

**Drug Side Effects Data Representation and Full Spectrum Inferencing using  
Knowledge Graphs in Intelligent Telehealth**

By  
Saravanan Jayaraman B.Tech, M.B.A

Submitted in partial fulfillment  
of the requirements for the degree of  
Doctor of Professional Studies  
in Computing

at

School of Computer Science and Information Systems

Pace University

October 2016

ProQuest Number: 10247923

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.



ProQuest 10247923

Published by ProQuest LLC (2016). Copyright of the Dissertation is held by the Author.

All rights reserved.

This work is protected against unauthorized copying under Title 17, United States Code  
Microform Edition © ProQuest LLC.

ProQuest LLC.  
789 East Eisenhower Parkway  
P.O. Box 1346  
Ann Arbor, MI 48106 - 1346

We hereby certify that this dissertation, submitted by **Saravanan Jayaraman** has successfully satisfied all the requirements for the degree of the Doctor of Professional Studies in Computing”

Lixin Tao  
Dr. Lixin Tao  
Chairperson of Dissertation Committee

October 11, 2016  
Date

Ronald J. Frank  
Dr. Ronald Frank  
Dissertation Committee Member

October 11, 2016  
Date

Meikang Qiu  
Dr. Meikang Qiu  
Dissertation Committee Member

October 11, 2016  
Date  
10/11/2016

Seidenberg School of Computer Science and Information Systems  
Pace University

## **Abstract**

### **Drug Side Effects Data Representation and Full Spectrum Inferencing using Knowledge Graphs in Intelligent Telehealth**

By  
Saravanan Jayaraman

Submitted in partial fulfillment  
of the requirements for the degree of  
Doctor of Professional Studies  
in Computing

October 2016

Drug adverse reaction data contains important constraints about side effects and conflict avoidance of component and compound drugs. We observe that many of these constraints are transitive in nature due to the relationship between drug and drug classes. Current drug side effects representations in XML does not have a proper knowledge representation mechanism to clearly specify all kinds of dependencies among the drug components and drugs. Even the recently introduced OWL based approach for medical drug side effects data representation still suffers from several shortcomings inherent to the OWL restrictions like using “is-a” relationship and usage of object property emulations.

In this research, we propose a model *Drug - Side Effects Representation And Inferencing (D -SERI)* built using Knowledge Graph (KG) and enhanced *PaceJena* to represent multiple custom relationships allowing domain experts to capture the transitive nature of the relations in an inference friendly way. The research also developed a concept demonstrator for checking out prescriptions to avoid complications. The research outcome shows that the proposed model allows the doctors and caregivers to derive

dynamic information about side-effects avoiding costly errors caused by human interpretation.

PREVIEW

## Acknowledgements

I would like to dedicate my dissertation work to my parents whose endless love and affection gave me the power to work harder everyday with determination and perseverance to complete the research.

I would like to thank my beloved wife Viji who stood with me in every step of the way and endured countless difficulties and challenges while supporting my family during my absence. Her encouraging words and motivation throughout was a constant energy booster for me during the DPS years.

I affectionately thank my children Lakshaanya and Siddesh whose curious questions and supportive behavior gave me the zeal to strive for excellence. I wish to see them successful in life in their own way and surpass their parent's accomplishments one day.

THANK YOU to my dissertation advisor Dr. Lixin Tao for believing in me and providing constant encouragement and showing the knack to breakdown the dissertation topic and the technique to solve complex research problems. I greatly benefited from his years of guidance experience and incredible counsel. I am and will always be forever GRATEFUL!

Thanks to Dr. Frank and Dr. Qiu for all the great advice, and for being on my dissertation committee.

Special Thanks to my fellow researcher Ning Jiang who offered expert technical assistance and code review with me.

Finally, I would like to sincerely acknowledge my cohorts from the class of 2016 for supporting each other and staying together during the DPS journey.

*“Dream is not what you see in Sleep, dream is something which does not let you sleep” – Dr A P J Abdul Kalam.* Like many others before me, I definitely had many sleepless nights to accomplish my dream. During this, I also realized that the Seeking higher level of knowledge is a journey not a destination and I intend to continue this journey throughout my life.

