D} = \emptyset$;
- $\mathcal{M} = \{m_1, m_2, ..., m_j\} \subset \mathbb{M}$ is a class set of medical conditions (e.g. diabetes, heart problem, lactation, pregnancy, etc.) that p is currently having, where $\mathbb{M}$ is the domain set of all medical conditions and $|\mathcal{M}| <= \theta_{\mathcal{M}}$. □

As shown in Figure 3.4, besides the biographic data of the patient, the presentation layer allows the user of the system to be aware of the drug profile of the patient, providing the healthcare professional crucial information to prescribe the appropriate drug to the patient.

The drugs which the dentist is going to prescribe are also stored in this layer. Such information is needed in the prediction layer for extraction of feature vectors. In order to maintain user-friendliness, which is crucial for clinical adoption of the system, it is important for this layer to present the results in an easy-to-understand format.

Based on the results transmitted from the prediction layer, the service at the presentation layer will then advise the user if the drug to be prescribed is safe for the patient. This approach allows the presentation layer to crystallise the results in a meaningful and friendly manner, so that prescription advice can be presented to the user. Hence, the presentation layer acts as an important supporting tool to the dentist in deciding whether the drug to be prescribed is safe for the patient.

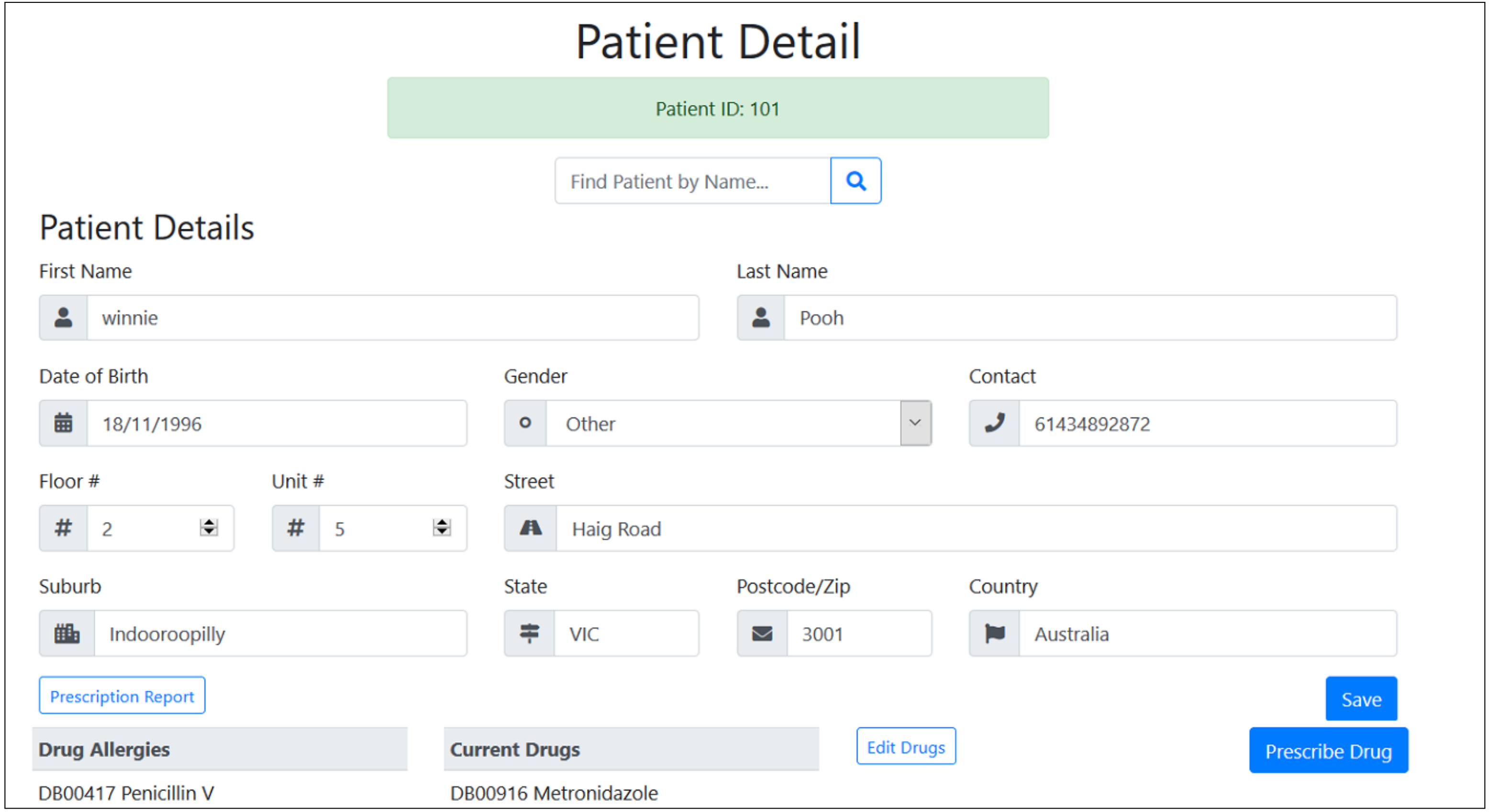
In order to deploy a CDSS within the three-layer conceptual framework, relevant and useful information has to be done on the textual description of the drugs stored in the drug repository. The next section describes how the knowledge base is built to facilitate decision making pertaining to drug prescription.

3.2 Constructing the Knowledge Base

Parsing and cleaning the information embedded within the bio-medical text is required to estabish a knowledge base for downstream application like a decision support system. The initial steps also align with the first two steps of a typical knowledge discovery process (Figure 3.5) [46]:

1. Selection
2. Pre-processing
3. Data transformation (into nodes, edges and tables)
4. Data mining (searching for features)
5. Evaluation
6. Visualisation
7. Decision making

These stages can be iterative where the output of one stage may indicate the need for the previous stage to be refined. Such iterative process within a knowledge discovery process ensures that the desired results at each stage can be obtained by fine-tuning the input to that stage [61, 82]. For example, in the evaluation stage, if the results are not satisfactory, the process of transforming the data can be amended.



Patient Detail

Patient ID: 101

Patient Details

First Name: winnie

Last Name: Pooh

Date of Birth: 18/11/1996

Gender: Other

Contact: 61434892872

Floor #: 2

Unit #: 5

Street: Haig Road

Suburb: Indooroopilly

State: VIC

Postcode/Zip: 3001

Country: Australia

Drug Allergies	Current Drugs
DB00417 Penicillin V	DB00916 Metronidazole

Figure 3.4: User interface

The selection stage involves the choice of database and the kinds of data to be used for the subsequent stages. Typically, knowledge bases that contain relevant data within the domain knowledge is identified. In order for the data to be transformed efficiently, stopwords are removed as they only serve to connect sentences and are not needed contextually. Stemming is also applied to the text to reduce words to their root form.

The knowledge base which provides the domain knowledge forms the backbone in this proposed system. In the context of the research, the domain knowledge will be the information pertaining to the drugs to be prescribed by the user of the system.

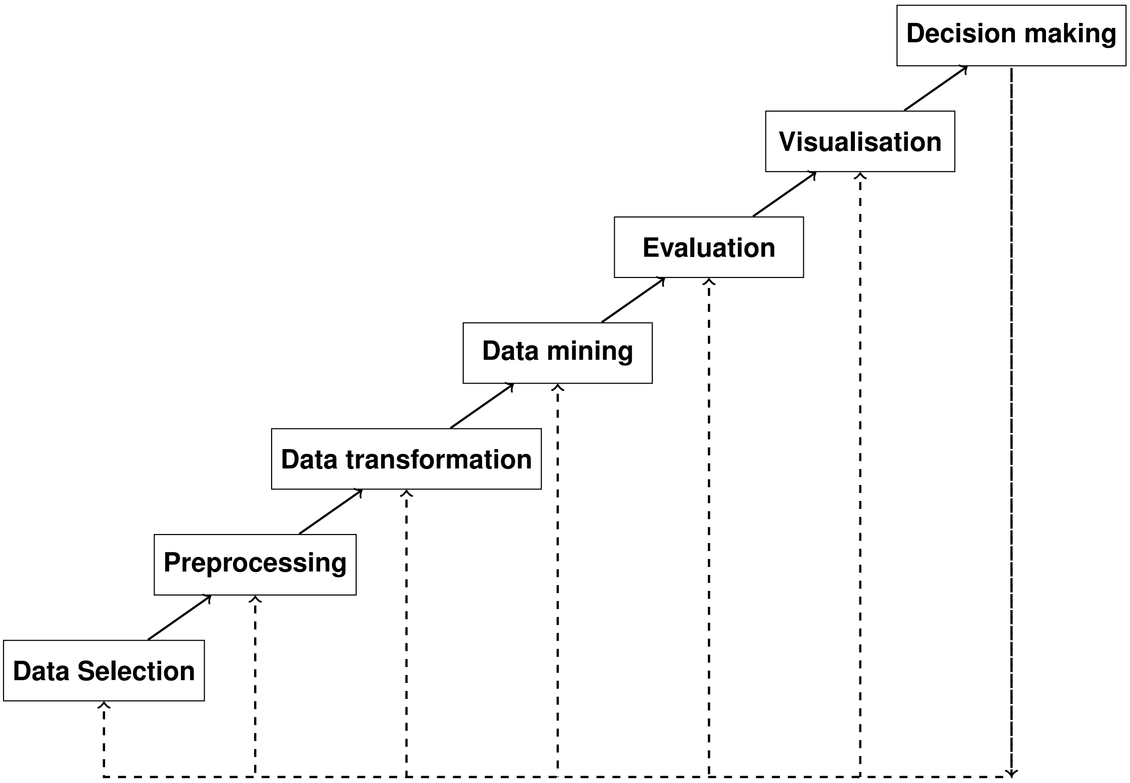


Figure 3.5: Typical knowledge discovery process

The selection of data sources together with pre-processing and data transformation forms the knowledge layer within the conceptual framework (Section 3.1). The layer contains important information relating to drug interactions.

The text comprises of a bag of words from which relevant information was extracted and used for the construction of a network of nodes and edges. Such tasks facilitate subsequent stages in the knowledge discovery process where feature vectors can be extracted and used for computing the similarity ratio of a drug pair.

In this study, the text from DrugBank is used as it is a richly richly annotated database with a wide spectrum of drugs and variety of identifiers including chemical structure and codes from different terminologies [5], and precedent use to build drug repositories on drug indications [110]. References are also provided for drugs that are indexed to other databases like the Kyoto Encyclopedia of Genes and Genomes, which is a collection of information on diseases and chemical substances useful for bio-informatics research and education. Many fields in this database are also hyper-linked to other resources including the RxList database which offers detailed and current pharmaceutical information on drugs useful for prescription and patient education. With a comprehensive corpus of information relevant to both end-users and professionals, information on drug action and drug interaction of FDA-approved drugs can be freely obtained [53].

The textual data that describes each drug in the taxonomy was extracted. An advantage of using DrugBank is that each drug is being described by different properties. Hence, in this study, different properties of each drug were stored for use in the experiment. Each drug is described from a different perspective to suit both patients (under the heading “Overview”) and healthcare professionals (under the heading “Professionals”) while information on side-effects are found under the heading “Side-effects”.

The collection of these drug properties then goes through the pre-processing stage. At this stage, stopwords were removed and words converted to their root form through stemming which enhanced the reliability of the data [124]. The processed data was then used by the different models to classify if a drug-pair is similar. The ability to store interactive drug pairs within a network of nodes and edges allowed the knowledge layer to be represented by a directed acyclic graph (DAG). Each drug was represented as a vertex on the graph. The edges that connect a pair of vertices show the interactions within the drug pair. The cluster of drugs that has adverse interactions with a given drug can be known from the drug taxonomy.

Figure 3.6 shows a subset of the major DDI in the drug taxonomy. Note that nodes in the taxonomy are connected to one another through arrows which indicate that an adverse interaction exists within the drug pair. Each node on the drug taxonomy consisted of a drug with its associated properties. Such a chain of DDI will form the backbone of a drug DAG. As shown in Figure 3.6, *capreomycin* adversely interacts with *adefovir* and *tenofovir*, which interacts with *ibuprofen* and *caffeine* respectively, showing that a given drug may interact adversely with more than one drug. Such information can be stored as a database for subsequent retrieval in building the models for the detection of adverse interactions of a drug pair. Table 3.3 shows the drugs that interacts with *ibuprofen* at different levels.

Major	Moderate	Minor
acetaminophen	atenolol	ampicillin
aspirin	azilsartan	calcium carbonate
adefovir	azithromycin	cascara sagrada
aluminum hydroxide	balsalazide	cimetidine
anisindione	bedaquiline	citric acid
apixaban	medoxomil	colchicine
		dasabuvir
		donepezil
		famotidine
		magnesium hydroxide
		sodium picosulfate

Table 3.3: Database of drug interaction

To facilitate the process of a personalised drug prescription system, a CDSS will be developed so that the user can make an informed choice on the drugs that is to be prescribed to the patient. With new drugs coming to the market every day, such a system will be very helpful to the dentist. As discussed in Section 2.1.2, the system must be user-friendly so that the user is willing to adopt it within their clinical workflow. Though there are many systems available, the system in this study is personalised to the medical profile of the patient. Drugs used within the dental clinic is only a mere fraction of the large variety of drugs available and used in a medical clinic, notwithstanding the fact that the dentist will still need to have knowledge of other drugs. Depending on the medical conditions of the patients, they may be taking other kinds of drugs which the dentist needs to be aware of, in order to avoid adverse interactions with subsequently prescribed drugs.

3.3 Mining Data from Text Corpus

In order to extract relevant information from the text describing each drug, it is necessary to examine how such information is being organised in the text corpus. There are many such bio-medical text corpus such as KEGG and DrugBank. By using text mining techniques through computer algorithms, information can be retrieved, processed and stored in a structured format to allow knowledge on DDI to be discovered. In the case of DrugBank,

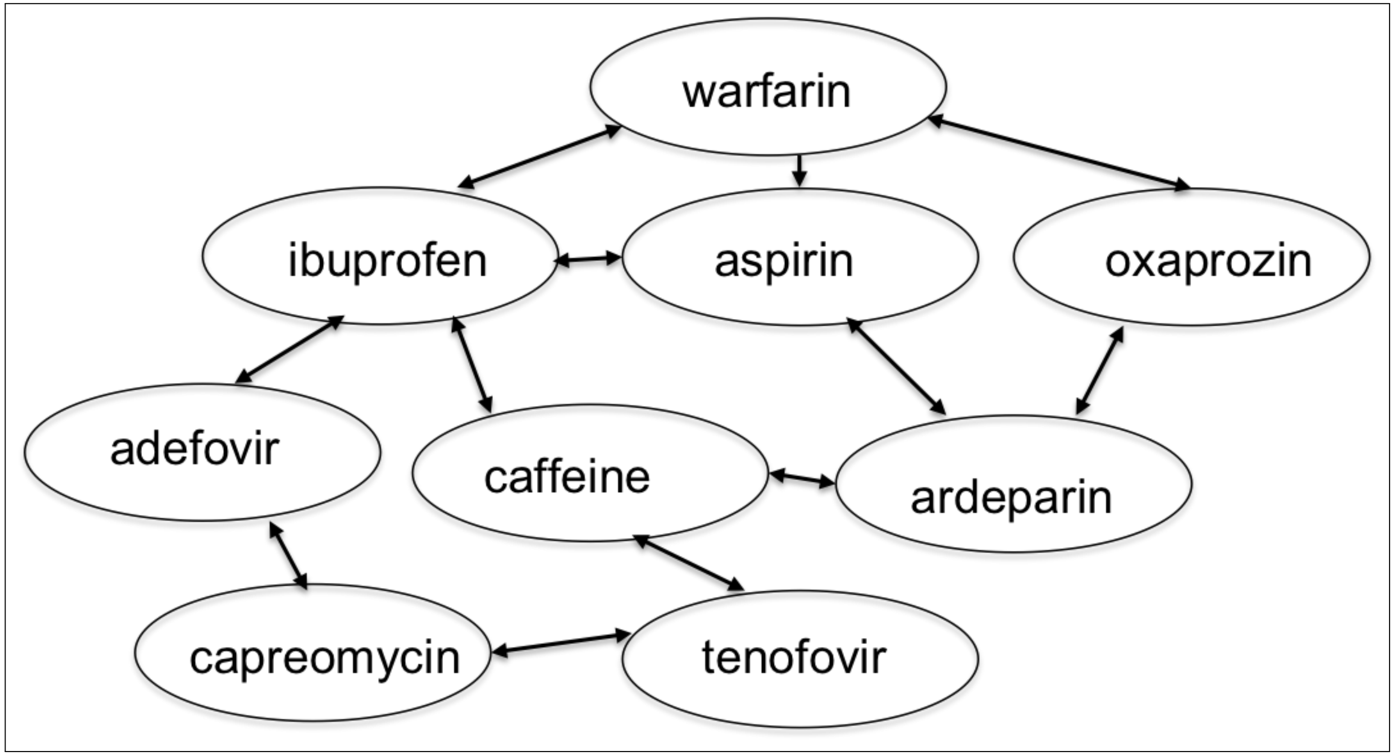


Figure 3.6: Subset of Drug Taxonomy

description of each drug comes under the headings “Overview”, “Professional”, “Side-effects” and “Interactions”. Although both bio-informatics and chem-informatics resources are contained in DrugBank, in the context of this research, only information pertaining to side-effects and adverse interactions are extracted and stored in a local database for the experiment.

The section on “Overview” describes the drugs using easy-to- understand language for lay people. The list here shows the information under “Overview” for the drug *warfarin*:

- *What is warfarin?*
- *Important information*
- *Before taking this medicine*
- *How should I take?*
- *What happens if I overdose?*
- *What should I avoid while taking warfarin?*
- *Warfarin side-effects?*
- *Warfarin dosing information?*
- *What other drugs will affect warfarin?*

Under Side-effects, information is separated into two portions, one for consumers and the other one for healthcare professionals. Under consumers, it lists out all the side-effects in terms of more common, less common and rare. As for healthcare professionals, information on side-effects ar categorised as “hepatic”, “gastrointestinal”, “hypersensitivity”, “dermatologic”, “respiratory”, “cardiovascular”, “metabolic” etc.

Under “Professional”, active and inactive ingredients, indications, dosage and warnings are listed. An image of the label is also included for most of the drugs in the database. Depending on the drug, certain information may be overlapped among the three properties. For example, critical side-effects can be mentioned under “Professional” though the details are listed under “Side-effects”. Text from these three properties of each drug are extracted and stored so that information can be obtained (see next section).

The data obtained on the properties (“Professional", “Overview", “Side-effects") of each drug from the bio-medical text is then transformed to a structured database of side-effects, adverse interactions for each drug as well as textual data of the three properties of each drug. Further information extracted from those three properties include the document frequency and term frequency for each drug. Figure 3.7 shows the transformed information obtained from the drug corpus DrugBank to be used for the experiment. The table of side-effects and table of adverse interactions together with the text on the three properties of each drug serve as a platform for the experiment.

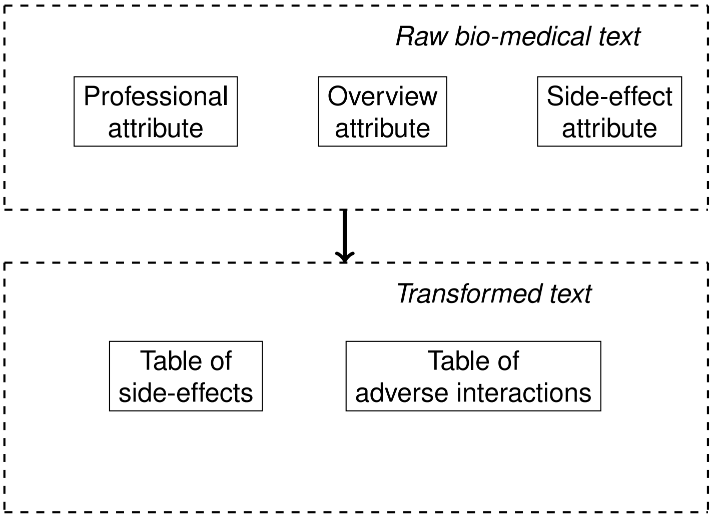


Figure 3.7: Transforming data into structured text

The collection of side-effects is organised in order of drugID where all the side-effects of each drug is stored, along with the intensity, from 0 to 5 (0 being the most common effects which require immediate medical attention and 5 which does not require immediate attention). The table is defined by three fields, drugBankID, the list of side-effects and the level of seriousness (Table 3.4). For example, in Table 3.5, diarrhea and hearing loss are common side-effects, but the former is of major effect and the other one is minor (Table 5.4 shows the codes for the different levels of side-effect).

Field	Type	Size
drugBankID	char	8
side_effect	char	30
level	char	1

Table 3.4: Structure of side-effect table

The other table of structured text mentioned in Figure 3.7 is the table of adverse interactions. Table 3.6 shows the structure of the interaction table after extracting the set of drugs that are in different levels of interaction with the subject drug. For example, Table

drugBankID	side_effect	level
DB01050	abdominal pain	0
DB01050	bloating	0
DB01050	diarrhea	0
DB01050	hearing loss	3
DB01050	nervousness	3

Table 3.5: Sample data for table of side-effects of Ibuprofen

3.7 shows a sample of drugs that are interactive with the drug Ibuprofen and the level of interaction.

Field	Type	Size
drugBankID	char	8
Interactive Drug	char	20
Level	char	1

Table 3.6: Structure of adverse interaction table

drugBankID	Interactive	Level
DB01050	warfarin	major
DB01050	cidofovoir	major
DB01050	leflunomid	major
DB01050	metformin	moderate
DB01050	famotidine	minor
DB01050	tacrine	minor

Table 3.7: Sample data for table of adverse interactions with ibuprofen

The next two sections explain the concept of similarity ratio within the context of this study and describes the different models used in discovering drug interactions.

3.4 Recognition of Drugs through Feature Extraction

After the raw bio-medical textual data has been processed and assembled into a structured database, the proposed models will then be able to determine if the drugs are similar through their textual description, and the extent of their similarity. This section describes the text processing methods by representing the drugs as feature vectors to allow the similarity of a drug pair to be determined.

To determine the similarity of a drug pair, data mining is done by classifying a drug pair as similar or dissimilar through extracting features of the drugs. The purpose of data mining is to discover useful information from the property of the drug. Just like mining for gold, the aim is to search the mass of ores systematically to look for the hidden treasure. The efficiency and accuracy of the search depends on the techniques used which have evolved over the years to increase the chance of discovering the gold in the most efficient manner.

Similarly, there are many methods for obtaining information on the description of a drug. Such information, or feature, of a drug pair can then be used to compute the similarity between them. This section introduces the frequency-based approach, the prediction-based approach and the network approach used in the models within the conceptual framework (See Table 3.1).

3.4.1 Bag of Words

The most basic method is the bag of words where an unordered list of individual words is extracted from the document. For example, take three simple documents c_1,c_2 and c_3 each with a single sentence as shown in Table 3.8.

Document	**Sentence**
Document c_1	Warfarin is used to treat or prevent blood clots in veins
Document c_2	Amoxicillin is used to treat infection caused by bacteria
Document c_3	Ibuprofen is used to treat pain caused by toothache

Table 3.8: Sample documents

Before counting the frequency of each word, some basic pre-processing of the text such as stop words removal and stemming is performed. By carrying out such tasks will result in eliminating common words (like “the” and “is”) mapping associated words to their root form (like “used” to “use”).

document term	c_1	c_2	c_3
warfarin	1	0	0
use	1	1	1
treat	1	1	1
prevent	1	0	0
blood	1	0	0
clot	1	0	0
vein	1	0	0
amoxicillin	0	1	0
infect	0	1	0
cause	0	1	1
bacteria	0	1	0
ibuprofen	0	0	1
pain	0	0	1
toothache	0	0	1

Table 3.9: Matrix of count vector

Table 3.9 shows the number of times each word occur in each document after pre-preprocessing is done. Such information can be represented as a matrix M of n rows by c columns. The column corresponds to the documents in the corpus and the rows correspond to the words in the vocabulary. If $b(w, c)$ is the number of times the word *w* occurs in document *c*, then each item $b_{w,d}$ can be represented as a matrix M as shown in Equation 3.1.

$$M = \begin{bmatrix} b_{0,0} & b_{0,1} & \cdots & b_{0,c-1} \\ b_{1,0} & b_{1,1} & \cdots & b_{1,c-1} \\ \vdots & & \ddots & \\ b_{n-1,0} & b_{n-1,1} & \cdots & b_{n-1,c-1} \end{bmatrix} \tag{3.1}$$

where *n* is the size of the vocabulary and *c* is the number of documents in the corpus. From Equation 3.1, $b(0, 0)$ refers to the number of occurrences for *warfarin* at the first row in document c_1. The word vector for *warfarin* $b_{0,*}$ can then be denoted as $[b(w_0, c_1), b(w_0, c2), b(w_0, c3)]$.

Although this is a straightforward way of coding the words, it does not capture the relationship among the words and hence the meaning behind the sentence. For example, “this is interesting” will be encoded in the same way as “is this interesting”. Another problem with this method is the inability to distinguish between important words and words such as ’is’ and ’the’ which also occur frequently.

3.4.2 Term Frequency Inverse Document Frequency (tf*idf)

An extension of the previous bag-of-words method is to consider word frequency not just in a single document but across the entire corpus. By doing this, words that occur frequently in all documents can be given less weight compared to words that occur frequently within a single document. Instead of just keeping track of the raw number of occurrences of each word in the count matrix as in Equation 3.1, *tf*idf* can be used where the frequency is normalised to take into account the number of occurrences throughout the corpus:

$$\begin{aligned} tf &= \frac{b(w,c)}{\sum\limits_{w' \in c} b(w',c)} \\ idf &= log\frac{|C|}{|\{c \in C : w \in c\}|} \end{aligned} \tag{3.2}$$

where $|C|$ is the total number of documents in the corpus and $|\{c \in C : w \in c\}|$ is the number of documents where the word *w* appears.

Thus if v represents the *tf*idf* vawaslue for a word w in the corpus, then

$$\begin{aligned} v &= tf * idf \\ &= \frac{b(w,c)}{\sum\limits_{w' \in c} b(w',c)} \times log\frac{|C|}{|\{c \in C : w \in c\}|} \end{aligned} \tag{3.3}$$

Using the same example as for bag-of-words, since *amoxicillin* occurs once in document c_2, the term frequency is $\frac{1}{6}$ which is 0.167 as there are 6 words in the document c_2. Since only document c_2 contains *amoxicillin* and there are 3 documents in the corpus, the $tf * idf$ for *amoxicillin* is $0.167 * log\frac{3}{1}$ which is 0.0795.

Compared to the word *cause* which also occurred once in c_2 but twice in the corpus, the $tf * idf$ for *cause* is $\frac{1}{6} * log\frac{3}{2}$. This goes to show that it is 2.7 times less important than the word *amoxicillin*.

This illustration shows that if a word in a document appears frequently in other documents across the corpus, it is likely to be less relevant to the document. On the other hand, the word will be of more relevance if it occurs more frequently within the document.

By using Equation 3.3, the *tf*idf* for each word in the textual description of drug d_i can be obtained. These values can be used in the model that uses *tf*idf* as feature vectors (Section 4.2).

If the words that describe drug d_i are $w_{1i}, w_{2i} \dots w_{ni}$ where n is the total number of words and their respective *tf*idf* from Equation 3.3 is $v_{1i} v_{2i} \dots v_{ni}$ then the feature vector of drug d_i is given by:

$$\overrightarrow{f_i} = \{(w_{1i}, v_{1i}), (w_{2i}, v_{2i}), \dots (w_{ni}, v_{ni})\} \tag{3.4}$$

Once the feature vector $\overrightarrow{f_j}$ is similarly obtained for drug d_j, the similarity ratio of drug pair d_i and d_j is:

$$Sim(d_i, d_j) = \frac{\sum\limits_{k=1}^{n} (v_{ki}) \times (v_{kj})}{\sqrt{\sum\limits_{k=1}^{n} (v_{ki})^2} \times \sqrt{\sum\limits_{k=1}^{n} (v_{kj})^2}} \tag{3.5}$$

Hence the term frequency of the words that describe each drug can be used to build feature vectors for determining if a drug pair is similar.

3.4.3 Co-occurrence

The concept of contextual distance is used in this approach to cater for the contextual meaning of words within a document. Instead of counting the occurrence of individual words like the previous methods, words within a distance is counted. Table 3.10 shows the occurrence matrix for the corpus of Table 3.8 representing the number of times the words occur together within the context distance of two words before and after the target word.

	Amoxicillin	use	treat	infect	cause	bacteria
Amoxicillin	0	1	0	0	0	0
use	1	0	1	0	0	0
treat	0	1	0	1	1	0
infect	0	0	1	0	1	0
cause	0	0	1	1	0	1
bacteria	0	0	0	0	1	0

Table 3.10: Co-occurrence matrix

Take for example the target word “cause” in document c_2. Within a context distance of two, each of these words “treat”, “infect”, “bacteria” occurred once as indicated in the row beginning with “cause” in Table 3.10. Thus, the row matrix for “cause” can be represented as $\begin{bmatrix} 0 & 0 & 1 & 1 & 0 & 1 \end{bmatrix}$. With each word in the textual description of the drugs in the bio-medical database being represented as a co-occurrence matrix, a knowledge base associated with each drug can be constructed. This will facilitate subsequent building of the drug model to determine their similarity ratio.

