School of Information Systems

Lipoprotein Ontology:
A Formal Representation of Lipoproteins

Meifania Monica Kohn

This thesis is presented for the Degree of

Doctor of Philosophy

of

Curtin University

July 2013

Declaration

To the best of my knowledge and belief, this thesis contains no material previously published by any other person except where due acknowledgement has been made.

This thesis contains no material which has been accepted for the award of any other degree or diploma in any university.

Meifania Monica Kohn

July 2013

Acknowledgements

First and foremost, I would like to thank my supervisor Dr Ponnie Clark. Without her constant encouragement and generous counsel, this thesis would not have come into fruition.

I would also like to extend my deepest gratitude to my co-supervisor Dr Michael Hecker for his immense support, valuable advice and utmost faith in me throughout my candidature. I am forever grateful to him for granting me the complete freedom in my scholarly pursuits, and yet at the same time challenging me in critical thinking and keeping me on the right track. He has done above and beyond his role by offering me the gift of comradeship, and I look forward to working alongside him as colleagues sometime in the future, given the opportunity.

I would like to thank my parents, Hendri and Evi Murtany, and brother, Reza Chen, for their unconditional love and support over the years. This thesis is dedicated to them; I would not be where I am without the sacrifices they've made.

Finally, I would like to thank my husband, Markus Kohn. He, who has been by my side every step of the way, cheering me on with the completion of each chapter, and believing in me even through the darkest period. For his unwavering love, kind understanding and boundless patience, I cannot be more grateful.

“The Semantic Web is not a separate Web but is an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation.”

Tim Berners-Lee, 2001

Table of Contents

List of Publications

Journals

Ooi EM, Watts GF, Chan DC, Chen MM, Nestel PJ, Sviridov D, Barrett PH. Dose-dependent effect of rosuvastatin on VLDL-apolipoprotein C-III kinetics in the metabolic syndrome. Diabetes Care. 2008 Aug; 31 (8): 1656-61.

Chan DC, Chen MM, Ooi EM, Watts GF. An ABC of apolipoprotein C-III: a clinically useful new cardiovascular risk factor? International journal of clinical practice 2008; 62 (5): 799-809.

Conferences

Chen MM, Hadzic M, "Towards a Methodology for Lipoprotein Ontology", in Proceedings of the 23rd IEEE international symposium on computer-based medical systems (CBMS 2010), Australia, 2010.

Chen MM, Hadzic M, "Lipoprotein ontology as a functional knowledge base", in Proceedings of the 22nd IEEE international symposium on computer-based medical systems (CBMS 2009), USA, 2009.

Hadzic M, Chen MM, Dillon TS, "Towards the Mental Health Ontology", in Proceedings of the IEEE international conference on bioinformatics and biomedicine (BIBM 2008), USA, pp. 284-288, 2008.

Hadzic M, Chen MM, Brouwer R, "A Modern Approach to Total Wellbeing", in Proceedings of the IEEE IT revolutions, Italy, 2008.

Book Chapters

Chen MM, Hadzic M, "A Framework for Lipoprotein Ontology: Towards Structured Representation of Lipoproteins", in Advances in Computational Biology, Springer, 2010.

List of Figures

List of Tables

List of Abbreviations

AHA	American Heart Association
Apo	Apolipoprotein
ATP III	Adult Treatment Panel III
BFO	Basic Formal Ontology
CETP	Cholesteryl ester transfer protein
DAML	DARPA Agent Markup Language
DL	Description Logics
ER	Entity-Relationship
FMA	Foundational Model of Anatomy
GO	Gene Ontology
HDL	High Density Lipoprotein
HL	Hepatic lipase
HMG-CoA	3-Hydroxy-3-methylglutaryl coenzyme A
ICD	International Classification of Diseases
IDF	International Diabetes Federation
IDL	Intermediate density lipoprotein
KR	Knowledge Representation
LCAT	Lecithin:cholesterol transferase
LDL	Low-density lipoprotein
LDL-R	Low-density lipoprotein receptor
Lp(a)	Lipoprotein(a)
LPL	Lipoprotein lipase
MeSH	Medical Subject Headings
NCEP	National Cholesterol Education Program
OBO	Open Biological and Biomedical Ontologies
OIL	Ontology Interchange Language
OWL	Web Ontology Language
RDF	Resource Descriptive Framework
RDFS	Resource Descriptive Framework Schema
SNOMED CT	Systematized Nomenclature Of Medicine Clinical Terms
UML	Unified Modelling Language
UMLS	Unified Medical Language System
WHO	World Health Organisation
XML	Extensible Markup Language

Summary of the Thesis

Lipoproteins play a crucial role in the regulation of biological and cellular functions in humans by serving as a mode of transport for the uptake, storage and metabolism of lipids. Correlation studies between various lipoproteins have indicated that the metabolism of plasma lipoproteins is complex and highly interrelated. Dysregulation in lipoprotein metabolism, also known as dyslipidaemia, has been found to be strongly correlated to various diseases such as cardiovascular disease, diabetes and hypertension, among others. Given the prominent role played by lipoproteins in these disease states, understanding lipoprotein management is a crucial treatment strategy.

The impact of dyslipidaemia on human health has led physiologists, biochemists, pharmacologists and clinicians to contribute a concerted effort in lipoprotein research. Although lipoprotein metabolism has been under intense investigation over the past 60 years, most of this information is contained in the form of natural language in various electronic journals and information sources. As the amount of experimental results, clinical data and scientific knowledge increases, there is a growing need to promote interoperability of these resources, support formal analyses, and utilise this knowledge that may lead to novel insights into physiological processes and hypothesis formulation. A particularly relevant end-result is translating research outputs from these endeavours into improvements in the treatments of dyslipidaemia.

The challenge is to represent and structure lipoprotein knowledge in an explicit way in order to produce a knowledge base that is amenable to scientific investigation and clinical application. Ontologies serve this purpose by providing a controlled vocabulary of well-defined concepts with specified relationships between them. They provide a mechanism for sharing a common vocabulary in a domain to support information exchange and are the basis for intelligent retrieval of information. This facilitates interoperability between research groups and information systems, which is essential in the lipoprotein area where there exists a multitude of research groups working disparately, in need of a common collaboration platform. Ontologies are becoming increasingly relevant in life sciences, as evident from the emergence of a number of biomedical ontologies. Some of these ontologies include the Gene Ontology, Protein Ontology, Lipid

Ontology, among many others. However, to our knowledge, an ontology specific to lipoproteins does not exist as yet.

In order to address the issues that have been outlined above, we have developed Lipoprotein Ontology, a structured formal framework which represents a concise and coherent picture of current, fundamental knowledge in the lipoprotein research domain. Lipoprotein Ontology consists of six sub-ontologies: *Classification*, *Metabolism*, *Pathophysiology*, *Aetiology*, *Treatment* and *Diagnostic Parameter*. Lipoprotein Ontology provides a formal representation for lipoprotein concepts and relationships that can be used to support the intelligent retrieval of information. Lipoprotein Ontology can also facilitate successful collaboration between research groups or software agents by providing a common platform of shared and agreed knowledge. Furthermore, Lipoprotein Ontology supports the inference of knowledge, which has tremendous potential in the development of tools for the diagnosis and treatment of dyslipidaemia.

Chapter 1. Introduction

1.1 Introduction

Lipids such as cholesterol, triglycerides and fatty acids are essential compounds for physiological function in humans. Due to their insolubility in aqueous medium, these fat-like substances form a complex with proteins (apolipoproteins) in order to be transported in blood through the circulation to where they are required. This lipid and protein complex is referred to as lipoproteins.

Dysregulation in lipoprotein metabolism, referred to as dyslipidaemia, has been found to be a key risk factor for cardiovascular disease, the leading cause of deaths worldwide (WHO, 2012). The prevalence of dyslipidaemia is increasing to epidemic proportions in industrialised nations as a consequence of sedentary lifestyles and high cholesterol/high-fat diet (WHO, 2012).

The impact of dyslipidaemia on human health has led physiologists, biochemists, pharmacologists and clinicians to address the complexities associated with the disruption of lipoprotein metabolism from different perspectives. As clinical data and experimental results from various research groups are constantly generated and added to the repository of literature, integrating all this information within a collective context proves to be a challenge.

This chapter begins with an overview of the lipoprotein research domain, and describes the lack of a formal framework for knowledge representation in the lipoprotein domain. We then discuss issues in the lipoprotein research domain which serve as motivation for the thesis, present the scope of the problem, followed by the research objectives. The chapter concludes with an outline of the thesis structure.

1.2 An Overview of Lipoprotein Research Domain

The lipoprotein transport system is critical for the supply, exchange and clearance of essential lipids in the body. Various lipoproteins, apolipoproteins, enzymes, transporters, and receptors in this system interact to maintain healthy body function. This delicate physiological balance can be impacted by a number of factors, including obesity, diet/nutrition, physical exercise and other factors such as smoking and alcohol consumption. Dysregulation in lipoprotein metabolism, known as dyslipidaemia, has been found to be significantly associated with various diseases such as cardiovascular disease, diabetes, and hypertension, among others. Treatment of dyslipidaemia is complex and varies from individual to individual according to the lipoprotein content. In order to present the lipoprotein domain in a coherent way, we have reviewed and organised lipoprotein knowledge from five following aspects: *Classification*, *Metabolism*, *Pathophysiology*, *Aetiology* and *Treatment*.

1.2.1 Classification

The basic lipoprotein structure comprises of a hydrophobic core of triglycerides and cholesteryl esters, surrounded by a hydrophilic outer layer of phospholipids, cholesterol and apolipoproteins. Lipoproteins are primarily classified into five major classes according to size and density, which are dependent on the amounts of lipid and protein components they contain (Alaupovic et al., 1971). Ranging from the largest to smallest in size, major lipoprotein fractions include chylomicrons, very low density lipoproteins (VLDL), intermediate density lipoproteins (IDL), low density lipoproteins (LDL) and high density lipoproteins (HDL) (Alaupovic et al., 1971).

1.2.2 Metabolism

Within the circulation, all lipoproteins are highly dynamic, changing in composition and structure as their lipid components and apolipoproteins are acquired and catabolised via various lipoprotein metabolism pathways. Eventually, lipoproteins are taken up and catabolised in the liver, kidney and peripheral tissues via receptor-mediated and other mechanisms. A clear review of lipoprotein metabolism in normal physiological condition is crucial to

understanding the concepts presented in this thesis as well as the changes in diseased states. For the purpose of this introduction, we briefly summarise the processes of lipoprotein physiology below, which will be further elucidated in Chapter 2.

There are three major pathways involved in lipoprotein metabolism, depending on whether the lipoproteins are composed of dietary lipids, whether they are synthesised in the liver or whether they involve the synthesis and transport of HDL.

The exogenous pathway involves the transport of dietary lipids (triglycerides and cholesterol) from the intestines to the liver and peripheral tissues. Following intestinal absorption, dietary triglycerides and cholesterol are combined with apoB-48, phospholipids and cholesterol to produce nascent chylomicrons. These chylomicron particles are then secreted via the lymphatic system into the circulation where they acquire apoC and apoE from other lipoproteins (Redgrave, 1983). Triglycerides are removed from the chylomicron particles by enzymatic action of the lipoprotein lipase (LPL), after which the chylomicron particles become smaller and are converted into chylomicron remnants, which are then removed from the circulation by hepatic apoE/apoB receptor (Redgrave, 1983).

In the endogenous pathway, lipids synthesised in the liver are transported from the liver to peripheral tissues. VLDL contains one molecule of apoB-100 and acquires cholesteryl esters, apoC and apoE from HDL. The lipolysis of VLDL is facilitated by apoC-II and inhibited by the apoC-III content of the lipoprotein particles (Shachter, 2001). The triglycerides in VLDL are lipolysed into free fatty acids by LPL as VLDL gets broken down to form IDL. The triglycerides in IDL are further lipolysed by hepatic lipase to form LDL or taken up by the liver via apoE-mediated binding with the LDL receptor (Chappell & Medh, 1998). LDL transport mainly cholesterol from the liver to cells of the body and are metabolised by apoB-100 mediated binding with the LDL receptor (Twisk et al., 2000).

The reverse cholesterol transport pathway involves the transport of excess cholesterol from the peripheral tissue to the liver via HDL. HDL is synthesised in

the liver and intestines and released into the blood-stream. HDL acts as cholesterol scavengers, acquiring free cholesterol from cell membranes and triglycerides from other lipoproteins and transporting excess cholesterol from the tissues back to the liver for excretion or re-utilisation (Sparks & Pritchard, 1989).

1.2.3 Pathophysiology

Dysregulation in lipoprotein metabolism, known as dyslipidaemia, has been identified to be an independent and major risk factor for cardiovascular disease (Lorenzo et al., 2007; Wilson et al., 2005). Dyslipidaemia is generally characterised by increased plasma triglycerides, increased levels of LDL cholesterol (commonly referred to as the "bad cholesterol"), and reduced HDL cholesterol (commonly referred to as the "good cholesterol") (Chan et al., 2004). Besides being a major risk factor for cardiovascular disease, dyslipidaemia has also been found to be present in other disease states such as the metabolic syndrome, diabetes and hypertension, among others. Although usually present concurrent with other symptoms such as obesity, insulin resistance, the importance of dyslipidaemia is that with proper management, it is an eminently remediable state. Major clinical trials have reported that manipulating lipoprotein levels such as lowering LDL cholesterol slowed the progression of atherosclerosis and decreased the incidence of cardiovascular events (Christakis et al., 1966). The reduction in the relative risk of cardiovascular events has been documented in patients with and without cardiovascular disease and also in patients with mild or severe dyslipidaemia. Thus, a better understanding of the pathophysiology of lipoprotein metabolism will help health care professionals to provide better care in the realm of dyslipidaemia management and improve clinical management strategies.

1.2.4 Aetiology of Dyslipidaemia

There are several contributing factors to dyslipidaemia. Lifestyle contributes to most cases of dyslipidaemia in adults. The most prominent cause in industrialised countries is a sedentary lifestyle with lack of physical exercise and excessive dietary intake of saturated fat and cholesterol. Other common causes include smoking, alcohol consumption and stress (butMarieb, 2000; Wright, 2002). Drug interactions may also contribute to lipoprotein dysregulation. Some

examples are corticosteroids (Fernández-Miranda et al., 1998; Hochberg & Petri, 1991) and beta-blockers (Feher et al., 1988). Genetic mutations may result in either the overproduction or defective clearance of LDL cholesterol and triglycerides, or the underproduction or excessive clearance of HDL (Clauss & Kwiterovich Jr., 2003).

1.2.5 Treatment of Dyslipidaemia

Treatment options to correct dyslipidaemia include lifestyle changes or pharmaceutical drugs such as statins, fibrates, bile acid sequestrants and cholesterol absorption inhibitors, or a combination of both. The first line approach is often lifestyle changes such as weight loss, exercise and dietary modification. Studies have found that the incidence of cardiovascular disease is lowered 25-60% when therapeutic agents which lower plasma lipids are administered to random, middle-aged men (Christakis et al., 1966). In light of this, tailoring the most appropriate dietary or drug therapy to individuals with specific lipoprotein profiles might be even more beneficial, especially when such treatment is administered earlier on. Combination therapy involves the use of dual or multiple lipid-regulating agents to treat lipoprotein abnormalities by targeting specific lipoproteins and utilising the complementary mechanisms of action of the different agents (Jacobson, 2001). This is a more effective treatment strategy where lipid-regulating monotherapy (e.g. statins or fibrates) may not provide adequate improvement in dyslipidaemia. However, there are some contraindications between different treatments. Although beneficial in correcting dyslipidaemia, the combinations of statins with fibrates or niacins have the potential for interactions that increase the risks of adverse effects, such as myositis and hepatotoxicity (Bays & Dujovne, 1998). Hence, treatment of lipoprotein dysregulation necessitates a thorough examination of lipoprotein profiles of specific individuals.

1.3 Lack of Formal Framework for Lipoprotein Knowledge

The lipoprotein research domain broadly covers the classification of lipoproteins, lipoprotein metabolism, lipoprotein dysregulation, causes as well as treatment of dyslipidaemia. Lipoproteins and their implication in health have been extensively documented and archived in PubMed over the past 60 years. As such, most of

this information is contained in the form of natural language in various electronic journals and information sources. This poses a challenge for researchers and health practitioners in retrieving and integrating relevant knowledge amongst massive quantities of heterogeneous information. In addition, as lipoprotein research rapidly becomes more advanced and complex with the discovery of new relevant components, associations and risk factors, it is becoming increasingly difficult even for lipoprotein experts to to assimilate and integrate the new information in the context of their existing knowledge.

The lack of formal framework for lipoprotein knowledge also undermines collaboration effort between different research groups. As lipoprotein research often involves researchers from various sub-disciplines of biomedical science such as physiology, biochemistry, pharmacology, etc., the lack of a common research platform may lead to duplication in research efforts, inconsistencies in research context, among others. As the biomedical community is moving towards a systems approach to solving research issues with practical implications and a strong emphasis towards improving health, having a formal framework for lipoproteins would provide a common platform for collaboration. In turn, this allows the effective integration of lipoprotein knowledge across different investigation avenues.

1.4 Motivation of Study

Several issues within the lipoprotein domain have been identified as the motivation for the thesis. These issues are discussed as follows.

1.4.1 Complex Metabolism Pathways

Correlation studies between various lipoproteins have indicated that the metabolism of plasma lipoproteins is complex and highly interrelated (Frayn, 2003) (Eisenberg, 1983). Understanding the impact of the dysregulation of lipoprotein metabolism necessitates knowledge of the synthesis, structure and metabolism of lipoproteins and their individual lipid components, including how they are incorporated, transported and trafficked within their respective lipoprotein classes. It is also important to note that these processes involve a

constant and dynamic flow and remodelling of particles where lipid molecules and apolipoproteins are gained and lost through highly complex pathways.

The complexity of lipoprotein metabolism pathway presents a need for a framework for lipoprotein knowledge representation. Modelling the relations between lipoprotein entities leads to a better understanding of the complex interrelationships between lipoprotein particles, receptors and enzymes necessary for functional lipid distribution in normal physiological conditions as well as in diseased states. Because the metabolism of the plasma lipoproteins is highly interrelated, one must consider each of the lipoproteins and their subclasses to optimise the complete lipid profile for different individuals.

1.4.2 Lipoprotein Dysregulation and its Implications on Disease

Lipoprotein dysregulation, known as dyslipidaemia, occurs as a consequence of alterations in the kinetics of lipoproteins. Dyslipidaemia has been found to be significantly associated with various diseases such as cardiovascular disease, diabetes and hypertension, among others. The key to alleviating this risk is a close examination lipoprotein metabolism and the changes in lipoprotein components associated with the various disease states.

1.4.3 Causes of Dyslipidaemia

As briefly outlined in Section 1.2.4, there are several contributing factors to dyslipidaemia. The challenge is to attempt to map these causes and link them to various lipoprotein disorders, in order to enable the extrapolation of the relations between these concepts into easily identifiable risk factors for quick identification.

1.4.4 Diagnosis and Treatment of Dyslipidaemia

Although several guidelines exist to define dyslipidaemia, diagnostic parameters vary among different individuals, according to gender and ethnicity. In addition, the diagnosis of dyslipidaemia also takes into consideration other factors such as blood pressure, waist circumference, etc. Thus, it can sometimes be

challenging for a health practitioner to interpret the correct diagnostic parameters for specific individuals and prescribe the appropriate treatment.