## Table of Contents

Abstract.....	iii
List of Tables.....	ix
List of Figures .....	x
Chapter 1 Introduction .....	1
1.1 Drug Side Effects Inference Problem.....	1
1.1.1 Side Effects Awareness .....	2
1.1.2 Growth of Drug Side Effects Data .....	6
1.1.3 Maintaining the Side Effects Data Current.....	6
1.1.4 Motivating Example .....	7
1.1.5 Benefits of dynamic Side Effect Inference .....	9
1.2 Representation of Drug Side Effects Data in Knowledge Graph .....	10
1.2.1 Knowledge Graph Usage in Healthcare .....	10
1.3 Problem statement.....	11
1.4 Proposed Solution Methodology.....	12
1.5 Expected Key Contributions.....	13
1.6 Dissertation Road Map.....	14
Chapter 2 Current State of Drug Adverse Relationships Data Representation.....	16
2.1 Industry wide Initiatives .....	16
2.1.1 MedWatch Initiative.....	16
2.1.2 EU-ADRS Data Initiative .....	19
2.1.3 Other Industry Wide Alternatives, SnowMed .....	21
2.2 DARs representation in XML.....	22

2.3	DARs Representation in OWL .....	24
2.4	Restrictions with Current Approaches .....	27
2.5	Comparative Study .....	29
2.6	Summary of findings .....	33
2.7	Key Tools and Methodology .....	33
2.7.1	Knowledge Graph .....	33
2.7.2	Usage of Knowledge graphs in Healthcare .....	34
2.7.3	RDF and RDFS .....	34
2.7.4	OWL and Custom Relationships .....	36
2.7.5	Apache Jena .....	36
2.7.6	Pace Jena .....	37
2.7.7	OWL VIZ .....	37
2.7.8	Protégé Tool with Pace Jena .....	37
2.7.9	Knowledge Association .....	38
2.7.10	OWL is a relationship .....	38
2.7.11	OWL part of relationship .....	39
2.7.12	Conclusion .....	40
Chapter 3	Solution Methodology .....	41
3.1	Knowledge Graph - Drug Side Effects Data .....	41
3.1.1	Knowledge Graph Usage in Healthcare .....	41
3.1.2	Knowledge Graph and Pace Jena Extension .....	42
3.2	Proposed Framework .....	44
3.2.1	DARs (Drug adverse relationships) using custom relationships .....	44
3.2.2	Syntax Definition .....	47
3.2.3	Knowledge Graph Development .....	47
3.2.4	Inferred Results .....	55
3.3	Solution Methodology - Outcome .....	57



3.4	Solution Methodology Files .....	58
3.4.1	Snippet of DrugSummary.owl .....	58
3.5	Conclusion .....	63
Chapter 4	Solution Implementation.....	64
4.1	Introduction.....	64
4.2	Implementation Details .....	64
4.3	Research Configuration and Data Capture .....	67
4.4	Research Equipment.....	67
4.5	Concept Demonstrator Web .....	67
4.5.1	Concept demonstrator design.....	68
4.5.2	Data download from public sources.....	70
4.5.3	Drug adverse data knowledge graph development .....	70
4.5.4	Parsing ontologies with an enhancement method - PaceJena Code.....	71
4.5.5	Protégé Web Application - Eclipse Java EE IDE .....	72
4.5.6	Trial Run.....	73
4.6	Key Benefits of the Concept Demonstrator.....	74
4.7	Summary.....	74
Chapter 5	Experimental Validation .....	75
5.1	Full Spectrum Drug Side Effect Inferencing – Trial Run 1 (Saxagliptin) .....	75
5.2	Full Spectrum Drug Side Effect Inferencing – Trial Run 2 (Alogliptin) .....	78
5.3	Experimental Results Analysis .....	80
5.4	Validate D-SERI model to larger set of drug domain data .....	81
5.5	Summary Findings and Concept Validation.....	87
5.6	Conclusion .....	87
Chapter 6	Conclusion and Future work .....	88
6.1	Conclusion of the dissertation.....	88
6.2	Future Work.....	88

Appendix A Abbreviations .....	89
Appendix B PaceJena.java getParentRelations() .....	92
Appendix C UploadHandler.java code .....	93
References .....	94

PREVIEW

## List of Tables

Table 1 Doctor's action during a patient visit .....	7
Table 2 Pros and Cons FAERS Data Model .....	18
Table 3 Summary of Known Ontology-Based Drug Knowledge Representation .....	31