3.4.4 Word Embedding

The previous three approaches were frequency based, using word frequency to build vectors. Those methods derived a matrix of numbers from the original corpus by counting the number of occurrences of the word, either individually or in association with its neighbouring words. These approaches are memory-intensive and an inefficient means of storing the vectors. The size of the vectors is the same as the size of the vocabulary, where a corpus of a million words will result in a matrix with a million numbers. On the other hand, a prediction-based approach will predict the probability of a word given the target word. Feature vectors created by this approach proved to be superior for tasks like word similarities and word analogies. This approach is used in one of the models for predicting the similarity of a drug pair. Explanation of this approach is described in Section 2.3.2.

In the current study, frequency-based and word-embedding approaches were used in deciding if a drug pair was suitable for prescription (Table 3.1). According to linguist JR Firth (as cited by Nguyen [96]), “you shall know a word by the company it keeps”. Similarly, related words and hence similar drugs can be known by finding similar words that describe the drugs.

When these words are converted into vectors, the similarity ratio of a drug pair can be easily obtained for comparison.

3.4.5 Common Paths in Adverse Network

Besides using word frequency and word embedding methods to extract features and represent the drugs, a path-oriented approach is also possible. Information about drug interactions can be represented as a network of nodes and edges. Such a network is a data structure with information about the nodes and their relationships. In mathematical terms, this data structure is described as a graph *G = (V,E)* with information on the nodes and the connections between the nodes [67]. In such a graph, the nodes are referred to as vertices and the connections as edges. The method is described in Section 2.3.1. A drug pair is highly similar if the number of common paths between them is large.

3.5 Summary

This chapter introduced the three-layer conceptual framework of the thesis.

The knowledge base within the knowledge layer is constructed based on the features obtained from the bio-medical corpus. Table 5.9 in Section 5.3.6 shows the number of features obtained from drugBank. This process also aligns with the first three stages of a typical knowledge discovery process. In order to obtain relevant information from the knowledge base, different approaches have been described, from finding out how often words occur to how likely those words would occur. These approaches in drug representations through vectors enable the discovery of drug interactions. Such information makes it possible to determine potential adverse interactions of a drug pair. The next chapter will describe the different approaches in obtaining feature vectors through the various models within the prediction layer.

Chapter 4

Discovering Drug Interactions through Data Mining

The previous chapter described how the knowledge base is built from information associated with drug properties available in DrugBank. The various forms of feature extraction from the text was also explained. The different feature extraction methods results in a corresponding model within the prediction layer. This chapter proposes the novel approaches within the prediction layer of the conceptual framework for discovering the possibility of a drug pair interaction:

- statistical model
- side-effect model
- adverse network model
- word embedding model

These models use data mining and evaluation which aims to discover patterns and meanings from the knowledge base. Obtaining feature vectors to compute similarity ratio then leads to the knowledge required for the user to decide if a drug is safe for prescription.

4.1 Cosine similarity for Vector Space Models

Cosine similarity was used to obtain the features of the drugs from their vector representations. The process of obtaining the cosine similarity starts with the dot product. Assuming $\overrightarrow{f_i}$ and $\overrightarrow{f_j}$ contains the respective set of feature vectors for drug d_i, d_j and $\{v_1, v_2..v_i\}$, $\{v_1, v_2...v_j\}$ are the respective values for the feature vectors $\overrightarrow{f_i}$, $\overrightarrow{f_j}$ stored in the respective array m, n.

Thus array m has items $\{v_1, v_2..v_i\}$ and array n has items $\{v_1, v_2..v_j\}$.

For illustration, assume an equal number of items in both feature vectors, i and j having the same value,

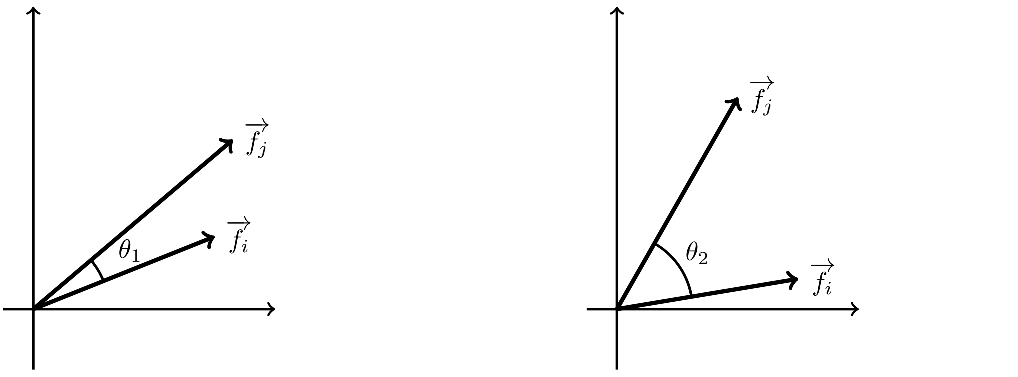
$$\begin{aligned} m &= \{v_{m1}, v_{m2} \dots v_{mi}\} \\ n &= \{v_{n1}, v_{n2} \dots v_{nj}\} \end{aligned} \tag{4.1}$$

Then the dot product will be

$$\begin{aligned} \overrightarrow{f_i} \cdot \overrightarrow{f_j} &= \sum_{p=1}^{j} v_{mp} v_{np} \\ &= v_{m1}v_{n1} + v_{m2}v_{n2} + \dots v_{mj}v_{nj} \end{aligned} \tag{4.2}$$

and the geometric definition of the dot product is given by

$$\overrightarrow{f_i} \cdot \overrightarrow{f_j} = |\overrightarrow{f_i}||\overrightarrow{f_j}| cos\theta \tag{4.3}$$



(a) More similar textual description

(b) Less similar textual description

Figure 4.1: Cosine similarity of documents

Re-arranging equation 4.3,

$$cos\theta = \frac{\overrightarrow{f_i}.\overrightarrow{f_j}}{|\overrightarrow{f_i}||\overrightarrow{f_j}|} \tag{4.4}$$

The angle θ represents the similarity between the two vectors $\overrightarrow{f_i}$ and $\overrightarrow{f_j}$. Depending on the model used during the experiment, various methods were used to obtain the feature vectors $\overrightarrow{f_i}$ and $\overrightarrow{f_j}$. For example, if feature vectors are based on the textual similarity of the documents that describe the drugs, then the feature vectors of two documents will be similar if both documents contain similar terms. In other words, the angle θ_1 between the two vectors $\overrightarrow{f_i}$ and $\overrightarrow{f_j}$ in the vector space will be small, as both are heading closely in the same direction (Figure 4.1a). Conversely, if the drug pair d_i and d_j contains more dissimilar terms, then their respective vectors $\overrightarrow{f_i}$ and $\overrightarrow{f_j}$ will be heading at a larger angle (Figure 4.1b) resulting in a smaller cosine similarity since cos θ_2 is lesser than cos θ_1 when θ_2 is larger than θ_1. In fact, if there are no common terms, the two vectors will be perpendicular to each other or orthogonal which results in zero similarity ratio since cos 90° is zero.

4.2 Discovering Drug Interactions Through Word Frequencies

During feature extraction in this model, noise was first removed from the unstructured textual corpus obtained from DrugBank. Such noise refers to paltry terms like 'the', 'a' which may affect the performance of the model. These stopwords were removed while similar words were stemmed with the help of Porter's algorithm [131]. The quote by Tobler (as cited in Sen [120]) that "Everything is related to everything else, but near things are more related than distant things" can be applied not just to spatial similarity but also to textual similarity. Since it is expected that similar drug pairs are described by more similar terms, the statistical values of term frequencies and inverse document frequency can be used to determine the similarity ratio of a drug pair.

Given that each drug has *k* terms each with their *tf*idf* computed, the task of the model within the prediction layer of the framework was to construct feature vectors for each attribute of the drug. This feature vector comprised of a set of pairs of keywords and their respective $tf * idf$.

The set of terms $\{t_1, t_2, t_3...t_n\}$ extracted for each drug to enable similarity to be computed consisted of a bag of unordered terms, defined as:

$$\begin{aligned} T_i^v &= \{t_1^v, t_2^v, ...t_x^v\} \\ T_i^p &= \{t_1^p, t_2^p, ...t_y^p\} \\ T_i^s &= \{t_1^s, t_2^s, ...t_z^s\} \end{aligned} \tag{4.5}$$

where T_i^v,T_i^p,T_i^s refer to the set of terms of drug d_i within the attribute “Overview”, “Professional” and “Side-effects” respectively and x, y, z are the respective number of terms.

It was then easy to obtain the term frequencies for each set of terms by constructing feature vectors:

$$\begin{aligned} \overrightarrow{f_i^v} &= \{(t_{1i}^v, v_{1i}^v), (t_{2i}^v, v_{2i}^v), ...(t_{xi}^v, v_{xi}^v)\} \\ \overrightarrow{f_i^p} &= \{(t_{1i}^p, v_{1i}^p), (t_{2i}^p, v_{2i}^p), ...(t_{yi}^p, v_{yi}^p)\} \\ \overrightarrow{f_i^s} &= \{(t_{1i}^s, v_{1i}^s), (t_{2i}^s, v_{2i}^s), ...(t_{zi}^s, v_{zi}^s)\} \end{aligned} \tag{4.6}$$

Similarity within a drug pair for an attribute was computed by comparing common terms within that attribute. For example, the similarity ratio $Sim(d_i, d_j)$ within the attribute “Professional” for drug d_i and drug d_j was obtained by comparing these two feature vectors, with each feature vector sorted in descending order of the size of the term frequency:

$$\overrightarrow{f_i^p} = \{(t_{1i}^p, v_{1i}^p), (t_{2i}^p, v_{2i}^p), ...(t_{ni}^p, v_{ni}^p)\} \tag{4.7}$$

such that $v_{ni}^p >= v_{(n+1)i}^p$

$$\overrightarrow{f_j^p} = \{(t_{1j}^p, v_{1j}^p), (t_{2j}^p, v_{2j}^p), ...(t_{nj}^p, v_{nj}^p)\} \tag{4.8}$$

such that $v_{nj}^p >= v_{(n+1)j}^p$ where n is the size of each feature vector.

Since the terms within the set T_i and T_j (Equation 4.5) are unordered, the corresponding values in the feature vectors f_i and f_j are also unordered (Equation 4.6). To ensure a consistent contribution of the term frequencies to the similarity ratio between a drug pair, the feature vectors are sorted to capture the highest n set of values of the vector for each drug.

From Equation 4.7 and Equation 4.8, the similarity ratio $Sim^p(i, j)$ of drug pair d_i and d_j

can be obtained.

$$Sim^p(d_i, d_j) = \frac{\sum_{k=1}^{n} (v_{ki}^p) \times (v_{kj}^p)}{\sqrt{\sum_{k=1}^{n} (v_{ki}^p)^2} \times \sqrt{\sum_{k=1}^{n} (v_{kj}^p)^2}} \tag{4.9}$$

The similarity ratio obtained from a drug pair was used to decide if the drug pair was similar. For example, if $Sim^v(i,j)$ is the similarity ratio between feature vectors $\overrightarrow{f_i^v}$ and $\overrightarrow{f_j^v}$ taken from drug property “Overview”, then the number of similar drug pairs that were correctly predicted as similar can be found by counting the number of similar pairs. The number of true positives and true negatives were then used to compute the F-score. A drug pair was considered to be similar if $Sim^v(i,j)$ was above a threshold value θ which occurs at F_{max}^v. Different cut-off points will result in different number of true positives which will influence the *F*-score. The threshold value is the value where the *F*-score is maximum.

Thus if $S(i,j)$ is used to represent the number of true positives which refer to those instances when the similarity ratio is above θ, then

$$S(i,j) = \begin{cases} 0, & \text{if } Sim^v(i,j) < \theta \\ 1, & \text{otherwise} \end{cases} \tag{4.10}$$

To obtain the overall gross similarity ratio, similarity ratio of the drug pair associated with each attribute ("Professional, “Overview” and “Side-effect”) is taken into consideration. Depending on how accurately the similarity of each drug pair was predicted, the similarity associated with each attribute was normalised by a factor depending on the F-score. Thus if F_{max}^v, F_{max}^p and F_{max}^s is the maximum F-score for drug attribute “Overview”, “Professional” and “Side-effect” respectively, then the weight w_1 against the similarity ratio for “Overview” is given by:

$$w_1 = \frac{F_{max}^v}{(F_{max}^v + F_{max}^p + F_{max}^s)} \tag{4.11}$$

w_2 and w_3 can also be calculated in a similar manner.
Thus the overall similarity ratio $Sim(p,q)$ for drug pair d_i and d_j is given by:

$$Sim(i,j) = w_1 * Sim^v(i,j) + w_2 * Sim^p(i,j) + w_3 * Sim^s(i,j) \tag{4.12}$$

where
Sim^v is the similarity ratio for the drug property from “Overview”
Sim^p is the similarity ratio for the drug property from “Professional”

Sim^s is the similarity ratio for the drug property from “Side-effects”

If the similarity ratio was above the threshold value, the service in the prediction Layer classified it as *similar*, else it was classified *dissimilar* (Equation 4.10).

4.3 Discovering Drug Interactions Through Side-effects

Using the same text corpus as the statistical model, the first step in this model was to extract the side-effects within the “Side-effect” attribute of each drug. These were stored in a knowledge base for easy retrieval during the experiments. In order to indicate the impact of the side-effects, the attribute was further categorised into major and minor categories, each of which can be common or rare.

Let $\mathbb{E}$ =$\{e_1, e_2 \ldots, e_r\}$ be the set of all possible side-effects and E_i, E_j are the set of side-effects of drug d_i and d_j respectively where $E_i \subset \mathbb{E}$, $E_j \subset \mathbb{E}$ and r= $|\mathbb{E}|$. Then a row matrix M can be used to represent the presence of the side-effects, depending on the position of the side-effects within the set of all side-effects $\mathbb{E}$:

$$M[i] = \begin{cases} 1, & \text{if } e_i \neq 0 \\ 0, & \text{otherwise} \end{cases} \tag{4.13}$$

Consider the case of a drug pair d_i and d_j with the set of side-effects E_i and E_j where $E_i = \{e_3, e_5\}$, $E_j = \{e_5, e_8\}$ and $e_3, e_5, e_8 \in \mathbb{E}$. Then we can have 2 row matrices M_i and M_j to represent the presence of the side-effects for the drug pair. The value of each element can be either 1 or 0. Thus $M_i[3]$ = 1, $M_i[5]$ = 1 and $M_j[5]$ = 1, $M_j[8]$ = 1. From M_i and M_j, similarity ratio $Sim(d_i, d_j)$ within the drug pair d_i and d_j can then be calculated. The drug pair can then be classified as similar or dissimilar if this ratio was above or below the threshold respectively (Equation 4.10). Hence, for a set of drug pairs $\{d(i_1, j_1), d(i_2, j_2)...d(i_n, j_n)\}$ that are supposed to be similar, the number of true positives which are drug pairs predicted as similar is given by:

$$\sum_{k=1}^{n} S(i_k, j_k) \tag{4.14}$$

4.4 Discovering Drug Interactions from Common Paths of Adverse Network

As described in Section 3.1, the drug taxonomy contains knowledge on drug interactions in the knowledge layer of the three-layer framework. Detailed attributes are extracted at the prediction layer where useful information can be obtained in relation to the set of drugs that interacts with each drug.

The performance of the model was based on the hypothesis that the larger set of common interacting drugs between a drug pair, the higher their similarity ratio. This lead to the development of the adverse network model to facilitate the finding of the number of common paths within a drug pair.

The approach introduced by Jeh *et al.* [58] in their model of measuring similarity based on theoretical foundations was adopted to represent all drugs as nodes in a network, enabling computation of their proximity in terms of the number of shared entities within the drug pair.

Assuming $\{d_{i1}, d_{i2}...d_{ik}\}$ refers to the set of interacting drugs of d_i, then $O(d_i)$ is known as the set of out-neighbours of d_i with d_i as the vertex node.

If drug d_i and d_j are represented as node i and node j respectively, then similarity within the drug pair will be given by:

$$Sim(d_i, d_j) = \frac{C1}{|O(d_i)||O(d_j)|} \sum_{i=1}^{|O(d_i)|} \sum_{j=1}^{|O(d_j)|} Sim(O_i(d_i), O_j(d_j)) \tag{4.15}$$

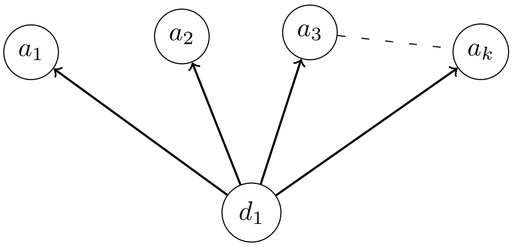


Figure 4.2: Graph of interactive drugs with drug d_1

Referring to Figure 4.2, if we consider the out-neighbours, drug d_1 will have a set of interacting drugs which can be denoted as $N^+(d_1)$ where the number of drugs $a_1, a_2, a_3...a_k$ that adversely interact with drug d_1 is $k = |N^+(d_1)|$. Individual drugs in $N^+(d_1)$ which interact with d_1 are then denoted by $N_i^+(d_1)(1 \leq i \leq |N^+(d_1)|)$.

The path connecting two nodes indicates the relationship within a drug pair. Adapting the notation from [125], the path from drug d_1 to set of interactive drugs $N^+(d_1)$ (denoted by A) with ratings r is denoted by $d_1 \xrightarrow{\text{r}} A$, which can also be written as $d_1(r)A$, where r is the relationship between d_1 and A. The relationship can be major, moderate or minor represented as 1,2 or 3 respectively. Thus, $d_1(1)A$ shows drug d_1 has a minor interaction with the drugs in set A.

Using the notation from the previous chapter, if $N^+(d_i)$ is found in A^r at position u, then the u^{th} column in matrix M^r will be updated as r, $M^r[u] \longleftarrow r$.

Assuming a major interaction with $r = 3$, if the interactive drug of d_1 occurs at position 5 and the interactive drugs of d_2 occurs at positions 2 and 5, then $M_1^3[5] = 3$, $M_2^3[2] = 3$ and $M_2^3[5] = 3$.

In this manner, the similarity ratio of the drug pair in the dataset can be computed from the pair of row matrices taken from M_1 and M_2, such that M_1= $\begin{bmatrix} a_1 & a_2 & \dots & a_p \end{bmatrix}$ and M_2= $\begin{bmatrix} b_1 & b_2 & \dots & b_p \end{bmatrix}$ where the value of each item in the matrix is:

$$\begin{aligned} a_u &= \begin{cases} r, & \text{if } N^+(d_1) == A^r[u] \\ 0, & \text{otherwise} \end{cases} \\ b_u &= \begin{cases} r, & \text{if } N^+(d_2) == A^r[u] \\ 0, & \text{otherwise} \end{cases} \end{aligned} \tag{4.16}$$

Hence for a drug pair d_i and d_j with interaction rating of r, the similarity ratio will be given by:

$$Sim^r(d_i, d_j) = \frac{\sum\limits_{i=1}^{p} a_i \times b_i}{\sqrt{\sum\limits_{i=1}^{p} a_i^2} \times \sqrt{\sum\limits_{i=1}^{p} b_i^2}} \tag{4.17}$$

Computation of the similarity ratio then facilitates drug interaction predictions.

4.5 Discovering Drug Interactions Through Word Embeddings

This framework used the skip-gram model which is commonly utilised for learning word embeddings by predicting context words given a target word. Features of context words were then extracted through word embeddings. This method of finding the similarity within a drug pair was used due to its increasing popularity in machine learning. Since it is expected that a higher set of common terms will be used to describe two drugs that are similar in functions, it was a sensible decision to measure the similarity between drugs in a drug pair by finding words that are most related to each of these drugs.

Such an approach to machine learning has already had major impact in many areas involving a large amount of data - such as medical imaging, speech recognition and natural language processing - and is very relevant considering the constant increase of drug-related bio-medical information [152]. Interest in word embedding has intensified with Mikolov *et al.*'s introduction of a simplified architecture, which eliminates the non-linear hidden layer, allowing training on much larger datasets than was previously possible [85].

Instead of using the set of interactive drugs as in the previous model, data from the text corpus was used in this model to compute the similarity within a drug pair. With the help of Word2Vec [85], tokens were then built by iterating through the sentences in the textual corpus, specifying parameters like minimum word frequencies and the size of the feature vectors. While Word2Vec is not strictly a deep neural network, the output vector that it produces in numerical format within the deep learning models can be easily understood by other deep networks making it very suitable for use in such works. Assuming w_c and w_t are the context word and target word, the goal of the skip-gram model is to maximise the log-likelihood of obtaining the output context word given the input target word, ie,.

$$J = logP(w_c|w_t) \tag{4.18}$$

where J is the objective function.

Suppose u_{wt} is a target embedding vector for w_t and v_{wc} is a context embedding vector for the context word w_c, then P, which is the conditional probability in the neural probabilistic language model [6] can be defined as:

$$P(w_c|w_t) = \frac{exp(v_{wc}^T u_{wt})}{\sum\limits_{w=1}^{W} exp(v_{wc}^T u_{wt})} \tag{4.19}$$

Taking the log on both sides of Equation 4.19 above,

$$logP(w_c|w_t) = log\frac{exp(v_{wc}^T u_{wt})}{\sum\limits_{w=1}^{W} exp(v_{wc}^T u_{wt})} \tag{4.20}$$

Since $J = logP$ (from Equation 4.18), Equation 4.20 becomes:

$$J = log\frac{exp(v_{wc}^T u_{wt})}{\sum\limits_{w=1}^{W} exp(v_{wc}^T u_{wt})} \tag{4.21}$$

$$J = logexp(v_{wc}^T u_{wt}) - log\sum_{w=1}^{W} exp(v_{wc}^T u_{wt}) \tag{4.22}$$

In order to avoid expensive computation of softmax for the whole vocabulary, negative sampling is commonly used as suggested by Mikolov *et al.* [86]. Then the objective function *J* becomes:

$$J' = \sum_{w_t,w_c \in D} logQ_\theta(D = 1|w_t, w_c) + \sum_{w_t,w_c \in D'} logQ_\theta(D = 0|w_t, w_c) \tag{4.23}$$

with the probability of w_t and w_c being observed is $Q_\theta(D = 1|w_t, w_c)$ and the probability of not being observed is $Q_\theta(D = 0|w_t, w_c)$, *D* and *D'* is the observed data and unobserved data respectively and θ word embeddings.

Once the text corpus has been trained by Word2Vec, the output vector for any name of a drug can be conveniently obtained through the skip-gram algorithm. As shown in Figure 2.10, two words before and after the context word was predicted through this algorithm.

The more frequently the combination of words occurred in the training sample, the more likely the word would be selected. If quick and brown occurred together more frequently than quick and black, then in the case where quick was chosen as an input word, brown would be selected as the nearest word. The layer size determined the size of the output feature vectors. Thus, if the vocabulary size of the corpus was k, and the number of terms in the text corpus was n, the input vector would be a single row vector $[1$ X $k]$ containing 0 at all positions within the vector except the n^{th} position which would be a 1. With a layer size of m, the size of the hidden layer used by Word2Vec is $[k$ X $m]$. In this way, the word vector produced for each word would be the product of the matrix $[1$ X $k]$ and $[k$ X $m]$ producing a single row vector of size m. With the set of neighbouring words that are related for each drug, word vectors can be constructed for each of these neighbouring words. The similarity ratio can be obtained to help discover drug interactions by comparing how similar the set of word vectors was for each drug. A higher similarity ratio indicates a higher chance of an adverse interaction within the drug pair.

4.6 Summary

The different methods used to search for features in the different models were described. With each drug represented in the vector space, knowledge about their interactions can be obtained. Various models have evolved through the study where drug interactions can be discovered. The statistical model and side-effect model use the *tf*idf* information on the text that describes the drugs. Another approach is to examine the number of common drugs between a drug pair in which have an adverse interaction. The larger the number of common paths between a drug pair, the higher the similarity ratio, which means the less chance of an adverse interaction. The word embedding approach is based on the expectation that a higher set of common terms are used to describe a pair of similar drugs. Similarity ratio of a drug pair is obtained from the vectors that represent the words most related to each drug.

These methods and the evaluation of their performances form the prediction layer of the three-layer conceptual framework introduced in Chapter 3. The next chapter describes the design of the models for the different approaches and implementation of the experiment to evaluate the performance.

Chapter 5

Empirical Experiments

The aim of this chapter is to describe both how the experiments were conducted and the data preparation process. This will enable an evaluation of the reliability of the novel approach in discovering the similarity of a drug pair. This process also falls under the testing stage of the scientific methodology (Figure 1.1) adopted in the research. The baseline models [138, 155] which serve as comparison for the subsequent results are also described in this chapter

5.1 Experimental Design

The purpose of the experiment was to evaluate the performance of the various models described in the previous chapter. The data used for all the models was sourced from the knowledge base.

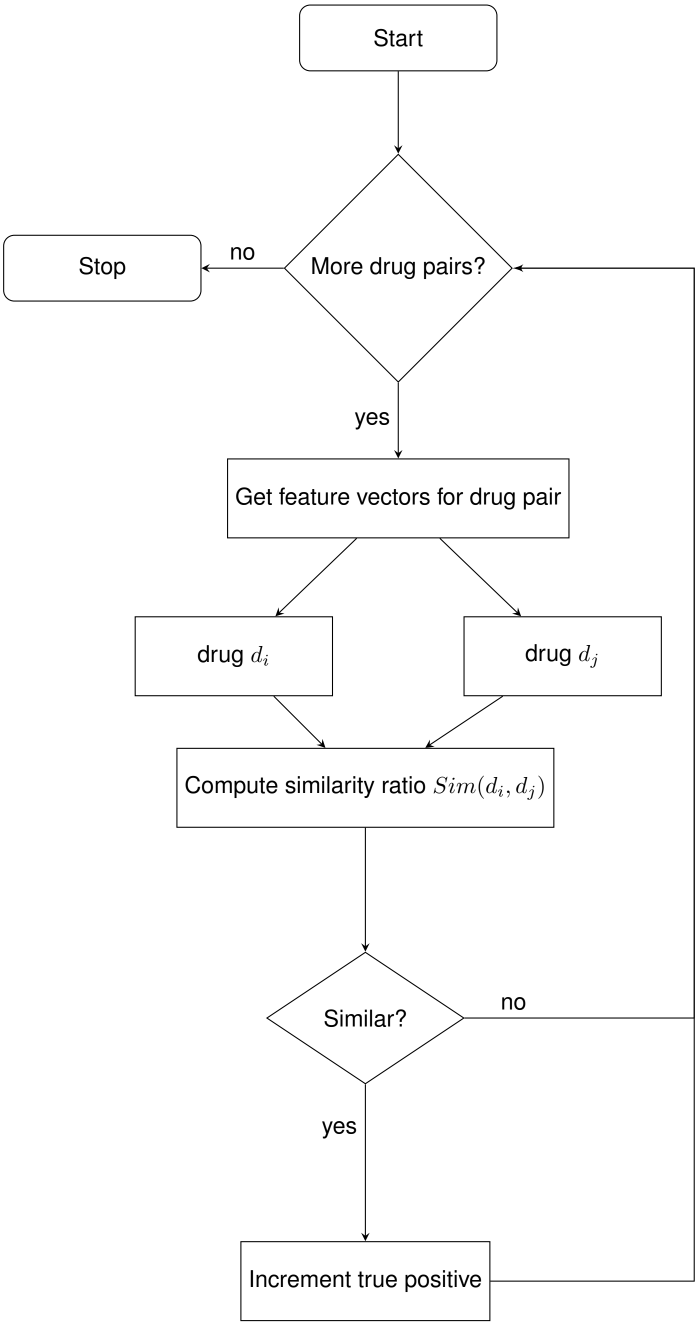


Figure 5.1: Experimental design to for measuring performance

Figure 5.1 shows the experiment flow culminating in the measurement of the model's performance. Feature vectors are extracted the positive and negative drug pairs. The way these vectors are extracted differs according to the type of model being used during the experiment. The similarity ratio of the drug pair is then calculated.

If the similarity ratio of the drug pair was above the threshold value, and the drug pair is supposed to be similar according to the ground truth specified in the knowledge base, then the number of true positives will be incremented. True negatives will be incremented if the similarity ratio is below the threshold value and the drug pair belongs to the dataset of drug pairs which are supposed to be dissimilar. The performance of the model can be evaluated by counting the number of true positives and true negatives in this manner.

5.1.1 Implementation of Statistical Model

In the statistical model, comparison was made between the feature vectors generated from the text corpus. The similarity of the drug pairs in the dataset was computed and compared against the ground truth. If the similarity ratio was above a certain threshold θ , the drug pair was classified as similar; otherwise, it was classified as dissimilar. If a particular drug pair was classified as similar and matched with the positive dataset, then it was a true positive; if it matched with the negative sample, then it was a true negative. The number of correct predictions is an indication of the model's performance. For convenience and ease of computation, recall and precision were used to gauge how well the prediction was made. Various values of θ was used as threshold values to decide if the drug pair was similar. The final value was chosen at the point when the model performed best as indicated by the F-score.

The F-score was then computed using the number of true positives and true negatives. The values for true positives and true negatives depend on the threshold value. Naturally, at a very low threshold value, the number of correct predictions within the positive sample dataset, which is the score for the true positive, will be high. Hence by varying the threshold value, different F-score were obtained since the F-score depends on the value of true positives. A drug pair was considered to be similar when the similarity ratio $Sim^v(i,j)$ was above a threshold value θ. A value of "1" was attached to $S(d_i, d_j)$ of a drug pair to specify that the similarity was above the threshold value θ and "0" otherwise. Hence, if $Sim(d_i, d_j)$ was the similarity ratio of drug pair d_i and d_j,

$$S(i,j) = \begin{cases} 0, & \text{if } Sim(d_i, d_j) < \theta \\ 1, & \text{otherwise} \end{cases} \tag{5.1}$$

The total number of $S(i,j)$ was the number of true positives for the experiment.