There are two issues with the treatment of dyslipidaemia which reinforce the need for a framework for lipoprotein knowledge representation.

Firstly, treatment of dyslipidaemia often involves the use of multiple lipid-regulating agents by targeting specific lipoproteins and utilising the complementary mechanisms of action of different agents. However, drug interactions may increase the risk to adverse effects and/or morbidity in certain individuals (Bays & Dujovne, 1998). These risks may be potentially reduced if these interactions can be mapped in a systematic way.

Secondly, studies on the interaction between the medical profession and the pharmaceutical industry suggest that health practitioners are affected by the influence of gifts and endorsements from the drug industry (Wazana, 2000). Bias was also found towards the sponsor's drugs in randomised clinical trials which undermine the results of research (Bero et al., 2007). These controversial issues present a need for an objective and unbiased representation of treatment pathways that are not tied to any sponsorship affiliation.

1.5 Scope of the Problem

Previously, we have discussed several challenges in the lipoprotein domain due to the lack of formal framework for lipoprotein knowledge. We attempt to address these challenges by developing a framework of lipoprotein concepts and their relationships to represent the lipoprotein research domain. The principal subject of interest is the lipoprotein research domain. One of the most significant issues identified in this thesis is to determine what it is that needs to be represented. Based on the challenges in the lipoprotein domain that have been identified, we will organise lipoprotein knowledge according to five key areas: *Classification*, *Metabolism*, *Pathophysiology*, *Aetiology* and *Treatment*. In order to represent these key areas, the research questions that need to be addressed are:

1. How can we model lipoprotein concepts and relationships such that they can be used as a representation of the lipoprotein research domain?

Specifically, how can we represent complex lipoprotein metabolism pathways in terms of concepts and relationships?

2. How can the model be used to aid knowledge integration and management in the lipoprotein domain? How can we use this framework to infer knowledge to aid diagnosis of lipoprotein dysregulation?
3. How can such a model facilitate collaboration between users? For instance, how can we develop this platform for easy transference of research output from different research groups?
4. How can we validate our framework such that it is consistent and easily extensible?

Due to the overwhelming amount of research available on lipoproteins, it is necessary to clarify at this stage that the focus of the lipoprotein knowledge representation framework is on breadth rather than depth; we limit ourselves to the most important lipoprotein concepts, especially focusing on lipoprotein metabolism pathways, the pathophysiology of lipoprotein dysregulation and its implication on health. Where discrepancy is found in the literature, the common approach is adopted and inconsistencies noted.

1.6 Research Objectives

The main objective of this thesis is to establish a framework for the concepts within the lipoprotein domain so that the heterogeneous and dispersed information in this field can be organised, related and formally represented in a meaningful way. In particular, the focus is to address specific challenges within the lipoprotein domain that has been outlined in Chapter 1.4. Therefore, the objectives of this thesis are aligned with the research questions identified in Chapter 1.5. This objective can be achieved through the following 4 sub-objectives:

- Objective 1. To develop an overall methodology framework towards the development of Lipoprotein Knowledge Representation. This is addressed in Chapter 5.

- Objective 2. To conceptualise lipoprotein concepts and structure the lipoprotein domain into six sub-domains as follows: describing the

classification of lipoproteins in healthy individuals (Classification), modelling the complex metabolism of lipoproteins through systematic relationships between lipoprotein concepts (Metabolism), transferring this approach to the classification of lipoproteins in patients with lipoprotein disorders (Pathophysiology), mapping the causes of dyslipidaemia (Aetiology), modelling treatment options for lipoprotein disorders (Treatment) and providing a consistent representation of diagnostic parameters for dyslipidaemia (Diagnostic Parameter). This is described in Chapter 6.

- Objective 3. To formalise the conceptualisation of lipoprotein knowledge in OWL representation language in the following sub-ontologies: *Classification*, *Metabolism*, *Pathophysiology*, *Aetiology*, *Treatment* and *Diagnostic Parameter*. This is detailed in Chapter 7.

- Objective 4. To evaluate Lipoprotein Ontology through the use of case studies, which have been developed to evaluate the robustness and consistency of the lipoprotein knowledge framework. These case studies will be discussed in Chapter 8.

1.7 Thesis Structure

Chapter 1. Introduction

In this chapter, we present an overview of the lipoprotein research domain and the lack of formal framework for lipoprotein knowledge representation. We then discuss the challenges in the lipoprotein domain, which serve as motivation for the thesis, present the scope of the problem as well as research objectives.

Chapter 2. Current State of Lipoprotein Research Domain

Chapter 2 provides a literature of the lipoprotein research domain. This section highlights the necessary lipoprotein concepts, and serves as the basis for developing the framework for lipoprotein knowledge representation. We then introduce some Knowledge Representation (KR) techniques in the biomedical domain.

Chapter 3. Problem Definition

The first section of Chapter 3 defines the main concepts in this thesis. We provide an overview of the problem, followed by the underlying research issues, from which we derive the key research questions. We then present our choice of research approach.

Chapter 4. Overview of the Solution

Chapter 4 presents an overview of the solution. We initially define ontology as a solution for the research issues identified in Chapter 3 and describe current applications of ontologies in the biomedical domain. We conclude this chapter by providing an overview of Lipoprotein Ontology and the components involved.

Chapter 5. Methodology

In Chapter 5, some of the commonly used ontology-building methodologies are reviewed. Based on the critical review of these methodologies, we develop our own methodology towards the development of Lipoprotein Ontology, and elaborate on the different stages of development in this section.

Chapter 6. Conceptual Framework of Lipoprotein Ontology

Chapter 6 outlines the conceptualisation and structure of Lipoprotein Ontology. Building on the literature review on lipoproteins presented in Chapter 2, we extract important concepts within the lipoprotein research domain arranged into the appropriate hierarchy or sub-ontologies.

Chapter 7. Formalisation of Lipoprotein Concepts

This chapter describes the formalisation of the lipoprotein concepts obtained in Chapter 6 in OWL by means of the ontology tool Protege 4.2. We present a number of visualisation figures to illustrate the different functionalities of Lipoprotein Ontology.

Chapter 8. Evaluation of Lipoprotein Ontology

Chapter 8 presents several case studies or scenarios which to evaluate and validate the correctness and completeness of Lipoprotein Ontology.

Chapter 9. Conclusion, Recapitulation and Future Work

Finally, Chapter 9 concludes this thesis. We present a summary of the thesis and review the objectives originally set out in the beginning of this research. We state the significance of this work and propose areas for future work.

1.8 Conclusion

In this chapter, we have provided an overview of the lipoprotein research domain. Lipoprotein knowledge is organised into the following categories: *Classification*, *Metabolism*, *Pathophysiology*, *Aetiology* and *Treatment*. We also discussed the need for a formal framework for the structured representation of lipoprotein knowledge and raised a number of challenges in the lipoprotein research domain. Subsequently, the scope of the problem was identified, followed by the research objectives, and an outline of the thesis structure.

References

Alaupovic, P., Lee, D. M., & McConathy, W. J. (1971). Studies on the composition and structure of plasma lipoproteins: Distribution of lipoprotein families in major density classes of normal human plasma lipoproteins *Biochimica et Biophysica 260*, 689-707.

Bays, H. E., & Dujovne, C. A. (1998). Drug interactions of lipid-altering drugs. *Drug Safety, 19*(5), 355–371.

Bero, L., Oostvogel, F., Bacchetti, P., & Lee, K. (2007). Factors associated with findings of published trials of drug–drug comparisons: why some statins appear more efficacious than others. *PLoS Med, 4*(6), e184.

Chappell, D. A., & Medh, J. D. (1998). Receptor-mediated mechanisms of lipoprotein remnant catabolism. *Progress in Lipid Research, 37*(6), 393–422.

Chan, D. C., Barrett, P. H. R., & Watts, G. F. (2004). Lipoprotein Kinetics in the Metabolic Syndrome: Pathophysiological and Therapeutic Lessons from Stable Isotope Studies. *Clinical Biochemistry Review, 25*, 31-48.

Christakis, G., Rinzler, S., Archer, M., et al. (1966). The anti-coronary club: A dietary approach to the prevention of coronary heart disease - a seven-year report. *American Journal of Public Health, 56*(2), 299–314.

Clauss, S. B., & Kwiterovich Jr., P. O. (2003). Genetic disorders of lipoprotein transport in children. *Progress in Pediatric Cardiology, 17*(2), 123-133.

Eisenberg, S. (1983). Lipoproteins and lipoprotein metabolism - a dynamic evaluation of the plasma fat transport system. *Journal of Molecular Medicine, 61*(3), 119-132.

Feher, M. D., Rains, S. G., Richmond, W., et al. (1988). Beta-blockers, lipoproteins and non-insulin dependent diabetes.

Fernández-Miranda, C., Guijarro, C., de la Calle, A., et al. (1998). Lipid abnormalities in stable liver transplant recipients-effects of cyclosporin, tacrolimus, and steroids. *Transplant International, 11*(2), 137-142.

Frayn, K. N. (2003). *Metabolic regulation: a human perspective*: Blackwell Publishing.

Hochberg, M. C., & Petri, M. (1991). The association of corticosteroid (CS) therapy with coronary heart disease (CHD) in patients with systemic lupus erythematosus (SLE): A meta-analysis. *Arthritis & Rheumatism, 34*, R24.

Jacobson, T. A. (2001). Combination lipid-altering therapy: an emerging treatment paradigm for 21st century. *Current Atherosclerosis Reports, 3*, 373-382.

Lorenzo, C., Williams, K., Hunt, K. J., et al. (2007). The National Cholesterol Education Program-Adult Treatment Panel III, International Diabetes Federation, and World Health Organization definitions of the metabolic syndrome as predictors of incident cardiovascular disease and diabetes. *Diabetes Care, 30*, 8-13.

Marieb, E. N. (2000). Nutrition, Metabolism and Body Temperature Regulation *Human Anatomy & Physiology*. USA: Benjamin Cummings.

Redgrave, T. G. (1983). Formation and metabolism of chylomicrons. *International Review of Physiology 28*, 103-130.

Shachter, N. S. (2001). Apolipoproteins C-I and C-III as important modulators of lipoprotein metabolism. *Current Opinion in Lipidology, 12*, 297-304.

Sparks, D. L., & Pritchard, P. H. (1989). Transfer of cholesteryl ester into high density lipoprotein by cholesteryl ester transfer protein: effect of HDL lipid and apoprotein content. *Journal of Lipid Research, 30*, 1491-1498.

Twisk, J., Gillian-Daniel, D. L., Tebon, A.,et al. (2000). The role of the LDL receptor in apolipoprotein B secretion. *Journal of Clinical Investigation, 105*(4), 521–532.

Wazana, A. (2000). Physicians and the pharmaceutical industry. Is a gift ever just a gift? *JAMA, 283*(3), 373-380.

WHO, World Health Organization. (2012). World Health Statistics 2012. France.

Wright, H. (2002). *A more excellent way, 5th Ed.*: Pleasant Valley Publications.

Chapter 2. Current State of Lipoprotein Research Domain

2.1 Introduction

This chapter begins with an introduction to lipoproteins, in terms of their structure, components and metabolism. Subsequently, the implications of lipoprotein dysregulation, also known as dyslipidaemia, on health are presented, followed by challenges in the diagnosis and treatment of dyslipidaemia. We then raise the issue of the nature of biomedical information and discuss KR techniques in the management of biomedical data.

2.2 Lipoproteins: Structure and Metabolism

Lipoproteins are soluble complexes of lipids and proteins, which serve as a mode of transport for the uptake, metabolism and storage of lipids in humans. Within the circulation, lipoproteins are in a state of constant dynamic flux. They undergo enzymatic reactions of their lipid components, facilitated and spontaneous lipid exchange, transfers of soluble apolipoproteins, as well as conformational changes of apolipoproteins in response to compositional changes. Eventually, lipoproteins are taken up and catabolised in the liver, kidney and peripheral tissues via receptor-mediated pathways and other mechanisms.

In order to gain a better understanding of lipoproteins and how they are implicated in human health, it is necessary to provide a clear outline of the lipoprotein structure and metabolism pathways. Due to the sheer amount of knowledge available on lipoprotein and lipid research, this literature review does not claim to be exhaustive of all lipoprotein knowledge. It does, however, attempt to provide a general overview of the lipoprotein research domain and highlights

the necessary concepts for modelling the domain and capture an accurate representation of lipoprotein knowledge.

2.2.1 Classification of Lipoprotein Entities

The basic lipoprotein structure is spherical and comprises of a hydrophobic core of triglycerides and cholesteryl esters, surrounded by a hydrophilic outer layer of phospholipids, cholesterol and apolipoproteins (Figure 1).

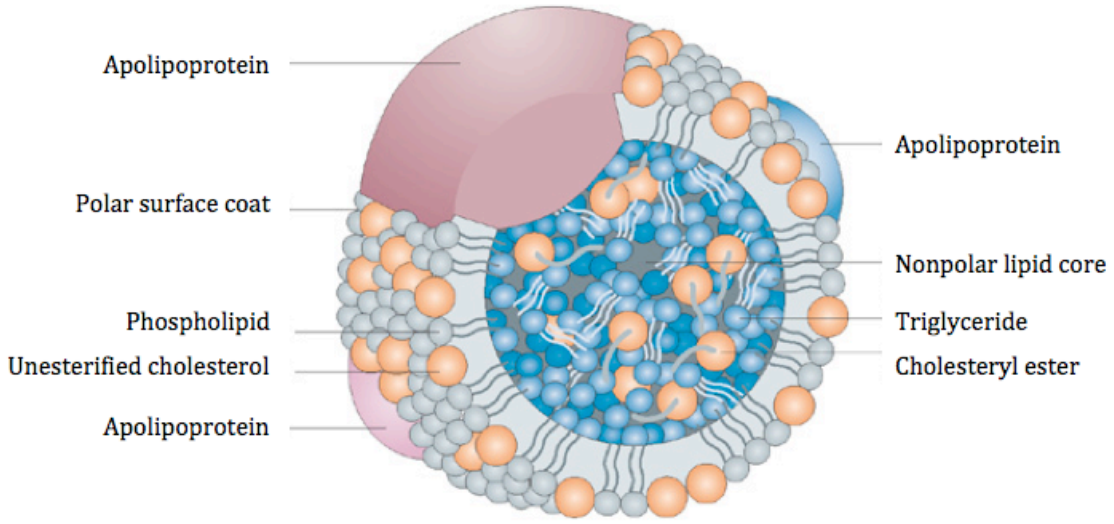


Figure 1. Basic lipoprotein structure (Wasan et al., 2008). The hydrophobic core is primarily composed of triglycerides and cholesterol esters. They are encased by a hydrophilic phospholipid monolayer. Apolipoproteins embedded in the phospholipid layer confer structural and functional properties to the molecule.

Lipoproteins are primarily classified into classes according to size, as well as their relative contents of protein and lipid that determine the densities and functions of these lipoprotein classes. Ranging from the largest to smallest in size, major lipoprotein fractions include chylomicrons, very low density lipoproteins (VLDL), intermediate density lipoproteins (IDL), low density lipoproteins (LDL) and high density lipoproteins (HDL). Chylomicrons contain approximately 1% protein while the protein content of HDL is around 50%. The total lipid content is inversely correlated with the density of the lipoproteins, chylomicrons being the least dense and HDL the most dense. Lipoprotein(a) (Lp(a)) is similar in content to LDL, except it has apolipoprotein(a) attached to it.

The lipid content and composition of the major lipoprotein classes are listed in Table 1. An enlarged version of this table is shown in Appendix A. While these values serve to categorise the different classes of lipoprotein fractions, it is important to note that the ratios of the lipids and protein components vary even within a given subclass of lipoproteins.

Table 1. Major lipoprotein classes: A comparison of density, size, electrophoretic mobility, percentage of total mass that is triglyceride (TG), cholesterol (C), cholesteryl ester (CE), phospholipid (Ph), protein (P) and apolipoproteins (apo) content.

Lipoprotein	Density (g/ml)	Size (nm)	Electrophoretic mobility	TG (%)	C (%)	CE (%)	Ph (%)	P (%)	Major apo	Other apo	Other constituents
Chylomicrons	<0.930	75-1200	Origin	80-95	1-3	2-4	3-6	1-2	B-48	A-I, A-II, A-IV, A-V, C-I, C-II, C-III, E	Vitamin A
Chylomicron remnants	0.930-1.006	30-80	Slow pre-β						B-48	E	Vitamin A
VLDL	0.930-1.006	30-80	Pre-β	45-65	4-8	16-22	15-20	6-10	B-100	A-I, A-II, A-IV, A-V, C-I, C-II, C-III, D, E	Vitamin A
IDL	1.006-1.019	25-35	Slow pre-β						B-100	E, C-I, C-II, C-III	Vitamin E
LDL	1.019-1.063	18-25	β	4-8	6-8	45-50	18-24	18-22	B-100	E, C-I, C-III, A-I	Vitamin E
HDL	1.063-1.210	5-12	α	2-7	3-5	15-20	26-32	45-55	A-I	A-II, A-IV, A-V, C-I, C-II, C-III, D, E	LCAT, CETP
Lp(a)	1.050-1.120	25	Pre-β						B-100	(a)	

Abbreviations: VLDL, very low density lipoprotein; IDL, intermediate density lipoprotein; LDL, low density lipoprotein; HDL, high density lipoprotein; Lp(a), lipoprotein (a).

The functions of the different lipoprotein classes are determined by their lipid and apolipoprotein components. Chylomicrons are synthesised in the intestines for the transport of dietary triglycerides to various tissues. VLDL are synthesised in the liver for the export of endogenous triglycerides, while IDL and LDL are derived from the metabolism of VLDL in circulation. The function of LDL is to deliver cholesteryl esters to the liver and peripheral tissues. HDL are synthesised and assembled in the liver and intestines or formed from the metabolism of other lipoproteins in circulation, and from cellular lipids at the cell membranes. HDL remove excess cholesterol from cells and transport it to liver and other tissues for metabolism and excretion (Chan et al., 2004b).

Because the lipoprotein divisions by density are arbitrary and vary somewhat within their own classes, they can also be classified according to their

electrophoretic mobility on agarose gels into α, pre-β and β lipoproteins, corresponding to HDL, VLDL and LDL lipoprotein fractions respectively. Chylomicrons remain at the electrophoretic origin (Snyder, 1977).

2.2.2 Lipoprotein Metabolism

Lipoprotein metabolism involves a constant, dynamic flux and remodelling of particles where lipid molecules and apolipoproteins are gained and lost through complex pathways involving the catabolism and exchange between particles. It is therefore essential to review individual lipoproteins in terms of structure (in terms of lipid composition and associated proteins), synthesis, catabolism, kinetics (how they interact with each other) and ability to be influenced by other metabolites (such as glucose, reactive oxygen species, etc.). This section introduces the key components of lipoproteins which are involved in the lipoprotein metabolism pathways, and subsequently describes them in the context of the various pathways they participate in. A summary of lipoprotein metabolism processes, along with the corresponding location and major components involved, is presented in Table 2 below. An enlarged version of this table is shown in Appendix B.

Table 2. Summary of lipoprotein metabolism processes, location and major components.