PREVIEW

## List of Figures

Figure 1 Aggressive Growth of side effects data in the last decade.....	3
Figure 2 Side effects data reporting by HCP vs Consumers.....	4
Figure 3 Prescription drug use in the past 30 days among adults aged 18 and over, United States, 1988–1994 and 2007–2010 .....	5
Figure 4 Prescription drug use in combinations among adults aged 18 and over .....	5
Figure 5 Saxagliptin drug class hierarchy [26] in knowledge graph.....	8
Figure 6 FDA Adverse Data reporting – data flow .....	17
Figure 7 EudraVigilance ICSR Data reporting .....	20
Figure 8 SNOWMED - standardized clinical terminology.....	21
Figure 9 FAERS data in xml representation. ....	24
Figure 10 Literature survey 1: Drug Ontology by Samson et al – quoted from drug ontology documentation.....	26
Figure 11 FAERS sample .....	28
Figure 12 Semantic Web Layer.....	35
Figure 13 Foundational knowledge representation - using custom OWL relations <i>takes</i> , <i>cause</i> and <i>partOf</i> in drug domain data.....	45
Figure 14 Knowledge Graph linked data model - seamless integration of Patient, Drugs, Side effects, Drug Class and Doctors .....	46
Figure 15 Protégé - Class Hierarchy definition.....	49
Figure 16 OWLViz - AntiDiabeticDrug.....	50
Figure 17 OWLViz – DrugClass.....	52
Figure 18 OWLViz – Drug .....	54
Figure 19 Adverse reactions caused by Saxagliptin drug with class DPP4Inhibitors and AntiDiabeticDrug .....	56
Figure 20 D-GPR – An Iterative approach to get full spectrum side effects using knowledge graph.....	65

Figure 21	Concept demonstrator web app .....	69
Figure 22	Drug ontology – UseCase Validation .....	71
Figure 23	getParentRelations – Extension to PaceJena .....	72
Figure 24	Eclipse IDE Project.....	73
Figure 25	Test Run – Home Page .....	76
Figure 26	Test Run – Outcome Page.....	77
Figure 27	Test Run – Home Page .....	79
Figure 28	Test Run – Outcome Page.....	80
Figure 29	Drug side effects identified by D-SERI (Glipitin drug class) .....	80
Figure 30	Drug side effects identified by D-SERI (Flozin drug class) .....	81
Figure 31	Knowledge graph with larger data set 1.....	83
Figure 32	Knowledge graph with larger data set 2 .....	84
Figure 33	Test Run – Show ALL Page.....	86

## **Chapter 1**

### **Introduction**

Study of Medical Drug side effects on Humans was first documented by the Greek Physician Hippocrates of Kos in 460 BC who first studied varied effects of Aspirin as a migraine to relieve pain and suffering on his patients. His followers from Hippocrates school of Medicine further continued the study of clinical study summing up the medical knowledge and wrote the everlasting drug side effect data on early human races. Along with these early studies by doctors in Graco-latino world, several similar advanced civilizations conducted studies in asymmetric manner setting up strong data spanning over centuries to be used by mankind.

#### **1.1 Drug Side Effects Inference Problem**

With the advent of the Web advancement with research and development, numerous domains have been remarkably driven by the emerging techniques, such as semantic web [1] [2] cloud computing [3][4][5], and big data[6]. This is even more relevant in the area of Drug development and monitoring in Telehealth [7] [8] where Domain experts depend upon these emerging technologies to encode knowledge. Recently, there is an explosion in the number of drugs approved for treatment and the effect they caused on human

population. In the US market alone, there are 2,000 medications, 94,450 health products and around 175,950 health packages with different active ingredients and these numbers are growing steadily every year. With these huge volume of drug combinations comes the challenge to have a relevant knowledge mechanism to capture the huge volume of side effects data and enable the patients and doctors a way to derive meanings about the potential outcome quickly. For example, when a doctor treats a patient with diabetic condition with any one of the drugs approved for treatment, he or she wants to make sure the benefit of prescribing the drug exceeds the side effects caused by the drug itself, in that case of side effect showing up, he or she wants to relate that to the prescribed drug and then quickly change the treatment option to suit the patient's responses to drugs. While there has been huge advances in the drug side effects data capturing globally including the nationally mandated options like MedWatch or EU drug watch and the availability of that information to public domain, the Information Technology processes behind have not been coping up due to several factors.