5.1.2 Implementation of Side-effect Model

In this experiment, all the side-effects were retrieved from the knowledge base constructed from DrugBank. In order to determine the similarity of a drug pair, a row matrix M where the elements in the matrix can be either a 1 or 0 was used (Equation 4.13). For example, within the global set of side-effects $\mathbb{E} = \{e_1, e_2 \ldots e_r\}$ (see Section 4.3), if side-effect “cough” for the drug *warfarin* occurred at location 87 (ie value of e_{87} was “cough”), then the 87th element of the row matrix M will be set to 1. The similarity ratio between the two drugs d_i and d_j was then obtained between the two row matrices M_i and M_j. Figure 5.2 shows the experimental flow for this model.

5.1.3 Implementation of Adverse Network Model

In the adverse network model, the aim of the experiment was to find the number of common paths within a drug pair. This was done by capturing the set of interactive drugs for each drug in the drug pair through the use of row matrices. The purpose of the matrix was to indicate the presence of all the out-going nodes of the drug.

At the beginning of the experiment, the contents of this matrix were initialised to 0. During the experiment, matrix M^r for drugs d_1 and d_2 was created to indicate the positional match with the adverse drugs for d_1 and d_2 with interaction rating r, where the number of columns in M^r = $|A^r|$. $|A^r|$ is the set of all the interacting drugs in the knowledge database. The ratings used to describe the relationship between the vertex node and the set of interactive drugs were major, moderate and minor.

Algorithm 1 shows the steps in obtaining the feature vectors to compute the similarity ratio.

With the similarity ratio results from the experiment, a threshold of θ = 0.5 was used to predict if the drug pair is similar. The threshold value of 0.5 was chosen as it is the default probability used by most classifiers [166]. A value of 0.5 or higher from the experiment was classified as similar, and a value lower than 0.5 was classified as dissimilar. The performance of the model can be evaluated by counting the number of correct predictions.

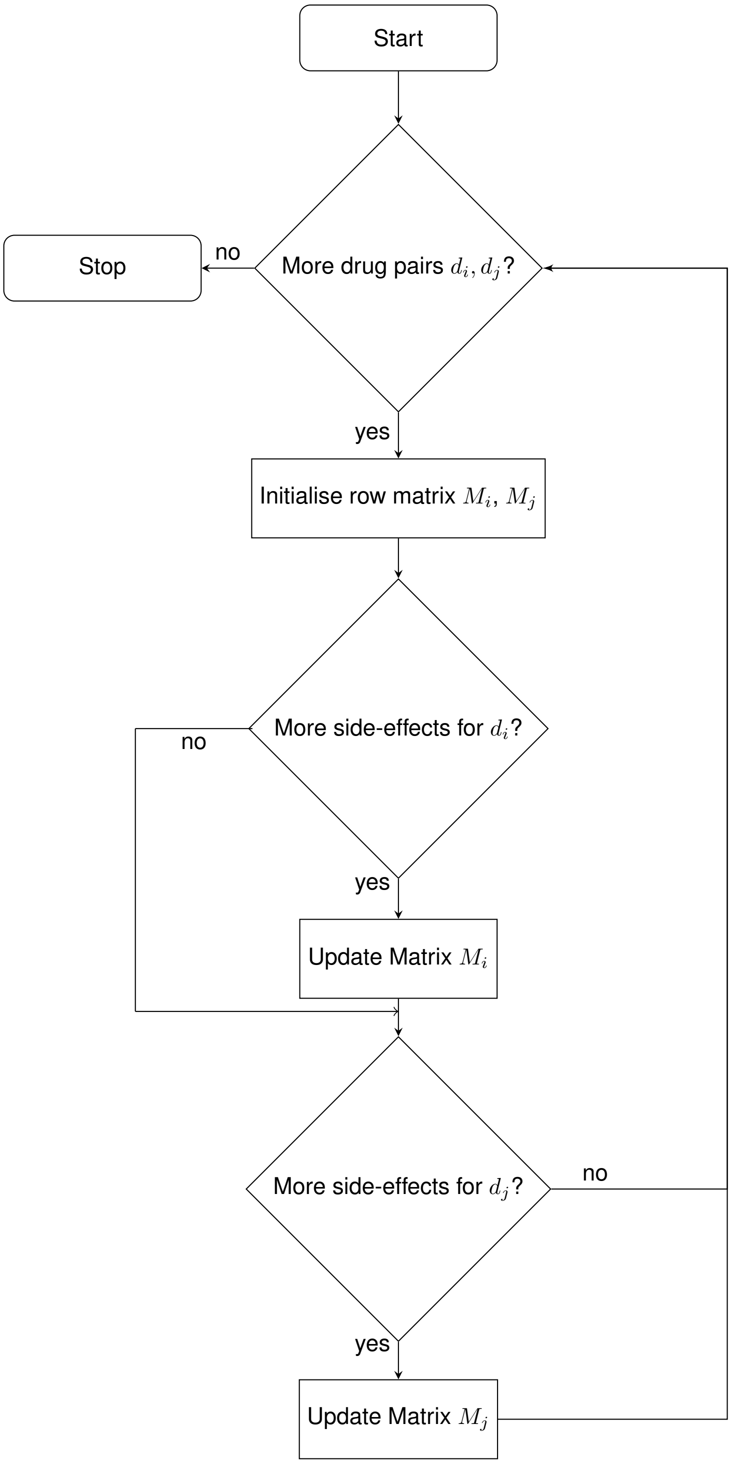

Start
Stop
no
More drug pairs d_i, d_j?
yes
Initialise row matrix M_i, M_j
More side-effects for d_i?
no
yes
Update Matrix M_i
More side-effects for d_j?
no
yes
Update Matrix M_j

Figure 5.2: Experimental design of side-effect model

```
input  : Let $\mathbb{D}$ be the drug corpus;
         $d_i, d_j$ be the drug pair;
         $M_i$ and $M_j$ be the row matrices;
output: Similarity ratio $Sim(i, j)$
while (for each drug that interacts with $d_i$) do
    get index where drug occurs within $\mathbb{D}$;
    update Matrix $M_i$;
end
while (for each drug that interacts with $d_j$) do
    get index where drug occurs within $\mathbb{D}$;
    update Matrix $M_j$;
end
Compute $Sim(i, j)$;
```

Algorithm 1: Getting similarity ratio of adverse network model

5.1.4 Implementation of Word Embedding Model

In the word embedding model, feature vectors were obtained through an artificial neural network approach. A predictive model was constructed for learning word embeddings from the raw corpus that described the properties of the drugs by using the skip-gram model from Word2Vec [85]. Word2Vec was a good fit for the context of the experiment since the problem domain aimed to extract related words to determine the extent of similarity from bio-medical text. Given a keyword, for example, the name of a drug, this method formulated a feature vector that best predicted a window of surrounding words occurring in some meaningful context. Such semantic similarity also conformed to the important criteria for selecting good word pairs [160].

When training the model, important parameters expected by Word2Vec were the word frequency, layer size and window size. The word frequency was the minimum number of times a word must appear in the corpus, the layer size was the number of desired features in the word vector, and the window size was the number of words before and after the target word to extract for the training sample. In order to observe the behaviour of this approach, the model was constructed with individual attributes of the drug ("Professional", "Overview" and "Side-effects") while varying the number of nearest neighbours. In the experiment, a number of keywords associated with the nearest neighbor of the drug name was retrieved from the model. Each keyword was represented by a word vector of numbers, the size of which depended on the layer size as explained in Section 4.5. For example, if d_{i1}, d_{i2}, d_{i3} were the three nearest keywords for a given drug d_i, a word vector would be obtained from the specified model by combining the three word vectors from the respective three keywords. Similarity ratio between each set of vectors produced from the keywords could then be computed.

To test the reliability of this model, the same dataset from the statistical model was used. However, word vectors were used instead of *tf * idf* to construct the feature vectors. As shown in Algorithm 2, to obtain the feature vector, nearest neighbours of each drug pair were converted into word vectors using Word2Vec utility. Similarity ratio between the pair of word vectors for each drug was then computed.

```
input : Let $\mathbb{D}$ be the drug corpus;
        $d_i, d_j$ be the drug pair;
output: Similarity ratio $Sim(d_i, d_j)$

while (for each attribute in $\mathbb{D}$ ) do
    set window size;
    set layer size;
    train the model using Word2Vec;
end
while (there are more drug pairs $d_i$, $d_j$) do
    get nearest neighbours;
    build word vectors;
    compute similarity ratio of $Sim(d_i, d_j);$
end
```

Algorithm 2: Experimental design of word-embedding model

5.2 Baseline Models

Comparison of the baseline model with other works highlighted how adoption of this novel approach resulted in superior performance. The work of [138] predicted DDI by parsing bio-medical text for syntactic and semantic information on biological entities such as induction and inhibition of enzymes by drugs. These relations were then mapped with the general knowledge about drug metabolism and interactions to derive the DDI.

Besides DrugBank, the work uses data from abstracts found in MEDLINE database, which has a large number of reputable source of citations from peer-reviewed biomedical text [107] and is open access [44]. The approach is based on the idea that the interaction between two drugs depends on the metabolism of the drug.

The enzyme acts as catalyst in the chemical reaction between drug d_i and drug d_j (Figure 5.3). However, drug d_i may inhibit the action of the enzyme, in turn decreasing the effect of drug d_j.

The methods in identifying drug interactions involved two phases: the extraction phase and the reasoning phase. In the extraction phase, information on sentence structure and biological entities was retrieved in order to obtain relationships within drug pairs. The reasoning phase then applied the extracted interactions to the logic rules in order to derive the interactions of the drug pair.

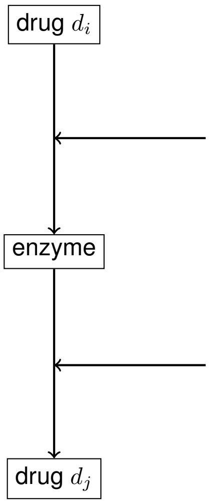


Figure 5.3: Drug interactions

In another work to predict interactions of a drug pair, Yan *et al.* developed various prediction models to leverage on text mining and statistical inference techniques [155]. One of the models

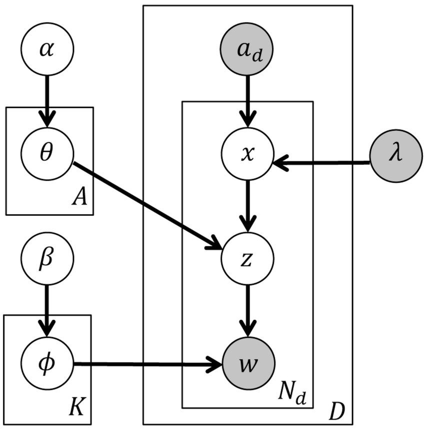


Figure 5.4: Drug-Entity model [155]

used included the popular DET model used to capture the relation between drugs and other entities. With plate notations [11], Figure 5.4 shows the generative process represented as a Bayesian model. A dummy document with subject section and content section was built for each drug found in the MEDLINE corpus[1], assuming the total number of drugs was *D*. Hence, the total number of documents was *D*, with each document *d* conveyed by the number of diseases. N_d refers to the total number of disease words *w* occurring in *d*. K was the total number of topics and a_d was the observed set of drugs, with A referring to the total number of drugs.

Each drug *x* and topic *z* was a probabilistic distribution over topics (parameterised by θ) and diseases (parameterised by φ) respectively. λ was the observable parameter which controlled the drugs sampling. Just like the current work, DrugBank was also used. However, one of the methods in their preparation of data was to represent each drug by a vector of drug targets. The values in each vector were either 1 or 0, depending on whether the drug target was associated with the given drug. In our work, we chose to construct feature vectors of *tf*idf* from textual information related to the properties of each drug.

The performance metrics Area under Curve (AUC) of Receiver Operating Characteristics (ROC) [33] used in our work was also adopted in this baseline model. A curve closer to the upper-left corner in the ROC, resulting in a higher area beneath the curve, indicated a better model performance.

	Tari [138]	**Yan** [155]	**Proposed Model**
Aim	Discover drug interaction	Predict drug interaction	Personalised drug prediction
Source	DrugBank and MeSH	DrugBank and MeSH	DrugBank
Method	Combine text mining and reasoning approach based on biological entities	Compose feature vectors based on names of disease and genes	Create feature vectors from textual drug description

Table 5.1: Baseline models

Table 5.1 shows a summary of the two baseline models to highlight the differences in methods compared to our model. A comparison of performance is made in Section 6.2. The two models were chosen as both works are associated with interaction between drug-pairs and uses similar drug repositories. Although the research was started in 2015, it is encouraging to note that the performance of a recent work [71] yields an overall F-score of 72% compared to our embedding model of 75%. Besides the baseline model, performance of other web-based

[1] https://www.ebscohost.com/nursing/products/medline-databases/medline

checkers were also tested against the prototype of this research. These are discussed in Section 6.2.3. Moreover, unlike the prototype created from this research, these online interaction checkers are unable to provide information about overall combination of multiple drugs [42].

For comparison with more recent works, another model has been included. This model, proposed by Zhang and Kordjamshidi (2018) uses dataset from Text Analysis Conference to predict if a drug-pair is in adverse relationship [161]. With the help of support vector machine and a medical dictionary, the model attempts to classify the precipitant drug from the training dataset. The predictions were classified as correct or incorrect by comparing the precipitant drug with the ground truth. An additional experiment is conducted with our model with the same dataset. The results are compared in Section 6.2.3.

5.3 Data Preparation

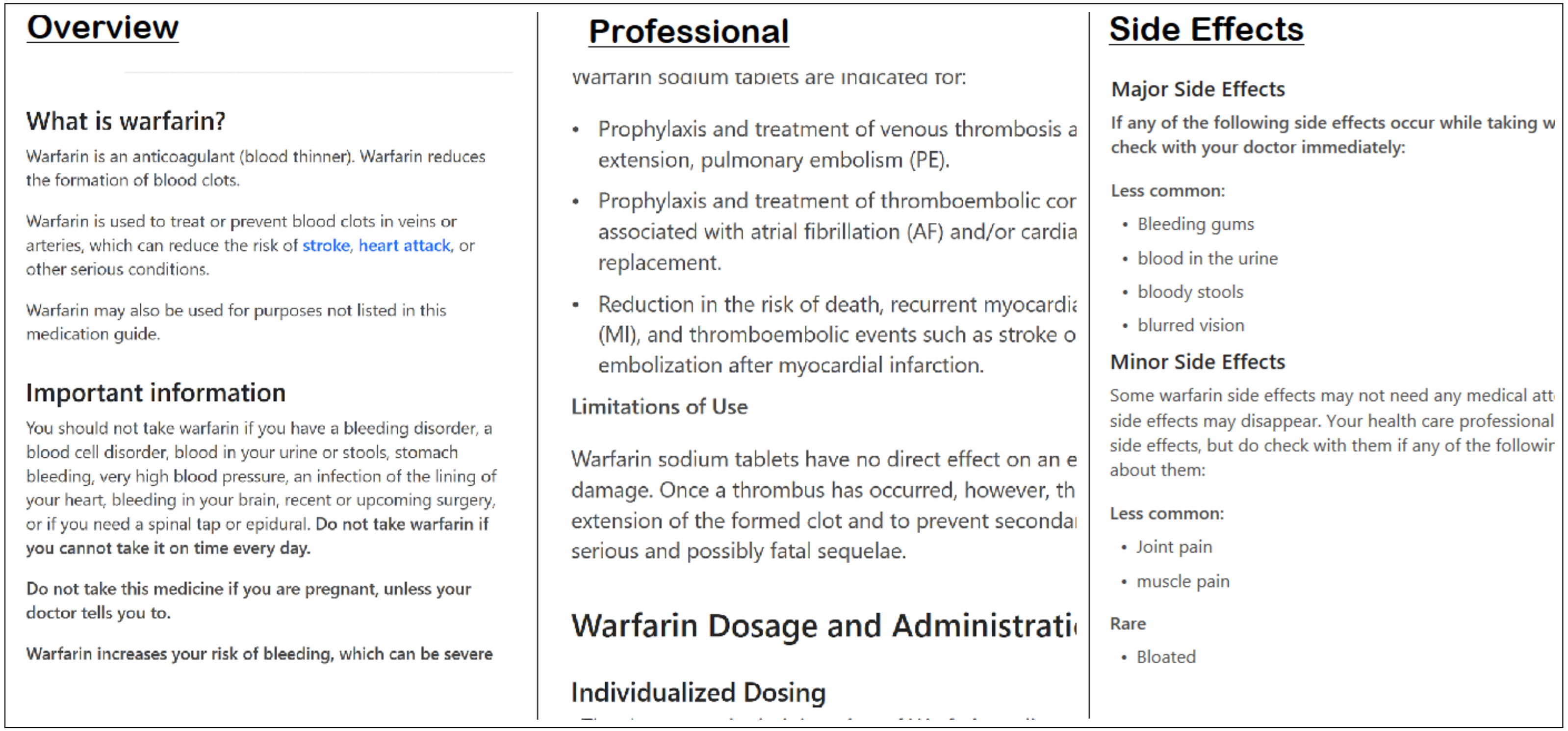
Experiments were conducted based on the data obtained from public source. As the first step in data preparation, data from DrugBank was downloaded and stored locally in a database. For convenience, the MySQL database was used[2].

To facilitate the ease of extracting relevant data from the drug taxonomy, the textual data was pre-processed. Stopwords such as pronouns, prepositions and conjunctions will be removed and similar words like run, running, runner will be stemmed by removing the suffixes, resulting in a reduced number of words to process.

5.3.1 Preparing Data for Statistical Model

The properties of interest were those related to interactions and side-effects. Data related to “Overview”, “Professional” and “Side-effects” was extracted from this database. Text mining was conducted on each of these properties in order to construct feature vectors for computing the similarity ratio between drugs. Figure 5.5 shows an extract of the three properties for the drug *warfarin*. The “Overview” section uses lay language to explain the drug effects, the appropriate dosage and areas to take note of while on the drug such as specific of food to

[2] As the thesis is not on system design and programming, the emphasis is to highlight the approach in extracting useful information in the data mining process and not the programming procedures.



Overview

What is warfarin?

Warfarin is an anticoagulant (blood thinner). Warfarin reduces the formation of blood clots.

Warfarin is used to treat or prevent blood clots in veins or arteries, which can reduce the risk of stroke, heart attack, or other serious conditions.

Warfarin may also be used for purposes not listed in this medication guide.

Important information

You should not take warfarin if you have a bleeding disorder, a blood cell disorder, blood in your urine or stools, stomach bleeding, very high blood pressure, an infection of the lining of your heart, bleeding in your brain, recent or upcoming surgery, or if you need a spinal tap or epidural. **Do not take warfarin if you cannot take it on time every day.**

Do not take this medicine if you are pregnant, unless your doctor tells you to.

Warfarin increases your risk of bleeding, which can be severe

Professional

Warfarin sodium tablets are indicated for:

- Prophylaxis and treatment of venous thrombosis a extension, pulmonary embolism (PE).
- Prophylaxis and treatment of thromboembolic cor associated with atrial fibrillation (AF) and/or cardia replacement.
- Reduction in the risk of death, recurrent myocardia (MI), and thromboembolic events such as stroke o embolization after myocardial infarction.

Limitations of Use

Warfarin sodium tablets have no direct effect on an e damage. Once a thrombus has occurred, however, th extension of the formed clot and to prevent seconda serious and possibly fatal sequelae.

Warfarin Dosage and Administrati

Individualized Dosing

Side Effects

Major Side Effects

If any of the following side effects occur while taking w check with your doctor immediately:

Less common:

- Bleeding gums
- blood in the urine
- bloody stools
- blurred vision

Minor Side Effects

Some warfarin side effects may not need any medical att side effects may disappear. Your health care professional side effects, but do check with them if any of the followir about them:

Less common:

- Joint pain
- muscle pain

Rare

- Bloated

Figure 5.5: Sample extract of properties of *warfarin*

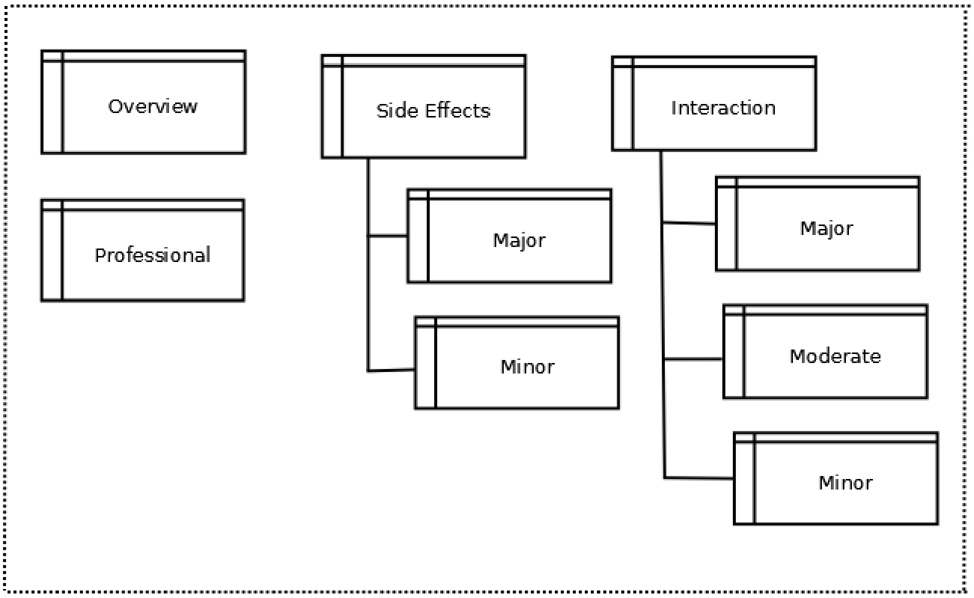


Figure 5.6: Organisation of the properties of dataset

avoid. For the professional, the attribute “Professional” has more in-depth description such as chemical structure, warnings and precautions as well as recommended dosage for various symptoms to achieve maximum effect. The “Side-Effects” attribute lists out the major and minor side-effects associated with the drug.

From the text of these attributes, feature vectors can be constructed from information like term frequency and used for comparison with other drugs. Figure 5.6 illustrates how the the attributes of each drug were structured for the purpose of this research. There were also subcategories for properties on “Side-effects” and “Interaction”. The textual data from these properties were mined in order to derive feature vectors.

Name of Field	**Example**
ID	DB00001
Name	Lepirudin
CAS Number	120993-53-5
Drug Type	BiotechDrug
Wikipedia ID	Lepirudin
rxlist link	http://www.rxlist.com/cgi/generic/lepirudin.htm
drug link	http://www.drugs.com/cdi/lepirudin.html

Table 5.2: Structure of dataset in experimental evaluation

From the basic structure of the dataset (Table 5.2), relevant knowledge on drug interactions and subsequently term frequency were obtained through data mining. This is done in the prediction layer, which computed the feature vectors and similarity ratio of drug pairs for subsequent processing during the experiment.

5.3.2 Preparing Data for Side-effect Model

As properties of a drug pair in terms of their side-effects were used in the experiment to build feature vectors, data mining was performed on the textual data from DrugBank to obtain the side-effects. Similar to the statistical model, the side-effects of each drug were stored in MySQL database for use during the experiment.

Level	Side-Effects
1	abdominal or stomach pain with cramping
2	arm, back, or jaw pain
1	bleeding gums
5	bloated
1	blood in the urine
1	bloody stools
4	muscle pain
2	nausea and vomiting
1	nosebleeds
1	unusual tiredness or weakness
2	upper right abdominal or stomach pain
2	vomiting of blood

Table 5.3: Sample side-effects of *warfarin*

Using *warfarin* as an example, a sample of the side-effects retrieved from the taxonomy stored locally for the experiment is shown in Table 5.3.

Level	Description
0	major,more common
1	major,less common
2	major,rare
3	minor,more common
4	minor,less common
5	minor,rare

Table 5.4: Levels of side-effects

The levels defined for this dataset ranged from common to rare as defined in Table 5.4.

5.3.3 Preparing Data for Adverse Network Model

To prepare the data for use during the experiment to test the adverse network model, all the drug interactions were extracted from the drug taxonomy and stored as described in the beginning of Section 5.3. The interactions extracted were “major”, “moderate” and “minor”. Unlike the categories for side-effects which were major and minor, the categories for interactions in this study had the additional “moderate” category to maximise the data available from DrugBank.

A sample of the drugs that interact with *warfarin* is shown in Table 5.5.

Property	**Interactive Drug**
Major	aspirin etodolac ibuprofen naproxen
Moderate	bacampicillin balsalazide corticorelin ethanol tramadol
Minor	acetaminophen methotrexate trazodone turmeric

Table 5.5: Sample drugs that interact with *warfarin*

5.3.4 Preparing Data for Word Embedding Model

The word embedding model was developed to allow for a comprehensive comparison of the different models in arriving at the similarity ratio of a drug pair. As the platform used for this model was Word2Vec, preparation of the data is slightly different from the previous three models.

Overview	**Professional**	**Side-effect**
dose	coumadin	headache
patient	anticoagul	effect
therapi	patient	report
recommend	clinic	hemorrhag
intraven	healthcar	pain

Table 5.6: Sample tokens for *warfarin*

In order to prepare data for the word embedding model, the textual data was used as an input file to Word2Vec. The raw data from the textual file of each drug was broken into atomic units or tokens and fed into Word2Vec. Using the drug *warfarin* as an example again, individual words or tokens extracted from the drug taxonomy is shown in Table 5.6 for each drug property. These textual files were the same files used for the preparation of data for the previous models taken from DrugBank.

Overview	**Professional**	**Side-effect**
154,645	196,352	53,644

Table 5.7: Size of dataset used for building Word2Vec model

Table 5.7 shows the number of tokens associated with each attribute of the drug ("Overview", "Professional" and "Side-effect").

Once the data was tokenised, a binary file was generated by Word2Vec for the experiment. In generating the binary file, different parameters were used for the sensitivity study and comparison of the performance. In the project, the two parameters used were the layer size and window size. Once the textual file was trained by Word2Vec with the binary file produced, feature vectors needed to compute the similarity ratio can be obtained by using built-in methods. Word vectors generated by Word2Vec were used for computing similarity ratio of a drug pair during the experiment. Table 5.8 shows the vectors generated by Word2Vec at a layer size of 8 for the sample words "patient", "clinic" and "pain". Note that the size of the vectors depends on the layer size used when training the model. In this case the layer size was 8.

patient	**clinic**	**pain**
-0.4644	-0.6514	0.1638
-0.029	0.1215	-0.1779
0.3551	0.0935	0.5075
0.5995	-0.0902	-0.5868
0.2739	0.6315	0.2253
0.4405	-0.1723	-0.1494
-0.1444	0.3328	0.5121
0.09	0.0696	-0.0655

Table 5.8: Sample word vectors from Word2Vec

5.3.5 TAC2018 Dataset

As new models are evolved towards the end of the research project, a new baseline model by Zhang and Kordjamshidi (2018) has been chosen to compare the approach adopted in this thesis [161]. For a fair comparison, the experiment used the data set from Text Analysis Conference (TAC) 2018 containing gold standard annotations from the National Library of Medicine and the U.S Food and Drug Administration which is used by the new baseline model. This dataset contains information on drug interaction for 22 drugs stored in XML format. As shown in Figure 5.7 which shows a sample extract of the dataset for the drug *Guanfacine*, the list of adverse interactive drugs are shown under the node 'precipitant' within the tag <LabelIteractions><LabelInteraction>.

5.3.6 Summary

Model	Property	Size of Data
Statistical	Overview	290,108
	Professional	598,903
	Side-effects	128,623
Side-Effect	Major	23,371
	Minor	12,946
Adverse Network	Major	48,344
	Moderate	306,915
	Minor	26,659

Table 5.9: Size of data for each model

The process of data preparation for each model was explained in this section. Table 5.9 shows a summary of the data size for the statistical model, side-effect model and adverse network model.

5.4 Performance Measuring Schemes

Precision, recall and F-score [139] were used to evaluate the model's performance. Precision (P) indicated how accurately the model predicted drug pairs as similar, while recall (R) indicated how accurately similar drug pairs were predicted. Accuracy was also used to measure the percentage of correct predictions combining both the similar and dissimilar predictions. Thus,

$$\text{P} = \frac{TP}{TP + FP} \tag{5.2}$$

$$\text{R} = \frac{TP}{TP + FN} \tag{5.3}$$

```
<?xml version="1.0" encoding="UTF–8"?>
–<Label setid="886e050c–dd22–4f35–ac3b–243f091125c3" drug="TENEX">
+<Text>
+<Sentences>
–<LabelInteractions>
        <LabelInteraction type="Pharmacokinetic interaction" effect="C54356"
                precipitant="microsomal enzyme inducer" precipitantCode="NO MAP"/>
        <LabelInteraction type="Pharmacokinetic interaction" effect="C54607"
                precipitant="microsomal enzyme inducer" precipitantCode="NO MAP"/>
        <LabelInteraction type="Pharmacokinetic interaction" effect="C54356"
                precipitant="phenobarbital" precipitantCode="N0000005893"/>
        <LabelInteraction type="Pharmacokinetic interaction" effect="C54607"
                precipitant="phenobarbital" precipitantCode="N0000005893"/>
        <LabelInteraction type="Pharmacokinetic interaction" effect="C54356"
                precipitant="phenytoin" precipitantCode="N0000006023"/>
</LabelInteractions>
</Label>
```

Figure 5.7: Sample XML extract of dataset from TAC2018

where TP is True Positive, FN is False Negative, and FP False Positive.