Lipoprotein	Location	Process	Major components
Chylomicron	Intestine	Assembly, secretion	Apo A, apo B, phospholipid, cholesterol, cholesteryl ester, triglyceride
	Lymph	Transfer	Apo C, apo E
	Plasma	Transfer	Apo C, cholesterol
	Endothelial cell	Hydrolysis	Triglyceride, phospholipid
		Transfer	Fatty acid, cholesterol, phospholipid, apo C, apo A, apo E
VLDL	Liver	Assembly, secretion	Apo B, apo C, phospholipid, cholesterol, cholesteryl ester, triglyceride
	Plasma	Transfer	Apo C, apo E, cholesteryl ester
	Endothelial cell	Hydrolysis	Triglyceride, phospholipid
		Transfer	Fatty acid, cholesterol, phospholipid, apo C, apo E
LDL	Plasma	Formation	From VLDL and chylomicrons
		Exchange	Cholesterol, phospholipid
	Peripheral tissues, liver	Uptake, catabolism, regulation	Cholesterol, cholesteryl ester
HDL	Liver	Assembly, secretion	Apo A, apo E, phospholipid, cholesterol
	Plasma	Formation	From chylomicrons
		Acyltransfer	Phosphatidylcholine, cholesterol, cholesteryl ester, apo C, apo E
		Transfer	Cholesteryl ester, apo C, apo E
		Exchange	Apo C, cholesterol, phospholipid
	Liver, peripheral tissues	Uptake, catabolism	Cholesterol, cholesteryl ester

Abbreviations: VLDL, very low density lipoprotein; IDL, intermediate density lipoprotein; LDL, low density lipoprotein; HDL, high density lipoprotein; apo, apolipoprotein.

A better understanding of lipoprotein metabolism and its pathophysiology is crucial to management of lipoprotein disorders. As plasma LDL concentration is positively correlated with the risk to atherosclerosis, it is important to look for the factors which cause the high concentrations of LDL particles. Such factors may include increased production of LDL or their slow removal from circulation. Lipoprotein metabolism is a highly complex and interrelated process. For instance, plasma LDL levels are determined not only by LDL receptors but also by the rate of VLDL synthesis, the activity of lipoprotein lipase, as well as other catabolic processes.

There are three major pathways involved in human lipoprotein metabolism in normal physiological conditions, illustrated and described briefly in Figure 2:

1. Exogenous pathway, the transport and metabolism of dietary lipids
2. Endogenous pathway, the transport of lipids synthesised in the liver
3. Reverse cholesterol transport, the synthesis and metabolism of HDL

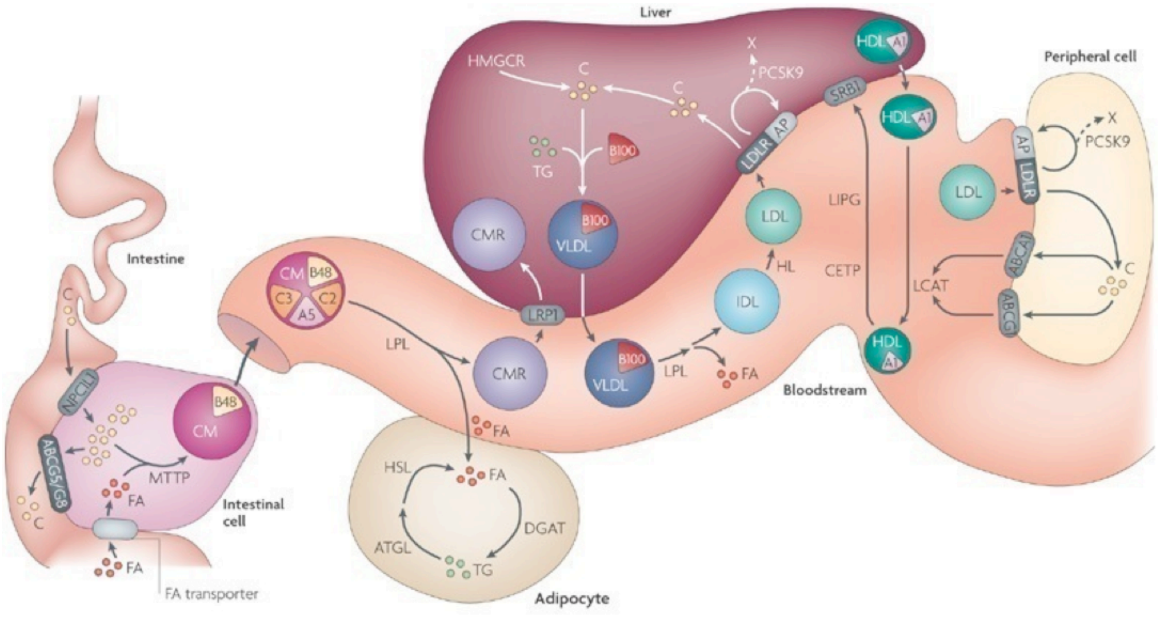


Figure 2. Overview of lipoprotein metabolism (Hegele, 2009). The three lipoprotein metabolism pathways are shown. The main lipids in lipoproteins are free cholesterol (C), cholesteryl ester (CE) and triglyceride (TG). In the *exogenous transport pathway,* hydrolysed dietary fats enter intestinal cells or enterocytes via fatty acid (FA) transporters TG is packaged with CE and the apolipoprotein (apo) B-48 (B48) into chylomicrons (CM) by microsomal TG-transfer protein (MTTP) through a vesicular pathway. CM, secreted via the lymphatic system, enter the vena

cava and circulate until they interact with lipoprotein lipase (LPL). CM contain apo, including apoA-I (not shown), apo A-II (not shown), apo A-IV (not shown), apo A-V (A5), Apo C-II (C2), Apo C-III (C3) and apo E (not shown). Some of the released free FA enter peripheral cells. In adipocytes, enzymes including acyl CoA:diacylglycerol acyltransferase (DGAT) resynthesise TG, which is hydrolysed by adipose TG lipase (ATGL) and hormone sensitive lipase (HSL). CM remnants (CMR) are taken up by hepatic LDL receptor (LDLR), in the absence of LDLR they are taken up by LDLR-related protein-1 (LRP1). In the *endogenous transport pathway*, TG is packaged with C and the apo B-100 (B100) into very low-density lipoprotein (VLDL) in the liver cells (hepatocytes); the TG contained in VLDL is hydrolysed by LPL, releasing FA and VLDL remnants (IDL) that are hydrolysed by hepatic lipase (HL), thereby yielding LDL. In LDL cholesterol metabolism, sterols in the intestinal lumen enter enterocytes via the Niemann-Pick C1-like 1 (NPC1L1) transporter and some are resecreted by heterodimeric ATP-binding cassette transporter G5/G8 (ABCG5/G8). In enterocytes, cholesterol is packaged with TG into CM. In hepatocytes, C is recycled or synthesized de novo, with 3-hydroxy-3-methylglutaryl coenzyme A reductase (HMGCR) being rate-limiting. LDL transports C from the liver to the peripheral tissues. LDL is endocytosed by peripheral cells and hepatocytes by LDLR, assisted by an adaptor protein (AP). Proprotein convertase subtilisin/kexin type 9 (PCSK9), when complexed to LDLR, short-circuits recycling of LDLR from the endosome, leading to its degradation (X). In the *reverse cholesterol transport pathway*, HDL, via APOA-I (A1), mediates reverse cholesterol transport by interacting with ATP-binding cassette A1 (ABCA1) and ABCG1 transporters on non-hepatic cells. Lecithincholesterol acyltransferase (LCAT) esterifies C so it can be used in HDL, which, after remodelling by cholesterol ester transfer protein (CETP) and by endothelial lipase (LIPG), enters hepatocytes via scavenger receptor class B type I (SR-B1).

In the following sections, we discuss the three lipoprotein metabolism pathways in further detail, with important concepts highlighted in bold (lipoprotein classes), italics (lipoprotein components) and underlined (processes) respectively.

2.2.2.1 Exogenous transport pathway

The exogenous pathway involves the transport of dietary lipids (*triglycerides* and *cholesterol*) from the intestines to the liver and peripheral tissues. After a meal, dietary *triglycerides* are <u>hydrolysed</u> to *monoglycerides* and *free fatty acids* which diffuse across the microvillus membrane across the intestines. *Triglycerides* are subsequently <u>resynthesised</u> in the enterocytes, combined with *cholesterol*, *apo B-48*, *apo A-I*, *apo A-II*, *apo A-IV* and *phospholipids* to produce nascent **chylomicrons**. These **chylomicron** particles are secreted via the lymphatic system into the circulation where they acquire *apoC-II* and *apoE* from plasma **HDL** (Redgrave, 2004). *Apo C-II* activates *lipoprotein lipase (LPL)* located on the surface of the endothelial cells in adipose and muscle tissues, where

triglycerides in **chylomicrons** are hydrolysed into *free fatty acids* and *glycerol* by the action of *LPL*. Liberated *glycerol* is returned to the liver and kidneys, where it is almost exclusively used to produce glycerol-3-phosphate, which can enter glycolysis or gluconeogenesis. The *free fatty acids* are transported into the adipose or muscle cells for storage or energy, or can be transported on albumin to other parts of the body. Adipocytes can also re-esterify *free fatty acids* to form *triglycerides*. As the *triglycerides* are hydrolysed into *free fatty acids*, **chylomicrons** progressively shrink to form **chylomicron remnants**. Throughout the catabolic process, a substantial portion of *phospholipid*, *apo A* and *apo C-II* is transferred to **HDL**. The loss of *apo C-II* prevents *LPL* from further degrading the chylomicron remnants. **Chylomicron remnants** contain primarily *cholesteryl esters*, *apo E* and *apo B-48*, and are transported through the endothelial cells lining the hepatic sinusoids, into the space of Disse. Here, **chylomicron remnants** are taken up by *LDL receptor* located on hepatocytes in the liver upon activation by *apo E*. While in the space of Disse, **chylomicron remnants** accumulate additional *apo E* that is secreted into the space and are taken up via the *chylomicron remnant receptors* which also require activation by *apo E*. In addition, chylomicron remnants can also be taken up via the hepatic *apo E/apo B-48 receptor* (Redgrave, 2004).

2.2.2.2 Endogenous transport pathway

In the endogenous pathway, hepatically synthesised lipids are transported from the liver to peripheral tissues (Mahley et al., 1984). Secreted via exocytosis from the liver, **VLDL** carry *triglycerides*, one molecule of *apo B-100*, and acquire *cholesteryl esters*, *apo C* and *apo E* from **HDL**. The hydrolysis of **VLDL** is facilitated by *apo C-II* and inhibited by the *apoC-III* content of the lipoprotein particles (Shachter, 2001). The *triglycerides* in **VLDL** are hydrolysed into *glycerol* and *free fatty acids* by *LPL*. *Free fatty acids* are transported into the adipose cells to be resynthesised into *triglycerides* and stored, or oxidised in the muscle tissues for energy. As **VLDL** get catabolised and lose most of their *triglycerides* to peripheral tissues, they transfer a substantial portion of *phospholipids* and *apo C* to **HDL**. Subsequently, some **VLDL** are returned to the liver, endocytosed and catabolised through *apo E* interaction with the *remnant receptor*, while others remain in the circulation as **IDL**, also known as **VLDL remnants**. At this point, the predominant remaining apolipoproteins associated

with **IDL** are *apoB-100* and *apo E.* The triglycerides in **IDL** are further hydrolysed by *hepatic lipase (HL)* and lose *apo E* to form **LDL**, or taken up by the liver via *apo B-100* and *apo E*-mediated binding with the *LDL receptor* (Chappell & Medh, 1998). **LDL** contain almost exclusively *apo B-100* and carries the majority of the *cholesterol* in the blood, amounting to 60 to 80 % of total *cholesterol*. Excess *cholesterol* may be deposited in artery walls, increasing the risk to atherosclerosis. After a plasma half-life of about 2 days, **LDL** bind to *LDL receptors* located on the plasma membrane of target cells coated with clathrin proteins (Brown & Goldstein, 1986). The uptake of **LDL** to *LDL receptors* is facilitated by *apo B-100* and occurs predominantly in the liver (75%), adrenals and adipose tissue. Once **LDL** binds to the *LDL receptor*, *clathrin* promotes endocytosis and rapidly dissociates the endosomal vesicle. **LDL** then dissociate from the receptors, and the latter are recycled to the cell surface. Upon fusing with a *lysosome*, **LDL** is then catabolised into its primary components – their *apolipoproteins* are degraded, *cholesterol esters* are hydrolysed to yield free *cholesterol*, and *cholesterol* is incorporated into plasma membranes as necessary, recycled into newly secreted lipoprotein particles, or degraded into bile acids. Excess *cholesterol* is re-esterified by *acyl-CoA-cholesterol acyltransferase (ACAT)*, for intracellular storage (Brown & Goldstein, 1986).

2.2.2.3 Reverse cholesterol transport

HDL-dependent lipid transport is often referred to as reverse cholesterol transport (Fielding & Fielding, 1995). It has been shown that **HDL** removes excess *cholesterol* within the arteries through this pathway; high levels of **HDL** are associated with decreased risk of heart disease (Ghali & Rodondi, 2009). **HDL** is synthesised in the liver and intestines and released into the blood-stream. Nascent **HDL** are disk-shaped but they become spherical as they acquire free *cholesterol* from cell membranes and *triglycerides* from other lipoproteins. **HDL** acts as cholesterol scavengers, transporting *cholesterol* from the tissues back to the liver for excretion or re-utilisation. The transfer of *cholesterol* from the cell membranes to **HDL** is mediated by *ATP-binding cassette protein G1 (ABCG1)*, and free *cholesterol* in **HDL** is esterified into *cholesteryl esters* by *lecithin:cholesterol acyltransferase (LCAT)* upon activation by *apo A-I*. Cholesterol-rich **HDL** return to the liver, where they bind to the *scavenger receptor class B type 1 (SR-B1)*. When **HDL** binds to *SR-B1*, the

cholesteryl esters of **HDL** are taken up by the hepatocytes through the caveolae while the **HDL** and *SR-B1* remain on the plasma membrane. *Cholesteryl esters* can also be transferred from **HDL** to **apoB-containing lipoproteins (VLDL, IDL and LDL)** by *cholesterol ester transfer protein (CETP)* in exchange for *triglycerides* (Tall et al., 1986). This process transforms **VLDL** and **IDL** into **LDL**. Thus, **HDL** has an important role in reverse cholesterol transport: firstly, specific **HDL** subclasses (preβ-HDL) function as primary cholesterol acceptors and are able to remove *cholesterol* from peripheral cells. In addition, after cholesterol esterification, *cholesterol esters* are transferred from **HDL** to **apoB-containing lipoproteins** in exchange for *triglycerides*, which can be removed from the circulation by *hepatic lipase*. As **HDL** grow in size, they acquire *apo E* which increases the binding affinity of **HDL** towards *hepatic receptors*, and are subsequently catabolised by the liver.

2.3 Impact of Dyslipidaemia on Health

Clinical and epidemiological studies have identified dyslipidaemia, to be a major and independent risk factor for cardiovascular disease (Isomaa et al., 2001). Dyslipidaemia is the term given to the dysregulation in lipoprotein metabolism; this dysregulation can be caused by various factors, which lead to increased plasma triglycerides, increased levels of LDL cholesterol (commonly referred to as the “bad cholesterol”), and reduced HDL cholesterol (commonly referred to as the “good cholesterol”) (Chan et al., 2004a; Chan et al., 2004b).

Dyslipidaemia has also been found to be present in other disease states such as the metabolic syndrome, diabetes and hypertension, among others (Alexander et al., 2003; Ninomiya et al., 2004). While the progression of each disease state is complex and dependent on many factors, managing abnormal lipoprotein levels is one of the key clinical targets in treating the associated diseases. Studies have found that the incidence of cardiovascular disease is lowered 25-60% when therapeutic agents which lower plasma lipids are administered to dyslipidaemic patients (Christakis et al., 1966). Studies have also found that with proper management, dyslipidaemia is an eminently remediable state. In light of this, tailoring the most appropriate lifestyle adjustment or drug therapy to individuals with specific lipoprotein profiles might be even more beneficial, especially when such treatment is administered earlier on.

In this section we focus on how changes in lipoprotein levels and their associated components alleviate the risk to developing certain disorders.

2.3.1 Key Components in Dyslipidaemia

Prior to describing the pathophysiology of lipoprotein dysregulation, it is essential to introduce the clinically relevant components in dyslipidaemia. To recapitulate the previous section, the basic lipoprotein structure comprises of a hydrophobic core of triglycerides and cholesteryl esters, surrounded by a hydrophilic outer layer of phospholipids, cholesterol and apolipoproteins. We have also mentioned previously that lipoprotein metabolism involve highly complex and dynamic processes which include the transfer, exchange, hydrolysis and catabolism of various lipid and apolipoprotein components between lipoprotein particles through the action of various receptors and enzymes. Even though each and every component is essential for optimal lipid distribution, in this section we highlight the key players involved in lipoprotein dysregulation.

2.3.1.1 Cholesterol

Cholesterol is one of the essential lipid components transported by lipoproteins, which has several critical physiological functions. It is an essential structural component of cell membranes required to establish proper membrane permeability and fluidity over a range of physiological temperatures (Brown & Goldstein, 1986). Cholesterol is also required in intracellular transport, cell signalling and nerve conduction (Brown & Goldstein, 1986). In addition, cholesterol is the precursor molecule in several biochemical pathways, and is implicated in the manufacture of vitamin D, bile acids, steroid hormones including cortisol and aldosterone, as well as sex hormones progesterone, oestrogen and testosterone, and their derivatives (Brown & Goldstein, 1986). Some research indicates cholesterol may also have antioxidant properties (Bandeali & Farmer, 2012).

Although cholesterol is important in human health, high plasma cholesterol has been found to be significantly associated with cardiovascular disease. It is important to note however, that it is not the total plasma cholesterol level that is

clinically relevant to developing the risk to cardiovascular disease, but rather the amount of cholesterol bound to the various classes of lipoproteins. We briefly clarify the distinction below.

LDL

Evidence suggests that the propensity towards developing cardiovascular disease substantially increases with elevated LDL cholesterol. Excessive LDLs lead to potentially lethal cholesterol deposits on the artery walls which increase the risk to atherosclerosis (Kannel et al., 1979).

Lp(a)

A positive correlation has been found between high plasma concentration of Lp(a) and the incidence of cardiovascular disease. Studies have found that elevated Lp(a) in plasma can double a man's risk to cardiovascular events before the age of 55 (Marieb, 2000). Lp(a) appears to promote plaque formation that stiffens the artery walls.

HDL

The risk to cardiovascular disease is inversely related to the concentration of HDL in blood; elevated levels of HDL have been found to be associated with a low incidence of cardiovascular disease (Natarajan et al., 2010). Further research into the role of HDL has shown that HDL containing apo A-I confers protective role in cardiovascular disease (Fruchart et al., 1994). In contrast, HDL containing apo A-II has been found to increase the risk of cardiovascular disease by promoting plaque formation (Fruchart et al., 1994).

2.3.1.3 Triglycerides

Triglycerides, the major lipid components of chylomicrons and VLDL, are the major type of dietary fat as well as the storage form of fat in the body. Research has found that high plasma triglycerides represent a high risk factor for cardiovascular disease in both men and women. Furthermore, patients with cardiovascular disease and elevated triglycerides have a higher risk of

premature death than patients with cardiovascular disease and normal triglyceride levels. Studies also suggest that high plasma triglyceride levels are associated with low HDL levels and increased amounts of small dense LDL, perhaps all resulting from abnormal breakdown of chylomicrons and VLDL.