#### *1.1.1 Side Effects Awareness*

The first is the explosion of the drug side effects data itself. Look at the growth of the side effects in the last 10 years growing exponentially. This data is extremely dynamic. Most of the Institutes and Research organizations spend enormous amount of time in just getting the data downloaded and synced periodically. As per the recent look on the MedWatch system there are 7 million data entries in the system. Figure 1 shows the continuous growth in the data. The interesting question to ask is what happens when a

new adverse event data is released and how it is synchronized with the previously stored data thereby staying relevant.

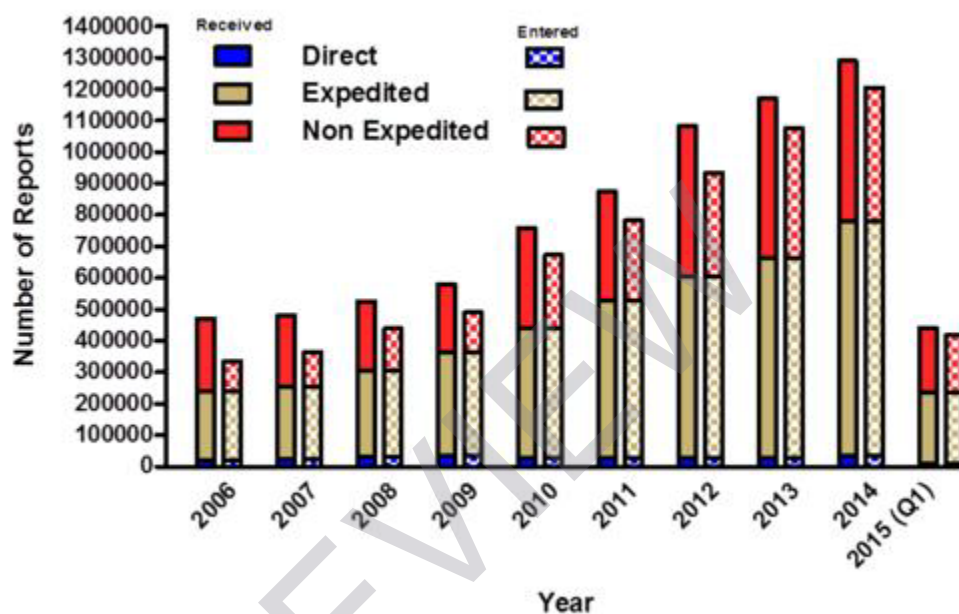
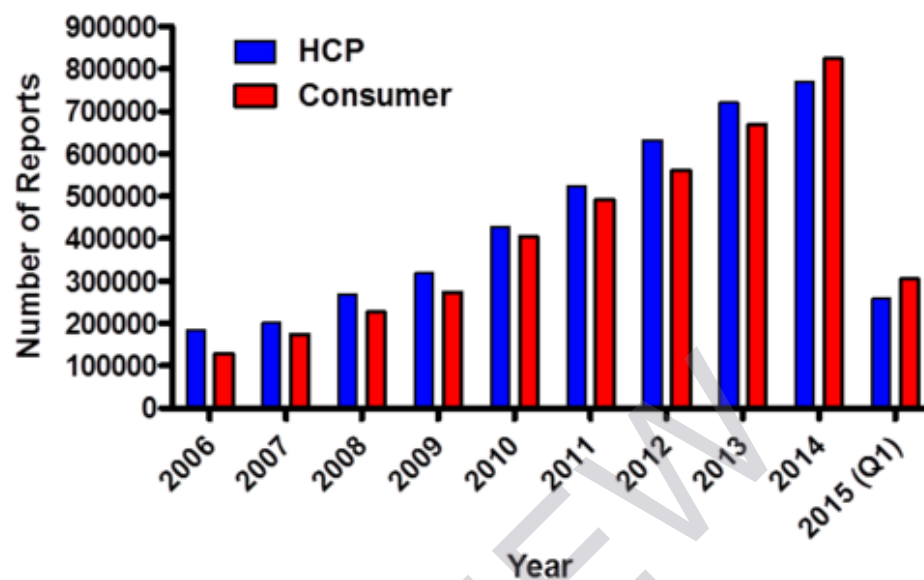