For example, if there were 100 possible drug pairs with adverse reactions and 150 without any adverse reactions, and only 60 out of the 70 predictions are accurate (TP=60, FP=10), then recall will be 60/(60+40). This is because only 60 were accurately predicted out of a total possible of 100 pairs, including the 40 pairs that were relevant but not selected (FN=40). Precision will be 60/(60+10) since only 60 records out of the 70 records predicted were accurate. F-score was based on the precision and recall:

$$F = \frac{2 * \text{P} * \text{R}}{\text{P} + \text{R}} \tag{5.4}$$

As seen from Equation 5.4, F-score represented by F depends on the precision and recall of the experiment which is not associated with the number of true negatives. As the datasets contain both positive and negative pairs, there is a need to measure how well the model can accurately predict those negative drug pairs, i.e. drugs that are in adverse interaction. Hence, ROC curves with computation of the area under the curve were used. Such plots have been extensively utilised to evaluate many systems including diagnostic systems, medical decision-making systems, and machine learning systems [139].

The ROC curve [33] is essentially a two-dimensional plot of the true positive rate (*tpr*) against the false positive rate (*fpr*). These values are given as:

$$\textit{tpr} = \frac{TP}{TP + FN} \tag{5.5}$$

$$\textit{fpr} = \frac{FP}{FP + TN} \tag{5.6}$$

ROC values take into account the true negative values, resulting in a more robust measuring scheme.

5.5 Summary

Chapter 5 addressed the design issues for conducting the experiments of the various models to determine their performance in predicting the similarity ratio of a drug pair. The implementation details including the baseline model were also explained. The process of data preparation and evaluation methodology for the different models were described. The next chapter presents the results from these experiments.

Chapter 6

Results and Discussions

Chapter 6 reports the results according to the experimental design explained in the previous chapter. The results support the hypothesis that similar drug pairs have a higher similarity ratio than dissimilar pairs. This explains the fact that attributes such as adverse interactions and side-effects of a drug can be used to construct feature vectors for computing similarity ratios. Moreover, paths linking the common drugs within the set of interacting drugs can also be used to arrive at the similarity ratio.

This represents a breakthrough in the design of CDSS in the context of drug prescription by mining feature vectors from textual data.

6.1 Experimental Results

The following sections report on the results obtained with the use of different models. The experiment used a dataset consisting of a set of positive sample drug pairs and a set of negative sample drug pairs. Drug pairs in the positive sample are supposed to be safe for consumption as they do not adversely interact with each other. Conversely, the negative sample consists of drug pairs that would be unsafe for prescription as they would adversely interact with one another. Using such drug pairs from both the positive and negative sample in the experiment, results were obtained by computing their similarity ratio based on the feature vectors. If the similarity ratio was above the threshold value, the drug pair is similar. A correct prediction of the results with the positive sample dataset yielded a true positive result.

6.1.1 Experiment Results from Statistical Model

In this model, the *tf*idf* was used in computing similarity ratio of each drug pair in the dataset. Table 6.1 shows the results for the pairs of drugs taken from the positive sample dataset.

Similarly, Table 6.2 shows the results obtained with the drug pairs from the negative dataset, where the drug pairs are supposed to be dissimilar.

Table 6.3 shows the results when similar and dissimilar datasets were applied to our model for attributes from “Overview”, “Professional” and “Side-effects”.

After computing the similarity ratio of drug pairs, different cut-off values of θ were used to decide if the drug pair is similar. For a given value of θ, the number of correct predictions for the dataset that was supposed to be similar (true positives) and dissimilar (true negatives) was counted. If the similarity ratio was above θ, it was considered “similar”, otherwise it was considered to be “dissimilar”. For example, from the “Professional” attributes, there were 13 and 23 correct predictions from the similar and dissimilar datasets respectively.

Table 6.4 shows the F-score obtained for a range of values for θ, applied for each of the drug properties “Overview”, “Professional” and “Side-effects”. For example, a θ of 0.45 is used as a threshold to compute the recall, precision and F-score for features gathered from the “Professional” attribute as the maximum value of F-score occurs at this value.

Figure 6.1 shows the recall, precision and F-score achieved with drug attributes gathered from “Overview”, “Professional” and “Side-effects”. As indicated in Figure 6.1, the recall rate of 96% was achieved from drug attributes obtained from “Side-effects”, showing that our model performed much better than other methods of prediction. In contrast, the work by [138] achieved 48.5% with predictions based on the inhibition properties of drugs in the knowledge base.

From the F-score of each attribute, a weightage was computed in proportion to the respective F_{max}. In our experiment, F_{max} for “Overview” was 0.6 and the total F_{max} for the three attributes was 1.93, so the $tf * idf$ for “Overview” will be weighted by a factor of 0.6/1.93

ID1	**Drug i**	**ID2**	**Drug j**	$\mathbf{Sim^v(i,j)}$	$\mathbf{Sim^p(i,j)}$	$\mathbf{Sim^s(i,j)}$
DB01060	Amoxicillin	DB01190	Clindamycin	0.964	0.804	0.996
DB00193	Tramadol	DB00316	Acetaminophen	0.873	0.976	0.985
DB00316	Acetaminophen	DB00318	Codeine	0.988	0.956	0.989
DB01050	Ibuprofen	DB00318	Codeine	0.898	0.86	0.989
DB01060	Amoxicillin	DB00916	Metronidazole	0.956	0.995	0.996
DB00916	Metronidazole	DB01190	Clindamycin	0.995	0.758	0.989
DB00784	Mefenamic acid	DB01050	Ibuprofen	0.943	0.86	0.966
DB00945	Acetylsalicylic acid	DB01060	Amoxicillin	0.991	0.894	0.99
DB00328	Indomethacin	DB00461	Nabumetone	0.983	0.882	0.997

Table 6.1: Sample results from positive sample

ID1	**Drug i**	**ID2**	**Drug j**	$\mathbf{Sim^{v}(i,j)}$	$\mathbf{Sim^{p}(i,j)}$	$\mathbf{Sim^{s}(i,j)}$
DB00945	Acetylsalicylic acid	DB00055	Drotrecogin alfa	0.932	0.935	0.832
DB00945	Acetylsalicylic acid	DB00078	Ibritumomab	0.831	0.893	0.832
DB01050	Ibuprofen	DB00078	Ibritumomab	0.849	0.908	0.952
DB01050	Ibuprofen	DB00300	Tenofovir	0.946	0.901	0.832
DB00945	Acetylsalicylic acid	DB00369	Cidofovir	0.931	0.865	0.832
DB01050	Ibuprofen	DB00369	Cidofovir	0.883	0.806	0.994
DB00945	Acetylsalicylic acid	DB00465	Ketorolac	0.948	0.927	0.832
DB01050	Ibuprofen	DB00465	Ketorolac	0.949	0.939	0.951
DB00945	Acetylsalicylic acid	DB00563	Methotrexate	0.984	0.894	0.832
DB01050	Ibuprofen	DB00563	Methotrexate	0.983	0.845	0.832

Table 6.2: Sample results from negative sample

	Overview	Professional	Side-effects	Normalised
Recall	0.79	0.54	0.96	0.68
Precision	0.79	0.93	0.53	0.70
F-score	0.79	0.68	0.69	0.69
True Positive	19	13	23	23
True Negative	19	23	4	56
False Positive	5	1	20	10
False Negative	5	11	1	11

Table 6.3: Results based on different attributes of the drug pairs

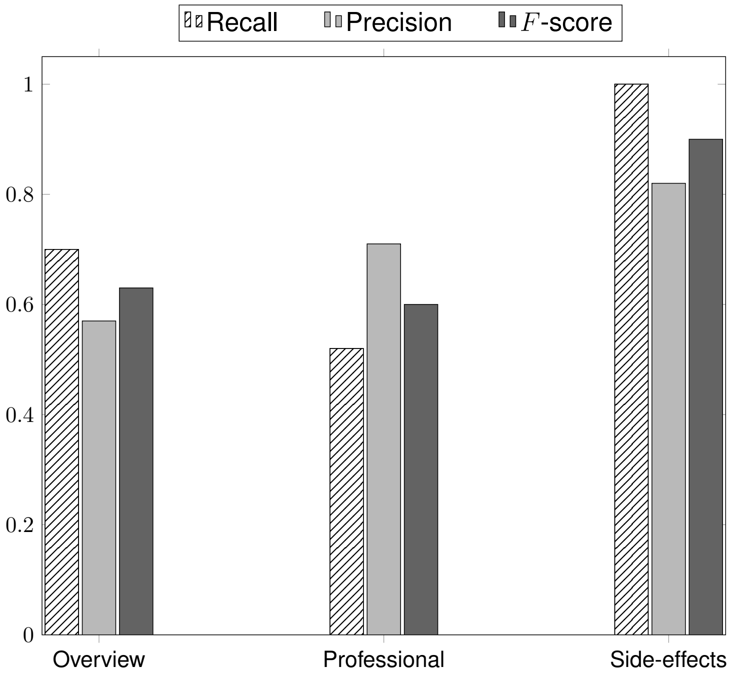


Figure 6.1: Performance comparison against different drug attributes

θ	**Overview**	**Professional**	**Side-effects**
0.45	0.60	0.60	0.67
0.48	0.61	0.56	0.66
0.50	0.60	0.57	0.67
0.53	0.56	0.51	0.68
0.55	0.53	0.52	0.69
0.58	0.51	0.54	0.69
0.60	0.54	0.52	0.68
0.63	0.56	0.46	0.73
0.65	0.53	0.44	0.70

Table 6.4: F-score at different threshold values of θ

Attribute	$\mathbf{F_{max}}$	**Weight**
Overview	0.60	0.31
Professional	0.60	0.31
Side-effect	0.73	0.38
Total	1.93	1.00

Table 6.5: Weights to normalise feature vectors

which is 0.31. The weights of the other attributes were computed in a similar manner and the values are shown in Table 6.5. By combining the normalised feature vectors for all the three attributes, an aggregated similar ratio was obtained for each drug pair.

In the same manner, different F-score values were obtained at different threshold levels by counting the number of true positives and true negatives produced from the model. The last column of Table 6.3 shows the results based on the aggregated similar ratio obtained from the normalised feature vectors. In terms of accuracy, the percentage of correct predictions combining both the similar and dissimilar predictions, our system comes out at 79% compared to 69% where drug predictions were based on the relationship between drug targets [155].

6.1.2 Experiment Results from Side-effect Model

Utilising the same dataset as the statistical model, the experiment was conducted using information on side-effects of the drug. Table 6.6 and Table 6.7 show the results for positive and negative drug pairs respectively. The feature vectors used for computing the similarity ratio were the row matrices as described in Section 4.3. Figure 6.2 shows the precision-recall curve and Figure 6.3 shows the ROC curve.

ID1	**Drug i**	**ID2**	**Drug j**	$\mathbf{Sim(i,j)}$
DB00331	Metformin	DB01211	Clarithromycin	0.254
DB00316	Acetaminophen	DB01050	Ibuprofen	0.165
DB00749	Etodolac	DB00784	Mefenam	0.28
DB00573	Fenoprofen	DB00605	Sulindac	0.385
DB00328	Indomethacin	DB00461	Nabumeton	0.262
DB01009	Ketoprofen	DB01050	Ibuprofen	0.487
DB00784	Mefenam	DB01050	Ibuprofen	0.427
DB00784	Mefenam	DB00788	Naproxen	0.502
DB00916	Metronidazole	DB01190	clindamycin	0.118
DB00788	Naproxen	DB00814	Meloxicam	0.404
DB00991	Oxaprozin	DB01009	Ketoprofen	0.388
DB00554	Piroxicam	DB00573	Fenoprofen	0.37
DB00605	Sulindac	DB00749	Etodolac	0.261
DB00500	Tolmetin	DB00554	Piroxicam	0.38
DB00193	Tramadol	DB00316	Acetaminophen	0.085
DB00207	Azithromycin	DB00438	Ceftazidime	0.13
DB00586	Diclofenac	DB00316	Acetaminophen	0.096

Table 6.6: Side-effect model: results from positive drug pairs

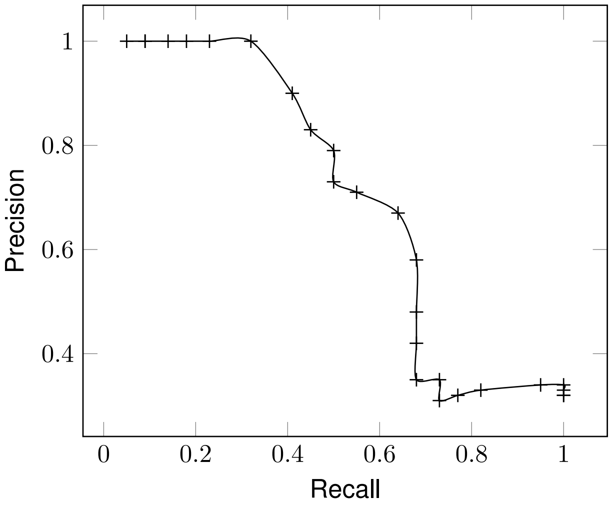


Figure 6.2: Side-effect model: precision-recall curve

ID1	Drug i	ID2	Drug j	Sim(i,j)
DB00945	Aspirin	DB00078	Ibritumomab	0.276
DB01050	Ibuprofen	DB00078	Ibritumomab	0.329
DB01050	Ibuprofen	DB00369	Cidofovir	0.155
DB01050	Ibuprofen	DB05528	Mipomersen	0.194
DB01050	Ibuprofen	DB08880	Teriflunomid	0.317
DB01050	Ibuprofen	DB08896	Regorafenib	0.203
DB01050	Ibuprofen	DB06605	Apixaban	0.281
DB00945	Aspirin	DB08880	Teriflunomid	0.242
DB01050	Ibuprofen	DB08901	Ponatinib	0.231
DB00945	Aspirin	DB00300	Tenofovir	0.055
DB00945	Aspirin	DB00563	Methotrexate	0.254
DB01050	Ibuprofen	DB00300	Tenofovir	0.038
DB01050	Ibuprofen	DB01254	Dasatinib	0.233
DB00788	Naproxen	DB06228	Rivaroxaban	0.082
DB00788	Naproxen	DB00563	Methotrexate	0.191
DB00788	Naproxen	DB00465	Ketorolac	0.027

Table 6.7: Side-effect model: results from negative drug pairs

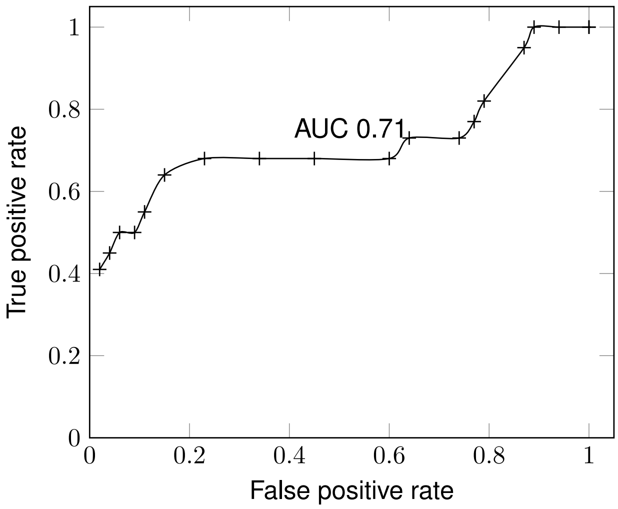


Figure 6.3: Side-effect model: ROC curve

From Figure 6.2, it can be seen that when the precision was high, the recall was low and vice-versa. Although most of the drug pairs selected as similar were correct, resulting in a high precision rate, a large portion of similar pairs from the positive dataset were not identified, resulting in a low recall rate. In fact, once the recall rate reached a low of around 30%, the precision rate hit a perfect value, which means there was no false positive drug pairs in the predicted set, though very few from the positive sample were identified correctly by the model. To factor in the true negatives, Figure 6.3 shows the ROC curve which is a plot of the true positive rate against the false positive rate. The closer the curve is to the top left, the better the performance is, as the area between the curve and the x-axis is higher.

6.1.3 Experiment Results from Adverse Network Model

With this model, the information on the set of drugs which adversely interact with the drug pair was represented on a graph. The nodes refer to the drug while the edges refer to the level of interaction, which is categorised as major, moderate and minor. In this experiment, the set of interactive drugs for each drug in the drug pair was compared by examining the number of common paths that link the common drugs together.

ID1	Drug i	ID2	Drug j	Major	Combined
DB00916	Metronidazole	DB01211	Clarithromycin	0.099	0.247
DB01060	Amoxicillin	DB00415	Ampicillin	0.9	0.782
DB01050	Ibuprofen	DB01060	Amoxicillin	0.188	0.138
DB00682	Warfarin	DB01035	Procainamide	0.155	0.237
DB00682	Warfarin	DB00390	Digoxin	0.054	0.363
DB00682	Warfarin	DB01076	Atorvastatin	0.166	0.45
DB00331	Metformin	DB01211	Clarithromycin	0.038	0.278
DB00316	Acetaminophen	DB01050	Ibuprofen	0.312	0.233
DB00945	Aspirin	DB01060	Amoxicillin	0.185	0.137
DB00749	Etodolac	DB00784	Mefenamic acid	0.996	0.977
DB00573	Fenoprofen	DB00605	Sulindac	0.978	0.987
DB01050	Ibuprofen	DB00328	Indomethacin	0.978	0.988
DB00554	Piroxicam	DB00573	Fenoprofen	0.996	0.975
DB00605	Sulindac	DB00749	Etodolac	0.978	0.992
DB00500	Tolmetin	DB00554	Piroxicam	0.996	0.977
DB00188	Bortezomib	DB00072	Trastuzumab	0.73	0.432
DB00515	Cisplatin	DB00531	Cyclophosphamide	0.653	0.549

Table 6.8: Adverse network model: sample results from similar drug pairs

Table 6.8 and Table 6.9 show the results of the experiment performed on the respective similar and dissimilar drug pairs. Figure 6.4 shows the ROC plot with the distribution of the F-scores shown in Figure 6.5. When finding the common paths between sets of interacting drugs, the model performed better if only the set of interacting drugs were limited to those of major interaction. When the set of major interacting drugs was combined with drugs at other levels of interaction (minor and moderate), the performance deteriorated due to noise introduced into the additional nodes. However, at very low threshold values below 0.2, the F-score is higher for the combined mode of drug interaction as the impact of noise was not prominent enough to affect the performance. Hence, more drug pairs are classified correctly when the levels of interaction are combined. It is also interesting to note from Figure 6.5 that the maximum F-score occurs at a threshold value of around 0.5. This implies that this value can be used to maximise the performance of prediction when classifying a drug pair as similar or dissimilar.

ID1	Drug i	ID2	Drug j	Major	Combined
DB01050	Ibuprofen	DB08827	Lomitapid	0.118	0.275
DB00945	Aspirin	DB06605	Pixaban	0.373	0.518
DB01050	Ibuprofen	DB08896	Regorafenib	0.5	0.493
DB01050	Ibuprofen	DB06605	Apixaban	0.358	0.441
DB00945	Aspirin	DB08880	Teriflunomid	0.157	0.301
DB01050	Ibuprofen	DB08901	Ponatinib	0.459	0.433
DB00945	Aspirin	DB08896	Regorafenib	0.468	0.565
DB00945	Aspirin	DB08901	Ponatinib	0.43	0.49
DB01050	Ibuprofen	DB01254	Dasatinib	0.319	0.399
DB00945	Aspirin	DB06228	Rivaroxaban	0.382	0.526
DB00945	Aspirin	DB01254	Dasatinib	0.33	0.45
DB00945	Aspirin	DB00864	Tacrolimu	0.143	0.373
DB01050	Ibuprofen	DB00864	Tacrolimu	0.128	0.32
DB00945	Aspirin	DB01050	Ibuprofen	0.869	0.802
DB00788	Naproxen	DB00300	Tenofovir	0.252	0.293
DB00316	Acetaminophen	DB08880	Teriflunomid	0.097	0.21

Table 6.9: Adverse network model: sample results from dissimilar drug pairs

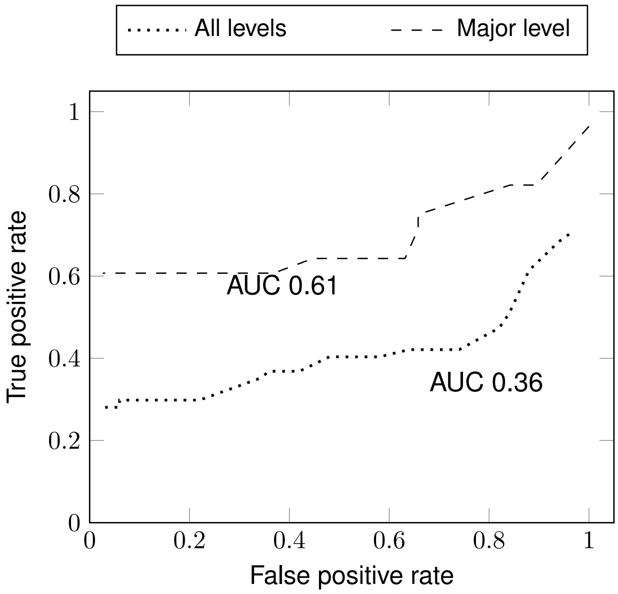


Figure 6.4: Adverse network model: ROC curve

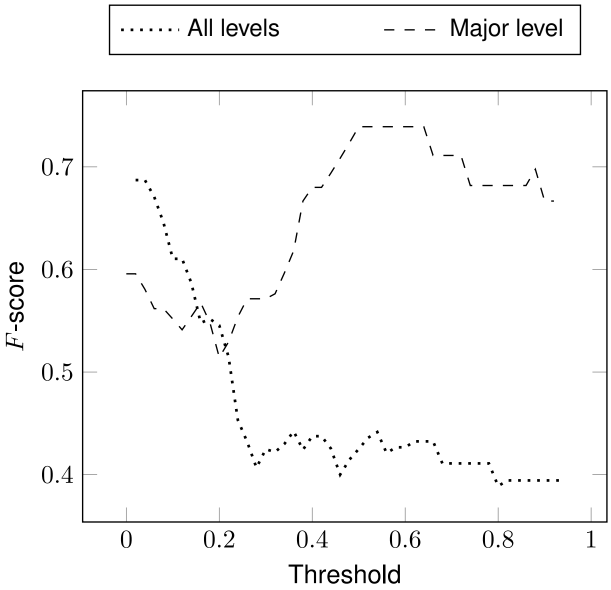


Figure 6.5: Adverse network model: F-score

6.1.4 Experiment Results from Word Embedding Model

Tables 6.10 and 6.11 show the sample results of similarity ratios for drug pairs that are similar and drug pairs that are dissimilar respectively with the word embedding model.

ID1	Drug i	ID2	Drug j	Sim(i,j)
DB00331	Metformin	DB01211	Clarithromycin	0.5599
DB00316	Acetaminophen	DB01050	Ibuprofen	0.6965
DB00573	Fenoprofen	DB00605	Sulindac	0.8529
DB00328	Indomethacin	DB00461	Nabumetone	0.7704
DB01009	Ketoprofen	DB01050	Ibuprofen	0.6855
DB00916	Metronidazole	DB01190	Clindamycin	0.8012
DB00788	Naproxen	DB00814	Meloxicam	0.5902
DB00991	Oxaprozin	DB01009	Ketoprofen	0.7573
DB00554	Piroxicam	DB00573	Fenoprofen	0.7875
DB00605	Sulindac	DB00749	Etodolac	0.8246
DB00500	Tolmetin	DB00554	Piroxicam	0.7985
DB00193	Tramadol	DB00316	Acetaminophen	0.6350
DB00586	Diclofenac	DB00316	Acetaminophen	0.7820

Table 6.10: Word embedding model: sample results from similar drug pairs

With the similarity ratios for the different drug pairs, the performance as to how accurately the model can predict the similarity of the drug pair is computed. At different threshold values of the similarity ratio, different F-scores were obtained. For example if the threshold chosen

ID1	Drug i	ID2	Drug j	Sim(i,j)
DB00945	Aspirin	DB00078	Ibritumomab	0.2750
DB01050	Ibuprofen	DB00078	Ibritumomab	0.1750
DB00945	Aspirin	DB00369	Cidofovir	0.5074
DB01050	Ibuprofen	DB05528	Mipomersen	0.2933
DB01050	Ibuprofen	DB08880	Teriflunomid	0.2372
DB00945	Aspirin	DB06605	Apixaban	0.4866
DB00945	Aspirin	DB00932	Tipranavir	0.6071
DB01050	Ibuprofen	DB08896	Regorafenib	0.2629
DB01050	Ibuprofen	DB06605	Apixaban	0.2754
DB00945	Aspirin	DB08880	Teriflunomid	0.3594
DB01050	Ibuprofen	DB08901	Ponatinib	0.2327
DB00945	Aspirin	DB00300	Tenofovir	0.5354
DB00945	Aspirin	DB00563	Methotrex	0.6750
DB01050	Ibuprofen	DB00300	Tenofovir	0.4575
DB01050	Ibuprofen	DB01254	Dasatinib	0.2426
DB00788	Naproxen	DB06228	Rivaroxaban	0.3980
DB00788	Naproxen	DB00465	Ketorolac	0.6367

Table 6.11: Word embedding model: sample results from dissimilar drug pairs

was 0.3, then the similarity ratios that were above 0.3 in Table 6.10 were considered true positive, while those below 0.3 in Table 6.11 were considered true negatives. The plot of true positive rate against false positive rate is shown in the next section to compare the results with other models. At a threshold of 0.5, Table 6.12 shows the confusion matrix which include the values for computing the F-score.

	Window Size=4	Window Size=2
Recall	0.85	0.88
Precision	0.67	0.93
F-score	0.75	0.64
True Positive	35	36
True Negative	31	5
False Positive	17	36
False Negative	6	5

Table 6.12: Performance of word embedding model at layer size=16

Experiment using TAC2018 dataset

Another experiment was conducted using dataset similar to the additional baseline model [161] to allow for a fairer comparison. Tables 6.13 and 6.14 show an extract of results from similar and dissimilar drug-pairs.

ID1	Drug i	ID2	Drug j	Sim(i,j)
DB01118	Amiodarone	DB00584	Enalapril	0.5282
DB00820	Tadalafil	DB00201	Caffeine	0.5007
DB00640	Adenosine	DB00502	Haloperidol	0.6297
DB00841	Dobutamine	DB00975	Dipyridamole	0.6954
DB00063	Eptifibatide	DB06402	Telavancin	0.6314
DB00187	Esmolol	DB00368	Norepinephrine	0.7024
DB08816	Ticagrelor	DB00208	Ticlopidine	0.8281
DB00765	Metyrosine	DB01203	Nadolrtan	0.8441
DB00727	Nitroglycerin	DB00091	Cyclosporine	0.5386
DB00177	Valsartan	DB00966	Telmisartan	0.8441
DB06228	Xarelto	DB01356	Lithium	0.5773

Table 6.13: Sample results from similar drug pairs

ID1	Drug i	ID2	Drug j	Sim(i,j)
DB00640	Adenosine	DB00277	Theophylline	0.1650
DB01118	Amiodarone	DB01356	Lithium	0.1421
DB00841	Dobutamine	DB00325	Nitroprusside	0.1447
DB00584	Enalapril	DB00877	Sirolimus	0.3890
DB00584	Enalapril	DB01590	Everolimus	0.2833
DB00063	Eptifibatide	DB01109	Heparin	0.1852
DB00695	Furosemide	DB00364	Sucralfate	0.1860
DB00695	Furosemide	DB00903	Ethacrynic acid	0.1460
DB00765	Metyrosine	DB00502	Haloperidol	0.1960
DB00727	Nitroglycerin	DB01109	Heparin	0.1960
DB08816	Ticagrelor	DB00227	Lovastatin	0.1465
DB08816	Ticagrelor	DB00641	Simvastatin	0.1587
DB06212	Tolvaptan	DB00035	Desmopressin	0.1049
DB00177	Valsartan	DB00384	Triamterene	0.1944
DB06228	Xarelto	DB01225	Enoxaparin	0.1420

Table 6.14: Sample results from dissimilar drug pairs

It can be seen that the similarity ratios for drug pairs in Table 6.13 are generally higher that those in Table 6.14. This supports the hypothesis that similar drug pairs have higher similar ratios. By classifying those drug pairs from 6.13 with similarity ratio above threshold value of 0.5 as 'similar', the number of true positives can be computed. Likewise, those drug pairs from Table 6.14 with similarity ratio above threshold value of 0.5 are classified as false positives. At this threshold value, the F-score of 0.85 was achieved. By varying the threshold values, different true positive values and false positive values can be obtained.