2.3.1.4 Apolipoprotein

Apolipoproteins are lipid-binding proteins which serve as enzyme modulators and receptor ligands that regulate the intravascular metabolism of lipoproteins and their tissue uptake. Apolipoproteins have the ability to change conformation to adjust to changing lipid contents, compositions and metabolic states of the lipoproteins. They also readily disassociate themselves from one lipoprotein and bind to another while in circulation; only apolipoprotein Bs, the principal protein components of LDL, VLDL and chylomicrons, maintain their association with the respective lipoprotein during the cycle of lipoprotein synthesis and metabolism. Table 3 lists the major apolipoproteins, their associated lipoprotein classes, molecular weight, concentration in plasma, synthesis site and function. An enlarged version of this table is shown in Appendix C.

Because apolipoproteins interact directly with various lipoprotein receptors and lipases, changes in apolipoprotein levels may therefore potentially have negative effects on lipoprotein metabolism.

Table 3. Major human apolipoproteins[a], associated lipoprotein classes, molecular weight, concentration in plasma, synthesis site and function.

Apolipoprotein[a]	Lipoprotein class[b]	Molecular weight (Da)	Concentration in plasma (mg/dl)	Synthesis site and function
Apo A-I	**HDL**, *chylomicron*	28,100	130	Secreted by the intestine, binds ABCA1 on macrophages, critical antioxidant protein of HDL, activates LCAT
Apo A-II	**HDL,** *chylomicron*	17,400	40	Synthesised by the liver, inhibits HL
Apo A-IV	**Chylomicron,** *VLDL, HDL*	44,500	15	Synthesised exclusively by the small intestine and the hypothalamus, inhibits food intake, activates LCAT*
Apo B-48	**Chylomicron**	512,000	250	Exclusively found in chylomicrons, secreted by the intestine and liver, derived from apoB-100 gene by RNA truncation
Apo B-100	**VLDL, IDL, LDL**	242,000		Major protein of LDL, secreted exclusively by the liver, binds to LDLr
Apo C-I	**VLDL**, **Chylomicron**, *HDL*	6,600	3	Synthesised mainly by the liver and the brain*, inhibits receptor-mediated uptake of TRL, inhibits CEPT, inhibits LPL-mediated hydrolysis of TG, activates LCAT*
Apo C-II	**VLDL**, **Chylomicron**, *HDL*	8,800	12	Synthesised mainly by the liver, activates LPL
Apo C-III	**VLDL**, **Chylomicron**, *HDL*	9,000	12	Synthesised mainly by the liver and the intestine*, normally binds to HDL, but binds to TRL in hypertriglyceridaemic patients, inhibits LPL, inhibits HL, inhibits apoB and apoE mediated binding of lipoproteins to LDLr, inhibits TRL binding to LSR, inhibits LCAT, stimulates CETP
Apo D	**HDL**	22,000	12	Expressed widely in human tissue, closely associated with LCAT
Apo E	**VLDL**, **HDL,** **Chylomicron**, *IDL*	34,200	7	Synthesised mainly by the liver and the brain, binds to LDL receptor, activates LCAT*
Apo (a)	**Lp(a)**	300-000-800,000	0.1-40	Disulfide bonded to apoB-100, forms a complex with LDL identified as Lp(a); strongly resembles plasminogen; may deliver cholesterol to sites of vascular injury, high risk association with premature CVD

Abbreviations: apo, apolipoprotein, VLDL, very low density lipoprotein; IDL, intermediate density lipoprotein; LDL, low density lipoprotein; HDL, high density lipoprotein; Lp(a), lipoprotein (a); ABCA1, ATP-binding cassete transporter; TG, triglycerides; TRL, triglyceride-rich lipoprotein; LPL, lipoprotein lipase; HL, hepatic lipase; LDLr, LDL receptor; LCAT, lecithin-cholesterol acyltransferase; CETP, cholesteryl ester transfer protein; LSR, lipolysis-stimulated lipoprotein receptor.

[a] Other minor apolipoproteins isolated from lipoprotein fractions include apo F, apo H, apo J, apo L, and apo M. They are present in plasma in low concentrations and do not have well-defined functions in lipoprotein metabolism

[b] In bold are the lipoprotein classes containing the highest proportion of the apolipoprotein; in italics are secondary lipoprotein classes

[*] To a lesser extent

2.3.2 Clinical and Metabolic Features of Dyslipidaemia

Lipid metabolism is a highly dynamic and interrelated process, leading to changes in plasma lipoprotein function and/or levels. Although lipoprotein concentrations in blood plasma are highly variable, depending on age, sex, feeding state, metabolic/hormonal state and disease state of individuals, a representative lipoprotein distribution of a fasting, healthy adult male in plasma

is approximately 0 mg/dl for chylomicrons, 150 mg/dl for VLDL, 410 mg/dl for LDL and 280 mg/dl for HDL.

The dysregulation of lipoprotein metabolism, known as dyslipidaemia, covers a wide range of lipid abnormalities, which increase the risk to cardiovascular disease. Atherosclerosis, a type of cardiovascular disease, is one of the most prevalent diseases leading to mortality in developed countries. It is initially caused by the deposition of cholesterol in blood vessels. Most of the cholesterol in blood is transported in the form of LDL particles into the arterial wall. In normal physiological conditions, cholesterol is removed by HDL and recirculated in plasma. Dysregulation in LDL metabolism leads to the increase in the residence time of LDL particles in blood, leading to their oxidation and deposition on the arterial wall, which is the site for atherogenesis. The inability of macrophages to regulate the uptake of oxidised LDL, in conjunction with insufficient removal of cholesterol by HDL particles, causes the accumulation of cholesterol and the formation of foam cells which gives rise to atherosclerosis (Kannel et al., 1979).

The processes by which lipids and lipoproteins contribute to the formation of atherosclerotic plaque and cardiovascular events continue to be an area of controversy and research; however the association between elevated levels of LDL cholesterol and increased risk to cardiovascular disease is indisputable (Kannel et al., 1979; Lorenzo et al., 2007). It is beyond the scope of this thesis to elaborate on the events that lead up to atherosclerosis; instead, we focus specifically on dyslipidaemia. In general, dyslipidaemia is characterised by increased levels of LDL cholesterol and VLDL cholesterol, elevated plasma triglycerides, reduced HDL cholesterol. These factors are strongly and independently associated with cardiovascular disease (Ginsberg & Stalenhoef, 2003). Each of these factors will be discussed briefly in the next section.

LDL Cholesterol

Concentrations of LDL cholesterol have been found to be predictive of the risk of cardiovascular disease over a wide age range (Kannel et al., 1979). Moreover, mortality rates from cardiovascular disease in different communities are directly and linearly correlated with plasma concentrations of cholesterol and LDL cholesterol (Lewis et al., 1978). Major clinical trials have reported that

manipulating lipoprotein levels such as lowering LDL cholesterol slowed the progression of atherosclerosis and decreased the incidence of cardiovascular events (Frick et al., 1987; Kannel et al., 1979). The reduction in the relative risk of cardiovascular events has been documented in patients with and without cardiovascular disease and also in patients with mild or severe dyslipidaemia (Kannel et al., 1979). Evidence have found that individuals with predominantly small dense LDL have a threefold risk of developing cardiovascular disease compared with those who have predominantly large LDL (Lamarche et al., 1997).

VLDL and Triglycerides

The elevated triglyceride concentration observed in dyslipidaemic patients is mainly a consequence of alterations in VLDL metabolism (Chan et al., 2004a; Chan et al., 2004b). Dysregulation of VLDL metabolism in hypertriglyceridaemia is predominantly caused by the changes in kinetics of various apolipoproteins present in plasma lipoproteins. This includes overproduction of plasma apoC-III (Batal et al., 2000; Cohn et al., 2004), overproduction of VLDL apoB-100 and decreased catabolism of apoB-containing particles (Chan et al., 2004a; Chan et al., 2004b), all of which have negative implications on lipid metabolism. In the insulin resistant state, increased levels of free fatty acids liberated from peripheral fat tissue stimulate the synthesis of triglyceride-rich VLDL particles by the liver. This process also lowers HDL cholesterol concentrations via stimulation of cholesteryl ester transfer protein (CETP) and the hepatic lipase (HL). In turn, LDL particles become triglyceride-enriched and undergo lipolysis by HL to form small dense LDL particles, leading to the accumulation of atherogenic particles in plasma (Barter et al., 2003).

HDL Cholesterol

The higher the HDL cholesterol concentration to the total cholesterol level, the lower the risk. Some factors known to influence cardiovascular risk can be related to HDL levels; for example, cigarette smoking lowers HDL, and the HDL level is higher in individuals who exercise regularly (Enger et al., 1977). Moderate alcohol intake was associated with increased HDL cholesterol and reduced cardiovascular risk (Pearson, 1996). Moreover, premenopausal women,

who have a lower incidence of cardiovascular disease compared to their male counterparts, have a higher concentration of HDL, due to the atherosclerotic-protective property of the female sex hormone, oestrogen (Kannel et al., 1976). After the production of oestrogen ceases at menopause, the incidence of cardiovascular events in women parallels that in men (Kannel et al., 1976).

LDL/HDL Ratio

Another predictor of cardiovascular disease is the inverse relationship between LDL and HDL. The correlation between LDL/HDL ratio with CVD has been studied extensively; these associations are strong, predictive and independent of other risk factors (Lewis, 1983). The plasma HDL-cholesterol/total cholesterol ratio is also one of the predictors of the risk to developing cardiovascular disease (Manninen et al., 1992).

Apo B

Recently, it was established that a high plasma apo B level combined with a high LDL cholesterol concentration, with or without hypertriglyceridaemia, is indicative of hyperapobetalipoproteinemia, which increases the risk to premature cardiovascular disease (Jellinger et al., 2012). The Apolipoprotein-Related Mortality Risk (AMORIS) study presents evidence that plasma apo B level may be equivalent or superior to LDL cholesterol, non-HDL cholesterol or other cholesterol ratios in predicting risk to cardiovascular disease (Walldius et al., 2001).

It is important to note that these risk factors do not exist exclusively on their own, but rather, they must be considered within the context of lipoprotein metabolism. For example, an increase in LDL cholesterol may be caused by the increased production of LDL, and/or their slow removal from the circulation, among other factors. It is therefore crucial to model these lipoprotein components in hierarchical and associative relations in order to present a clear elucidation of complex lipoprotein metabolism pathways.

2.3.3 Dyslipidaemia and the Metabolic Syndrome

Dyslipidaemia in the metabolic syndrome is often clustered with other factors such as insulin resistance, impaired glucose regulation, visceral obesity and hypertension (NCEP, 2001; IDF, 2005; WHO, 1998). Studies have shown that the clustering of these metabolic abnormalities co-occur beyond chance (Aizawa et al., 2006), and that when grouped together, they increase the risk to cardiovascular disease, type 2 diabetes and chronic kidney disease (Isomaa et al., 2001; Lakka et al., 2002). The prevalence of metabolic syndrome has varied significantly amongst different studies, presumably due to the lack of accepted criteria for the definition of the syndrome (Cameron et al., 2004). Numerous definitions have therefore been established to diagnose "metabolic syndrome" (Table 4). An enlarged version of this table is shown in Appendix D.

Table 4. Clinical definitions of the metabolic syndrome.

	WHO 1998	NCEP/ATP III 2001; updated 2004	IDF 2005
Criteria	Hyperglycemia/insulin resistance plus two or more of four criteria	Three or more of five criteria	Central obesity plus two or more of four criteria
Central Obesity	* *Waist/hip ratio:* > 0.9 (men) > 0.85 (women) * *And/or body mass index:* > 30 kg/m2	Waist circumference: * *Caucasian*: ≥ 102 cm (men) ≥ 88 cm (women) * *Asian*: ≥ 90 cm (men) ≥ 80 cm (women) * Lower cut-offs (≥ 94 cm [men], ≥ 80 cm [women]) for some non-Asian adults with strong genetic predisposition to insulin resistance	Waist circumference: * *Europid, African, Mediterranean and Middle Eastern (Arab)*: ≥ 94 cm (men) ≥ 80 cm (women) * *South Asian, Chinese, South/Central American*: ≥ 90 cm (men) ≥ 80 cm (women) * *Japanese*: ≥ 85 cm (men) ≥ 90 cm (women)
Hyperglycemia	Insulin resistance, diabetes, impaired fasting glucose, impaired glucose tolerance	* *Fasting plasma glucose level:* ≥ 5.6 mmol/L * *Or current drug treatment for elevated glucose level*	* *Fasting plasma glucose level:* ≥ 5.6 mmol/L * *Or previous diagnosis of type 2 diabetes*
Dyslipidemia	* *Triglyceride levels:* ≥ 1.7 mmol/L **And/or HDL-cholesterol level:* < 0.9 mmol/L (men), < 1.0 mmol/L (women)	* *Triglyceride levels:* ≥ 1.7 mmol/L * *HDL-cholesterol level:* < 1.0 mmol/L (men), < 1.3 mmol/L (women) * *Or current drug treatment for hypertriglyceridemia/low HDL-cholesterol level*	* *Triglyceride levels:* ≥ 1.7 mmol/L * *HDL-cholesterol level:* < 0.9 mmol/L (men), < 1.1 mmol/L (women) * *Or current drug treatment for hypertriglyceridemia/low HDL-cholesterol level*
Hypertension	* *Blood pressure (BP):* ≥ 140/90 mmHg	* *Systolic BP:* ≥ 130 mmHg * *Diastolic BP:* ≥ 85 mmHg * *Or current drug therapy for known hypertension*	* *Systolic BP:* ≥ 130 mmHg * *Diastolic BP:* ≥ 85 mmHg * *Or current drug therapy for known hypertension*
Others	Microalbuminuria: * *Urinary albumin excretion rate > 20 µg/min* * *Or urinary albumin/creatinine ratio > 3.5 mg/mmol*		

Abbreviations: WHO, World Health Organization; NCEP/ATP III, National Cholesterol Education Program Adult Treatment Panel III; IDF, International Diabetes Foundation.

The first operational definition of metabolic syndrome was proposed by the World Health Organization in 1998 (WHO, 1998). WHO states that the main criterion is impaired glucose regulation, and two or more other components, such as abdominal obesity, dyslipidaemia, hypertension and microalbuminuria. The National Cholesterol Education Program Expert Panel (NCEP/ATP III) describes metabolic syndrome as the occurrence of three or more criteria such as obesity, hyperglycaemia, dyslipidaemia and hypertension. Although the above two definitions include the same core criteria, the cut-off points for each component and the mandatory requirements differ, which leads to a difficulty in assessing metabolic syndrome within different populations (NCEP, 2001). To resolve this issue, the International Diabetes Federation (IDF) proposed a global definition of metabolic syndrome in (IDF, 2005). The IDF definition has increased waist circumference as a key requirement, stating central obesity as the main feature plus two or more criteria such as hyperglycaemia, dyslipidaemia and hypertension. Ethnic-specific cut-off points were also introduced for waist measurements due to racial differences between level of adiposity and risk to morbidities, allowing for a more accurate diagnosis of metabolic syndrome.

Currently, the two definitions that are most commonly used are the IDF and NCEP/ATP III definitions, although some practices still adhere to the WHO definition. An update of the NCEP/ATP III guidelines, renamed NCEP/ATP IV are still in progress, and expected to be released in 2013 (Martin et al., 2012).

2.4 Management of Dyslipidaemia

Although a number of guidelines currently exists on the appropriate lipid/lipoprotein levels, the management of dyslipidaemia remains a challenge. Much of this challenge lies in the interpretation of abnormal lipid and lipoprotein levels in the context of traditional risk factors such as family history or smoking and/or other newer risk factors. To have a better understanding of the management of dyslipidaemia, we initially discuss the guidelines which define the dyslipidaemic condition, the causes (aetiology) of lipoprotein dysregulation as well as treatment options for dyslipidaemia.

2.4.1 Diagnostic Parameters

In clinical practice, dyslipidaemia can be diagnosed by measuring the plasma levels of total cholesterol, triglycerides and individual lipoproteins and comparing these levels to established guidelines with a strong emphasis towards cardiovascular risk. The most frequently used guideline for dyslipidaemia management is the National Cholesterol Education Program guideline presented by the Third Report of the Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (NCEP/ATP III) (NCEP, 2001). This guideline is currently being updated to its fourth version (NCEP/ATP IV) and is expected to be released in 2013 (Martin et al., 2012). Many other guidelines exist, such as the International Diabetes Foundation guidelines (IDF, 2005), the World Health Organization guidelines (WHO, 1998), and more recently, the American Association of Clinical Endocrinologists (AACE) 2012 guidelines (Jellinger et al., 2012) and the American Heart Association (AHA) 2011 guidelines (Miller et al., 2011).

Although these guidelines exist to assist in the diagnosis of dyslipidaemia, all of them also take into consideration other factors such as blood pressure, waist circumference, as well as other lipid markers in cardiovascular risk assessment. Therefore, establishing accurate guidelines for achieving “ideal” lipid values remain to be a very contentious area of research. As it is beyond the scope of this research to discuss the correctness of each of these guidelines, we review the literature and base our choice of optimal lipid values on three of the most commonly used classifications for dyslipidaemia and their associated risk level towards developing cardiovascular disease, the NCEP/ATP III, AACE and AHA guidelines. Table 5 below summarises the values for various lipids that indicate low to high risk for cardiovascular disease; standard fasting blood tests for cholesterol and lipid profiles will include values for total cholesterol, HDL cholesterol, LDL cholesterol, triglycerides and apo B level. An enlarged version of this table is shown in Appendix E. For the purpose of this thesis, we focus specifically on the lipid profile values, although for future work, it would be very interesting to factor in other risk factors such as blood pressure, and lifestyle choices such as smoking and alcohol consumption.

Table 5. Classification of risk factors of various lipids according to LDL-cholesterol, total plasma cholesterol, total plasma triglycerides, HDL cholesterol and plasma apolipoprotein B levels.

	NCEP/ATP III 2001, updated 2004	AACE 2012		AHA 2001
LDL-Cholesterol (mg/dL)				
< 100	Optimal	Optimal		Optimal
100 - 129	Near optimal	Near optimal		Near optimal
130 - 159	Borderline high	Borderline high		Borderline high
160 - 189	High	High		High
≥ 190	Very high	Very high		Very high
Total Cholesterol (mg/dL)				
< 200	Optimal	Optimal		Optimal
200 - 239	Borderline high	Borderline high		Borderline high
≥ 240	High	High		High
Total Triglyceride (mg/dL)				
< 100				Optimal
< 150	Normal	Normal		Normal
150 - 199	Borderline high	Borderline high		Borderline high
200 - 499	High	High		High
≥ 500	Very high	Very high		Very high
HDL-Cholesterol (mg/dL)		**Men**	**Women**	
< 40	High	High	High	High
> 60	Optimal	Optimal	Optimal	Optimal
Apo B (ml/dL)				
< 80		Optimal (for patients with CVD or those with diabetes plus ≥ 1 CVD risk factor)		
< 90		Optimal (for patients at risk for CVD, including those with diabetes)		

Abbreviations: NCEP/ATP III, National Cholesterol Education Program Adult Treatment Panel III; AACE, American Association of Clinical Endocrinologists; AHA, American Heart Association

2.4.2 Aetiology of Dyslipidaemia

There are several contributing factors to dyslipidaemia. Dyslipidaemia can be the result of lifestyle, the interaction between genetic predisposition and environmental factors, or other diseases, or a combination of both. Dyslipidaemia which occur due to genetic causes are referred to as primary dyslipidaemia, while secondary dyslipidaemia arise due to other underlying causes such as lifestyle, or other disease states.