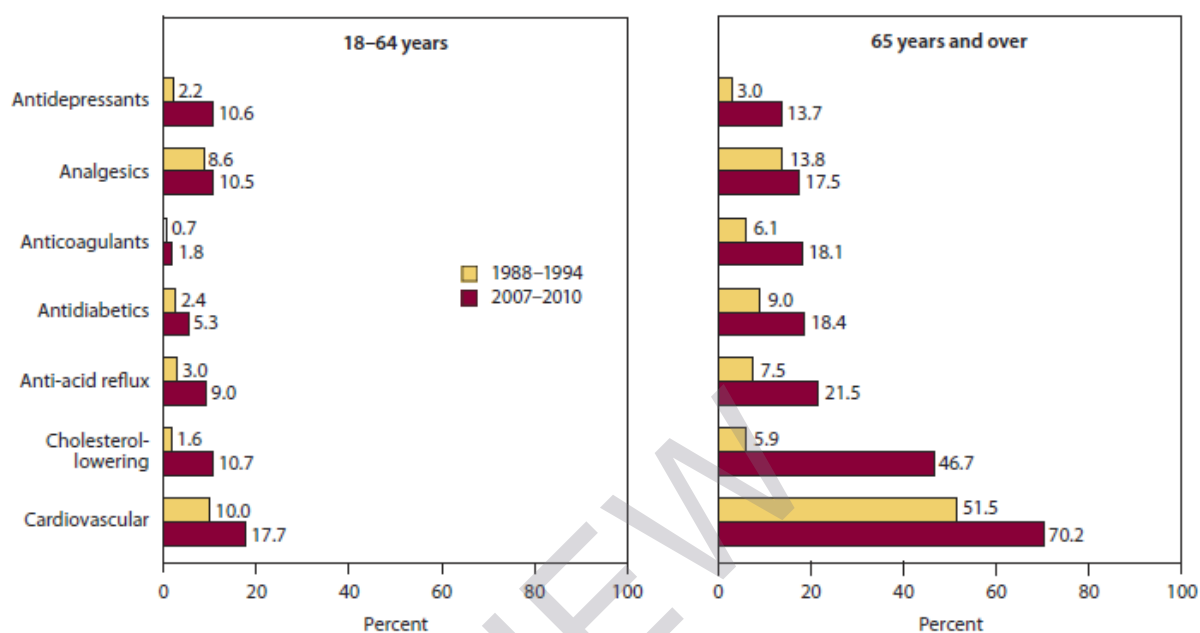
Figure 1 Aggressive Growth of side effects data in the last decade