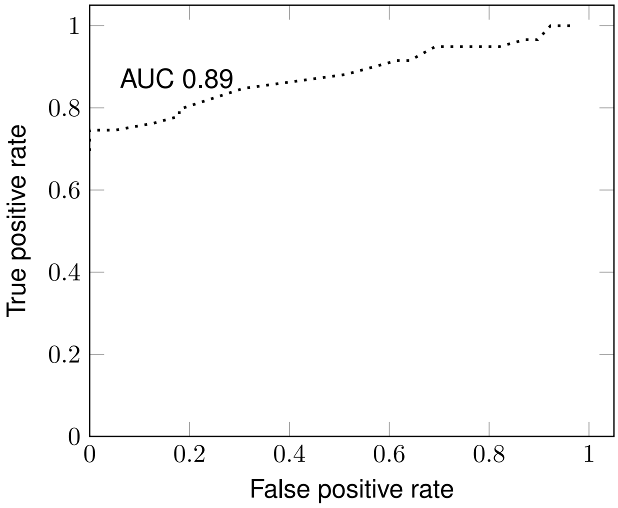


Figure 6.6: Word embedding model with TAC2018 dataset: ROC Curve

A plot of the true positive rate against the false positive rate is shown in Figure 6.6 which shows an AUC value of 0.89%.

6.2 Discussions

The experimental results support the hypothesis that similar drug pairs have a higher similarity ratio compared to dissimilar pairs. This theoretical finding has been applied in clinical systems to benefit healthcare professionals during drug prescription. Such a CDSS is the result of the experimental findings through the various models. The discussion that follows compares the performance between the various models and also with the baseline models.

6.2.1 Comparing Adverse Network Model with Word Embedding model

Table 6.15 shows the results obtained from individual models by running the experiment with the two sets of drug pairs. The word embedding model had a higher F-score in predicting positive drug pairs, hence leading to the higher recall rate of 0.85 compared to 0.61 for the adverse network model. In contrast, it had a lower precision rate (0.67 against 0.94), which measured the fraction of positive records that were accurately predicted. This was due to an increase in the false predictions (number of false positives). As the true positives increase, the number of positive pairs that were not correctly predicted (false negatives) decreases, resulting in an increase in the recall rate. When common paths for those drugs in adverse interaction with the original set of interactive drugs were included, the F-score dropped drastically compared to the case when only the original set of interactive drugs were considered. As expected, the performance deteriorated when additional attributes of adverse interactions, such as minor and moderate interactions, were introduced. However, due to fewer possible paths when only major interactions are considered, the threshold occurs sooner, where beyond that, there were no true positives obtained in the experiment, which explains the unavailability of the F-score when the cut-off was over 0.6. With the word embedding model, F-score was at a maximum at a layer size of 16. Performance deteriorated when the layer size was decreased since important information from the drug corpus was lost. Window size also affected the F-score. Since the number of words before and after the target word was decreased, the quality of the training model was adversely affected, hence the drop in performance with a smaller window size.

Threshold	**Adverse Network**				**Word Embedding**		
	r=1 (Major)	r=1 (All)	r=2 (Major)	r=2 (All)	w=2 L=16	w=4 L=16	w=4 L=8
0.1	0.55	0.61	0.52	0.70	0.67	0.63	0.65
0.2	0.51	0.55	0.45	0.59	0.67	0.64	0.65
0.3	0.57	0.42	0.41	0.49	0.67	0.71	0.66
0.4	0.68	0.44	0.40	0.44	0.66	0.74	0.68
0.5	0.74	0.43	0.47	0.36	0.64	0.75	0.71
0.6	0.74	0.43	-	0.35	0.61	0.71	0.70
0.7	0.71	0.41	-	0.38	0.51	0.62	0.74
0.8	0.68	0.39	-	0.34	0.31	0.36	0.60
0.9	0.67	0.39	-	-	-	-	-

Table 6.15: F-score distribution

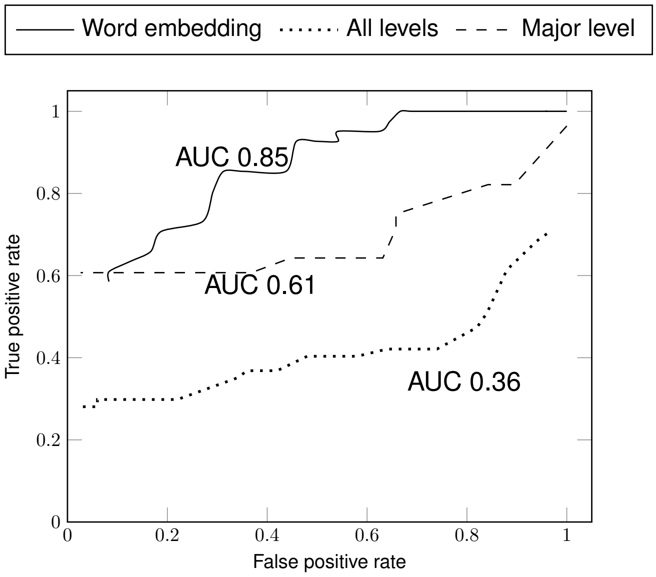


Figure 6.7: AUC for adverse network and word embedding models

Since precision does not factor in the correct negative predictions within the drug pairs, (true negatives), we attempt to assess this performance by plotting the true positive rate *tpr* against the false positive rate *fpr* to obtain the receiver operating characteristic (ROC) curve [33]. With this plot, the area under curve (AUC) can be used to further determine the performance of the model in a more comprehensive manner. A higher AUC indicated a better performance [19]. The AUC for the word embedding model was 0.85 compared with that for the network model which is 0.61 (Figure 6.7). When minor and moderate interactions were also included in considering the number of common paths within the drug pair, it was noted from the ROC that the AUC was less than 0.5. This is due to the noise introduced into the experiment with the additional paths, which does not aid performance.

6.2.2 Comparing Statistical Model with Word Embedding model

Comparison of the statistical and the word embedding approach at different threshold values showed that the latter performed better (Figure 6.8). An interesting comparison between the results obtained from the statistical approach and word embedding approach is noted. With the same dataset, the F-score obtained from the statistical approach was 0.52 compared with 0.68 using the deep learning approach. At different threshold values of θ, results are obtained and compared for the two different approaches. With the common dataset used for both methods, results show that the Word2Vec approach performed better than the statistical approach (Figure 6.8). This was expected as the former approach in computing similarity was

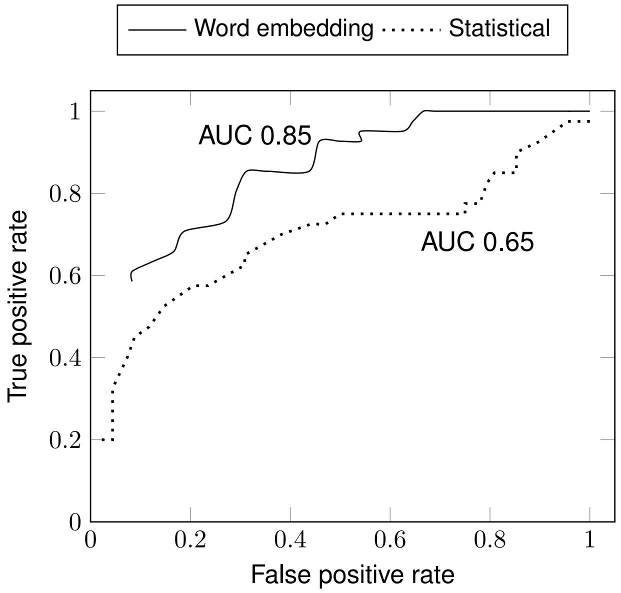


Figure 6.8: Performance comparison of statistical and word embedding models

to gather the term frequency from the text describing the drugs. In the latter approach, feature vectors used to find the similarity were obtained from closely related words. To illustrate the conceptual framework of this study, the same model can be used to decide if the drug is suitable for prescription. Based on the overall similarity from the three properties of the drug pair, the system can detect if the drug is similar to the drugs that the patient is allergic to. This approach highlights the usefulness of our framework where knowledge generated from the prediction layer can be applied to the presentation layer and become useful to the user, in this case, as a decision support tool for the healthcare professional. This novel strategy is in line with the aim of the study: allowing us to support the dentist with the right prescription by ensuring the drug is not in adverse relationship with the drugs the patient is taking, and also dissimilar to the drugs the patient is allergic to.

6.2.3 Comparing with Baseline Models

In terms of accuracy, which indicates the percentage of correct predictions taking into consideration both the similar and dissimilar predictions, the adverse network model achieved 82% when only immediate neighbours with major interactions are considered. This is superior to the baseline model from Tari(2010) which is 77.7% [138] and Yan *et al.*(2013) which achieved 69% [155]. One factor that adversely affected the performance of these baseline models was the way information was retrieved from the textual description associated with the drug pair. As information used by the baseline models included genetic structures which very often are embedded in tables and figures, these could not be easily detected [87] and thus affected the results of the study. This is aggravated with the models attempting to parse for semantic information on biological entities like induction of enzymes. MeSH is not used in the experiment as it does not consistently link drugs to diseases and other conditions [123]. Nevertheless, to ensure a fair comparison, an experiment with the official dataset from Text Analysis Conference (TAC) 2018, which contains gold standard annotations from the National Library of Medicine and the U.S Food and Drug Administration[1], is used for the word-embedding model. When compared to the baseline model by Zhang [161] which yielded a F-score of 0.73, our work achieved a F-score of 0.85. It can also be observed that with a dedicated dataset from TAC2018, the AUC improved from 85% to 89% (Figure 6.6).

Using the top ten drug interactive pairs from a work done by Rohani *et al.* [114] does not produce satisfactory result when using online checkers. For example, MIMS (Monthly Index of Medical Specialties), a classic interactive checker which doctors and dentists have been using for over half a century, was not able to report on any interactions for the ten pairs of drugs in Table 6.16.

Drug	ID	Drug	ID
DB00642	Pemetrexed	DB01331	Cefoxitin
DB00642	Pemetrexed	DB01060	Amoxicillin
DB00633	Dexmedetomidine	DB01183	Naloxone
DB00633	Dexmedetomidine	DB00361	Vinorelbine
DB00535	Cefdinir	DB00373	Timolol
DB01236	Sevoflurane	DB01586	Ursodeoxycholic acid
DB01236	Sevoflurane	DB00415	Ampicillin
DB00742	Mannitol	DB00441	Gemcitabine
DB00585	Nizatidine	DB01577	Methamphetamine
DB01136	Carvedilol	DB00952	Naratriptan

Table 6.16: Top ten DDI's from [114]

The same goes with Medscape[2], another common online drug interaction checker. In contrast, the prototype predicted a similarity ratio of less than 60% for the majority of the ten pairs. As a lower ratio shows a higher chance of adverse interaction, this shows the attractive performance of our prototype.

Although the approach adopted in the experiment performs better even with similar dataset used by the baseline model by [161], it will be interesting to compare the performance

[1] https://bionlp.nlm.nih.gov/tac2018druginteractions/trainingFiles.zip (Accessed 16 Jul 2020).

[2] https://reference.medscape.com/drug-interactionchecker (Accessed 14 Jul 2020).

when other datasets are used for the experiment. As mentioned under Future Work (See Section 8.3), comparison of performance can be done in future with other common datasets like PubMed.

6.3 Sensitivity Study

Experimental parameters were varied to find the combination that yielded the best performance. These parameters included the proximity distance from the root node and the relationship between the nodes in the adverse network model. Word size and layer size were also varied in the word embedding model. As shown in Table 6.17, the adverse network model performed best by only considering the major interaction between nodes in the immediate neighbourhood of each drug in the drug pair. Table 6.18 shows the performance of the word embedding model with varying window sizes and layer sizes.

Property	Promixity	Recall	Precision	Accuracy	F-score
Major only	1	0.61	0.94	0.82	0.74
	2	0.34	0.75	0.60	0.47
Combined	1	0.30	0.74	0.57	0.43
	2	0.34	0.38	0.37	0.36

Table 6.17: Effect of proximity and nodes properties on performance of the adverse network model

Window size	Layer size	Recall	Precision	Accuracy	F-score
2	8	1.00	0.49	0.52	0.66
2	16	0.98	0.49	0.53	0.66
4	8	0.98	0.56	0.63	0.71
4	16	0.85	0.67	0.74	0.75

Table 6.18: Influence on performance by training parameters

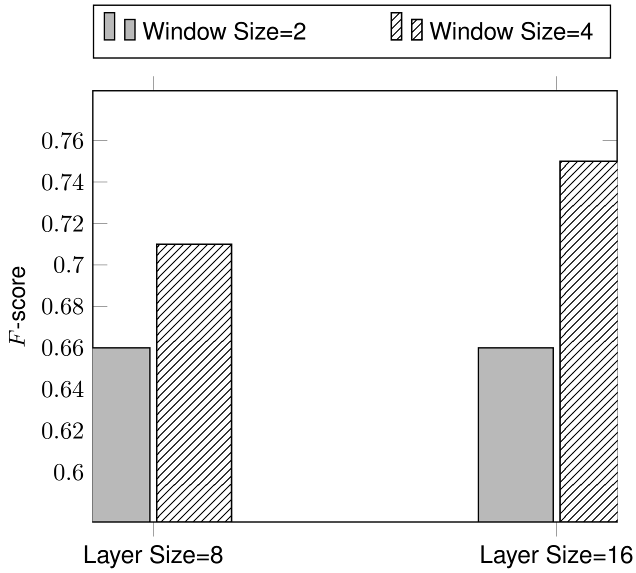


Figure 6.9: Effect of Word2Vec parameters on performance

Changing the window size affected the performance significantly. This is shown in Figure 6.9. Since a smaller number of words before and after the target word was used during training, it is expected that the probability of a word-match with the drug pair during the experiment would be lower, hence the drop in performance. Changing the layer size had minimal impact on the performance. The model performed best at a window size of 4 and a layer size of 16.

6.4 Summary

Chapter 6 presented and discussed the experimental results for the various models in extracting feature vectors to determine the similarity of a drug pair. The results were obtained by running the experiment with positive and negative drug pairs. The models in the experiment used information from the knowledge layer to build feature vectors to enable the similarity ratio to be computed. The model that used the word embedding method for building feature vectors performed best with a F-score of 0.75. This was due to the use of the relationships between words instead of word frequency for computing the similarity ratio of the drug pair. It was also found that the performance of the network model was sensitive to the number of paths used.

The experimental results support the hypothesis that similar drug pairs have a higher similarity ratio than dissimilar pairs. These findings have already made an impact on the creation of CDSS which has contributed to the efficient prescription of drugs by healthcare professionals. The next chapter illustrates such a system by combining the clinical history of the patient in terms of drug allergies and current medications.

Chapter 7

Clinical Deployment

As mentioned in Section 1.3, the aim of the research is to assist the healthcare professional to ensure that the drug to be prescribed is safe for the patient in terms of adverse interactions and allergies. Chapter 7 shows how the framework described in the thesis is being implemented for clinical use. To ensure the user finds it relevant and useful, additional features are presented to the user depending on the specific drugs being prescribed. Hence, some of the interactive behaviours of common drugs used in the dental clinic that influence the design of the prototype will also be described. The results of a survey on dentists regarding ease of use of the system will also be reported.

7.1 Design of Prototype

In connection with the three-layer framework, the prototype system is a means to allow the results of the research models to be visualised and implemented for use by the healthcare professional to prescribe drugs based on the current drugs that the patient is taking and the drug allergies that the patient may have. It is assumed that the healthcare professional does not prescribe drugs that the patient is currently taking. Hence, the system to be described belongs to the presentation layer within the conceptual framework described in Section 3.1. Within the patient profile, medical condition is excluded as it is not as common and crucial compared to the drugs that the patient is taking and drugs that the patient has allergy. Hence only the first two tuples are taken into consideration. This prototype is a web-based product to be used within a clinical environment. The users are healthcare professionals and clinic assistants. Since this is a data-centric product, a database is required to store the data. Based on the results from the empirical experiments reported in Chapter 6, the word embedding model (Section 4.5) was employed to build the feature vectors from the knowledge base. This model was used within the prediction algorithm to determine the similarity ratio of a drug pair. As described in Chapter 4, similarity ratio of the drug pair was obtained from the vectors that represent words most related to each drug.

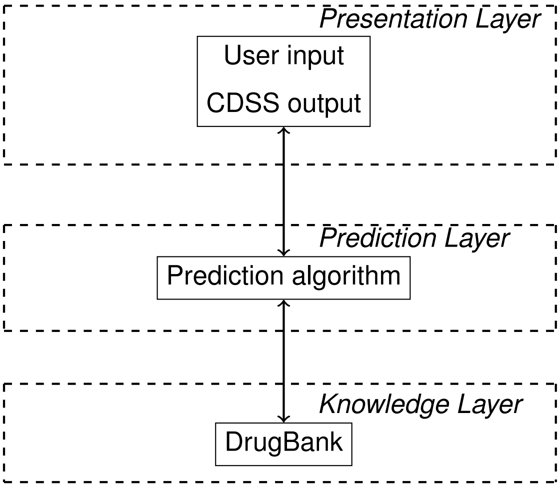


Figure 7.1: Mapping of prototype to the three-layer framework

The main engine of the prototype lies in the prediction algorithm as shown in Algorithm 3. The prediction algorithm, which resides within the prediction layer (Figure 7.1), presents the results to the presentation layer. At the same time, it also requires information from the presentation layer associated with the drugs to be prescribed, the drugs that the patient is taking and the drugs that the patient is allergic to. Besides the information related to the drugs which is stored in DrugBank in the knowledge layer, additional information on usual dosages prescribed when using the prototype is also stored (Table 7.1). Such information will increase user-friendliness and adoption of the system. A more detailed description of common drugs used in the dental clinic is explained in Section 7.2.

In order for the prediction algorithm to determine if the drug pair is safe for prescription, similarity ratio was computed based on the feature vectors generated from the word embedding model (Section 4.5). These feature vectors were obtained in the same manner as described in Section 5.3.4. With the vectors obtained for each drug pair, similarity ratio can be computed

Drug	**Qty**	**Unit**	**Dosage**	**Dosage** unit	**Frequency**	**Duration**	**Qualifier**	**Unit** per dosage
Amoxicillin	30	capsules	250	mg	3 times daily	5 days		2
Paracetamol	10	tablets	500	mg	3 times daily		when needed	1
Metronidazole	15	tablets	200	mg	3 times daily	5 days		1
Erythromycin	30	tablets	250	mg	3 times daily	5 days		1
					daily		needed	
Ibuprofen	20	tablets	200	mg	3 times daily		when needed	1
Clarithromycin	5	tablets	500	mg	Once daily	5 days		1
					daily		needed	

Table 7.1: Common drugs prescribed in a dental clinic

for the drug pair.

Referring to Algorithm 3, there were two crucial tests to determine if the drug is safe for prescription.

```
input  : Let $D^p$ be the set of prescription drugs;
         $d_j$ be the medicine to be prescribed by the dentist;
         $d_i$ is the medicine the patient is currently taking;
         $d_g$ is the medicine the patient is allergic to;
output: Recommended prescription

Let $\beta$ be flag for drug allergy;
Let $\delta$ be flag for adverse relationship of drug pair;
Let $\theta$ be threshold of similarity for drug pair $d_i$ and $d_j$;
Initialise $\delta$ to false and $\beta$ to false;
Create default candidate set that belongs to same class as $drug_j$;
Let recommended drug $d_r$ be drug from candidate set;
while $\delta$ is false or $\beta$ is $false$ do
    $d_j \leftarrow d_r$;
    while there are more drugs in drug allergy set do
        if $d_j$ belongs to same class $d_g$ then
            $\beta$ = $false$;
            break from loop;
        end
    end
    while there are more drugs in drug taking set do
        find $Sim(d_i, d_j)$;
        if $Sim(d_i, d_j) \leq \theta$ then
            $\delta$ = $false$;
            break from loop;
        end
    end
    if ($\delta$ == $false$ ) or ($\beta$ == $false$) then
        get $d_r$ from next drug in candidate set;
    end
    if there are no more items in candidate set then
        break and exit from testing for $\delta$,$\beta$;
    end
end
display recommended drug $d_r$.
```

Algorithm 3: Prediction Algorithm for personalised prescription support

The first test was to check if the drug to be prescribed belongs to the same group of drugs which the patient is allergic to. To do so, the algorithm used the information in the knowledge layer to determine which group the drug to be prescribed belongs to.

The second test used the word embedding model to check the similarity between the drug to be prescribed and the drug that patient is currently taking. If there was an adverse drug interaction, a candidate list of drugs with a similarity ratio above the threshold specified by the user was displayed. This condition uses the hypothesis of the research which is supported by the results that a similar drug pairs has a higher similarity ratio. Hence the drug pairs below the threshold value were deemed to be in an adverse relationship and were not be shown to the user. The process repeats until a drug was found to satisfy both tests. The algorithm terminated once the list of candidate drugs was exhausted or a drug was found.

In order to enhance user adoption, the prototype incorporated additional features within the presentation layer to alert the user if certain conditions of the drug were met. For example, if a painkiller was prescribed for a child or an elderly patient, the system gave an alert of the dosage limitation to avoid over dosage. The following section attempts to highlight some of the features associated with the common drugs used in the dental clinic.

7.2 Common Drugs in Dental Clinics

The number of drugs to be prescribed within a CDSS in a dental clinic is small compared to a medical clinic. However, it is still important for the dentist to be aware of the different groups of drugs available and the way the system suggests and responds to the user's query. This will ensure that the results from the system can be understood in a better perspective.

The approach taken in the design of the system considered the patient's current medications and drug allergies. Additional information related to common drugs used in the dental clinic were included in the CDSS to maximise the usability of the system. These can be grouped into the following four main classes [54]:

- antimicrobials: antibiotics used to treat infections that may result after dental treatment; e.g. penicillin, metronidazole, clindamycin
- analgesics: relief of toothache or pain following dental treatment, e.g. ibuprofen, aspirin
- corticosteroids: relief of oral discomfort and swelling, e.g. betamethasone
- anxiolytics: management of anxiety during dental treatment, e.g. diazepam and temazepam

Name	ID	Class
Anti-bacterial drugs		
beta lactams - penicillin		
Dicloxacillin	DB00485	Penicillinase resistant penicillins
Amoxicillin	DB01060	Aminopenicillins
Ampicillin	DB00415	Aminopenicillin
beta lactams - cephalosporins		
Cephalexin	DB00567	First generation cephalosporins
lincosamides		
Clindamycin	DB01190	Lincomycin derivatives
Lincomycin	DB01627	Lincomycin derivatives
nitroimidazoles		
Metronidazole	DB00916	Amebicides
Tinidazole	DB00911	Amebicides
tetracyclines		
Doxycycline	DB00254	Miscellaneous antimalarials
Anti-fungal drugs		
Fluconazole	DB00196	Azole antifungals
Itraconazole	DB01167	Azole antifungals
Miconazole	DB01110	Topical antifungals
Amphotericin B	DB00681	Polyenes
Nystatin	DB00646	Polyenes
Anti-viral drugs		
Penciclovir	DB00299	Topical antivirals
Famciclovir	DB00426	Purine nucleosides
Aciclovir	DB00787	Topical antihyper
Valaciclovir	DB00577	Purine nucleosides

Table 7.2: Common antimicrobial drugs

7.2.1 Antimicrobials

Antimicrobials are commonly used after surgical procedures such as placement of dental implants and gum treatment [133]. While active treatment is usually required when patients have an infection, the dentist will still prescribe this drug to provide relief from pain due to the infections pending the design of the treatment plan. It should be noted that adverse drug interaction may take place as a result of unnecessary prescription of antimicrobial drug, depending on what other drug the patient is taking.

Table 7.2 shows common antimicrobial drugs used by the dentist. They consist of anti-bacterial drugs, anti-fungal drugs and anti-viral drugs. Groups within anti-bacterial drugs are:

- beta lactams
- glycopeptides

- lincosamides
- nitroimidazoles
- tetracyclines

Penicillin and its variants (e.g. amoxicillin) is a very common drug prescribed [62], and is also a very common drug allergy.

M	Microbiology guides therapy wherever possible
I	Indications are evidence based
N	Narrowest spectrum required
D	Dosage appropriate to the site and type of infection
M	Minimise duration of therapy
E	Ensure mono-therapy in most cases

Table 7.3: Antimicrobial creed

It is thus given special attention in the decision support system. Users are alerted to the protocol of prescribing antibiotics commonly known as the MINDME creed (Table 7.3) [1], which involves choosing the antibiotics with the narrowest spectrum for an appropriate of time at an appropriate dosage.

It should be noted that penicillin belongs to the *beta lactams* group. Hence in the prototype, when the search engine encountered penicillin as a drug allergy, all other drugs belonging to the *beta lactams* group should also be avoided. Properties of these drugs in the knowledge base were also manually annotated to ensure higher accuracy when processing drug rankings in the CDSS.

Other important messages generated by the system were:

- Amoxicillin and clavulanate should not be prescribed together as it can caused diarrhea which occur more frequently than with amoxicillin alone. But combining amoxicillin and clavulanate is sometimes preferred over amoxicillin alone for stronger infections because clavulanate prevents the bacteria inactivating the amoxicillin.
- Clindamycin should not be given intravenously unless necessary to avoid arrhythmias as it is well absorbed orally.
- Roxithromycin is rarely used unless patient is allergic to penicillin.

[1] https://www.nps.org.au/news/antibiotic-resistance-in-australia-here-and-now (Accessed 12 Mar 2019).

- Doxycycline is the preferred drug within the *tetrayclines* group.
- Avoid applying topical aciclovir and penciclovir to mucous membranes as they may irritate the patient

Since adverse effects on the elderly is common with this group of drugs, such additional feature is incorporated in the prototype design described in the previous section. This is crucial as safe prescription is essential in preventing morbidity and mortality [38]. Since it is common for geriatric patients to be taking more than one kind of drugs, the prototype will issue a warning when patients above 65 years old are prescribed anti-bacterial drugs.

7.2.2 Analgesics

Analgesics are used for pain relief, which is an important consideration in the clinical workflow of a dental clinic [109]. Three common groups in this class are:

Non-steroidal anti-inflammatory drugs (NSAIDs) This group of painkiller drugs is the most commonly used drugs for pain relief. Since the majority of toothache is inflammatory in origin, NSAIDs is a popular choice to treat acute dental pain. Hence, in the deployment of the CDSS, priority is given to this group of drugs in the ranking list when suggesting alternative drugs. Since the effect is dose-related, a warning is also given to ensure only the required dosage is presribed. The need to prescribe this drug should be assessed carefully especially among the elderly as they have a higher risk to adverse effects. Hence attention is drawn to the user of the system whenever this group of drug is prescribed for the elderly.

Paracetamol Paracetamol is generally considered safe, and often used to reduce the required dosage of NSAIDs, thus minimising the adverse effects that comes with NSAIDs.

Since this is a common drug, the system will give a warning to the dentist to ensure that the prescription of this drug does not result in a overdose, which can be life threatening. The optimal dosage for children below 12 years old is 15mg/kg every four to six hours and not exceeding 60mg per day. For adults, the recommended dosage is 500mg to 1000mg every four to six hours with no more than 4g in 24 hours[2].

In reviews covering 72 randomised controlled trials, majority of those trials involving up to

[2] Australian Department of Health Therapeutic Goods Administration https://www.tga.gov.au/community-qa/recommended-paracetamol-doses (Accessed 15 May 2019).

System	Adverese effects
Gastrointestinal	Constipation Nausea Vomiting
Cutaneous	Pruritus Sweating
Neurologic	Sedation/fatigue Headache Delirium/confusion Clouded vision Dizziness
Autonomic	Xerostomia Bladder dysfunction (eg, urinary retention) postural hypotension

Table 7.4: Adverse effect of opioids [7, 102]

Name	ID	Class
Codeine	DB00318	Antitussives
Acetaminophen	DB00316	Miscellaneous analgesics
Ibuprofen	DB01050	Nonsteroidal anti-inflammatory agents
Oxycodone	DB00497	Narcotic analgesics
Morphine	DB00295	Narcotic analgesics
Tramadol	DB00193	Narcotic analgesics

Table 7.5: Common analgesic drugs

100 children, conclusions were either unclear, inconclusive or there was no opinion regarding the safety and efficacy of paracetamol and ibuprofen [112]. Thus to avoid potential adverse effects due to overdose, it is best to take precautions when assessing the need for prescribing this drug.

Opioids Opioids are used to treat severe pain, especially pain after surgery. The common drugs used in the dental clinic within this group include codeine, morphine, oxycodone and tramadol. An alert to the user will be triggered once the dentist prescribe this drug to ensure a balance between pain management and the clinical burden of prescribing this drug. In fact they are the most commonly misused prescription drugs according to a report from the Substance Abuse and Mental Health Service Administration [91]. A list of side-effects is shown in Table 7.4, and is made available by the system whenever the dentist prescribes drugs belonging to this group.

Table 7.5 shows the common painkillers used in the dental clinic.

7.2.3 Corticosteroids

Corticosteroids are used extensively in managing many oral diseases due to their excellent anti-inflammatory and immuno-modulatory effects [119].

They have a significant impact on postoperative pain reduction at 4 to 6 hours and 12 hours following endodontic treatment [93]. Although this drug can be used in the tooth, topically on the oral mucosa or systemically, intra-dental or topical intra-oral use is more common to minimise adverse effects [106], as well as for ease of administration and cost effectiveness [16].

Name	ID	Strength
High potency		
Betamethasone	DB00443	0.1%
Mometasone	DB00764	0.1%
Moderate potency		
Triamcinolone acetonideMethylprednisolone aceponate	DB00959	0.1%
Clobetasone	DB01013	0.05%
Desonide	DB01260	0.05%
Low potency		
Hydrocortisone acetate	DB00741	0.5-1%

Table 7.6: Properties of topical corticosteriods ointment used on the oral mucosa [130]

Given the potent effect of corticosteroids, the system will show a help message (Table 7.6) to the user once any drug within this group is recommended by the prototype. This is to warn the dentist of the associated risks so that a well-informed decision can be made to weigh the benefits of the therapeutic power against the potential risks.