Lifestyle contributes to most cases of dyslipidaemia in adults. The most prominent cause in industrialised countries is a sedentary lifestyle with lack of physical exercise and excessive dietary intake of saturated fat and cholesterol. Other common causes include smoking, alcohol consumption and stress (Marieb, 2000; Wright, 2002).

Genetic mutations may result in either the overproduction or defective clearance of LDL cholesterol and triglycerides, or the underproduction or excessive clearance of HDL (Clauss & Kwiterovich Jr., 2003).

Drug interactions may also contribute to lipoprotein dysregulation. Some examples are corticosteroids (Fernández-Miranda et al., 1998; Hochberg & Petri, 1991) and beta-blockers (Feher et al., 1988).

2.4.2.1 Lifestyle

Lifestyle contributes mainly to the development of dyslipidaemia, referred to as secondary dyslipidaemia, due to the increasing trend of high fat/high cholesterol diet in the industrialised world. It has been established that high intake of saturated fat is significantly associated with the elevation in blood cholesterol levels which increase the risk to cardiovascular disease (Gordon, 1988; Wissler & Vesselinovitch, 1975). Further studies indicate that trans fatty acids are also significantly associated with cardiovascular risk, more so than saturated fats (Hu et al., 1997). Reductions in cardiovascular rates have been noted by replacing saturated and trans fat with a combination of poly- and mono-unsaturated fat than reducing the overall fat intake, with the benefits of polyunsaturated fat appearing to be the strongest (Hu et al., 1997). Epidemiological studies comparing the rates of cardiovascular disease of migrants from low- to high-risk areas indicate that these migrants adopt the rates of the new area (Robertson et al., 1977). This evidence indicate strongly that the significant differences in rates are not due to genetic factors, but rather, environmental factors and are therefore potentially modifiable.

The consumption of alcohol is associated with either a protective or a negative effect on the level of circulating LDL. Low-level alcohol consumption, particularly red wines which contain the antioxidant resveratrol, appear to be beneficial with respect to cardiovascular health. The consumption of resveratrol, a compound found largely in the skin of red grapes and other foods, is associated with a reduced risk of cardiovascular disease (Vidavalur et al., 2006). One major effect of resveratrol in the blood is the prevention of oxidation of LDLs. This has a protective effect towards the development of cardiovascular disease as oxidised LDLs are significantly associated with atherosclerosis.

Conversely, excess alcohol consumption is associated with increasing the level of triglycerides which lead to the development of fatty liver (Di Castelnuovo et

al., 2006). This in turn impairs the ability of the liver to take up LDL via the LDL receptor resulting in increased LDL in the circulation (Di Castelnuovo et al., 2006). Clearly a reduction in alcohol consumption will have a significant impact on overall cardiovascular and hepatic function.

Cigarette smoking has also been found to lower HDL cholesterol, which increases the risk to developing cardiovascular disease (Enger et al., 1977).

2.4.2.2 Genetic

Primary dyslipidaemia is caused by a number of hereditary factors, which result in abnormal quantities of blood fats due to a lack of certain critical components involved in lipoprotein metabolism. The most common form of primary dyslipidaemia is hyperlipidaemia, can be further defined into various subtypes according to the types of lipids which are elevated, such as hypercholesterolaemia (elevated cholesterol), hypertriglyceridaemia (elevated triglycerides) and combined hyperlipidaemia (a mix of both elevated cholesterol and triglycerides). The life long elevation of LDL concentrations in familial hypercholesterolemia is associated with a several-fold excess risk of premature cardiovascular disease (Stone et al., 1974). On the other hand, cardiovascular disease is found to be rare in a familial syndrome characterised by high HDL and low LDL concentrations (Glueck et al., 1975) (Glueck et al., 1976).

Primary dyslipidaemia can be classified into various types according to the Fredrickson classification through plasma analysis of lipoprotein patterns (Fredrickson & Lees, 1965). Table 6 is a summary of primary dyslipidaemia classified under the Fredrickson classification system according to the type of disorder with respect to its common name (synonym), type of defect, electrophoresis pattern, lipoprotein abnormality, concentrations of plasma cholesterol and triglycerides, clinical features as well as recommended treatment (Beaumont et al., 1970; Fredrickson & Lees, 1965; Levy & Fredrickson, 1968). An enlarged version of this table is shown in Appendix F.

Table 6. Fredrickson/WHO classification of primary hyperlipidaemia according to type with respect to its synonym, type of defect, electrophoresis patter, lipoprotein abnormality, concentrations of plasma cholesterol and triglycerides, clinical features as well as recommended treatment.

Type	Synonym	Defect	Electrophoresis Pattern	Lipoprotein Abnormality	Plasma Cholesterol concentration	Plasma Triglyceride concentration	Clinical Features	Treatment
I	Familial hyperchylomicronaemia	↓ LDL Defective ApoC-II	↑ Chylomicron zone	↑ Chylomicron	Normal to ↑	↑↑↑↑	Pancreatitis, lipemia, skin eruption, xanthoma, hepatosplenomegaly	Diet
IIa	Familial hypercholesterolaemia	↓ LDL receptor	↑ Beta lipoprotein zone	↑ LDL	↑↑	Normal	Xanthelasma, arcus senilis, tendon xanthoma	Cholestyramine, cholestipol, statin, niacin
IIb	Familial combined hypercholesterolaemia	↓ LDL receptor ↑ apoB	↑ Pre-beta and beta-lipoprotein zone	↑ LDL ↑ VLDL	↑↑	↑↑		Statin, niacin, fibrate
III	Familial dysbetalipoproteinaemia	Defective ApoE2 synthesis	Beta band is broad in beta band zone	↑ IDL	↑↑	↑↑↑	Tubo-eruptive xanthoma, palmar xanthoma	Statin, fibrate
IV	Familial hyperlipaemia	↑ VLDL production ↓ catabolism	↑ pre-beta lipoprotein zone	↑ VLDL	Normal to ↑	↑↑		Statin, niacin, fibrate
V	Endogenous hypertriglyceridaemia	↑ VLDL production ↓ LPL	↑ pre-beta lipoprotein and chylomicron zone	↑ VLDL ↑ Chylomicron	↑ to ↑↑	↑↑↑↑		Niacin, fibrate

Abbreviations: VLDL, very low density lipoprotein; IDL, intermediate density lipoprotein; LDL, low density lipoprotein.

The Fredrickson classification system has since been adopted by the World Health Organisation (WHO) to provide guidelines with respect to the diagnosis and treatment of various types of primary dyslipidaemias (Fredrickson, 1971). Although the Fredrickson/WHO classification system remains the most widely accepted classification system for primary dyslipidaemia, it is by no means the most comprehensive classification of lipoprotein disorders. As the Fredrickson classification relies on lipoprotein pattern recognition rather than the genetic basis of the condition, patients with the same genetic defect may fall into different groups or may change groups as the disease progresses or changes with treatment.

Recent discoveries of the genetic determinants of variation in plasma lipoprotein levels in humans and animal models have enhanced existing knowledge and how they have identified new avenues for investigation. Individual variations in the levels and structures of lipoproteins in plasma have been found to be strongly associated to hereditary influences through classical linkage analysis, association studies and animal models. Within the past decade, a small but important group of genetic determinants of lipoprotein variation was uncovered by resequencing genomic DNA from individuals with extreme lipoprotein phenotypes (Cohen et al., 2004). More recently, genome-wide association

studies have implicated common variants in numerous loci and genes as being the genetic influences underlying the variation that is closer to the population median. For example, familial hypercholesterolaemia may be caused by any one of over 800 different mutations of the LDL receptor gene (Villéger et al., 2002). Mutations of the apolipoprotein B gene have also been found to produce an identical syndrome (Myant, 1993).

As these recent advances in genetic research are paving the pathway towards new discovery of novel pathways and genetic basis of lipoprotein dysregulation, it is becoming increasingly relevant to structure these complex findings within the context of existing knowledge. At this stage, it is suffice to classify primary dyslipidaemia according to the Fredrickson/WHO Classification, as it is beyond the scope of this thesis to explore the intricacies of these genetic causes towards primary dyslipidaemia. Although they still need to be investigated further, the biological implications of these recent genetic advances may be far-reaching as they have the potential to broaden our understanding of basic metabolic pathways and improve classification, diagnosis and treatment strategies.

2.4.2.3 Disease States

In addition to the above genetic causes of blood fat disorders, a number of acquired conditions can raise lipoprotein levels.

Diabetes Mellitus is classified under Type I (insulin-dependent) diabetes and Type II (non insulin-dependent) diabetes. Type I diabetes provides a clear understanding of the relationship between diabetes, insulin deficiency and lipid metabolism. Hypertriglyceridaemia and reduced HDL commonly occur in this condition, although these can be easily treated with replacement of insulin (Ginsberg, 1996). The lipid abnormalities in Type II diabetes, however, cannot be easily managed with insulin replacement. These abnormalities include elevated triglycerides and reduced HDL cholesterol. In addition, there is also an increased amount of small dense LDL which have atherogenic properties (Goldberg, 2001). This effect is amplified by obesity (Brunzell & Hokanson,

1998). Dyslipidaemia observed in Type II diabetes is often found prior to developing the disorder, and is associated with abnormalities in insulin action (referred to as insulin resistance) (Goldberg, 2001). In support of this hypothesis, some thiazides which improve insulin action leads to significant improvements in the lipid profiles of Type II diabetic patients (Ginsberg et al., 1999).

Hypothyroidism is a common cause of lipid abnormalities as thyroid hormones influence all major metabolic pathways, including lipid metabolism. With specific regard to lipid metabolism, thyroid hormones regulate the synthesis, mobilisation and catabolism of lipids (Pucci et al., 2000). Hypothyroidism is a condition where the thyroid hormone is lacking. Most hypothyroid patients have high total cholesterol and LDL cholesterol, while some have high levels of triglycerides and IDL (Lithell et al., 1981). In addition, one study has documented and classified 295 hypothyroid patients according to the Fredrickson classification using plasma analysis of lipoprotein patterns (O'Brien et al., 1993).

Chronic renal failure in kidney disease primarily causes the dysregulation of HDL and triglyceride-rich lipoprotein metabolism (Vaziri, 2006). This leads to hypertriglyceridaemia, reflected in elevated plasma levels of VLDL, IDL and chylomicron remnants.

Chronic liver disease can also raise or lower the blood fats depending on the classification and severity. In a study comparing the different types of liver disease such as chronic hepatitis, liver cirrhosis, hepatocellular carcinoma and metastatic liver disease, it was found that different lipid abnormalities were present in each liver disease (Ooi et al., 2005). In chronic hepatitis, liver cirrhosis, and hepatocellular carcinoma, the cholesterol and triglyceride levels decreased while the LDL triglycerides increased, whereas metastatic liver cancer patients showed a lower HDL level but higher levels of other parameters (Ooi et al., 2005).

2.4.2.4 Drug Interaction

Certain medications have been found to elevate lipid levels. Because some of these medications are used to treat heart disease, which often occur concurrently with dyslipidaemia, it is necessary to reevaluate their usefulness. Some of these medications are:

- Thiazides, medication used to treat high blood pressure, can raise both cholesterol and triglycerides
- Beta-blockers, another class of medication used to treat high blood pressure, cortisone-like drugs, and oestrogen can raise triglycerides
- Progesterone, the pregnancy hormone, raises cholesterol

For the purpose of this thesis, the mechanism of these drugs will not be discussed in great detail, but will be classified under the general umbrella of the concept Drug Interaction.

2.4.3 Treatment of Dyslipidaemia

Treatment options to correct dyslipidaemia include drugs such as statins, fibrates, bile acid sequestrants and cholesterol absorption inhibitors, as well as lifestyle changes, or a combination of both. Combination therapy involves the use of dual or multiple lipid-regulating agents to treat lipoprotein abnormalities by targeting specific lipoproteins and utilising the complementary mechanisms of action of the different agents (Jacobson, 2001). This is a more effective treatment strategy where lipid-regulating monotherapy (e.g. statins or fibrates) may not provide adequate improvement in dyslipidaemia. However, there are some contraindications between different treatments. Although beneficial in correcting dyslipidaemia, the combinations of statins with fibrates or niacins have the potential for interactions that increase the risks of adverse effects, such as myositis and hepatotoxicity (Bays & Dujovne, 1998). Hence, treatment of lipoprotein dysregulation warrants a thorough examination of lipoprotein profiles of specific individuals.

2.4.3.1 Lifestyle Modification

Therapeutic lifestyle interventions are the first line approach to managing dyslipidaemia. These include lifestyle changes such as dietary modification, weight loss and physical exercise.

Varying the intake of dietary fatty acids may alter the total plasma cholesterol levels by influencing the mechanisms for cholesterol balance. The plasma cholesterol level tends to be raised by the ingestion of saturated fatty acids found predominantly in animal fats and tropical plant oils such as palm oil and coconut oil (Cameron et al., 1988). These fatty acids stimulate the synthesis of cholesterol and inhibit its conversion to bile salts. On the other hand, polyunsaturated fatty acids, the predominant fatty acids of most plants, tend to reduce plasma cholesterol levels by enhancing the elimination of both cholesterol and cholesterol-derived bile salts in the faeces. Furthermore, dietary soluble fibre supplements have also been shown to reduce the level of plasma cholesterol, in particular LDL cholesterol, by physically interfering with its absorption from the intestine. Thus, dietary manipulations can reduce the cholesterol-related risk of cardiovascular disease, but not just by reducing cholesterol intake.

Fish consumption, a rich source of n-3 fatty acids, was shown to effectively raise HDL cholesterol and reduce triglycerides by up to 40% in overweight hypertensive individuals (Mori et al., 2000). Increasing evidence also suggest that fish oil consumption protects against coronary disease (Kris-Etherton et al., 2002).

Antioxidants inhibit the oxidation of toxic LDL and is inversely correlated with cardiovascular risk. Recent studies have demonstrated that oxidised LDL is more likely than nonoxidised LDL to promote the development of atherosclerotic plaques. In related investigations, antioxidant vitamins that prevent LDL oxidation, such as vitamin A, vitamin E and vitamin C, have been shown to slow plaque deposition.

Additionally, the consumption of red wines which contain the antioxidant resveratrol, was associated with increased HDL cholesterol and reduced

cardiovascular risk (Pearson, 1996).

Dietary supplements that lower LDL cholesterol levels include fibre supplements and products containing plant sterols (phytosterols). Plant sterols reduce LDL cholesterol by 10%-15% by inhibiting cholesterol incorporation into micelles, decreasing the absorption of total cholesterol (St-Onge et al., 2003).

Evidence suggests that increased physical activity is associated with decreased plasma triglycerides and increased HDL cholesterol levels in individuals with the metabolic syndrome (Carroll & Dudfield, 2004)

These lifestyle modifications need to be promoted as first line therapies for individuals with dyslipidaemia. More effort to integrate educational and lifestyle intervention into the regular care of dyslipidaemic patients is essential. Success with such interventions will limit the need for pharmacotherapy and may provide added benefits if drug therapies are employed.

2.4.3.2 Pharmacotherapy

In addition to lifestyle changes, several lipid-regulating agents may be used to improve dyslipidaemia, described briefly in the following section.

Table 7 summarises these agents with respect to their effects on lipid/lipoprotein levels, as well as side effects and contraindications (NCEP, 2001). An enlarged version of this table is shown in Appendix G. It is important to note that this list is not fully exhaustive of all treatments available for improvement of lipid levels. For the purpose of this thesis, we have chosen to represent these drug classes that are commonly used to treat dyslipidaemia.

Table 7. Pharmacotherapy effects on lipoprotein metabolism.

Drug Class	Agents and Daily Doses	Lipid/Lipoprotein Effects		Side Effects	Contraindications
Statins	Lovastatin (20-80 mg) Pravastatin (20-40 mg) Simvastatin (20-80 mg) Fluvastatin (20-80 mg) Atorvastatin (10-80 mg) Cerivastatin (0.4-0.8 mg)	LDL HDL TG	↓ 18-55% ↑ 5-15% ↓ 7-30%	Myopathy Increased liver enzymes	*Absolute:* • Active or chronic liver disease *Relative:* • Concomitant use of certain drugs*
Fibrates	Gemfibrozil (600 mg) Fenofibrate (200 mg) Clofibrate (1000 mg)	LDL HDL TG	↓ 5-20% ↑ 10-20% ↓ 20-50%	Dyspepsia Gallstones Myopathy	*Absolute:* • Severe renal disease • Severe hepatic disease
Bile acid sequestrants	Cholestyramine (4-16 g) Colestipol (5-20 g) Colesevelam (2.6-3.8 g)	LDL HDL TG	↓ 15-30% ↑ 3-5% No change	Gastrointestinal distress Constipation Decreased absorption of other drugs	*Absolute:* • Dysbetalipoproteinemia • TG >400 mg/dL *Relative:* • TG >200 mg/dL
Nicotinic acid	Immediate release (crystalline) (1.5-3 gm) Extended release (Niaspan®) (1-2 g) Sustained release (1-2 g)	LDL HDL TG	↓ 5-25% ↑ 15-35% ↓ 20-50%	Flushing Hyperglycaemia Hyperuricaemia (gout) Upper GI distress Hepatotoxicity	*Absolute:* • Chronic liver disease • Severe gout *Relative:* • Diabetes • Hyperuricemia • Peptic ulcer disease

Abbreviations: LDL, low density lipoprotein; HDL, high density lipoprotein; TG, triglycerides.

2.4.3.2.1 Statins

Statins are possibly the treatment of choice for reducing LDL cholesterol as they demonstrably reduce cardiovascular mortality. Statins inhibit 3-Hydroxy-3-methylglutaryl coenzyme A (HMG-CoA) reductase, a key enzyme in cholesterol synthesis, leading to up-regulation of LDL receptors and increased LDL clearance (Goldstein & Brown, 1987) (Horton et al., 2002). They reduce LDL cholesterol by up to 60%, produce small increases in HDL and modest decreases in triglycerides through the reduction of VLDL (Jones et al., 2003). Recent clinical trials have demonstrated that statins can decrease cardiovascular events, irrespective of the initial level of cholesterol (Scandinavian Simvastatin Survival Group, 1994; Jones et al., 2003).

2.4.3.2.2 Fibrates

Fibrates reduce triglycerides by about 50% and significantly lowers VLDL content. Clinical studies have shown that fibrates can reduce cardiovascular events in high- risk subjects (Frick et al., 1987; The Diabetes Atherosclerosis Intervention Study, 2001). The mechanisms of action of fibrates on lipoprotein metabolism have been elucidated in various experimental studies (Berge & Moller, 2002). Fibrates are often used concurrently with statin therapy, although

they have also been used as monotherapy agents, especially for patients who do not respond to statin therapy or are intolerant of statins (Abourbih et al., 2009).