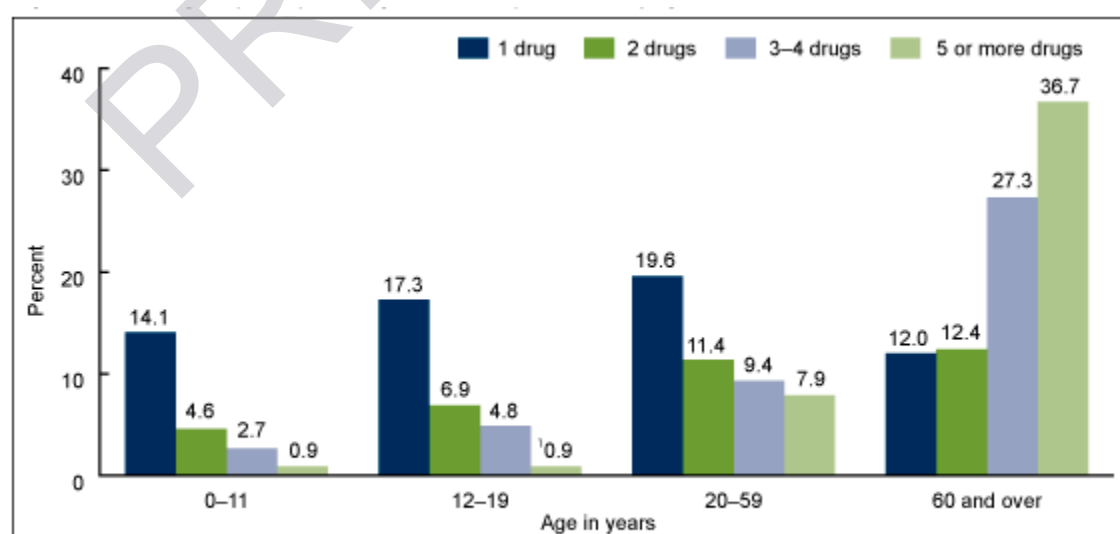


**Figure 2 Side effects data reporting by HCP vs Consumers**

A recent study by Mayo Clinic Survey [9] points that nearly 7 in 10 Americans take some form of Prescription drugs, and medication errors are surprisingly common and costly to the nation [10]. As per a study conducted by NIH.gov in a North Indian City [27], only 33% of patients knew about the side effects produced by the concerned drug, and only 15.68% knew how to recognize them. This is also confirmed by the CDC report where 70% of adults over 65 years take prescription drugs and 1 in 4 senior Americans take more than 3 prescription drugs.



**Figure 3 Prescription drug use in the past 30 days among adults aged 18 and over, United States, 1988–1994 and 2007–2010**



<sup>1</sup>Estimate is unstable; the relative standard error is greater than 30%.  
SOURCE: CDC/NCHS, National Health and Nutrition Examination Survey.

**Figure 4 Prescription drug use in combinations among adults aged 18 and over**

### *1.1.2 Growth of Drug Side Effects Data*

As per the CDC study the usage of prescription drugs increased by 10% last decade and the use of multiple prescription drugs increased by 20% and the use of 5 or more drugs increased by 70%. As pointed by the same CDC source the usage of prescription drug is directly related to availability of regular healthcare and as more and more people have access to regular healthcare the usage of prescription drugs is expected to grow linear as the necessity to treat multiple diseases at the same time especially on the senior population. This shows the problem with the data representation is not a one-time issues rather a long term issue requiring clarity and newer approaches from knowledge scientists.

### *1.1.3 Maintaining the Side Effects Data Current*

Whenever newer treatment options are available in the market, healthcare providers prefer to prescribe the advanced treatment option to their patients so they get the advanced therapeutic benefits and often doctors are looking for a better way to understand the knowledge about these new drugs to help treat the patients.

Here is one scenario. A patient suffering from uncontrolled diabetes is looking to utilize some of the newer drugs in the category. The doctor, while aware of the potential side effects of a new drug is looking forward to understand side effects of the drug better so that it does not conflict with the current treatment option for the patient. Here the doctor relies on several types of information like FDA's MedWatch data [20], or any

other data provided by the manufacturer or labelling information. While these information provide some levels of details about the drug side effects, the studies have pointed that the study of the drug side effects is one of the complex task to perform by a common man or physicians requiring enormous levels of investigation which is simply impossible due to lack of time or resources. According to the recent study by Harvard medical analytic group the knowledge mechanism of the drug data domain is not catching up with the growth pace of the data itself due to the restrictions on the current format and the lack of validation mechanisms.

#### *1.1.4 Motivating Example*

This section shows a real time example where the proposed methodology can make a difference in the drug side effects data representation.

Table 1 exhibits different scenarios where the research can help doctors to avoid human interpretation errors.

**Table 1 Doctor's action during a patient visit**

<b>Case</b>	<b>Patient observation for a drug</b>	<b>Doctor's action</b>
1	Patient does not report side effect with drug	Continue to prescribe the drug
2	Patient reports side effects but it's not known with drug	Continue to prescribe the drug while reporting new side effects.

3	Patient reports side effects and its known with the drug	STOP the drug and look for alternatives.
---	--	--

While Doctors prescribe drugs during routine visits, they always strive to make sure the benefits of prescribing the drug outweighs the risk caused by side effects. Here they always seek in-depth knowledge about the side effects keeping in mind the ultimate safety of the patient's life at the forefront of the drug prescription strategy. We take three typical scenarios often faced by doctors for this dissertation. There are several other variations of the scenarios are possible but for the case study perspective we restrict it to the three choices.