In addition, the user will be prompted with these important points to consider when prescribing corticosteroids:

- To avoid over-dosage, start with the lowest potency at the lowest dose and as infrequent as possible.
- When using corticosteroids, avoid cotton tips and fibrous materials to minimise damage to fragile atrophic mucosa. It is best applied with the pad of a washed finger.
- As absorption is considerably rapid, there is no need to avoid food and drink for a prolonged period.

7.2.4 Anxiolytics

Anxiolytics have the ability to relieve anxiety due to its hypnotic effect.

Common anxiolytics in the dental clinic are diazepam and temazepam. These two drugs belong to the benzodiazepines group where prescription is not uncommon in dental clinics [32]. Table 7.7 shows some of the drugs within this group as extracted from Drugbank. The user will be alerted with these messages when any drug in this group is prescribed:

Name	ID	Class
Diazepam	DB00829	Benzodiazepine anticonvulsants
Oxazepam	DB00842	Benzodiazepines
Temazepam	DB00231	Benzodiazepines

Table 7.7: Common anxiolytics

- This drug may affect the ability to drive and operate machinery safely.
- Avoid prescribing this drug to patients with myasthenia gravis or severe respiratory or hepatic impairment.

7.3 Point-of-Care Scenario

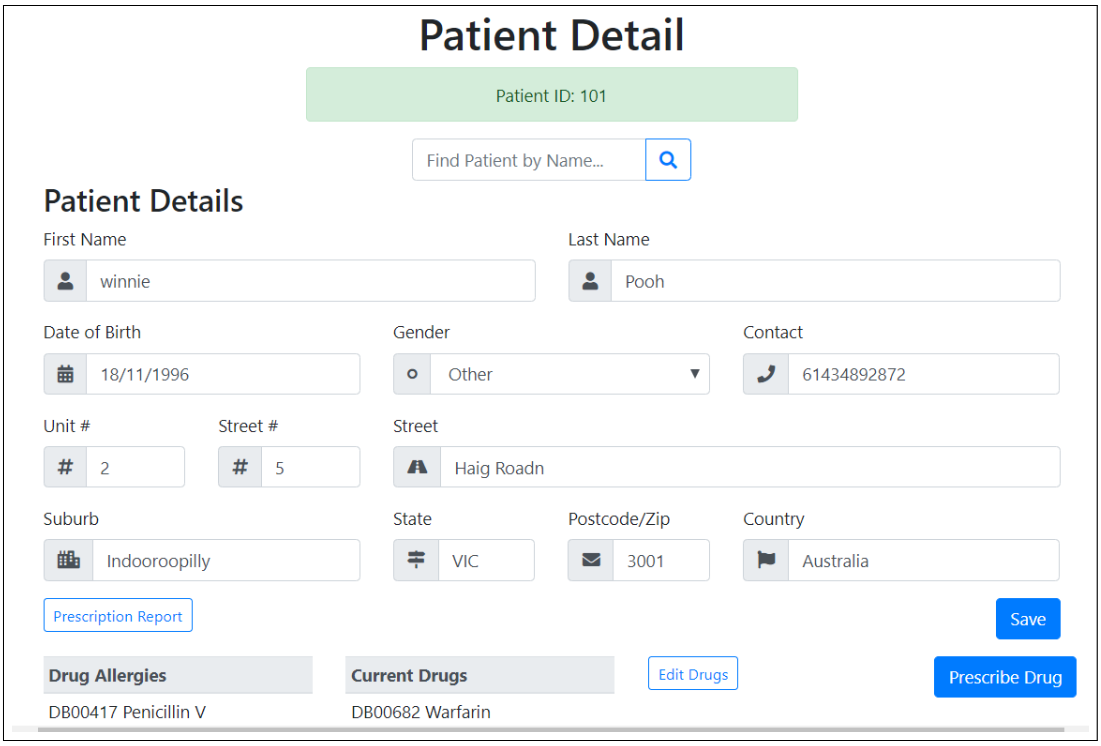
In this scenario, a typical clinical flow is described to illustrate how the prototype can be adopted and used in a dental clinic. It is assumed that the system does not claim to treat the patient's medical condition - it only attempts to check for possible side-effects of the drug to be prescribed with the condition. For example, if the patient has a cardiovascular condition, the system will not consider the therapeutic effect of the drug to be prescribed, but it should consider the adverse effects it may have in the setting of the cardiovascular condition. The system in this scenario also assumes that the patient does not have a cross-allergy to the drugs they are currently taking. This is a safe assumption since the fact that patient can attend for dental treatment shows that the patient can function normally and is not impaired by the adverse effects of the drugs. Furthermore, the drugs that the patient is currently taking is assumed to be prescribed by a medical doctor who should already have considered the patient's medical conditions and known drug allergies.

Since the deployment based on the results of the study is a decision support system, the user has the liberty to overwrite the system's suggestions as the function of the system is limited to assisting the user in checking for possible adverse reactions between the drug to be prescribed and the drugs that patient is currently taking. The system assumes that information regarding the patient's current medications and drug allergies is accurate and up-to-date.

The scenario comes in two different events which are common within a typical clinical flow of a dental clinic: patient registration and drug prescription.

7.3.1 Patient Registration

Consider when a patient attends for consultation. Before any treatment is performed, the dentist updates the patient's profile in regards to the medical conditions, the drugs the patient is currently taking and the drugs the patient is allergic to. There are no changes to the previous record with regards to drug allergies (*penicillin)*; however, the patient has recently been diagnosed with a heart condition, meriting an update of the profile with regards to the medical conditions and the drugs the patient is currently taking (*warfarin*).



Patient Detail

Patient ID: 101

Patient Details

First Name: winnie

Last Name: Pooh

Date of Birth: 18/11/1996

Gender: Other

Contact: 61434892872

Unit #: 2

Street #: 5

Street: Haig Roadn

Suburb: Indooroopilly

State: VIC

Postcode/Zip: 3001

Country: Australia

Drug Allergies	Current Drugs
DB00417 Penicillin V	DB00682 Warfarin

Figure 7.2: Registering patient into the system

Once the user is signed into the system, the user can either register a new patient or search for the records of an existing patient (Figure 7.2). The patient's drug profile is shown beneath the demographic record. The user can either edit the records or click on the **Prescribe Drug** button to begin prescribing a drug for the patient. Note that in this case the patient is taking *warfarin* and is allergic to *penicillin*.

7.3.2 Drug Prescription

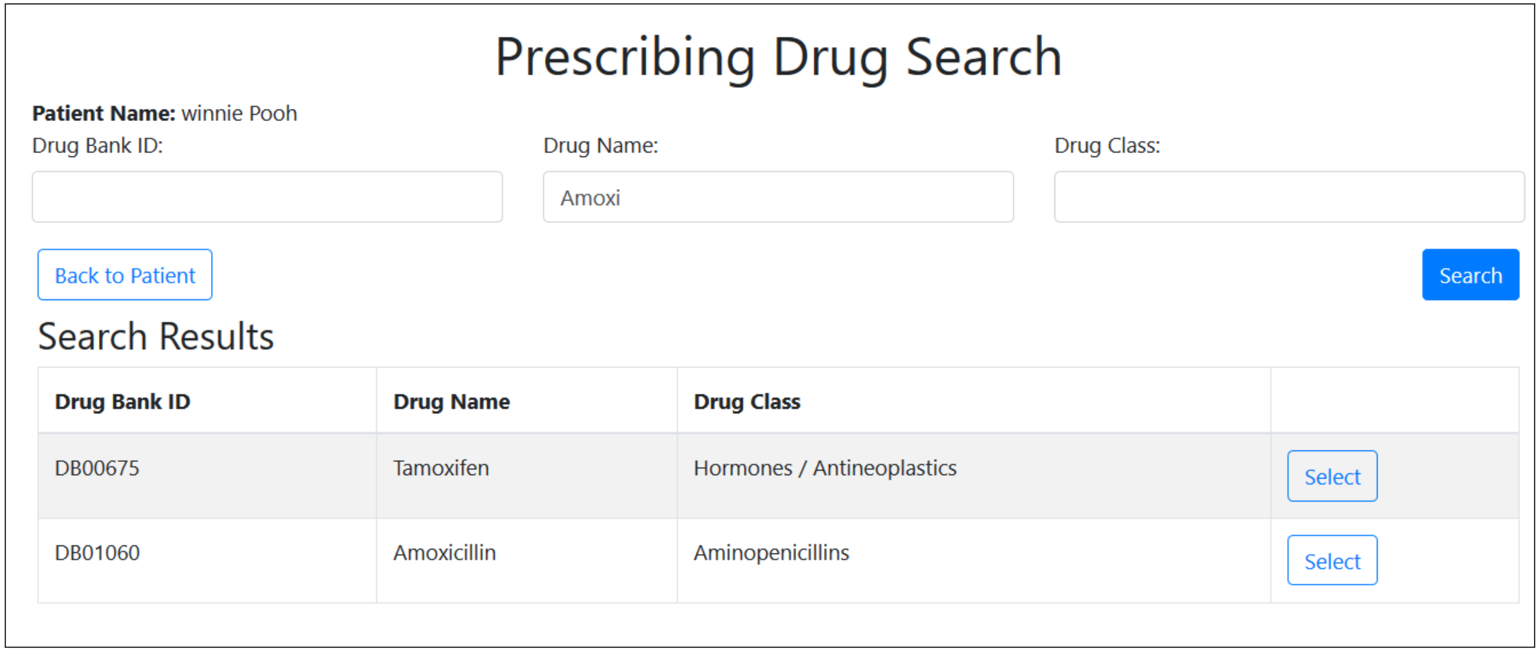
During treatment, it is decided that the patient requires an antibiotic. In this case, the dentist may consider the commonly prescribed *amoxicillin*.

At the drug prescription screen (Figure 7.3), the user may enter any part of the drug name and the system will display all the drugs with names containing the string of words entered by the user. For example, if the user enters *Amoxi*, the drug *amoxicillin* will appear. As indicated in Algorithm 3, not only should *amoxicillin* be dissimilar to *penicillin*, it should also not adversely interact with *warfarin*. The next screen (Figure 7.4) shows the list of suggested drugs which take into consideration the patient's current medications and drug allergies. The number in the list of suggested drugs will increase as the threshold for similarity is lowered. In this example, the drugs are at least 70% similar to the prescribed drug *amoxicillin* and 50% similar to *warfarin*, the drug that the patient is currently taking.

On the other hand, if the dentist is considering the prescription of *penciclovir*, the model will evaluate and produce another list of suggested drugs. Notice that the threshold for the similarity ratio with *warfarin* is now raised to 90% to reduce the number of suggest ed drugs (Figure 7.5). On receiving the suggestion of the alternative drug, it is then for the dentist to decide whether this is an appropriate drug to prescribe after further consideration of the duration and dosage of the patient's current drugs.

7.4 Dentists' Adoption

One of the unique features of this research is the clinical translation where the conceptual framework is deployed as a CDSS for drug prescription. Dentists at the clinic were involved during the design stage of the prototype. They have also been using it whenever they need to query on the status of the prescription. This section describes the usefulness of the system from the dentist's point of view.



Prescribing Drug Search

Patient Name: winnie Pooh

Drug Bank ID:

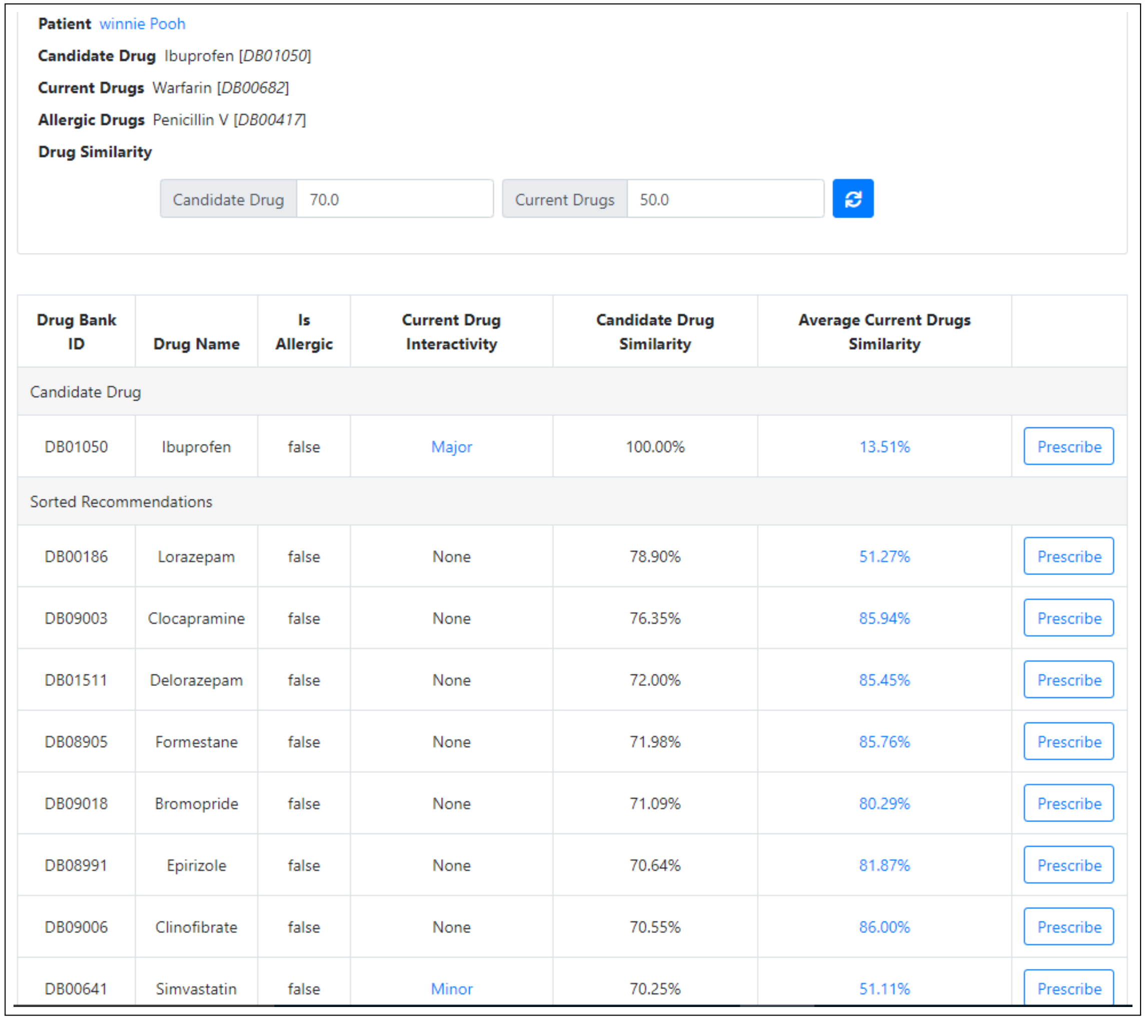
Drug Name: Amoxi

Drug Class:

Search Results

Drug Bank ID	Drug Name	Drug Class	
DB00675	Tamoxifen	Hormones / Antineoplastics	Select
DB01060	Amoxicillin	Aminopenicillins	Select

Figure 7.3: Prescribing drug based on patient's profile



Patient winnie Pooh

Candidate Drug Ibuprofen [*DB01050*]

Current Drugs Warfarin [*DB00682*]

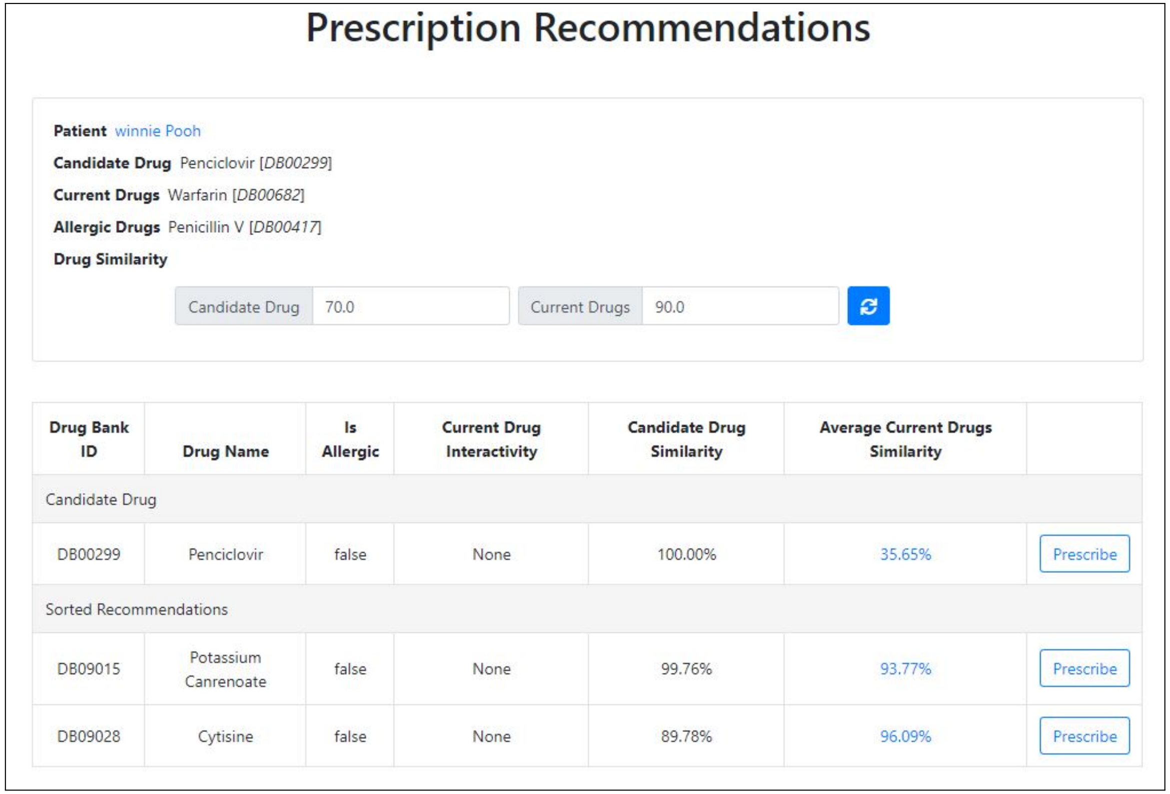
Allergic Drugs Penicillin V [*DB00417*]

Drug Similarity

Candidate Drug 70.0 Current Drugs 50.0

Drug Bank ID	Drug Name	Is Allergic	Current Drug Interactivity	Candidate Drug Similarity	Average Current Drugs Similarity	
Candidate Drug						
DB01050	Ibuprofen	false	Major	100.00%	13.51%	Prescribe
Sorted Recommendations						
DB00186	Lorazepam	false	None	78.90%	51.27%	Prescribe
DB09003	Clocapramine	false	None	76.35%	85.94%	Prescribe
DB01511	Delorazepam	false	None	72.00%	85.45%	Prescribe
DB08905	Formestane	false	None	71.98%	85.76%	Prescribe
DB09018	Bromopride	false	None	71.09%	80.29%	Prescribe
DB08991	Epirizole	false	None	70.64%	81.87%	Prescribe
DB09006	Clinofibrate	false	None	70.55%	86.00%	Prescribe
DB00641	Simvastatin	false	Minor	70.25%	51.11%	Prescribe

Figure 7.4: Candidate list of drugs for *ibuprofen*



Prescription Recommendations

Patient winnie Pooh

Candidate Drug Penciclovir [*DB00299*]

Current Drugs Warfarin [*DB00682*]

Allergic Drugs Penicillin V [*DB00417*]

Drug Similarity

Candidate Drug 70.0

Current Drugs 90.0

Drug Bank ID	Drug Name	Is Allergic	Current Drug Interactivity	Candidate Drug Similarity	Average Current Drugs Similarity	
Candidate Drug						
DB00299	Penciclovir	false	None	100.00%	35.65%	Prescribe
Sorted Recommendations						
DB09015	Potassium Canrenoate	false	None	99.76%	93.77%	Prescribe
DB09028	Cytisine	false	None	89.78%	96.09%	Prescribe

Figure 7.5: Candidate list of drugs for *penciclovoir*

A survey was performed with six dentists, with questions pertaining to:

- Perceived Usefulness
- Perceived Ease of Use
- User Satisfaction
- Attributes of Usability

Participants were aware of their obligations, benefits and risks as outlined in Appendix A.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Perceived Usefulness	0	2	7	19	7
Perceived Ease of Use	0	2	3	12	19
User Satisfaction	0	1	8	14	13
Attributes of Usability	0	6	4	10	10

Table 7.8: Responses to CDSS Survey

They were given a list of questions where they have to respond on a Likert scale [134] (1 being “Strongly Disagree” to 5 being “Strongly Agree”). The full list of questions is listed in Appendix B. The survey was conducted according to ethical guidelines approved by the ethics committee[3], University of Southern Queensland.

Table 7.8 summarises the responses for each category. Most of the dentists surveyed welcomed the idea of a CDSS considering the patient’s current medications and drug allergies. Negative responses related to suggestions for improvement of the user interface due to navigation difficulties.

Responses grouping the “Strongly Disagree/Disagree” responses as ’Negative’ and “Agree/Strongly Agree ” responses as ’Positive’ are shown in Figure 7.6. The majority are positive with the features of the CDSS.

[3] Human Research Ethics Approval Number: H19REA262

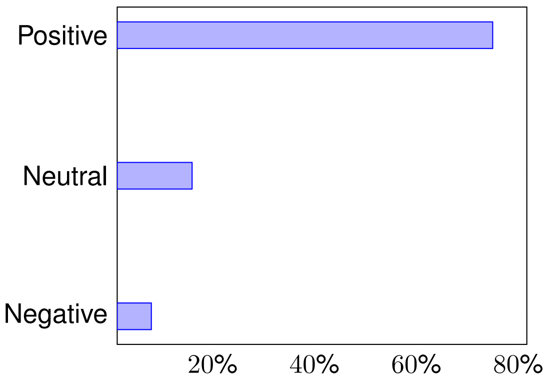


Figure 7.6: Summary of response

The positive response is expected since the prototype provided users with additional information on drug prescription. The user is no longer at the mercy of online tools to search if a drug is safe for prescription. Instead of going to random sites on the web to search if a drug is safe, this prototype allows the user to get more information on the similarity ratio between the drug to be prescribed and the current drug that the patient is taking. It also ensures that the patient does not have allergy to the drug that is prescribed. Those online checker programs will just merely check the interaction between two drugs. Hence this prototype has a positive impact on the clinical workflow, increasing the efficiency and meeting the prescription needs of the dentist at point-of-care. Besides the superior of the performance of the prototype when compared with other online checkers describe in Section 6.2.3, a special feature of the prototype is also the ability to adjust the number of candidate drugs for the user's consideration. Such features are not found in any of the five common DDI software programs, notwithstanding the fact that none of these common programs are ideal [65]. Moreover, the prototype developed as a result of this research has the ability to check for interaction with multiple drugs. This is especially useful with the growing number of patients who are taking multiple drugs. However, there is no such resource currently available for such situations [42].

7.5 Prototype Design - Mobile Learning

Another example of clinical application for the three-layer conceptual model is in the area of mobile learning associated with drug prescription. The design of this proposed mobile learning application aimed to make learning about drug prescription enjoyable and engaging. It can be easily developed and deployed on a wide variety of portable devices and platforms.

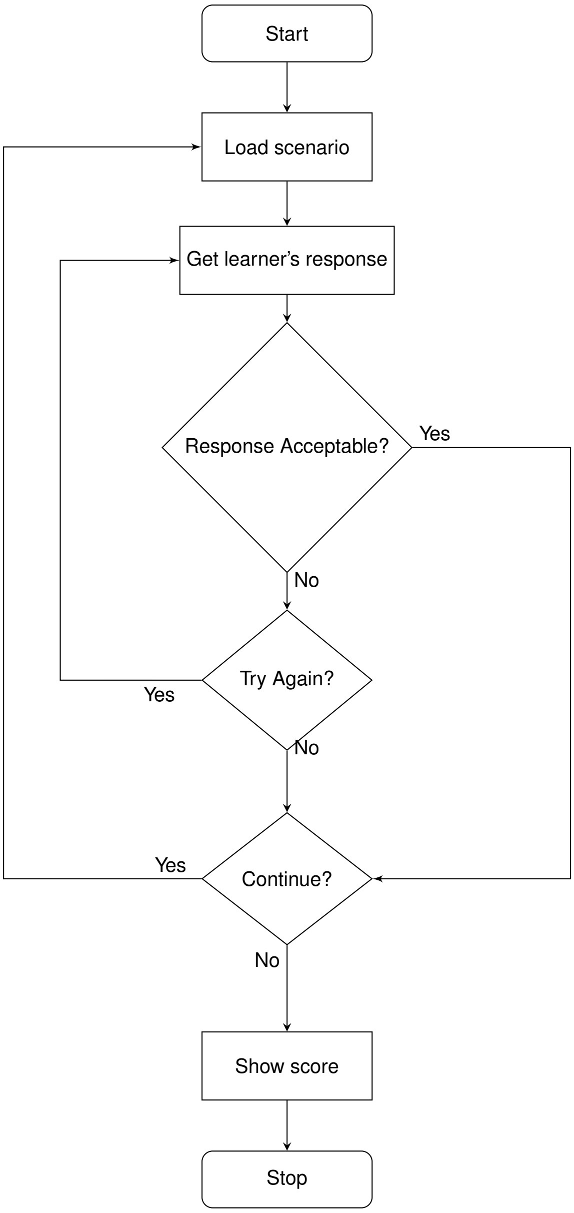
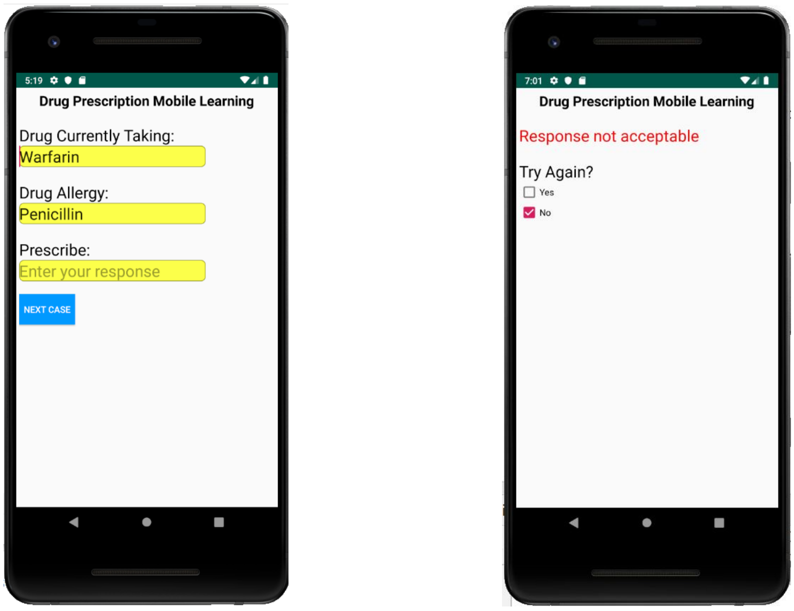


Figure 7.7: System flow

The architecture of the prototype followed the CDSS described in the previous section. As shown in Figure 7.1 the main processing engine resided in the prediction layer.

In the case of the mobile learning application, the prediction algorithm consists of the system flow shown in Figure 7.7.

The algorithm communicates with the user through the presentation layer of the three-layer framework. The user will be presented with a scenario and a prompt to enter a drug for



5:19

Drug Prescription Mobile Learning

Drug Currently Taking: Warfarin

Drug Allergy: Penicillin

Prescribe: Enter your response

(a) Scenario for user to response

7:01

Drug Prescription Mobile Learning

Response not acceptable

Try Again?

□ Yes

☑ No

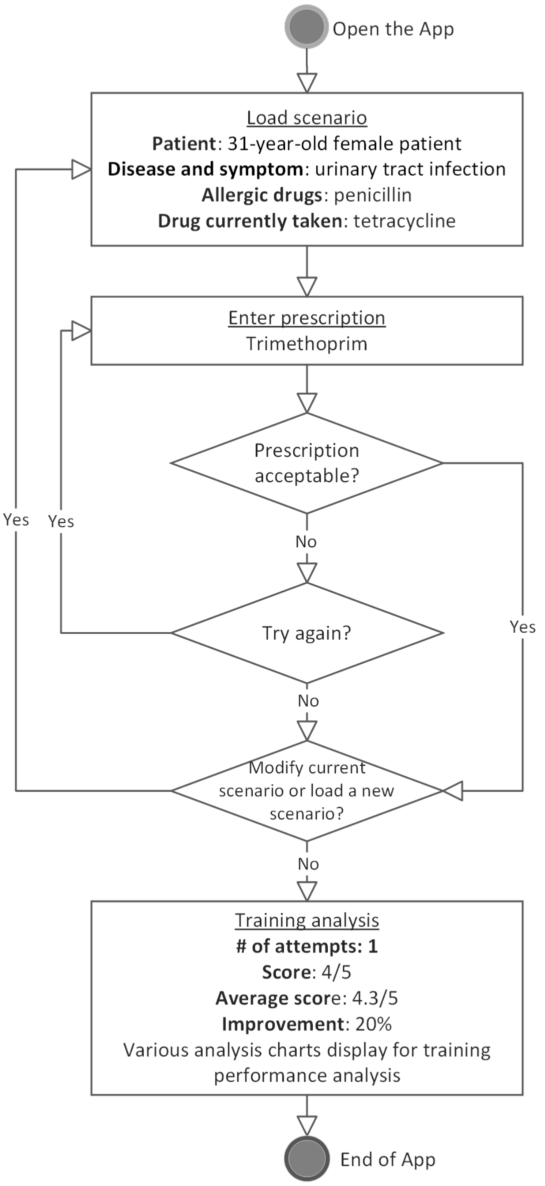
(b) Error display

Figure 7.8: User mobile interface

prescription (Figure 7.8a). Based on the user response, the processing engine uses feature vectors generated from the word-embedding model to determine if the drug to be prescribed is acceptable (Section 4.5). With these feature vectors, the mobile application will compute the similarity ratio and decide if the drug is appropriate for prescription. If the response is inappropriate, the user can try again (Figure 7.8b). They can also choose to skip to another scenario.