2.4.3.2.3 Nicotinic Acid

Nicotinic acid (niacin) blocks the metabolism of fats in adipose tissue, leading to a decrease in free fatty acids in the blood and decreased hepatic secretion of VLDL and cholesterol (Duffield et al., 1983). Niacin also increases HDL by promoting HDL production and inhibiting HDL clearance (Duffield et al., 1983). It must be noted however, that despite improvements in lipid and lipoprotein concentrations, treatment with nicotinic acid leads to the worsening of hyperglycaemia and the development of hyperuricaemia, and therefore cannot be used as a first line therapy for patients with type II diabetes mellitus (Garg & Grundy, 1990).

2.4.3.2.4 Bile Acid Sequestrants

Bile acid sequestrants have also been proven to reduce cardiovascular mortality. They prevent the reabsorption of intestinal bile acid, forcing the up-regulation of hepatic LDL receptors to recruit circulating cholesterol for bile synthesis, leading to a decrease in LDL levels (Goldfine, 2008).

Bile salts are the principal catabolic end products of cholesterol and increase in hepatic bile acid synthesis is accompanied by enhanced cholesterol catabolism. A simple method for increasing bile acid synthesis is to prevent bile acid reabsorption from the intestine. Consequently, drugs have been developed that will reduce the availability of bile acids for intestinal absorption. These agents also impair the function of bile acids within the intestinal lumen and thus can lead to triglyceride malabsorption.

Colestyramine is a bile acid sequestrant that causes increased excretion of these compounds in the faeces (Hashim & Van Itallie, 1965). When used in low doses (13mg/day), colestryamine does not induce steatorrhaea in normal subjects but after high doses (30mg/day) significant malabsorption of fat was

observed. The steatorrhaea could be reversed by the administration of polysorbate (Shuster, Spoto, & Jacobs, 1970).

2.4.3.3 Combination Therapy

Combination therapy involves the use of dual or multiple lipid-regulating agents to treat lipoprotein abnormalities by targeting specific lipoproteins and utilising the complementary mechanisms of action of the different agents (Jacobson, 2001)

For example, statin-fibrate, statin-niacin, statin-ezetimibe or ezetimibe-fibrate may further optimise the lipid profile of subjects with the metabolic syndrome. In addition, a combination of fibrates with either metformin or thiazolidinediones (TZD) which treats insulin resistance in metabolic syndrome may be beneficial as it would simultaneously address both dyslipidaemia and insulin resistance. Additionally, combined alpha and gamma peroxisome proliferator-activated receptors (PPAR) agonists can also simultaneously improve insulin resistance, glucose intolerance, elevated triglycerides and low HDL cholesterol levels (Pourcet et al., 2006).

This is a more effective treatment strategy where lipid-regulating monotherapy (e.g. statins or fibrates) may not provide adequate improvement in dyslipidaemia. However, there are some contraindications between different treatments. Although beneficial in correcting dyslipidaemia, the combinations of statins with fibrates or niacins have the potential for interactions that increase the risks of adverse effects, such as myositis and hepatotoxicity (Bays & Dujovne, 1998). Hence, treatment of lipoprotein dysregulation warrants a thorough examination of lipoprotein profiles of specific individuals.

2.5 The Nature of Biomedical Information

Discovery in life sciences typically begins with the study of cells, the fundamental building blocks of living systems. Although the process of investigating the structure and function of individual components remains a core in biomedical research, the emphasis of research naturally shifts from cells to systems as the repository of knowledge increases. Understanding biological

systems and how they function require not just the knowledge of individual parts, but how they interact with one another. These metabolic processes are often complex and highly dynamic, involving many different components, which undergo physiological changes in response to various regulators and inhibitors. As much as these pathways have been extensively reviewed in the literature by different sources, it is virtually impossible to elucidate the complete extent of the interaction of components in biological processes. Therefore, an integrated view of biology requires the retrieval of respective knowledge from the literature or from various and often heterogeneous databases.

As the degree of complexity of the biological processes under study grows, it is becoming increasingly challenging for domain experts to integrate and link this information within context. This is in part due to the sheer volume of information available, but also because of the loose, heterogeneous nature of biomedical data. In addition to the complex nature of biological processes described above, the types of research output in the biomedical domain are highly diverse. They range from literature publications, laboratory records, epidemiological studies, biomedical images, to various experimental techniques as diverse as qualitative and quantitative assay, light and electron microscopy, mass spectrometry, among others. Moreover, as biological experiments are conducted on different individuals in different physiological states and across different species, the research output also varies accordingly. For instance, the structure and function of organs vary across age and gender, in normal and diseased states, and across species.

In context of these challenges, it is becoming necessary to structure biomedical information such that biological concepts can be accurately represented and the appropriate relationships between these concepts established clearly. This framework will be represented in a format that is widely accepted for the purpose of interoperability between different research groups and also research domains.

2.6 Knowledge Representation Techniques

Bioinformatics is a broad discipline which utilises Information and Communications Technology (ICT) in response to the research problems concerning the nature of biological data in the biomedical science domain. Although the adoption of ICT is relatively recent in biomedical science, knowledge engineering has been extensively researched in computer science. KR techniques have been developed to manage the information explosion by structuring a complex domain systematically and relating the concepts within the domain. In this section, we review KR techniques currently used in the biomedical domain.

2.6.1 Conceptual Schema and Information Modelling Approaches

The advent of high-throughput technologies in biological research has led to a prominent increase in the amount of biological data being generated. In an attempt to manage these resources, large centralised repositories for data have been established such as UniProt KB and Genbank, to manage protein and nucleotide sequences respectively. However, managing large quantities of data is a challenge in itself. An important aspect of data management is the clear representation of the available data. This can be done by modelling the domain of interest through conceptual schemas. Conceptual schemas are the most fundamental basis of an information system and provide the structure or grammar of the given domain (Halpin, 2001). The main challenge lies in modelling the domain clearly and precisely. Two main information modelling approaches in bioinformatics are discussed: Entity-Relationship Modelling and Object-Oriented Modelling.

2.6.1.1 Entity-Relationship Modelling

Entity-Relationship (ER) modelling is one of the most widely used approaches for data modelling. There are many different versions of ER but in general, ER models describe the world in terms of entities that have attributes and participate in relationships. These entities and relationships are illustrated in terms of symbols, notations and lines within an ER diagram (Bornberg-Bauer & Paton, 2002). In Figure 3 below, entity types are encased within a rectangle: **Enzyme**,

Protein, **DNA**, **Reaction** and **Bipolymer**. The attributes of the entity type **Bipolymer** are depicted in ovals: **accno** (accession number), **name**, **species** and **sequence**.

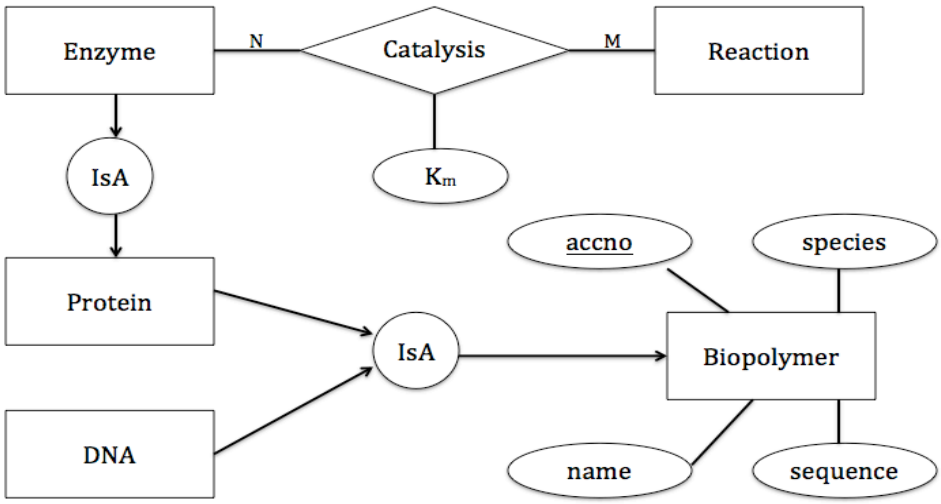


Figure 3. ER notation for some biological concepts. Every enzyme is a protein, which is depicted by the IsA relationship between Enzyme and Protein. Both Protein and DNA are kinds of Biopolymer, also depicted by the IsA relationship. Reaction is related to Enzyme via many-to-many relationship.

Hierarchical relationships between entities can be represented by arrows from the more specialised type to the more general type through a circle containing **IsA**: both **Protein** and **DNA** are shown to be kinds of **Biopolymer**. These relationships have two principal roles. Firstly, the properties of a supertype are inherited by its subtypes, thereby leading to more concise models. For example, the attributes of **Biopolymer** are inherited by **Protein** and **DNA** through the **IsA** relationship. Secondly, the **IsA** relationship makes subsumption relationships explicit. For example, every instance of **Enzyme** is an instance of **Protein**, and every instance of **Protein** is an instance of **Biopolymer**.

Any relationship other than the IsA relationship between two entities is depicted by a rhombus between the related entity types: the **Catalysis** relationship between **Enzyme** and **Reaction** indicates reactions catalysed by enzymes. For example, an instance of an enzyme may catalyse many reactions and at the same time be catalysed by many other enzymes. This many-to-many relationship is depicted by the **M** and **N** in the figure, which denote the number of

participants in the relationship. This cardinality value can be left blank, or specified to a particular value.

Based on the attributes described above, the ER model offers an advantage in representing the domain in an objective way. An example of a database which is based on the ER model is PRINTS, a database for fingerprints. Through this model, the relationship between sequences, their similarity and function is easily represented, thus enabling the efficient and accurate prediction of fingerprint match. Although the ER model offers an objective representation of the domain that is independent of the application platform, it can be limited in its description which may result in missing constraints. In addition, the notation is not conducive to the validation process with domain experts compared to natural language (Halpin, 2001).

2.6.1.2 Object-Oriented Modelling

Object-oriented modelling is an approach that encapsulates both data and behaviour within objects. Although used mainly in object-oriented programs, it can also be used as the basis for database models. Out of many object-oriented approaches, the most popular is Unified Modelling Language (UML). UML class diagrams are used to specify operations, attributes as well as associations, and can be seen as an extended version of ER (Halpin, 2001). Figure 4 shows an example of a UML class diagram for protein structure. **Protein** is the topmost class with attributes: **name**, **pdbcode** and **molecularWeight**. Each **Protein** consists of one or more **Chain**s, which can be made of a type of **Residue** or a type of **SecondaryStructureElement (Bornberg-Bauer & Paton, 2002).**

The class **Residue** provides information about the primary structure with attributes **name**, **position** within the **Chain**, as well as tertiary structure information modelled using the class **Coordinates**. The class ***SecondaryStructureElement*** depicted in italics is an abstract class which has no instances, but play an organisational role in the diagram by serving as the superclass for the classes **Loop**, **Helix** and **Strand**.

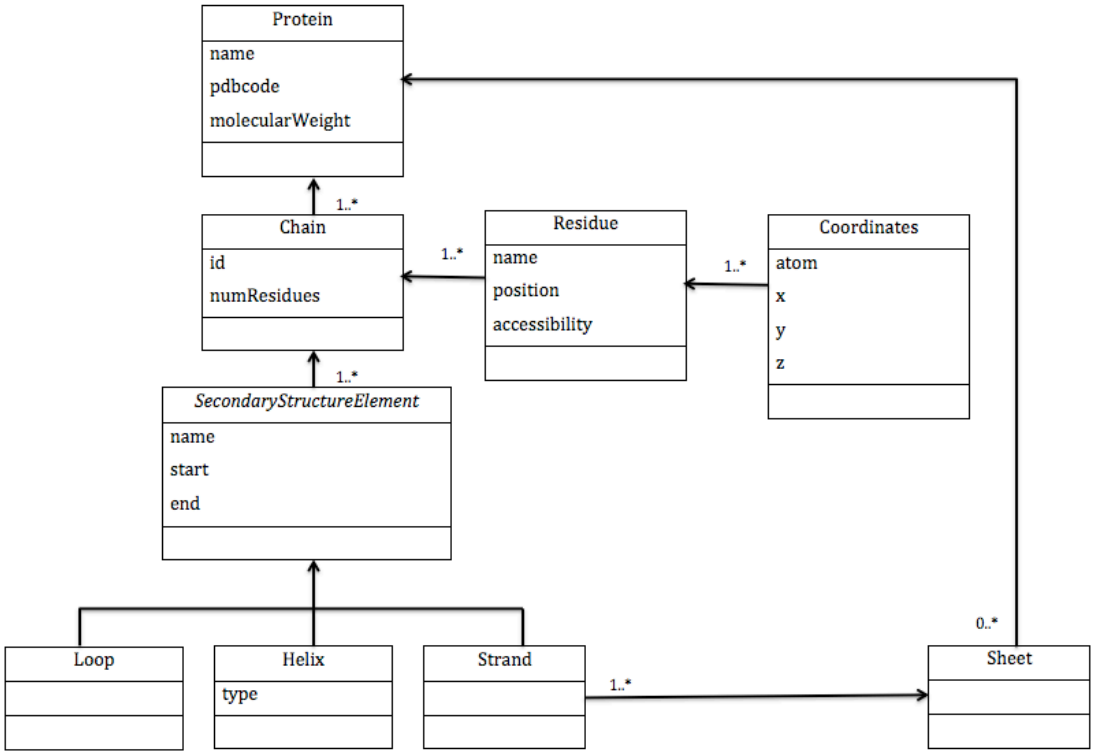


Figure 4. UML class diagram for protein structure.

Like the ER model, the UML class diagrams are independent of the application platform. However, the UML class diagram differs from the ER model in having ternary associations instead of binary. In addition, UML does not require conceptual identification schemes for its entities; instead, entity instances are identified by internal object identifiers (Halpin, 2001). The UML class diagram has been successful in describing sequence information for fully sequenced eukaryotic genomes (Hu et al., 1998). In addition, the UML class diagram has also been implemented using the POET database, which makes use of Java class definitions at the same time. A disadvantage to UML is that there are currently no standard notations for some attributes. Furthermore, validation of UML class diagrams with domain experts are still rife with issues (Halpin, 2001).

2.6.2 Controlled Vocabularies

Controlled vocabularies define a set of standardised terms for labelling entities for specific purposes, such as indexing the literature or annotating gene functions, among others. Biomedical literature often contains many synonymous terms, acronyms and abbreviations which refer to the same concept. For

example, in different texts, “high density lipoproteins” is sometimes referred to as “HDL” or “α lipoproteins” or “apoA containing lipoproteins”. The inconsistent labelling of terms poses a challenge for the integration of information from various heterogeneous resources and databases. By providing a standardised term for each entity it represents, controlled terminology allows the labelling of biomedical entities in a consistent way.

An example of a controlled vocabulary is the Medical Subject Headings (MeSH), created by the National Library of Medicine for the indexing of biomedical literature and the MEDLINE/PubMed article database (Lowe & Barnett, 1994). MeSH provides a standard set of term used to describe the main topics covered in papers, the species studies, funding source, and other attributes. The standard names provided by MeSH are particularly useful for text processing, extraction, and classification (Rubin et al., 2007). Articles can be annotated with MeSH terms and mapped to various databases to retrieve similar content.

2.6.2 Ontologies

Although information models and controlled vocabularies offer many advantages in terms of structuring information, they have limitations with respect to inferring knowledge. In recent years, ontologies have become a topic of interest in the knowledge engineering community. Ontology is defined in the literature as a “formal, explicit specification of a shared conceptualisation” (Studer et al., 1998). Conceptualisation refers to an abstract model of some phenomenon in the world by having identified the relevant concepts of that phenomenon. Explicit means that the type of concepts used, and the constraints on their use are explicitly defined. Formal refers to the fact that the ontology should be machine-readable. Shared reflects the notion that an ontology captures consensual knowledge.

Applications of ontologies are becoming particularly prevalent in biomedical science as more scientists are starting to adopt ontology-based s to model their research domain (Bard & Rhee, 2004) (Stevens et al., 2000). In contrast to information models which operate under a Closed World Assumption (that is, models are developed using predefined information), ontologies operate under an Open World Assumption which supports the inference of new knowledge through automatic reasoning. In addition, most ontologies are based on formal

semantics (Description Logics or DL) and articulated by OWL as the specification language. Not only are ontologies often regarded for conceptual analysis and domain modelling, they are also used to analyse the meaning of an object in the world, or a particular domain, and provide a formal specification to describe the object (Guarino, 1998).

One advantage of ontologies over controlled vocabularies and terminological systems is to support reasoning. The formal structure and rules of inference provided by logic may be coupled with the properties of the relations among things in an ontology in order to draw inferences. While controlled vocabularies are typically created with a specific purpose in mind, ontologies aim at representing what exists independent of any specific use. Both can be shared, however ontologies can often be reused, sometimes in widely differing applications from the ones for which they were originally designed.

The distinguishing feature of an ontology is to explicitly represent various concepts and relationships between them in formal categories and classification systems. This way, the ontology can be used as a core component in knowledge-based information systems in order to aid information retrieval, automatic text processing, content management, inference of new knowledge. In addition, ontology facilitates interoperability between computer systems by providing a common language used in software agents, and can therefore be considered to be the basis for which diverse applications within a knowledge domain can be managed. It is important to note that the process of ontology development varies depending on factors such as the purpose of the ontology as well as the subjective viewpoint of the ontology developer (McCray, 2006).

2.7 Conclusion

In this chapter, we presented the current state of the lipoprotein research domain. We introduced various lipoprotein concepts, including the different lipoprotein classes, components and metabolism pathways. The implications of dyslipidaemia on health were discussed, followed by challenges in the diagnosis and treatment of dyslipidaemia. We then described the nature of biomedical information and reviewed some of the KR techniques currently used in the biomedical domain.

References

Abourbih, S., Filion, K. B., Joseph, L., et al. (2009). Effect of fibrates on lipid profiles and cardiovascular outcomes: a systematic review. *The American Journal of Medicine, 122*(10), 962.e961-968.

Alexander, C. M., Landsman, P. B., Teutsch, S. M., & Haffner, S. M. (2003). NCEP-defined metabolic syndrome, diabetes, and prevalence of coronary heart disease among NHANES III participants age 50 years and older. *Diabetes, 52*, 1210-1214.

Bandeali, S., & Farmer, J. (2012). High-density lipoprotein and atherosclerosis: the role of antioxidant activity. *Current Atherosclerosis Reports, 14*(2), 101-107.

Bard, J., & Rhee, S. (2004). Ontologies in biology: design, applications and future challenges. *Nature Review: Genetics, 5*, 213-222.

Batal, R., Tremblay, M., Barrett, P. H. R., et al. (2000). Plasma kinetics of apoC-III and apoE in normolipidemic and hypertriglyceridemic subjects. *Journal of Lipid Research, 41*, 706-718.

Beaumont, J. L., Carlson, L. A., Cooper, G. R., Fejfar, Z., Fredrickson, D. S., & Strasser, T. (1970). Classification of hyperlipidaemias and hyperlipoproteinaemias. *Bulletin World Health Organization, 43*(6), 891-915.

Berge, J., & Moller, D. E. (2002). The mechanisms of action of PPARs. *Annual Review of Medicine, 53*, 409-435.

Bornberg-Bauer, E., & Paton, N. W. (2002). Conceptual data modelling for bioinformatics. *Briefings in Bioinformatics, 3*(2), 166-180.

Brown, M. S., & Goldstein, J. L. (1986). A receptor-mediated pathway for cholesterol homeostasis. *Science, 232*, 34-47.

Brunzell, J. D., & Hokanson, J. E. (1998). Dyslipidemia of central obesity and insulin resistance. *Diabetes Care, 22*(3), C10-13.