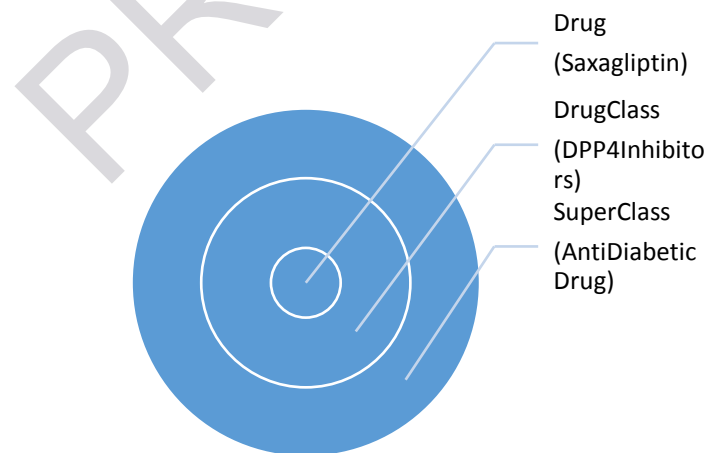


Figure 5 Saxagliptin drug class hierarchy [26] in knowledge graph

*Saxagliptin* drug is used to treat *diabetes* and they are newest treatment options for patients who are not responding well for other diabetic treatment options. While Saxagliptin----causes--- set of side effects (|A| - *Abdominal Pain, Motor Dysfunction, Hyperhidrosis, malaise, Nasal Congestion, Increased Blood Sugar, Arrhythmia, rash, Cerebro Vascular Accident* ) which is well known, doctors often find it that it does not represent the full list possible side effects. In this case, Saxagliptin's parent class DPP4Inhibitors (gliptin) causes its own set of common side effects – ( |B| - *nausea,diarrhea,stomach pain,headache,runny nose,sore throat,pancreatitis,severe join pain*).

When Doctors look for side effects caused by a drug against reported by patents, they often rely on the direct side effect list |A| as primary source as the full spectrum of the side effects is not available to them due to several reasons. This could cause them to overlook side effects like *pancreatitis* which is in |B|. Studies show that while emulations with object properties are used to capture these additional relations they often cause other problems like high cost of maintenance on data modelers.

### 1.1.5 Benefits of dynamic Side Effect Inference

Linking the component and compound drugs using the proposed knowledge graph based approach allows the domain experts to capture the full spectrum side effects of the drug ( |A| + |B| ) by including all possible side effects while reducing syntax burden to knowledge modelers compared with any other workarounds like object properties.

Such a dynamic data representation model will also provide a full spectrum side effects to the doctors and patient helping them immensely to benefit to either adapt newer treatment models without fear or just to choose a suitable treatment model beneficial to the patient. Even when the knowledge about the drug side effects is available the current format makes it harder medical drug side effects is neither flexible nor suitable in the full spectrum nuisances of the drug side effects data drug being a chemical component.

## **1.2 Representation of Drug Side Effects Data in Knowledge Graph**

A Knowledge graph describes the concepts in the domain and also the relationships that hold between those concepts. Different knowledge graph languages provide different facilities. It makes it possible for concepts to be defined as well as described. Complex concepts can therefore be built up in definitions out of simpler concepts.

### *1.2.1 Knowledge Graph Usage in Healthcare*

The usage of Knowledge Graph is well accepted especially in the semantic web [16][17] area as a primary way of disseminating the information to users or machines even though it's still evolving in the medical domain. For example, Google's knowledge vault [50] is enriched with information about 570 million objects of data and 18 billion facts making the world's largest public knowledge graph vault.

In Knowledge graph, classes are interpreted as sets that contain individuals. They are described using formal (mathematical) descriptions that state precisely the requirements for membership of the class. Unlike traditional approaches where the focus is storage of

the data with less consideration of the timely interpretation or reasoning, the primary goal of the knowledge graph is to enable timely retrieval of the knowledge in this case the use by Doctors or Patients to retrieve time sensitive data. The key once again is the modular ability of the knowledge graph to extend and grow making it an ideal option to store drug adverse data and making it highly suitable for capturing drug side effects data due to the dynamic nature of the domain.

### **1.3 Problem statement**

Drug adverse reaction data contains important constraints about side-effects and conflict avoidance of component and compound drug. These are critically important in checking out prescriptions to avoid complications. Although MedWatch FAERS drug data are in XML, it doesn't have a proper knowledge representation mechanism to clearly specify all kinds of dependencies among the drug components and drugs. Therefore one has to depend on human interpretation to check prescriptions which can be error-prone. The newly introduced OWL based approach for medical drug data representation still suffers from several shortcomings inherent to the OWL restrictions like using "is-a" relationship and usage of object property based workarounds losing the clarity and dynamic relationship building expected by domain experts to represent knowledge. It's often difficult for the domain experts to process and derive meaningful information quickly for patients or doctors due to the following reasons.