7.6 Mobile Learning Scenario

Most, if not all, medical students have access to a mobile device such as a smart-phone or a tablet. Hence, the mobile learning application can be easily applied for use by educators to train students in enhancing their knowledge on drug interactions and prescription. The educator can design different combinations of medical conditions, current medications and drug allergies which are stored in the patient-dependent medical profile within the presentation layer. The scenario below shows how mobile learning on smart devices can be engaging and easy [2]. A flow chart of the vignette is shown in Figure 7.9, illustrating a walk-through of the mobile application by a medical student.



Open the App
Load scenario
Patient: 31-year-old female patient
Disease and symptom: urinary tract infection
Allergic drugs: penicillin
Drug currently taken: tetracycline
Enter prescription
Trimethoprim
Prescription acceptable?
Yes
Yes
No
Try again?
Yes
No
Modify current scenario or load a new scenario?
No
Training analysis
# of attempts: 1
Score: 4/5
Average score: 4.3/5
Improvement: 20%
Various analysis charts display for training performance analysis
End of App

Figure 7.9: Flowchart of learning vignette

Winnie is a 24-year-old medical student at a reputable university in Queensland. She is on her daily train journey to class, and has just finished catching up on yesterday's lecture recordings. One of the lecturers had strongly recommended a mobile learning application on the student portal.

Being a motivated student preparing for her examinations in a week, this thought flashed through her mind:

"*Well, let's see how useful this app is. I guess I'm all caught up on study for now, so I have* *some time to give this a go.*"

She downloads the application and opens it. There is a range of topics available. Since she had just finished listening to a pharmacology lecture, she chooses to revise that.

In the first scenario, a 31-year-old female patient presents with a simple urinary tract infection. The relevant medical history consists of moderate acne, for which she is taking a tetracycline medication, and a penicillin allergy. The task for this scenario is to prescribe an appropriate antibiotic for the urinary tract infection.

Winnie thought back to the lecture she had just finished listening to, where there was heavy emphasis on careful prescription of drugs to prevent adverse outcomes. Based on that, her answer to this scenario should be a drug that is not in the penicillin class, since this patient has a penicillin allergy. She enters "trimethoprim" and bingo! The screen goes "ping!" in approval. A pop-up explains that this is within acceptable range for prescription because *trimethoprim* is in a different class to penicillin -

"*yes, I thought of that!*" -

and has a low probability of interacting with drugs from the tetracycline class -

"*oops, I totally didn't consider her acne medication.*"

Just at this moment, another "ping" sounds - she has arrived at the university station. With a pity, she terminated the practice and took a quick glance at the final screen of the app on the mobile – she secured 4 marks out of 5 in the play, and had an average score of 4.3 out of 5 so far, marking a 20% improvement since last week. A bar chart also displayed, illustrating the improvements she had made along with the last few weeks.

"*Not bad!*", she thought,

"*Just need to work a bit harder!*"

She winded up her earphones and popped her phone in her bag. Getting the right answer has given her a nice confidence boost to start off the day, and it's also reminded her about how she can improve on her thought processes for drug prescription. Maybe she would give the app another go on the ride home.

7.7 Summary

This chapter has illustrated how the three-layer framework is being adopted for use in the CDSS as proposed in the thesis. Two prototypes were described: one to aid drug prescription in the dental clinic at point-of-care, and the other for mobile learning of drug prescription. The main processing engine is part of the prediction layer where it obtains the user response from the presentation layer. The results are also communicated to the user via the presentation layer. In order to enhance the usability of the CDSS, interactive behaviours of some common drugs used in the dental clinic have also been described. A typical clinical scenario was also included to illustrate how the system is being deployed in the dental clinic at point-of-care. The results of the survey indicated that the dentists found the prototype easy to use.

Chapter 8

Conclusions

With the methodology and research problem introduced in Chapter 1, the thesis has devised the novel approach of the 3-layer framework in predicting drug similarity, resulting in the deployment of a CDSS which dentists have found useful and easy to use. The entire research process from conceptual framework to experiments are summarised in this chapter. The impact of the research findings are also discussed together with some suggestions for future work.

8.1 Summary of the Research

The number of drugs in drug databases is constantly expanding with novel drugs appearing on the market each year. As no healthcare professional can be expected to recall all the drugs available, let alone potential drug-drug interactions, problems can arise when prescribing drugs to patients, especially those with multiple medical conditions taking multiple medications.

Examining possible solutions by performing a survey of existing decision support systems found that ease of use was a significant factor in increasing adoption rate by users. At the same time, the CDSS should integrate with the patient's existing medical profile for enhanced efficiency in drug prescription at point-of-care.

In order to create an appropriate CDSS, a conceptual three-layer framework was introduced in the thesis. The innovative framework enables a system to be easily designed and implemented. Each layer can be developed and maintained as an independent tier. The proposed approach also ensures the interfacing between each layer conforms to standards to enable smooth linkage between them. The knowledge layer stores the domain knowledge which describes the properties of the drugs. In order to create the knowledge layer, raw data from bio-medical corpus is selected, pre-processed and transformed into structured tables. These processes align with the beginning stages of a typical knowledge discovery process. With the raw data processed, different approaches have been introduced to represent the drugs. These approaches include finding out how frequent the words occur (*tf*idf*) and how likely those words would occur. These approaches in drug representations enable the discovery of drug interactions. Such information makes it possible to determine potential adverse interactions of a drug pair.

The prediction layer then performs the extraction of feature vectors to allow the similarity of a drug pair to be computed. Four methods of feature vector extraction are described in the thesis: the statistical model, the side-effect model, the adverse network model and the word embedding model. These models use data mining and evaluation which aims to discover patterns and meanings from the knowledge base. The statistical model and side-effect model use *tf*idf* information on the text that describes the drugs. The adverse network model examines the number of common drugs between a drug pair where there is an adverse interaction. The larger the number of common paths between a drug pair, the higher the similarity ratio, which means the less chance of an adverse interaction. The word embedding model is based on the expectation that a higher set of common terms are used to describe a pair of similar drugs. Similarity ratio of a drug pair is obtained from the vectors that represent the words most related to each drug. These models can be readily applied to a CDSS to assist dentist in their drug prescription at point-of-care.

Empirical experiments were conducted to evaluate the performance of each model. From the results, the best approach was used for deployment at the presentation layer. The evaluation and visualisation process also align with the last stages of the knowledge discovery process. In the presentation layer, the drug to be prescribed is presented to the user, with consideration of the drugs the patient is taking and the drugs the patient is allergic to.

The word embedding model residing within the prediction layer of the conceptual framework proposed in the thesis has also been used in the deployment of a CDSS. The deployment allow the results of the research to be visualised and implemented for use by the healthcare professional to prescribe drugs based on the current drugs that the patient is taking and the drug allergies that the patient may have. To enhance the usability of the CDSS, additional alerts are incorporated into the system which consider some of the special features of common drugs used in the dental clinic. A survey on dentist regarding the ease of use of the system shows that a majority are positive with the features of the CDSS.

The three-layer framework in this research can also be easily extended for use in medical clinics, and to supplement information relating to new drugs in pharmaceutical research. The findings of the research that similar drug pairs have higher similarity ratio has also been applied to the development of a mobile application to facilitate medical education. Medical students can use such applications on their smartphone any time and anywhere to practice drug prescription for different patient profiles. Both learners and educators will benefit from the engaging and enriching nature of such an application.

8.2 Contributions and Significance

This thesis has made several theoretical and practical contributions. Traditionally, chemical structures were used in studying and analysing interactions between drugs. The approach used in the thesis uses information extracted from bio-medical corpus to build feature vectors. The discovery of DDI by performing data mining on drug repositories has contributed to strategic insights into new research model to enhance the field of pharmacovigilance. With feature vectors obtained from the drug repository, similarity ratios are computed for each drug pair and can be used to determine how interactive the drug pair is. The three-layer innovative approach allows multiple models to be adopted in the extraction of information and building of feature vectors. Through data mining method to discover the semantics of the text, the feature vectors are obtained for computing the similarity ratio of a drug pair. The empirical results show better performance than related baseline models using similar drug corpus. The findings have high impact on pharmaceutical research relating to the use of word embedding for discovering drug properties. The discovery of DDI obtained through feature vectors has an impact on the design of efficient computational workflow for *in silico* drug repurposing.

With the word embedding approach, personalised CDSS has been implemented for use in dental clinics. Experimental results support the hypothesis that similar drug pairs have a higher similarity ratio compared to dissimilar pairs. While creating the different models, this thesis has demonstrated that paths linking the common set of interacting drugs of a drug pair can also be used to build feature vectors to arrive at a similarity ratio for the drug pair.

The findings of the research have also been presented at major workshops and conferences and published in peer-reviewed journals.

These theoretical contributions also lead to practical contributions in terms of possible implementations for any user needing to prescribe drugs. Although there are many systems that provide support in terms of drug interactions, the framework provided in the thesis enables the consideration of the patient's medical profile, hence allowing the implementation of a personalised decision support system for use in dental clinics. A user-friendly interface will allow more dentists to adopt this system to assist them in drug prescription within their clinical work-flow. The outcome of the knowledge discovery on drug interactions has provided a platform for further research on data mining and machine learning methods within the medical domain which will transform the clinical work-flow for the healthcare professional at point-of-care.

8.3 Future Work

A number of areas for future work have been identified. The methodology adopted in the current work provides a heuristic approach where feature vectors are extracted for deciding if a drug pair is similar. Although text-mining approaches using the network model and word embedding models are used in the prediction layer, other extensions to the experiment in this research can be explored. One of these tools is GloVe[1], an approach combining the local word embedding method of Word2Vec with the global statistics of matrix factorisation techniques.

Although both medical and dental practitioners need to deal with the same pool of patients who take multiple drugs, the number of drugs to be prescribed by the dentist is much lesser compared to the medical practitioner. Hence, in terms of implementation, the design of the CDSS can be easily amended to cater to an extended set of prescribed drugs, allowing the system to be used within medical clinics as well.

Besides using different prediction methods, different knowledge resources can be used to conduct the experiment as more open-source repositories are made available. Since the theoretical approaches used within the prediction layer of the three-layer framework is based

[1] https://nlp.stanford.edu/projects/glove/ (Accessed 12 Jan 2020).

on DrugBank, it will be interesting to conduct the experiment with alternative resources such as PubMed[2] and compare the results to evaluate if it is more efficient. The breakthrough approach of prediction of an adverse interaction of a drug pair as discovered in the current work through data mining has set the direction for future research in prediction of high-order drug-drug interaction prediction.

8.4 Overall Conclusion

The study explores the novel use of various data mining approaches to obtain feature vectors for the purpose of determining the similarity ratios of drug pairs. This results in a significant contribution relating to the design of personalised clinical decision support systems for drug prescription.

[2] http://bio.nlplab.org (Accessed 12 Jan 2020).

References

[1] J. E. Adams, E. L. Mounib, and A. Shabo. IT-enabled personalized healthcare: Improving the science of health promotion and care delivery. http://www.ictliteracy.info/rf.pdf/IT-enabled_healthcare.pdf, 2010. [Online; accessed 19-Jan-2019].

[2] M. Aljohani and T. Alam. Design an m-learning framework for smart learning in ad hoc network of android devices. In *2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, pages 1–5, 2015. http://doi.org/10.1109/ICCIC.2015.7435817

[3] A. Appadurai. *Modernity al large: cultural dimensions of globalization*. University of Minnesota Press, Minnesota, 1996.

[4] T. Ayer, J. Chhatwal, O. Alagoz, J. Charles E. Kahn, R. W. Woods, and E. S. Burnside. Comparison of logistic regression and artificial neural network models in breast cancer risk estimation. *Radio Graphics*, 30(1):13–22, 2010. http://doi.org/10.1148/rg.301095057

[5] S. Ayvaz, J. Horn, O. Hassanzadeh, Q. Zhu, J. Stan, N. P. Tatonetti, S. Vilar, M. Brochhausen, M. Samwald, M. Rastegar-Mojarad, M. Dumontier, and R. D. Boyce. Toward a complete dataset of drug-drug interaction information from publicly available sources. *Biomedical Informatics*, 55:206–217, 2015. http://doi.org/10.1016/j.jbi.2015.04.006

[6] Y. Bengio, H. Schwenk, J.-S. Senecal, F. Morin, and J.-L. Gauvain. Neural probabilistic language models. In D. Holmes and L. Jain, editors, *Innovations in Machine Learning*, pages 137–186. Springer International Publishing, Berlin, Heidelberg, 2006.

[7] R. Benyamin, A. Trescot, S. Datta, and R. Buenaventura. Opioid complications and side effects. *Pain Physician*, 11(2):S105–20, 2008.

[8] M. Bessani, E. Lins, A. Delbem, and C. Maciel. Construction of a clinical decision support system for dental caries management using Bayesian Networks. In *Brazilian Congress on Biomedical Engineering*, pages 517–520, 2014. http://doi.org/10.13140/RG.2.1.4698.3205

[9] A. Bhatia and R. Singh. Using bayesian network as decision making system tool for deciding treatment plan for dental caries. *Journal of Academia and Industrial Research*, 2(2):93–96, 2013.

[10] J. Bjorne and T. Salakoski. Biomedical event extraction using convolutional neural networks and dependency parsing. In *Proceedings of the BioNLP 2018 workshop*, pages 98–108. Association for Computational Linguistics, 2018. http://www.doi.org/10.18653/v1/W18-2311

[11] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[12] B. Bokharaeian, A. Diaz, and H. Chitsaz. Enhancing extraction of drug-drug interaction from literature using neutral candidates, negation, and clause dependency. *PLoS ONE*, 11(10):1–20, 10 2016. http://doi.org/10.1371/journal.pone.0163480

[13] J. Bouaud, V. Koutkias, and Section Editors for the IMIA Yearbook Section on Decision Support. Computerized clinical decision support: Contributions from 2014. *Yearbook of Medical Informatics*, 24(1):119–124, 2015. http://doi.org/10.15265/IY-2015-036

[14] Q. Bui, P. Sloot, E. vanMulligen, and J. Kors. A novel feature-based approach to extract drug-drug interactions from biomedical text. *BioInformatics*, 30(23):3365–3371, 2014.10.1093/bioinformatics/btu557 http://doi.org/10.1093/bioinformatics/btu557

[15] Y. Cai, C.-m. Au Yeung, and H.-f. Leung. Knowledge representation on the web. In *Fuzzy Computational Ontologies in Contexts: Formal Models of Knowledge Representation with Membership Degree and Typicality of Objects, and Their Applications*, pages 15–21. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[16] M. Carbone, E. Goss, M. Carrozzo, S. Castellano, D. Conrotto, R. Broccoletti, and S. Gandolfo. Systemic and topical corticosteroid treatment of oral lichen planus: A comparative study with long-term follow-up. *Journal of oral pathology and medicine*, 32:323–329, 2003. http://doi.org/10.1034/j.1600-0714.2003.00173.x

[17] A. Casillas, A. Perez, M. Oronoz, K. Gojenola, and S. Santiso. Learning to extract adverse drug reaction events from electronic health records in spanish. *Expert Systems with Applications*, 61:235–245, 2016. http://doi.org/10.1016/j.eswa.2016.05.034

[18] R. Cederberg, M. Walji, and J. Valenza. Electronic health records in dentistry: Clinical challenges and ethical issues. In S. Kumar, editor, *Teledentistry*, Health Informatics, pages 1–12. Springer International Publishing, 2015. http://doi.org/10.1007/978-3-319-08973-7_1

[19] L. Chen, B. Fang, Z. Shang, and Y. Tang. Tackling class overlap and imbalance problems in software defect prediction. *Software Quality Journal*, 26(1):97–125, Mar 2018. http://doi.org/10.1007/s11219-016-9342-6

[20] D. Cirillo and A. Valencia. Big data analytics for personalized medicine. *Current Opinion in Biotechnology*, 58:161–167, 2019. http://doi.org/10.1016/j.copbio.2019.03.004

[21] Clinical Decision Support System Consortium. B. Middleton. https://www.ncbi.nlm.nih.gov/pubmed/19745260, 2009. [Online; accessed 19-Apr-2019].

[22] C. Crenner. Introduction of the blood pressure cuff into U.S. medical practice: technology and skilled practice. *Ann Intern Med*, 128(6):488–493, 1998. http://doi.org/10.7326/0003-4819-128-6-199803150-00010

[23] F. Davis, R. Bagozzi, and P. Warshaw. User acceptance of information technology: a comparison of two theories. *Management Science*, 36(8):982–1003, 1989. http://doi.org/10.1287/mnsc.35.8.982

[24] S. Dechanont, S. Maphanta, B. Butthum, and C. Kongkaew. Hospital admissions/visits associated with drug-drug interactions: a systematic review and meta-analysis. *Pharmacoepidemiology and Drug Safety*, 23(5):489–497, 2014. http://doi.org/10.1002/pds.3592

[25] D. Demner-Fushman, W. W. Chapman, and C. J. McDonald. What can natural language processing do for clinical decision support? *Journal of Biomedical Informatics*, 42(5):760–772, 2009. http://doi.org/10.1016/j.jbi.2009.08.007

[26] S. Devaraj, S. K. Sharma, D. J. Fausto, S. Viernes, and H. Kharrazi. Barriers and facilitators to clinical decision support systems adoption: A systematic review. *Journal of Business Administration Research*, 3(2):36–53, 2014. http://doi.org/10.5430/jbar.v3n2p36

[27] S. Dhanya, F. Shobeir, and G. Lise. A probabilistic approach for collective similarity-based drug-drug interaction prediction. *Bioinformatics*, 32:3175–3182, 2016. http://doi.org/10.1093/bioinformatics/btw342

[28] G. J. Dhiman, K. T. Amber, and K. W. Goodman. Comparative outcome studies of clinical decision support software: limitations to the practice of evidence-based system acquisition. *Journal of the American Medical Informatics Association*,22(e1):e13–e20, 2015. http://doi.org/10.1093/jamia/ocu033

[29] O. Dympna, P. Fraccaro, E. Carson, and P. Weller. Decision time for clinical decision support systems. *Clinical Medicine*, 14(4):338–341, 2014. http://doi.org/10.7861/clinmedicine.14-4-338

[30] S. H. El-Sappagh and S. El-Masri. A distributed clinical decision support system architecture. *Journal of King Saud University - 'Computer and Information Sciences*, 26(1):69–78, 2014. http://doi.org/10.1016/j.jksuci.2013.03.005

[31] G. Eng, A. Chen, T. Vess, and G. Ginsburg. Genome technologies and personalized dental medicine. *Oral Diseases*, 18(3):223–235, 2012. http://doi.org/10.1111/j.1601-0825.2011.01876.x

[32] A. Farag, J. York, M. Finkelman, and B. Desai. Prescription of potentially inappropriate medications in geriatric patients: Data from a single dental institution. *Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology*, 128(1):e6–e12, 2019. https://doi.org/10.1016/j.oooo.2019.01.075

[33] T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006. https://doi.org/10.1016/j.patrec.2005.10.010

[34] R. Ferdousi, R. Safdari, and Y. Omidi. Computational prediction of drug-drug interactions based on drugs functional similarities. *Journal of Biomedical Informatics*, 70:54–64, 2017. http://doi.org/10.1016/j.jbi.2017.04.021

[35] P. Fraccaroa, D. O'Sullivana, P. Plastirasa, H. O'Sullivanb, C. Dentonec, A. DiBiagioe, and P. Wellera. Behind the screens: Clinical decision support methodologies - a review. *Health Policy and Technology*, 4(1):29–38, 2014. https://doi.org/10.1016/j.hlpt.2014.10.001

[36] M.-P. Gagnon, E. K. Ghandour, P. K. Talla, D. Simonyan, G. Godin, M. Labrecque, M. Ouimet, and M. Rousseau. Electronic health record acceptance by physicians:

Testing an integrated theoretical model. *Journal of Biomedical Informatics*, 48(0):17–27, 2014. http://doi.org/10.1016/j.jbi.2013.10.010

[37] I. Garcia, R. Kuska, and M. J. Somerman. Expanding the foundation for personalized medicine: implications and challenges for dentistry. *Journal of dental research*, 92(7):S3–S10, 2013. http://doi.org/10.1177/0022034513487209

[38] N. D. Giarratano A, Green SE. Review of antimicrobial use and considerations in the elderly population. *Clinical Interventions in Aging*, 13:657–667, 2018. http://doi.org/10.2147/CIA.S133640

[39] G. Ginsburg and J. McCarthy. Personalized medicine: revolutionizing drug discovery and patient care. *Trends Biotechnol*, 19(12):491–496, 2001. http://doi.org/10.1016/S0167-7799(01)01814-5

[40] W. Goh, X. Tao, J. Zhang, and J. Yong. A study of drug interaction for personalised decision support in dental clinics. In *2015 IEEE/WIC/ACM Workshop Proceedings on International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)* , volume 1, pages 88–91, 2015. http://doi.org/10.1109/WI-IAT.2015.28

[41] O. Gottesman and H. Kuivaniemi. The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. https://pubmed.ncbi.nlm.nih.gov/23743551/, 2013. [Online; accessed 19-Dec-2018].

[42] L. Grannell. Drug interaction resources: mind the gaps. *Australian prescriber*, 43(1):18–23, 2020. http://doi.org/10.18773/austprescr.2020.005

[43] E. Grossi. Medical concepts related to individual risk are better explained with “plausibility” rather than “probability”. *BMC Cardiovascular Disorders*, 5(1):1–4, 2005. http://doi.org/10.1186/1471-2261-5-31

[44] H. Gurulingappa, A. Mateen, A. Roberts, L. Toldo, J. Fluck, M. Hofmann-Apitius, and L. Tolda. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *Journal of Biomedical Informatics*, 45(5):885–892, 2012. https://doi.org/10.1016/j.jbi.2012.04.008

[45] F. Hammann and J. Drewe. Data mining for potential adverse drug-drug interactions. *Expert Opinion on Drug Metabolism & Toxicology*, 10(5):665–671, 2014. http://doi.org/10.1517/17425255.2014.894507

[46] J. Han, M. Kamber, and J. Pei. *Data Mining Concepts and Techniques*. Morgan Kaufmann Publishers, 2012. http://doi.org/10.1016/C2009-0-61819-5

[47] K. Hedna, M. L. Andersson, H. Gyllensten, S. Hagg, and Y. Bottiger. Clinical relevance of alerts from a decision support system, pharao, for drug safety assessment in the older adults. *Biomed Central Geriatrics*, 19(164), 2019. http://doi.org/10.1186/s12877-019-1179-y

[48] P. J. Helmons, B. O. Suijkerbuijk, P. V. N. Panday, and J. G. Kosterink. Drug-drug interaction checking assisted by clinical decision support: a return on investment analysis. *Journal of the American Medical Informatics Association*, 22(4):764 – 772, 2015. http://doi.org/10.1093/jamia/ocu010

[49] HL7 International. HL 7 Standards. http://www.hl7.org/, 2014. [Online; accessed 19-Jan-2018].

[50] A. Holzinger and I. Jurisica. Knowledge discovery and data mining in biomedical informatics: The future is in integrative, interactive machine learning solutions. In A. Holzinger and I. Jurisica, editors, *Interactive Knowledge Discovery and Data Mining in Biomedical Informatics*, chapter 1, pages 1–18. Springer-Verlag, 2014. http://doi.org/10.1007/978-3-662-43968-5_1

[51] D. Horgan, M. Jansen, L. Leyens, J. A. Lal, R. Sudbrak, E. Hackenitz, U. Bubhoff, W. Ballensiefen, and A. Brand. An index of barriers for the implementation of personalised medicine and pharmacogenomics in europe. *Public Health Genomic*, 17:287–298, 2014. http://doi.org/10.1159/000368034

[52] J. Horsky, S. Phansalkar, A. Desai, D. Bell, and B. Middleton. Design of decision support interventions for medication prescribing. *International Journal of Medical Informatics*, 82(6):492–503, 2013. http://doi.org/10.1016/j.ijmedinf.2013.02.003

[53] S. Huang, F. Mu, F. Li, W. Wang, W. Zhang, L. Lei, Y. Ma, and J. Wan. Systematic elucidation of the potential mechanism of erzhi pill against drug-induced liver injury via network pharmacology approach. *Evidence-Based Complementary and Alternative Medicine*, 2020:1–16, 2020. http://doi.org/10.1155/2020/6219432

[54] K. Huxhagen and M. Mccullough. Therapeutic guidelines: Oral and dental. *Australian Prescriber*, 34:63–66, 2011. http://doi.org/10.18773/austprescr.2011.040

[55] International Health Terminology Standards Development Organisation. SNOMED International. http://www.ihtsdo.org, 2014. [Online; accessed 19-Dec-2019].

[56] E. Ippoliti, F. Sterpetti, and T. Nickles. *Models and Inferences in Science*. Springer Publishing Company, Incorporated, 1st edition, 2016.

[57] K. Jeddi, M. Alborzi, and R. Radfar. A Decision Support System for a new product specifications selection: Using fuzzy QFD and ANN. *International Journal of Innovative Technology and Research*, 2(1), 2014.

[58] G. Jeh and J. Widom. Simrank: A measure of structural-context similarity. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, pages 538–543, New York, NY, USA, 2002. http://doi.org/10.1145/775047.775126

[59] J. Johnson and L. Cavallari. Pharmacogenetics and cardiovascular disease implications for personalized medicine. *Pharmacol Rev*, 65(3):987–1009, 2013. http://doi.org/10.1124/pr.112.007252

[60] S. G. Johnson, P. Jacobson, S. M. Wolf, K. K. Sinha, and D. Yee. Generalizable architectures and principles of informatics for scalable personalized and precision medicine decision support. In D. Holmes and L. Jain, editors, *Personalized and Precision Medicine Informatics. Health Informatics*. Springer, Cham, 2020. http://doi.org/10.1007/978-3-030-18626-5_18

[61] C. Kamath. *Scientific Data Mining - A Practical Perspective*. Society for Industrial and Applied Mathematics, 01 2009. http://doi.org/10.1137/1.9780898717693

[62] P. Kamolratanakul and P. Jansisyanont. A review of antibiotic prophylaxis protocols in oral and maxillofacial surgery. *Journal of Oral and Maxillofacial Surgery, Medicine, and Pathology*, 30(5):395–404, 2018. http://doi.org/10.1016/j.ajoms.2018.03.008

[63] S. Kang, P. Kang, T. Ko, S. Cho, S. jin Rhee, and K.-S. Yu. An efficient and effective ensemble of support vector machines for anti-diabetic drug failure prediction. *Expert Systems with Applications*, 42(9):4265–4273, 2015. https://doi.org/10.1016/j.eswa.2015.01.042

[64] M. Khalilfa. Clinical decision support: Strategies for success. *Procedia Computer Science*, 37:422–427, 2014. https://doi.org/10.1016/j.procs.2014.08.063

[65] R. Kheshti, M. Aalipour, and S. Namazi. A comparison of five common drug-drug interaction software programs regarding accuracy and comprehensiveness. *Journal of research in pharmacy practice*, 5:257–263, 2016. http://doi.org/10.4103/2279-042X.192461

[66] J. A. Kim, I. Cho, and N. Kim. CDSS Service Integration Reference Model. *Advanced Science and Technology Letters*, 24:9–12, 2013.

[67] J. Kralj, M. Robnik-Sikonja, and N. Lavra. Netsdm: Semantic data mining with network analysis. *Journal of Machine Learning Research*, 20:1–50, 2019.

[68] S. Lee, J. Yang, and J. Han. Development of a decision making system for selection of dental implant abutments based on the fuzzy cognitive map. *Expert Systems with Applications*, 39(14):11564–11575, 2012. `https://doi.org/10.1016/j.eswa.2012.04.032`

[69] A. Li, Q. Zang, D. Sun, and M. Wang. A text feature-based approach for literature mining of lncrna-protein interactions. *Neurocomput.*, 206(C):73–80, Sept. 2016. `http://doi.org/10.1016/j.neucom.2015.11.110`

[70] C. Liu. A group decision-making method with fuzzy set theory and genetic algorithms in quality function deployment. *Quality & Quantity*, 44(6):1175–1189, 2009. `http://doi.org/10.1007/s11135-009-9304-1`

[71] J. Liu, Z. Huang, F. Ren, and L. Hua. Drug-drug interaction extraction based on transfer weight matrix and memory network. *IEEE Access*, PP:1–1, 07 2019. `http://doi.org/10.1109/ACCESS.2019.2930641`

[72] J. Liu, J. C. Wyatt, and D. G. Altman. Decision tools in health care: focus on the problem, not the solution. *BMC Med Inform Decis Mak*, 6(4):4–4, 2006. `http://doi.org/10.1186/1472-6947-6-4`

[73] S. Liu, B. Tang, Q. Chen, and X. Wang. Drug-drug interaction extraction via convolutional neural networks. *Computational and Mathematical Methods in Medicine*, 2016:1–8, 2016. `http://doi.org/10.1155/2016/6918381`

[74] J. Lu, G. Zhang, and D. Ruan. Intelligent multi-criteria fuzzy group decision-making for situation assessments. *Soft Computing*, 12(3):289–299, 2007. `http://doi.org/10.1007/s00500-007-0197-4`

[75] J. Lu, G. Zhang, and D. Ruan. *Multi-objective Group Decision Making: Methods, Software and Applications With Fuzzy Set Techniques*. Imperial College Press, London,UK, 2007. `http://doi.org/10.1142/p505`

[76] J. Lu, G. Zhang, and F. Wu. Web-based multi-criteria group decision support system with linguistic term processing function. *IEEE Intelligent Informatics Bulletin*, 5(1):34–43, 2005.