Cameron, J. A., Shaw, J. E., & Zimmet, P. Z. (2004). The metabolic syndrome: prevalence in worldwide populations. *Endocrinology and metabolism clinics of North America, 33*(2), 351-375.

Carroll, S., & Dudfield, M. (2004). What is the relationship between exercise and metabolic abnormalities? A review of the metabolic syndrome. *Sports Medicine, 34*(6), 371-418.

Chan, D. C., Barrett, P. H. R., & Watts, G. F. (2004a). Lipoprotein kinetics in the metabolic syndrome: Pathophysiological and therapeutic lessons from stable isotope studies. *Clinical Biochemistry Review, 25*, 31-48.

Chan, D. C., Barrett, P. H. R., & Watts, G. F. (2004b). Lipoprotein transport in the metabolic syndrome: methodological aspects of stable isotope kinetic studies. *Clinical Science, 107*(221-232).

Chappell, D. A., & Medh, J. D. (1998). Receptor-mediated mechanisms of lipoprotein remnant catabolism. *Progress in Lipid Research, 37*(6), 393–422.

Christakis, G., Rinzler, S., Archer, M.,et al. (1966). The anti-coronary club: A dietary approach to the prevention of coronary heart disease – a seven-year report. *American Journal of Public Health, 56*(2), 299–314.

Clauss, S. B., & Kwiterovich Jr., P. O. (2003). Genetic disorders of lipoprotein transport in children. *Progress in Pediatric Cardiology, 17*(2), 123-133.

Cohen, J. C., Kiss, R. S., Pertsemlidis, A., et al. (2004). Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science 305*, 869-872.

Cohn, J. S., Tremblay, M., Batal, R., et al. (2004). Increased apoC-III production is a characteristic feature of patients with hypertriglyceridemia. *Atherosclerosis, 177*, 137-145.

Duffield, R., Lewis, B., Miller, N., et al. (1983). Treatment of hyperlipidaemia retards progression of symptomatic femoral atherosclerosis. A randomised controlled trial. *The Lancet, 322*(8351), 639-642.

Enger, S. C., Herbjørnsen, K., Erikssen, J., & Fretland, A. (1977). High density lipoproteins (HDL) and physical activity: the influence of physical exercise, age and smoking on HDL-cholesterol and the HDL-/total cholesterol ratio.

Feher, M. D., Rains, S. G., Richmond, W., et al. (1988). Beta-blockers, lipoproteins and non-insulin dependent diabetes. *Postgraduate Medical Journal, 64*, 926-930.

Fernández-Miranda, C., Guijarro, C., de la Calle, A., Loinaz, C., Gonzalez-Pinto, I., & Gómez-Izquierdo, T. (1998). Lipid abnormalities in stable liver transplant recipients-effects of cyclosporin, tacrolimus, and steroids. *Transplant International, 11*(2), 137-142.

Fielding, C. J., & Fielding, P. E. (1995). Molecular physiology of reverse cholesterol transport. *Journal of Lipid Research*, 36, 211–228.

Fredrickson, D. S. (1971). An international classification of hyperlipidemias and hyperlipoproteinemias. *Annals of Internal Medicine, 75*(471-472).

Fredrickson, D. S., & Lees, R. S. (1965). A system for phenotyping hyperlipoproteinemia. *Circulation, 31*(3), 321-327.

Frick, M. H., Elo, O., Haapa, K., Heinonen, O. P., et al. (1987). Helsinki Heart Study: primary prevention trial with gemfibrozil in middle-aged men with dyslipidemia. Safety of treatment, changes in risk factors, and incidence of coronary heart disease. *New England Journal of Medicine, 317*, 1237-1245.

Fruchart, J. C., De Geteire, C., Delfly, B., & Castro, G. R. (1994). Apolipoprotein A-I-containing particles and reverse cholesterol transport: evidence for connection between cholesterol efflux and atherosclerosis risk. *Atherosclerosis, 110*, S35-39.

Ghali, W. A., & Rodondi, N. (2009). HDL cholesterol and cardiovascular risk. *British Medical Journal*. 338, a3065.

Ginsberg, H. N. (1996). Diabetic dyslipidemia: basic mechanisms underlying the common hypertriglyceridemia and low HDL cholesterol levels. *Diabetes, 45*(3), S27-S30.

Ginsberg, H. N., & Stalenhoef, A. F. H. (2003). The metabolic syndrome: targeting dyslipidaemia to reduce coronary risk. *Journal of Cardiovascular Risk, 10*, 121-128.

Goldfine, A. B. (2008). Modulating LDL cholesterol and glucose in patients with type 2 diabetes mellitus: targeting the bile acid pathway. *Current Opinion in Lipidology, 23*(5), 502-511.

Glueck, C., Fallat, R., Millett, F., Gartside, P., & Elston, R. (1975). Familial hyper-alpha-lipoproteinemia: studies in eighteen kindreds. *Metabolism*, 24(11), 1243-1265.

Glueck, C., Gartside, P., Fallat, R., Sielski, J., & Steiner, P. (1976). Longevity syndromes: familial hypobeta and familial hyperalpha lipoproteinemia. *The Journal of Laboratory and Clinical Medicine*, 88(6), 941-957.

Goldberg, I. J. (2001). Diabetic dyslipidemia: causes and consequences. *The Journal of Clinical Endocrinology and Metabolism, 86*(3), 965-971.

Gordon, T. (1988). The diet-heart idea. Outline of a history. *American Journal of Epidemiology, 127*, 220-225.

Guarino, N. (1998). *Formal ontology and information systems.* Paper presented at the Formal Ontology in Information Systems.

Halpin, T. (2001). *Information Modelling and Relational Databases*. USA: Morgan Kaufmann Publishers.

Hashim, S. A., & Van Itallie, T. B. (1965). Cholestyramine resin therapy for hypercholesteremia: Clinical and metabolic studies. *JAMA, 192*(4), 289-293.

Hochberg, M. C., & Petri, M. (1991). The association of corticosteroid (CS) therapy with coronary heart disease (CHD) in patients with systemic lupus erythematosus (SLE): A meta-analysis. *Arthritis & Rheumatism, 34*, R24.

Horton, J. D., Goldstein, J. L., & Brown, M. S. (2002). SREBPs: activators of the complete program of cholesterol and fatty acid synthesis in the liver. *Journal of Clinical Investigation, 109*, 1125-1131.

Hu, F. B., Stampfer, M. J., Manson, J. E., et al. (1997). Dietary fat intake and the risk of coronary heart disease in women.

Hu, J., Mungall, C., Nicholson, D., & Archibald, A. L. (1998). Design and implementation of a corba-based genome mapping system prototype. *Bioinformatics, 14*(2), 112-120.

IDF, International Diabetes Federation. (2005). The IDF consensus worldwide definition of the metabolic syndrome.

Isomaa, B., Almgren, P., Tuomi, T., et al. (2001). Cardiovascular morbidity and mortality associated with the metabolic syndrome. *Diabetes Care, 24*, 683-689.

Jacobson, T. A. (2001). Combination lipid-altering therapy: an emerging treatment paradigm for 21st century. *Current Atherosclerosis Reports, 3*, 373-382.

Jellinger, P. S., Smith, D. A., Mehta, A. E., et al. (2012). American Association of Clinical Endocrinologists' guidelines for management of dyslipidemia and prevention of atherosclerosis. *Endocrine Practice, 18*(1), 1-78.

Jones, P. H., Davidson, M. H., Stein, E. A., et al. (2003). Comparison of the efficacy and safety of rosuvastatin versus atorvastatin, simvastatin, and pravastatin across doses (STELLAR Trial). *American Journal of Cardiology, 92*, 152-160.

Kannel, W. B., Hjortland, M. C., McNamara, P. M., & Gordon, T. (1976). Menopause and risk of cardiovascular disease: the Framingham study. *Annals of Internal Medicine, 85*(4), 447-452.

Kannel, W. B., Castelli, W. P., & Gordon, T. (1979). Cholesterol in the prediction of atherosclerotic disease. New perspectives based on the Framingham study. *Annals of Internal Medicine, 90*(1), 85-91.

Kris-Etherton, P. M., Harris, W. S., & Appel, L. J. (2002). Fish consumption, fish oil, omega-3 fatty acids, and cardiovascular disease. *Circulation, 106*, 2747–2757.

Lamarche, B., Tchernof, A., Moorjani, S., et al. (1997). Small, dense low-density lipoprotein particles as a predictor of the risk of ischemic heart Disease in men. *Circulation, 95*, 69-75.

Lewis, B., Chait, A., & Sigurdsson, G. (1978). Serum lipoproteins in four European communities: a quantitative comparison. *European Journal of Clinical Investigation, 8*(3), 165-173.

Lewis, B. (1983). The lipoproteins: predictors, protectors, and pathogens. *British Medical Journal, 287*, 1161–1164.

Levy, R. I., & Fredrickson, D. S. (1968). Diagnosis and management of hyperlipoproteinemia. *American Journal of Cardiology, 22*, 576-583.

Lithell, H., Boberg, J., Hellsing, K., et al. (1981). Serum lipoprotein and apolipoprotein concentrations and tissue lipoprotein-lipase activity in overt and subclinical hypothyroidism: the effect of substitution therapy. *European Journal of Clinical Investigation, 11*(1), 3-10.

Lorenzo, C., Williams, K., Hunt, K. J., et al. (2007). The National Cholesterol Education Program-Adult Treatment Panel III, International Diabetes Federation, and World Health Organization definitions of the metabolic syndrome as predictors of incident cardiovascular disease and diabetes. *Diabetes Care, 30*, 8-13.

Mahley, R. W., Innerarity, T. L., Rall Jr., S. C., & Weisgraber, K. H. (1984). Plasma lipoproteins: apolipoprotein structure and function. *Journal of Lipid Research, 25*, 1277-1294.

Manninen, V., Tenkanen, L., Koshinen, P., et al. (1992). Joint effects of serum triglyceride and LDL cholesterol and HDL cholesterol concentrations on coronary heart disease risk in the Helsinki Heart Study: implications for treatment. *Circulation, 85*, 37-45.

Marieb, E. N. (2000). Nutrition, Metabolism and Body Temperature Regulation *Human Anatomy & Physiology*. USA: Benjamin Cummings.

Martin, S. S., Metkus, T. S., Horne, A., et al. (2012). Waiting for the National Cholesterol Education Program Adult Treatment Panel IV Guidelines, and in the meantime, some challenges and recommendations. *The American Journal of Cardiology, 110*(2), 307-313.

McCray, A. (2006). Conceptualizing the world: lessons from history. *Journal of Biomedical Informatics, 39*, 267-273.

Miller, M., Stone, N. J., Ballantyne, C., et al. (2011). Triglycerides and Cardiovascular Disease: A Scientific Statement From the American Heart Association. *Circulation, 123*, 2292-2333.

Mori, T. A., Watts, G. F., Burke, V., et al. (2000). Differential effects of eicosapentaenoic acid and docosahexaenoic acid on vascular reactivity of the forearm microcirculation in hyperlipidemic overweight men. *Circulation, 102*, 1264–1269.

Myant, N. B. (1993). Familial defective apolipoprotein B-100: a review, including some comparisons with familial hypercholesterolaemia. *Atherosclerosis, 104*, 1-18.

Natarajan, P., Ray, K. K., & Cannon, C. P. (2010). High-density lipoprotein and coronary heart disease: current and future therapies. *Journal of the American College of Cardiology, 55*(13), 1283-1299.

NCEP, Expert Panel on Detection and Treatment of High Blood Cholesterol in Adults. (2001). Executive Summary of the Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III). *JAMA, 285*, 2486-2497.

O'Brien, T., Dinneen, S. F., O'Brien, P. C., & Palumbo, P. J. (1993). Hyperlipidemia in patients with primary and secondary hypothyroidism. *Mayo Clinic Proceedings, 68*(9), 860-866.

Ooi, K., Shiraki, K., Sakurai, Y., et al. (2005). Clinical significance of abnormal lipoprotein patterns in liver diseases. *International Journal of Molecular Medicine, 15*(4), 655-660.

Pearson, T. A. (1996). Alcohol and Heart Disease. *Circulation, 94*, 3023-3025.

Pucci, E., Chiovato, L., & Pinchera, A. (2000). Thyroid and lipid metabolism. *International Journal of Obesity and Related Metabolic Disorders, 24*(2), S109-112.

Pourcet, B., Fruchart, J.-C., Staels, B., & Glineur, C. (2006). Selective PPAR modulators, dual and pan PPAR agonists: multimodal drugs for the treatment of type 2 diabetes and atherosclerosis. *Expert Opinion on Emerging Drugs, 11*(3), 379-401.

Redgrave, T. G. (2004). Chylomicron metabolism. *Biochemical Society Transactions 32*(1), 79-82.

Robertson, T. L., Kato, H., Rhoads, G. G., et al. (1977). Epidemiologic studies of coronary heart disease and stroke in Japanese men living in Japan, Hawaii and California: incidence of myocardial infarction and death from coronary heart disease. *American Journal of Cardiology, 39*(239-243).

Rubin, D. L., Shah, N. H., & Noy, N. F. (2007). Biomedical ontologies: a functional perspective. *Briefings in Bioinformatics, 9*(1), 75-90.

Scandinavian Simvastatin Survival Group. (1994). Randomised trial of cholesterol lowering in 4444 patients with coronary heart disease: the Scandinavian Simvastatin Survival Study (4S). *Lancet, 344*(1383-1389).

Shachter, N. S. (2001). Apolipoproteins C-I and C-III as important modulators of lipoprotein metabolism. *Current Opinion in Lipidology, 12*, 297-304.

Shuster, F., Spoto, R. C., & Jacobs, M. N. (1970). Cholestyramine and polysorbate-80 in the treatment of cholerheic enteropathy. *The American Journal of Digestive Diseases, 15*(4), 353-358.

Snyder, F. E. (1977). Lipid metabolism in mammals. In F. Snyder (Ed.), *Lipid metabolism in mammals* (Vol. 1). New York: Plenum Press.

Stevens, R., Goble, C. A., & Bechhofer, S. (2000). Ontology-based knowledge representation for bioinformatics. *Briefings in Bioinformatics, 1*(4), 398-414.

Stone, N. J., Levy, R. I., Fredrickson, D. S., & Verter, J. (1974). Coronary Artery Disease in 116 Kindred with Familial Type II Hyperlipoproteinemia. *Circulation, 49*, 476-488.

St-Onge, M.-P., Lamarche, B., Mauger, J.-F., & Jones, P. H. (2003). Consumption of a functional oil rich in phytosterols and medium-chain triglyceride oil improves plasma lipid profiles in men. *Journal of Nutrition, 133*, 1815-1820.

Studer, R., Benjamines, V, & Fensel, D. (1998). Knowledge Engineering: Principles and Methods. *Data & Knowledge Engineering, 25*, 161-197.

Tall, A., Sammett, D. & Granot, E. (1986). Mechanisms of enhanced cholesteryl ester transfer from high density lipoproteins to apolipoprotein B-containing lipoproteins during alimentary lipemia. *Journal of Clinical Investigation, 77*(4), 1163-1172.

The Diabetes Atherosclerosis Intervention Study. (2001). Effect of fenofibrate on progression of coronary-artery disease in type 2 diabetes: a randomised study. *Lancet, 357*, 905-910.

Vaziri, N. D. (2006). Dyslipidemia of chronic renal failure: the nature, mechanisms, and potential consequences. *American Journal of Physiology – Renal Physiology, 290*(2), F262-F272

Vidavalur, R., Otani, H., Singal, P. K., & Maulik, N. (2006). Significance of wine and resveratrol in cardiovascular disease: French paradox revisited. *Experimental and Clinical Cardiology 11*(3), 217–225.

Villéger, L., Abifadel, M., Allard, D., et al. (2002). The UMD-LDLR database: additions to the software and 490 new entries to the database. *Human Mutation, 20*, 81-87.

Walldius, G., Jungner, I., Holme, I., et al. (2001). High apolipoprotein B, low apolipoprotein A-I, and improvement in the prediction of fatal myocardial infarction (AMORIS study): a prospective study. *The Lancet, 358*(9298), 2026 – 2033.
WHO, World Health Organization. (1998). Definitions, Diagnosis and Classification of Diabetes Mellitus and its Complications. Report of a WHO Consultation, Department of Noncommunicable Disease Surveillance. Geneva, Switzerland.
Wissler, R. W., & Vesselinovitch, D. (1975). The effects of feeding various dietary fats on the development and regression of hypercholesterolemia and atherosclerosis. *Advances in Experimental Medicine and Biology, 60*, 65-76.
Wright, H. (2002). *A more excellent way, 5th Ed.*: Pleasant Valley Publications.

Chapter 3. Problem Definition

3.1 Introduction

In Chapter 1, we provided an overview of lipoproteins, raised a number of issues within the lipoprotein research domain and highlighted the need for a formal framework for the management of lipoprotein knowledge. The lipoprotein research domain was extensively reviewed in Chapter 2, followed by a discussion on KR techniques in the biomedical domain. In this chapter, we first define the key concepts which will appear throughout this thesis, and discuss problems within the lipoprotein domain which motivate us to conduct this work. The underlying research issues are identified, from which we derive some key research questions which will be addressed in the thesis. Finally, the choice of research approach to address these research questions will be presented.

3.2 Concept Definition

The main objective of this thesis is to address KR issues in the lipoprotein research domain. In order to introduce the main concepts within the lipoprotein research domain and to avoid ambiguity, the key concepts in this thesis will be defined in this section. First and foremost, we define our domain of study.

Lipoprotein Domain Concept

Definition: Lipoprotein domain concept is defined as a set of lipoprotein concepts and relations. In order to represent the lipoprotein domain knowledge in a coherent way, these concepts and relations are organised in a hierarchical structure, under five distinct superclasses: *Classification*, *Metabolism*, *Pathophysiology*, *Aetiology* and *Treatment*. These superclasses are further defined as follows.

3.2.1 Classification

Definition: Classification is defined as the classification of Lipoprotein Entity. Lipoprotein Entity is defined as a soluble complex of lipids and proteins. Depending on the lipid and protein content, lipoprotein entities are classified accordingly into various subclasses. Examples of Lipoprotein Entity are: Chylomicron, High Density Lipoprotein.

3.2.2 Metabolism

Definition: Lipoprotein Metabolism is defined as the chemical processes which involve the synthesis and/or catabolism of Lipoprotein Entity. In this thesis, we classify “things” associated lipoprotein metabolism into three classes: Physical Entity, Occurring Entity, and Participant Role. These subclasses will be further defined in Chapter 4.

3.2.3 Pathophysiology

Definition: Pathophysiology is defined as the pathophysiology related to the dysregulation in lipoprotein disorders. Pathophysiology can be further categorised into two classes: Disorder and Symptom. These subclasses will be further defined in Chapter 4.

3.2.4 Aetiology

Definition: Aetiology is defined as the causes of lipoprotein disorder. Aetiology can be further categorised into three classes: Lifestyle, Genetic and Drug Interaction. These subclasses will be further defined in Chapter 4.

3.2.5 Treatment

Definition: Treatment is defined as the treatment of lipoprotein disorder. Treatment can be further categorised into three classes: Lifestyle Change, Drug and Combination Therapy. These subclasses will be further defined in Chapter 4.

3.2.6 Diagnostic Parameter

Definition: Diagnostic Parameter is defined as the requirement of diagnosis specific to the characteristics of an individual.