[77] J. Lu, Y. Zhu, X. Zeng, L. Koehl, J. Ma, and G. Zhang. *AI 2008: Advances in Artificial Intelligence: 21st Australasian Joint Conference on Artificial Intelligence Auckland, New Zealand, December 1-5, 2008. Proceedings*, chapter A Fuzzy Decision Support System for Garment New Product Development, pages 532–543. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008. https://doi.org/10.1007/978-3-540-89378-3

[78] J. Ma, J. Lu, and G. Zhang. Decider: A fuzzy multi-criteria group decision support system. *Knowledge-Based Systems*, 23(1):23–31, 2010. http://doi.org/10.1016/j.knosys.2009.07.006 Special Issue on Intelligent Decision Support and Warning Systems.

[79] V. K. Mago, N. Bhatia, A. Bhatia, and A. Mago. Clinical decision support system for dental treatment. *Journal of Computational Science*, 3(5):254–261, 2012. http://doi.org/10.1016/j.jocs.2012.01.008

[80] S. Majid, S. F. S, and L. B. Adopting evidence-based practice in clinical decision making: nurses perceptions, knowledge, and barriers. *Journal of the Medical Library Association*, 99(3):229–236, 2011. http://doi.org/10.3163/1536-5050.99.3.010

[81] M. Mamatela. An empirical study of the technological, organisational and environmental factors influencing south african medical enterprises' propensity to adopt electronic health technologies, 2014. http://hdl.handle.net/10539/15126, 2014. [Online; accessed 21-Dec-2019].

[82] G. Mariscal, O. Marban, and C. Fernandez. A survey of data mining and knowledge discovery process models and methodologies. *The Knowledge Engineering Review*, 25(2):137–166, 2010.

[83] H. Mengash and A. Brodsky. A group package recommender based on learning group preferences, multi-criteria decision analysis, and voting. *EURO Journal on Decision Processes*, 3(3):275–304, 2015. http://doi.org/10.1007/s40070-015-0048-y

[84] B. Middleton, M. Bloomrosen, M. Dente, B. Hashmat, R. Koppel, J. Overhage, T. Payne, S. Rosenbloom, C. Weaver, and J. Zhang. Enhancing patient safety and quality of care by improving the usability of electronic health record systems: Recommendations from amia. *Journal of the American Medical Informatics Association*, 20(E1):e2–e8, 2013. http://doi.org/10.1136/amiajnl-2012-001458

[85] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. In Y. Bengio and Y. LeCun, editors, *1st International*

Conference on Learning Representations, 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings, 2013.

[86] T. Mikolov, W. Yih, and G. Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia, 2013. Association for Computational Linguistics.

[87] N. Milosevic, C. Gregson, R. Hernandez, and G. Nenadic. A framework for information extraction from tables in biomedical literature. *International Journal on Document Analysis and Recognition*, 22:55–78, 2019. `http://doi.org/10.1007/s10032-019-00317-0`

[88] L. G. M. Mulder-Wildemors, M. Heringa, A. Floor-Schreudering, P. A. F. Jansen, and M. L. Bouvy. Reducing inappropriate drug use in older patients by use of clinical decision support in community pharmacy: A mixed-methods evaluation. *Drugs and Aging*, 37:115–123, 2019. `http://doi.org/10.1007/s40266-019-00728-y`

[89] M. Musen, B. Middleton, and R. Greenes. Clinical decision-support systems. In E. H. Shortliffe and J. J. Cimino, editors, *Biomedical Informatics*, pages 643–674. Springer London, 2014.

[90] M. A. Musen. Decision-support systems. In E. Shortliffe, L. Perreault, G. Wiederhold, and L. Fagan, editors, *Medical Informatics: Computer Applications in Health Care and Biomedicine*, pages 573–609. Springer-Berlag, New York, 2nd edition, 2001.

[91] B. Nack, S. Haas, and J. Portnof. Opioid use disorder in dental patients: The latest on how to identify, treat, refer and apply laws and regulations in your practice. *Anesthesia Progres*, 64(3):178–187, 2017. `https://doi.org/10.2344/anpr-64-03-09`

[92] P. Nambisan. EMR Adoption among Office Based Physicians and Practices: Impact of Peer-to-Peer Interactions, Peer Support and Online Forums. In *2014 47th Hawaii International Conference on System Sciences*, pages 2733–2740, 2014. `http://doi.org/10.1109/HICSS.2014.343`

[93] R. Nath, A. Daneshmand, D. Sizemore, J. Guo, and R. Enciso. Efficacy of corticosteroids for postoperative endodontic pain: A systematic review and meta-analysis. *Journal of Dental Anesthesia and Pain Medicine*, 18(4):205–221, 2018. `http://doi.org/10.17245/jdapm.2018.18.4.205`

[94] M. Newman. Clinical decision support complements evidence-based decision making in dental practice. *Journal of Evidence Based Dental Practice*, 7(1):1–5, 2007. http://doi.org/10.1016/j.jebdp.2006.12.016

[95] L. Nguyen, E. Bellucci, and L. T. Nguyen. Electronic health records implementation: An evaluation of information system impact and contingency factors. *International Journal of Medical Informatics*, 83(11):779–796, 2014. http://doi.org/10.1016/j.ijmedinf.2014.06.011

[96] T. Nguyen, N. Le, Q. Ho, D. Phan, and Y. Ou. Using word embedding technique to efficiently represent protein sequences for identifying substrate specificities of transporters. *Analytical Biochemistry*, 577:73–81, 2019. http://doi.org/10.1016/j.ab.2019.04.011

[97] S. Nobre, L.-O. Eriksson, and R. Trubins. The use of decision support systems in forest management: Analysis of forsys country reports. *Forests*, 7(3):72, 2016. http://doi.org/10.3390/f7030072

[98] W. J. Orlikowski. The duality of technology: Rethinking the concept of technology in organizations. *Organization Science*, 3(3):398–427, 1992. http://doi.org/10.1287/orsc.3.3.398

[99] J. Osheroff, J. M. Teich, B. Middleton, E. Steen, A. Wright, and D. Detmer. A roadmap for national action on clinical decision support. *Journal of the American Medical Informatics Association*, 14(2):141–145, 2007. http://doi.org/10.1197/jamia.M2334

[100] J. Ovretveit, T. Scott, T. Rundall, S. Shortell, and M. Brommels. Improving quality through effective implementation of information technology in health care. *International Journal for Quality in Health Care*, 19(5):259–266, 2007. http://doi.org/10.1093/intqhc/mzm031

[101] G. Palma, M.-E. Vidal, and L. Raschid. Drug-target interaction prediction using semantic similarity and edge partitioning. In P. Mika, T. Tudorache, A. Bernstein, C. Welty, C. Knoblock, D. Vrandečić, P. Groth, N. Noy, K. Janowicz, and C. Goble, editors, *The Semantic Web - ISWC 2014*, pages 131–146, Cham, 2014. Springer International Publishing. http://doi.org/10.1007/978-3-319-11964-9_9

[102] M. Papaleontiou, C. Henderson, B. Turner, A. Moore, Y. Olkhovskaya, L. Amanfo, and C. Reid. Outcomes associated with opioid use in the treatment of chronic noncancer pain in older adults: a systematic review and meta-analysis. *Journal of the American*

Geriatrics Society, 58(7):1353–1369, 2010. http://doi.org/10.1111/j.1532-5415.2010.02920.x

[103] G. Papantonopoulos, K. Takahashi, T. Bountis, and B. G. Loos. Artificial neural networks for the diagnosis of aggressive periodontitis trained by immunologic parameters. *PLoS ONE*, 9(3):e89757, 2014. https://doi.org/10.1371/journal.pone.0089757

[104] S. G. Park, S. Lee, M.-K. Kim, and H.-G. Kim. The use of ontology in dental restorative treatment decision support system. In *Proceedings of the 2010 Conference on Formal Ontology in Information Systems*, pages 172–181, Amsterdam, The Netherlands, The Netherlands, 2010. IOS Press. http://doi.org/10.3233/978-1-60750-535-8-172

[105] S. G. Park, S. Lee, M.-K. Kim, and H.-G. Kim. Shared decision support system on dental restoration. *Expert Systems with Applications*, 39(14):11775–11781, 2012. https://doi.org/10.1016/j.eswa.2012.04.074

[106] S.-Y. Park, H.-J. Lee, S.-H. Kim, S.-B. Kim, Y.-H. Choi, Y.-K. Kim, and P.-Y. Yun. Factors affecting treatment outcomes in patients with oral lichen planus lesions: a retrospective study of 113 cases. *Journal of Periodontal and Implant Science*, 48(4):213–223, 2018. http://doi.org/10.5051/jpis.2018.48.4.213

[107] T. Pham, X. Tao, J. Zhang, J. Yong, X. Zhou, and R. Gururajan. MeKG: Building a Medical Knowledge Graph by Data Mining from MEDLINE. In P. Liang, V. Goel, and C. Shan, editors, *Brain Informatics*, pages 159–168, Cham, 2019. Springer International Publishing. http://doi.org/10.1007/978-3-030-37078-7_16

[108] M. Pota, M. Esposito, and G. De Pietro. Fuzzy partitioning for clinical dsss using statistical information transformed into possibility-based knowledge. *Knowledge-Based Systems*, 67:1–15, Sept. 2014. http://doi.org/10.1016/j.knosys.2014.06.021

[109] R. A. Dionne, S. M. Gordon, and S. A. Cooper. *Use of Ibuprofen in Dentistry*, chapter 8, pages 346–362. John Wiley & Sons, Ltd, 2015. https://doi.org/10.1002/9781118743614.ch8

[110] R. Khare, C. Wei, and Z. Lu. Automatic extraction of drug indications from fda drug labels. *AMIA Annual Symposium Proceedings*, pages 787–794, 2014.

[111] A. Rad, I. Amin, M. Rahim, and H. Kolivand. Computer-aided dental caries detection system from x-ray images. In S. Phon-Amnuaisuk and T. W. Au, editors, *Computational Intelligence in Information Systems*, volume 331 of *Advances in Intelligent Systems and*

Computing, pages 233–243. Springer International Publishing, 2015. http://doi.org/10.1007/978-3-319-13153-5_23

[112] M. Radman, A. Babic, E. Runjic, A. J. Kadic, M. Jeric, L. Moja, and L. Puljak. Revisiting established medicines: an overview of systematic reviews about ibuprofen and paracetamol for treating pain in children. *European Journal of Pain*, 23(6):1071–1082, 2019.

[113] E. Roger. *Diffusion of innovations*. The Free Press: New York, 4th edition, 1995.

[114] N. Rohani and C. Eslahchi. Drug-drug interaction predicting by neural network using integrated similarity. *Scientific Reports*, 9(13645), 2019. http://doi.org/10.1038/s41598-019-50121-3

[115] D. Ruan, J. Lu, E. Laes, G. Zhang, J. Ma, and G. Meskens. Multi-criteria Group Decision Support with Linguistic Variables in Long-term Scenarios for Belgian Energy Policy. *Journal of Universal Computer Science*, 16:103–120, 2010. http://doi.org/10.3217/jucs-016-01-0103

[116] W. Rudman, S. Hart-hester, W. Jones, N. Caputo, and M. Madison. Integrating medical and dental records: a new frontier in health information management. *Journal of American Health Information Management Association*, 81(10):36–39, 2010.

[117] D. L. Sackett, W. S. Richardson, W. Rosenberg, and R. B. Haynes. *Evidence-Based Medicine: How to Practice and Teach EBM*. Churchill Livingstone, Edinburgh,United Kingdom, 1997. https://doi.org/10.1177/088506660101600307

[118] L. Sadighpour, S. M. M. Rezaei, M. Paknejad, F. Jafary, and P. Aslani. The application of an artificial neural network to support decision making in edentulous maxillary implant prostheses. *Journal of Research and Practice in Dentistry*, 2014:i1–10, 2014. http://doi.org/10.5171/2014.369025

[119] J. Sanghavi and A. Aditya. Applications of corticosteroids in dentistry. *Journal of Dental and Allied Sciences*, 4(1):19–24, 2015. http://doi.org/10.4103/2277-4696.167533

[120] S. Sen, A. B. Swoap, Q. Li, B. Boatman, I. Dippenaar, R. Gold, M. Ngo, S. Pujol, B. Jackson, and B. Hecht. Cartograph: Unlocking spatial visualization through semantic enhancement. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, IUI '17, pages 179–190, New York, NY, USA, 2017. http://doi.org/10.1145/3025171.3025233

[121] T. Shah, F. Rabhi, and P. Ray. Investigating an ontology-based approach for big data analysis of inter-dependent medical and oral health conditions. *Cluster Computing*, 18(1):351–367, 2015. http://doi.org/10.1007/s10586-014-0406-8

[122] T. Shah, F. Rabhi, P. Ray, and K. Taylor. A guiding framework for ontology reuse in the biomedical domain. In *2014 47th Hawaii International Conference on System Sciences (HICSS)*, pages 2878–2887, 2014. http://doi.org/10.1109/HICSS.2014.360

[123] M. Sharp. Toward a comprehensive drug ontology: extraction of drug-indication relations from diverse information sources. *Journal of Biomedical Semantics*, 8(2), 2017. http://doi.org/10.1186/s13326-016-0110-0

[124] S. Shastri and V. Mansotra. Kdd-based decision making: A conceptual framework model for maternal health and child immunization databases. In *Advances in Computer Communication and Computational Sciences*, pages 243–253, 2019. http://doi.org/10.1007/978-981-13-6861-5_21

[125] C. Shi, Z. Zhang, P. Luo, P. S. Yu, Y. Yue, and B. Wu. Semantic path based personalized recommendation on weighted heterogeneous information networks. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, pages 453–462, New York, NY, USA, 2015. http://doi.org/10.1145/2806416.2806528

[126] R. Shibl, M. Lawley, and J. Debuse. Factors influencing decision support system acceptance. *Decision Support Systems*, 54(2):953–961, 2013. https://doi.org/10.1016/j.dss.2012.09.018

[127] A. Simoes, A. Correia, T. Marques, and R. Figueiredo. Preliminary study of the clinical applicaiton of the cds system oradii in a univerisyt dental clinic. In T. Jorge and N. Jorge, editors, *Computational vision and medical image processing*, pages 93–96. Taylor & Francis Group, London, 2012.

[128] A. M. Sirajuddin, J. A. Osheroff, D. F. Sittig, J. Chuo, F. Velasco, and D. A. Collins. Implementation pearls from a new guidebook on improving medication use and outcomes with clinical decision supportvement imperatives. *Journal of Healthcare Information Management*, 23(4):38–45, 2009.

[129] P. R. Smart and M. Sadraie. Applications and uses of dental ontologies. In *Proceedings of the 2012 IADIS International Conference*, pages 499–504, 2012.

[130] F. Spada, T. Barnes, and K. Greive. Comparative safety and efficacy of topical mometasone furoate with other topical corticosteroids analysis. *Australasian Journal of Dermatology*, 59(3), 2018. http://doi.org/10.1111/ajd.12762

[131] K. Sparck Jones and P. Willett, editors. *Readings in Information Retrieval*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1997.

[132] W. W. Stead and H. S. Lin. *Computational Technology for Effective Health Care: Immediate Steps and Strategic Directions*. National Academies Press, 2009. http://doi.org/10.17226/12572

[133] K. Suda, H. Henschel, and U. Patel. Use of antibiotic prophylaxis for tooth extractions, dental implants, and periodontal surgical procedures. *Open Forum Infectious Diseases*, 5(1):250–254, 2018. http://doi.org/10.1093/ofid/ofx250

[134] G. M. Sullivan and A. R. Artino. Analyzing and interpreting data from likert-type scaless. *Journal of graduate medical education*, 5(4):541–542, 2013. http://doi.org/10.4300/JGME-5-4-18

[135] R. Sutton, D. Pincock, D. Baumgart, D. Sadowski, R. Fedorak, and K. Kroeker. An overview of clinical decision support systems: benefits, risks,and strategies for success. *Nature Partner Journals Digital Medicine*, 3(17), 2020. http://doi.org/10.1038/s41746-020-0221-y

[136] T. Takagi and M. Sugeno. Fuzzy identification of systems and its applications to modeling and control. *IEEE Transactions on Systems, Man and Cyberbstems*, 15(1):116–132, 1985. http://doi.org/10.1109/TSMC.1985.6313399

[137] D. Tang, F. Wei, N. Yang, M. Zhou, T. Liu, and B. Qin. Learning sentiment-specific word embedding for twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1555–1565, Baltimore, Maryland, June 2014. http://doi.org/10.3115/v1/P14-1146

[138] L. Tari, S. Anwar, S. Liang, J. Cai, and C. Baral. Discovering drug-drug interactions: a text-mining and reasoning approach based on properties of drug metabolism. *Bioinformatics*, 26(18):547–553, 2010.

[139] A. Tharwat. Classification assessment methods. *Applied Computing and Informatics*, 2018. http://doi.org/10.1016/j.aci.2018.08.003

[140] T. Thyvalikakath, M. Dziabiak, R. Johnson, M. Torres-Urquidy, J. Yabes, T. Schleyer, and A. Acharya. Designing clinical data presentation in electronic dental records using cognitive task analysis methods. In *2012 IEEE Second International Conference on Healthcare Informatics, Imaging and Systems Biology*, pages 74–74, 2012. http://doi.org/10.1109/HISB.2012.24

[141] T. P. Thyvalikakath, M. P. Dziabiak, R. Johnson, M. H. Torres-Urquidy, A. Acharya, J. Yabes, and T. K. Schleyer. Advancing cognitive engineering methods to support user interface design for electronic health records. *International Journal of Medical Informatics*, 83(4):292–302, 2014. http://doi.org/10.1016/j.ijmedinf.2014.01.007

[142] M. M. Van der Zande, R. C. Gorter, and D. Wismeijer. Dental practitioners and a digital future: an initial exploration of barriers and incentives to adopting digital technologies. *British dental journal*, 215(11):E21, 2013. http://doi.org/10.1038/sj.bdj.2013.1146

[143] V. Varonen, T. Kortteisto, and M. Kaila. What may help or hinder the implementation of computerized decision support systems: a focus group study with physicians. *Health Policy and Technology*, 25(3):162–167, 2008. http://doi.org/10.1093/fampra/cmn020

[144] V. Venkatest, M. Morris, G. Davis, and F. Davis. User acceptance of information technology: towards a unified view. *MIS Quarterly*, 27(3):425–478, 2003. http://doi.org/10.2307/30036540

[145] P. M. Victor Sudrez-Paniagua, Isabel Segura-Bedmar. Exploring convolutional neural networks for drug-drug interaction extraction. *Database*, 2017:bax019, 2017. http://doi.org/10.1093/database/bax019

[146] K. Vikram and F. Karjodkar. Decision support systems in dental decision making: An introduction. *Journal of Evidence Based Dental Practice*, 9(2):73–76, 2009. http://doi.org/10.1016/j.jebdp.2009.03.003

[147] M. F. Walji, D. Taylor, J. R. L. II, and J. A. Valenza. Factors influencing implementation and outcomes of a dental electronic patient record system. *Journal of Dental Education*, 73(5):589–600, 2009.

[148] C. Wang, Y. Song, H. Li, Y. Sun, M. Zhang, and J. Han. Distant meta-path similarities for text-based heterogeneous information networks. In *Proceedings of the 2017 ACM on*

Conference on Information and Knowledge Management, pages 1629–1638, New York, NY, USA, 2017. http://doi.org/10.1145/3132847.3133029

[149] Y. Wang, S. Liu, M. Rastegar-Mojarad, L. Wang, F. Shen, F. Liu, and H. Liu. Dependency and amr embeddings for drug-drug interaction extraction from biomedical literature. In *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology,and Health Informatics*, ACM-BCB '17, pages 36–43, New York, NY, USA, 2017. http://doi.org/10.1145/3107411.3107426

[150] B. M. Welch, S. R. Loya, K. Eilbeck, and K. Kawamoto. A proposed clinical decision support architecture capable of supporting whole genome sequence information. *Journal of Personalised Medicine*, 4(2):176–199, 2014.

[151] S. White. Decision-support systems in dentistry. *Journal of Dental Education*, 60(1):47–63, 1996.

[152] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2016.

[153] A. P. Wright, A. T. Wright, A. B. McCoy, and D. F. Sittig. The use of sequential pattern mining to predict next prescribed medications. *Journal of Biomedical Informatics*, 53:73–80, 2015. http://doi.org/10.1016/j.jbi.2014.09.003

[154] H.-Y. Wu, C.-W. Chiang, and L. Li. Text mining for drug-drug interaction. *Methods in molecular biology*, 1159:47–75, 2014.

[155] S. Yan, X. Jiang, and Y. Chen. Text mining driven drug-drug interaction detection. *2013 IEEE International Conference on Bioinformatics and Biomedicine*, pages 349–354, 2013. http://doi.org/10.1109/BIBM.2013.6732517

[156] Z. Yi, S. Li, J. Yu, Y. Tan, Q. Wu, H. Yuan, and T. Wang. Drug-drug interaction extraction via recurrent neural network with multiple attention layers. In G. Cong, W.-C. Peng, W. E. Zhang, C. Li, and A. Sun, editors, *Advanced Data Mining and Applications*, pages 554–556, 2017. http://doi.org/10.1007/978-3-319-69179-4_39

[157] X. Zeng, Z. Jia, Z. He, W. Chen, X. Lu, H. Duan, and H. Li. Measure clinical drug-drug similarity using electronic medical records. *International Journal of Medical Informatics*, 124:97–103, 2019. http://doi.org/10.1016/j.ijmedinf.2019.02.003

[158] H. Zhang, Q. Hu, Y. Yao, and Q. Wang. Design and implementation of a knowledge engineering-based dental diagnostic expert system. In *2009 WRI World Congress on*

Computer Science and Information Engineering, volume 5, pages 362–366, 2009. http://doi.org/10.1109/CSIE.2009.589

[159] X. Zhang, J. Zhao, and Y. LeCun. Character-level convolutional networks for text classification. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 649–657. Curran Associates, Inc., 2015.

[160] Y. Zhang, A. Jatowt, and K. Tanaka. Towards understanding word embeddings: Automatically explaining similarity of terms. *2016 IEEE International Conference on Big Data (Big Data)*, pages 823–832, 2016. http://doi.org/10.1109/BigData.2016.7840675

[161] Y. Zhang and P. Kordjamshidi. Pe-tu participation at tac 2018 drug-drug interaction extraction from drug labels. In *Proceedings of the Eleventh Text Analysis Conference, 13-14 Nov 2018, Maryland, USA*, 2018.

[162] Z. Zhao, Z. Yang, LingLuo, H. Lin, and JianWang. Drug drug interaction extraction from biomedical literature using syntax convolutional neural network. *Bioinformatics*, 32(22):3444–3453, 2016. http://doi.org/10.1093/bioinformatics/btw486

[163] G. Zhong, L.-N. Wang, X. Ling, and J. Dong. An overview on data representation learning: From traditional feature learning to recent deep learning. *The Journal of Finance and Data Science*, 2(4):265–278, 2016. http://doi.org/10.1016/j.jfds.2017.05.001

[164] Y. Zhu, E. Yan, and F. Wang. Semantic relatedness and similarity of biomedical terms: examining the effects of recency, size, and section of biomedical publications on the performance of word2vec. *BMC Medical Informatics and Decision Making*, 17(1):95, Jul 2017. http://doi.org/10.1186/s12911-017-0498-1

[165] J. Y. Zong, J. Leese, A. Klemm, E. C. Sayre, J. Memetovic, J. M. Esdaile, and L. C. Li. Rheumatologist's views and perceived barriers to using patient decision aids in clinical practice. *Arthritis Care & Research*, 67(10):1463–1470, 2015. http://doi.org/10.1002/acr.22605

[166] Q. Zou, S. Xie, Z. Lin, M. Wu, and Y. Ju. Finding the best classification threshold in imbalanced classification. *Big Data Research*, 5:2–8, 2016. http://doi.org/doi.org/10.1016/j.bdr.2015.12.001

Appendix A

Participation Information Sheet

Participant Information for USQ
Research Project Questionnaire

Project Details

Title of Project: Survey on Clinical Decision Support System
Human Research Ethics Approval Number: H19REA262

Research Team Contact Details

Principal Investigator:	Supervisor Details
Mr Wee Goh	Dr Xiaohui Tao
email: w0077925@umail.usq.edu.au	xtao@usq.edu.au
Tel: +65 63447747	Tel: +61 7 4631 1576

Description

This evaluation is to gauge the usefulness and ease of use of a proposed clinical decision support system to assist dental health professionals at point-of-care in the prescription of medicine to patients. It is part of a research project in collaboration with the University of Southern Queensland. The research team requests your assistance as you will be the potential user of such a system in the dental clinic.

Participation

Your participation will involve completion of about 20 survey questions pertaining to the perceived usefulness, usability and user satisfaction. You will need to choose on a scale of 1 to 5 (1 being “Strongly Disagree” and 5 being “Strongly Agree” and will take less than 10 minutes of your time. For example, in this question, you will circle one of the numbers 1,2,3,4 or 5:
Using the system in my job would increase my productivity: 1 2 3 4 5

Your participation in this project is entirely voluntary. If you do not wish to take part, you are not obliged to. If you decide to take part and later change your mind, you are free to withdraw from the project at any stage. If you wish to withdraw the response that you have submitted, kindly inform me.
Your decision whether you take part, do not take part, or to take part and then withdraw, will in no way impact your current or future relationship with the University of Southern Queensland.

Expected Benefits

It is expected that the response collected will be used for possible upgrade of the system. Trying out the system in order to answer the survey will also enable you to get familiar with the functions and features provided by the system.

Risk

There are no risks associated with this participation except the occasional interruption from patients or phone calls. In such a case, you can resume at another time at your convenience.

Privacy and Confidentiality

All comments and responses will be treated confidentially unless required by law. Identity of participants in the evaluation is not required and will not be requested. The responses from all the participants will be compiled to gauge the overall usefulness of the system. The extent of the usefulness and reliability of the system may be reported in journals, conferences and thesis.
You may also request a copy of the survey results where the overall score will be computed Any data collected as a part of this project will be stored securely as per University of Southern Queensland’s Research Data Management policy.

Consent to Participate

The return of the completed questionnaire is accepted as an indication of your consent to participate in this project.

Questions or Further Information about the Project

Please refer to the Research Team Contact Details at the top of the form to have any questions answered or to request further information about this project.

Concerns or Complaints Regarding the Conduct of the Project

If you have any concerns or complaints about the ethical conduct of the project, you may contact the University of Southern Queensland Manager of Research Integrity and Ethics on +61 7 4631 1839 or email researchintegrity@usq.edu.au. The Manager of Research Integrity and Ethics is not connected with the research project and can facilitate a resolution to your concern in an unbiased manner.

Thank you for taking the time to help with this research project. Please keep this sheet for your information.

Appendix B

Survey Questions

Dear Participant, this evaluation is to gauge the usefulness and ease of use of a proposed clinical decision support system to assist dental health professionals at point-of-care in the prescription of medicine to patients. It is part of a research project in collaboration with the University of Southern Queensland

Evaluation Procedures

1. Start the software system.
2. Register a new patient.
3. Specify the drug allergy and current drugs that the patient may be taking.
4. Experiment with the system by specifying the drug allergy and/or current drugs that the patient is currently taking.
5. Evaluate the usefulness of the system by prescribing a drug that the patient usually needs, like antibiotics.
6. Mark your response on the paper on a scale of 1 to 5, 1 being “Strongly Disagree” and 5 being “Strongly Agree”.
7. Place the completed responses on the out tray beside the computer.

Results
The responses from all the dentists will be compiled and the results will be made available through conferences and seminars

Privacy
Identity of participants in the evaluation is not required and will not be requested

Risk There are no risks associated with this participation except the occasional interruption from patients or phone calls. In such a case, you can resume at another time at your convenience

		1	2	3	4	5
	Perceived Usefulness					
1	Using the system in my job would enable me to accomplish tasks more quickly					
2	Using the system would improve my job performance					
3	Using the system in my job would increase my productivity					
4	Using the system would enhance my effectiveness on the job					
5	Using the system would make it easier to do my job					
6	I find the system useful when I need to check if the drug to be prescribed is appropriate					
	Perceived Ease of Use					
1	Learning to operate the system is easy for me					
2	I find it easy to get the system to do what I want it to do					
3	My interaction with the system is clear and understandable					
4	I find the system to be flexible to interact with					
5	It would be easy for me to become skillful at using the system					
6	I find the system easy to use					
	User Satisfaction					
1	I am satisfied with the way the system advises on the drug that I prescribe					
2	I feel confident in using the system					
3	I find it easy to share the information with my patients					
4	I can get the results quickly					
5	The system enhances the quality of care for my patients					
6	It would be easy for me to become skilful at using the system					
	Attributes of Usability					
1	It is easy to interact with the drug prescription system					
2	The features enable me to decide if the drug is appropriate for my patient					
3	I find it easy to specify the medical profile of the patient					
4	I found the various functions in this system were well integrated					
5	I think that I would like to use this system if made available					