3.3 Problem Overview: Lack of Framework for Lipoprotein Concepts

The past 60 years of lipoprotein research has yielded a great deal of information regarding the structure and metabolism of lipoproteins, the implication of lipoprotein dysregulation on health, causes of dyslipidaemia as well as its management and treatment. Despite the vast amount of research available, a large number of the world's population still suffers from lipoprotein disorders. The issue with dyslipidaemia is that it covers a wide variety of clinical findings and underlying disorders which involve a range of causes and processes. The lack of an inclusive framework for lipoprotein related concepts has 2 implications: research and clinical, described below.

3.3.1 Research Implications

The semantics of biomedical information are usually not explicitly stated but implicitly expressed in biomedical literature. This is because the definitions of concepts within the domain are often well understood among domain experts, thus eliminating the need for semantics to be explicitly defined. Biomedical literature often contains many synonymous terms, acronyms and abbreviations which refer to the same concept. Although domain experts are expected to be familiar with important lipoprotein concepts, the inconsistent representation of terms can sometimes present a challenge in the retrieval and integration of information. In addition, the biomedical community is composed of many different units of small, focused research groups investigating different aspects of a common domain, that would benefit from sharing a common platform for the consistent representation of the research domain. For example, in the lipoprotein research domain, physiologists may specialise in investigating the processes of lipoprotein metabolism, pharmacologists may be interested in examining the effect of certain drugs towards the reduction of a certain lipoprotein class, epidemiologists may be concerned about longitudinal studies

on the dyslipidaemic factors (such as increased LDL levels) which contribute to developing the risk to cardiovascular disease. A framework for lipoprotein knowledge representation provides unambiguous references to domain concepts. This is particularly important when there are tens of thousands of concepts, including information about genes, diseases, chemicals and organisms. Having unambiguous terms is essential for organising information generated in different laboratories. The standardisation of terms is also essential for locating publications; having unambiguous names makes retrieving appropriate information much easier.

3.3.2 Clinical Implications

The clinical implications of this research may be far-reaching and have tremendous potential in opening new possibilities for the development of decision-based support systems for the diagnosis and treatment of dyslipidaemia. Often, researchers and clinicians want to obtain inclusive information about lipoproteins, such as the effect of statins on the different lipoprotein parameters, I.e. LDL, HDL, apolipoprotein C-III, etc., in order to prescribe a treatment that would best suit individuals with specific lipid profiles. Using current information resources and clinical manuals, it is hardly likely that one will find all aspects of the different classes of statins in one place. An inclusive framework for lipoprotein knowledge representation brings together as much information as possible under predefined lipoprotein concepts and their relations. This framework can potentially serve as the basis for the development of software agents or applications which can utilise the structured knowledge to derive and infer information for diagnostic and treatment recommendation purposes.

3.4 Motivation of Study

Currently there does not exist a formal framework for the conceptualisation and classification of lipoprotein-related information, to our knowledge. There is an increasing need to create standardised systems for the formal representation of lipoprotein knowledge, in order to facilitate the collaboration among a community of domain experts. It is not envisaged that such systems could overtake the role of human experts; rather, they can alleviate information integration issues and

aid in decision-making. Therefore, this thesis aims to design a formal conceptualisation framework for the description, organisation and classification of lipoprotein-related information. We discuss specific challenges in the lipoprotein domain in the following section that serve as the motivation for this work.

3.4.1 Modelling Complex Metabolism Pathways

Lipoprotein metabolism is complex, highly interrelated and involves many different pathways under certain physiological conditions (Frayn, 2003) (Eisenberg, 1983). Understanding the impact of the dysregulation of lipoprotein metabolism necessitates knowledge of the synthesis, structure and metabolism of lipoproteins and their individual lipid components, including how they are incorporated, transported and trafficked within their respective lipoprotein classes. It is also important to note that these processes involve a constant and dynamic flow and remodelling of particles where lipid molecules and apolipoproteins are gained and lost through highly complex pathways.

Numerous studies have been carried out investigating various aspects of lipoprotein metabolism, and stored within a large number of information sources. This information can be overwhelming even amongst lipoprotein experts as new discoveries are made every day. Currently there is no classification system in the field of lipoproteins, which offers an inclusive view of links and interconnections between lipoprotein concepts. This has led to difficulties in obtaining an integrated view of lipoprotein knowledge.

The complexity of lipoprotein metabolism pathway presents a need for a framework for lipoprotein knowledge representation. Modelling the relations between lipoprotein entities leads to a better understanding of the complex interrelationships between lipoprotein particles, receptors and enzymes necessary for functional lipid distribution in normal physiological conditions as well as in diseased states. Because the metabolism of the plasma lipoproteins is highly interrelated, one must consider each of the lipoproteins and their subclasses to optimise the complete lipid profile for different individuals.

3.4.2 Lipoprotein Dysregulation and its Implications on Disease

Lipoprotein dysregulation, known as dyslipidaemia, covers a wide range of lipid abnormalities. Dyslipidaemia occurs as a consequence of alterations in the kinetics of lipoproteins, and has been found to be significantly associated with various diseases such as cardiovascular disease, diabetes and hypertension, among others. Therefore, the key to alleviating this risk is a close examination lipoprotein metabolism and the changes in lipoprotein components associated with the various disease states.

Although the processes by which lipids and lipoproteins contribute to cardiovascular events are complex and may also be dependent on other external factors and/or disease states (e.g. hypertension), studies have established the significant correlation between certain lipid and lipoprotein abnormalities and increased risk to cardiovascular disease (Frick et al., 1987; Kannel et al., 1979). Therefore, presenting these abnormalities in terms of the lipoprotein components involved in dyslipidaemia in relation to the risk allows a better understanding of these associations in context of one another.

In addition, dyslipidaemia observed in the metabolic syndrome is often clustered with other factors such as insulin resistance, impaired glucose regulation, visceral obesity and hypertension (NCEP, 2001; IDF, 2005; WHO, 1998). Although it is beyond the scope of this thesis to describe the pathophysiology of each of these risk factors, the framework can be extended to include these factors for future work.

3.4.3 Determining the Causes of Dyslipidaemia

Another challenging issue in the lipoprotein research domain is determining the causes of dyslipidaemia. Dyslipidaemia can be broadly classified into two categories: primary and secondary dyslipidaemias.

Primary dyslipidaemia is relatively more straightforward as it is defined to be lipid and lipoprotein abnormalities caused by genetic factors. Although at this stage it is beyond the scope of this thesis to delve into the specific genetic defects associated with certain dyslipidaemic conditions, the framework can be

extended in the future to include gene concepts from other closely-related ontologies such as the Gene Ontology.

The aetiology of secondary dyslipidaemia is much more complex, as it takes into account various environmental factors and other disease states or disorders, or a combination of either. The challenge is to attempt to map these causes and link them to various lipoprotein disorders, in order to enable the extrapolation of the relations between these concepts into easily identifiable risk factors for quick identification. Again, it must be reiterated that although these aetiology concepts will be mapped to various lipoprotein components and disorders, it is beyond the scope of this thesis to explore the mechanisms of these pathways. Hence, we will be representing aetiology concepts as stand-alone concepts with the possibility of extending these concepts and merging them with other neighbouring ontologies, for future work. For example, the concept of "Diet" under the subclass "Lifestyle" can be merged with concepts from Nutrition Ontology in order to trace the effects of dietary concepts such as "Saturated Fatty Acids" on various lipoprotein components.

3.4.4 Issues in the Diagnosis and Treatment of Dyslipidaemia

In clinical practice, dyslipidaemia can be diagnosed by measuring the plasma levels of total cholesterol, triglycerides and individual lipoproteins and comparing these levels to established guidelines with a strong objective towards reducing cardiovascular risk. Therefore, the diagnosis and treatment of dyslipidaemic patients is a prominent component of cardiovascular care. Although guidelines exist to define dyslipidaemia, diagnostic parameters vary among different individuals, according to gender and ethnicity. In addition, the diagnosis of dyslipidaemia also takes into consideration other factors such as blood pressure, waist circumference, etc. Thus, it can sometimes be challenging for a health practitioner to interpret the correct diagnostic parameters for specific individuals and prescribe the appropriate treatment.

We have identified two issues with the treatment of dyslipidamia which reinforce the need for a framework for lipoprotein knowledge representation.

Firstly, treatment of dyslipidaemia often involves the use of multiple lipid-regulating agents by targeting specific lipoproteins and utilising the complementary mechanisms of action of different agents. However, drug interactions may increase the risk to adverse effects and/or morbidity in certain individuals (Bays & Dujovne, 1998). These risks may be potentially reduced if these interactions can be mapped in a systematic way.

Secondly, controversy exists over the interaction between the medical profession and the pharmaceutical industry. Studies have found that the interaction between health practitioners and the pharmaceutical companies appears to affect prescribing and professional behaviour due to the influence of gifts and various forms of sponsorship and endorsements from the drug industry (Wazana, 2000). In addition, it was also found that the type of sponsorship for randomised controlled trials of statins was strongly linked to the results of those studies (Bero et al., 2007). Head-to-head comparisons of statins with other drugs are more likely to report conclusions that are favourable to the sponsor's drug compared to comparison drugs. This could be due to various reasons such as using lower dosages for the competitor drugs or selective reporting of results. However, regardless of the reasons, the implications suggest that the evidence base relating to statins may be substantially biased. Therefore, these controversial issues present a need for an objective and unbiased representation of treatment pathways that are not tied to any sponsorship affiliation. Modelling the direct relations between treatment with respect to their actions on the associated lipoprotein classes can minimise this bias to a certain extent.

3.5 Underlying Research Issues

In the previous section we have identified the main problems associated with the lack of a framework for lipoprotein-related concepts. In order to resolve these challenges, we explore the underlying research issues described below.

3.5.1 Information Explosion and Unstructured Domain Knowledge

Since the advancement of technology and the emergence of the World Wide Web, modern biomedical research has evolved to be information-intensive. To this effect, the amount of published literature has increased exponentially. Up to 1980, the search term “lipoprotein” generated 16,129 results in PubMed, compared to 69,232 in 2000; currently, as of 2013, the same term fetches approximately 150,000 articles. This information explosion clearly poses a challenge for researchers to retrieve and integrate knowledge relevant to their research.

The premise of biomedical research primarily involves identifying biological components (concepts), and investigating the interaction (relations) between these components in both normal physiological and diseased states. These pathways are highly complex and interrelated, involving many different components. As the amount of published literature grows at a tremendous rate daily, integrating newly discovered components and pathways in context to previously known knowledge is a challenge that is beyond the expertise of a single domain expert. For example, a search of the literature for causes of lipoprotein dysregulation may return 1,203 articles. Not only is it challenging for a single researcher to go through each of these 1,203 articles, but it would be practically impossible to manually represent the information contained in these articles within a structured framework of concepts and relations.

As such, there is a pressing need to structure these concepts and relations in a formal way in order to provide contextual meaning to the available knowledge. By providing a vocabulary of well-defined terms with specified relationships between them, the formal conceptual model alleviates the issues that researchers face in extracting and analysing these concepts. In addition, this framework also makes this knowledge amenable to automatic processing by computer systems.

3.5.2 Heterogeneity of Biomedical Information

Another underlying issue of this work is the proliferation of research output being represented in the form of natural language, from numerous information

sources, all of which function autonomously. This autonomy has led to different information sources having different contents and dissimilar formats, and is collectively referred to as information heterogeneity.

Information heterogeneity can be classified into three broad types: syntactic heterogeneity, schematic heterogeneity and semantic heterogeneity (Sheth, 1998). Syntactic heterogeneity refers to the use of different representation languages or models. Schematic heterogeneity refers to schematic or structural differences between different information systems. Semantic heterogeneity results from the lack of well-defined semantics or meaning of information items.

The emergence of markup languages such as XML (Extensible Markup Language) has resolved the problem of syntactic heterogeneity, and to a certain extent, schematic heterogeneity. By defining a set of rules for encoding documents in a format that is readable by both humans and computer systems, XML supports integration at the syntactic and schematic levels.

Schematic heterogeneity results as a consequence from the proliferation of a variety of data, ranging from structured databases to unstructured (textual or visual) data. Techniques which have been successful in alleviating schematic heterogeneity problems include, but are not exclusive to, the use of object modelling standards such as UML (Unified Modelling Language) as well as RDF (Resource Descriptive Framework) for general purpose description of information systems (Sheth, 1998).

Semantic heterogeneity remains to be one of the most challenging issues in interoperability (Kashyap & Sheth, 1996). For the purpose of this thesis, we will discuss semantic heterogeneity problems specifically within the lipoprotein domain as follows:

3.5.2.1 Inconsistent Terminologies

Biomedical literature often contains many synonymous terms, acronyms and abbreviations which refer to the same concept. For example, in the lipoprotein research domain, “high density lipoproteins” is sometimes referred to as “HDL” or “α lipoproteins” or “apoA containing lipoproteins”. It is challenging to integrate

information in a consistent way when the biological relevance of the same entity is labeled differently in different resources.

3.5.2.2 Differences in Scope and Granularity

Due to the importance of lipoproteins in the regulation of biological and cellular functions in humans, as well as the impact of lipoprotein dysregulation on health, lipoprotein research involves many sub-domains of biomedical science such as physiology, biochemistry, pharmacology and medicine. For example a physiologist may specialise in studying the kinetics of lipoprotein metabolism, whereas a biochemist may focus on the various components involved in lipoprotein metabolism pathways, a pharmacologist may investigate the action of a drug on a specific lipoprotein component, and a health practitioner may be more concerned with parameters of lipoprotein components which indicate risk to developing cardiovascular disease. The differences in scope and granularity of lipoprotein research can lead to challenges in the meaningful representation of lipoprotein knowledge. As such, there is a need to structure concepts and relations in a system of hierarchy, where lower level, more specific concepts can be generalised to broader, high level concepts, in order to present the domain knowledge in a systematic way.

3.5.2.3 Differences in Research Output

In addition to the complex nature of biological processes, the types of research methodologies and output in the biomedical domain are highly diverse. They range from literature publications, laboratory records, epidemiological studies, biomedical images, to various experimental techniques as diverse as qualitative and quantitative assay, light and electron microscopy, mass spectrometry, among others. Moreover, as biological experiments are conducted on different individuals in different physiological states and across different species, the research output also varies accordingly. For instance, the structure and function of organs vary across age and gender, in normal and diseased states, and across species. This presents a challenge towards the integration of these research findings in a meaningful way.

3.5.3 Information Integration

One of the most challenging issues in the biomedical domain is analysing and integrating the rapidly expanding information generated from various sub-domains – from physiology to pharmacology to clinical care. As the amount of experimental data and scientific knowledge increases, there is a need to promote the interoperability of these resources and provide a platform where information can be shared and utilised efficiently.

The effective integration of information is particularly important in translating research outputs into applications in the management and treatment of diseases. However, researchers often face difficulties in retrieving relevant literature embedded in a diverse range of text or electronic resources. Moreover, the effective and efficient retrieval of particular information from one single information resource through a key word based search engine is a difficult, if not impossible process. For example, a search for the term LDL in PubMed generates more than 31,000 results. This diversity of results, by nature, incorporates both relevant and irrelevant data from various research works, as discussed in the problem of information heterogeneity in the previous section. Such a widespread compilation of data about LDL is highly unlikely to offer associations, interrelations, similarities and differences between related concepts and theories although most of these studies have investigated the same particle, shared many metabolism pathways with each other. Therefore, in order to integrate the knowledge available from diverse, autonomous information sources, these important concepts and their relations must be structured in a shared, common manner through a formalised framework.

Defining the semantics of lipoprotein concepts and relations in a formalised framework also allows automation of knowledge retrieval and integration. By formalising these concepts, software agents are able to analyse and elicit the desired information embedded within various sources in a precise and integrative manner. At this stage, we must emphasise that developing these software agents is beyond the scope of this thesis. However, this framework can serve the underlying basis for the design of semantic search engines in the lipoprotein research domain.

3.5.4 Inference of Knowledge

Traditionally, knowledge from science is derived from observations, which are used to form hypotheses. Scientists seek to prove or disprove these hypotheses through experiments, which are documented in natural language and communicated with other scientists through manuscripts. This knowledge is subsequently used to make inferences about other uncharacterised observations, which lead to the formation of new hypotheses. Eventually, a collection of proven hypotheses encompasses the scientific paradigm or domain knowledge.

As modern experimental techniques are rapidly expanding the knowledge base in biomedical science, the lack of formal framework for lipoprotein concepts and relations is also proving to be a challenge in the development of new hypotheses. The investigation of new components or pathways is generally dependent on an integrated view of the domain, as it involves many different interrelated factors. However, the size of the existing knowledge base has become too large even for experienced researchers to extract the complete and relevant information that they need. As such, critical information that may have been embedded in obscure experiments may be missed.

There is an increasing need to manage knowledge in a structured form, in order for the vast range of web-accessible information to be more effectively exploited by both humans and automated tools. Structuring concepts and relations in a formal way enables the automatic processing and inference of knowledge which might have otherwise been missed by human capabilities. This can then potentially lead to novel insights into biological processes and hypothesis formulation.

3.6 Key Research Questions

One of the most significant issues identified in this thesis is to determine what it is that needs to be represented. Based on the challenges in the lipoprotein domain that have been identified, we will organise lipoprotein knowledge according to five key areas: *Classification*, *Metabolism*, *Pathophysiology*,

Aetiology and *Treatment*. In order to represent these key areas, the research questions that need to be addressed are:

1. How can we model lipoprotein concepts and relationships such that they can be used as a representation of the lipoprotein research domain? Specifically, how can we represent complex lipoprotein metabolism pathways in terms of concepts and relationships?

2. How can the model be used to aid knowledge integration and management in the lipoprotein domain? How can we use this framework to infer knowledge to aid diagnosis of lipoprotein dysregulation?

3. How can such a model facilitate collaboration between users? For instance, how can we develop this underlying framework to enable the efficient transference of research output from different research groups?

4. How can we validate our framework such that it is consistent and easily extensible?

3.7 Choice of Research Approach

Based on the literature review of KR techniques in Chapter 2, we propose the ontology approach to solve the issues defined in this thesis. Ontologies provide well-defined meaning to Web-accessible information through formal, structured vocabularies of concepts and relationships between them. These common frameworks facilitate the integration of knowledge in a machine-processable format, in order to allow information to be shared and used effectively by researchers as well as automatically through software tools (Cuenca Grau et al., 2008). The use of ontologies has become increasingly common in the biomedical domain, as domain experts have adopted ontologies as a method for representing the domain knowledge in a systematic way, as well as to support knowledge discovery and interoperability among different research groups. However, to this date, to our knowledge there does not exist an ontological framework for the lipoprotein research domain.

We propose Lipoprotein Ontology as our choice of framework to address the issues that have been defined in this chapter. This thesis will focus on the development and validation of an ontological framework for lipoprotein knowledge representation. The objective is to present lipoprotein concepts and their relationships in a hierarchical and associative manner. Our main research approach follows the category of the science and engineering based research methodology, which focuses on a systematic, investigative approach to problem solving (Galliers, 1992). There are three broad steps to this approach, illustrated in detail in Figure 5:

- Conceptual level: The creation of new concepts through systematic literature review
- Perceptual level: The formulation of new methodology and new framework design through implementation
- Practical level: The evaluation of this framework through real world examples, such as case studies

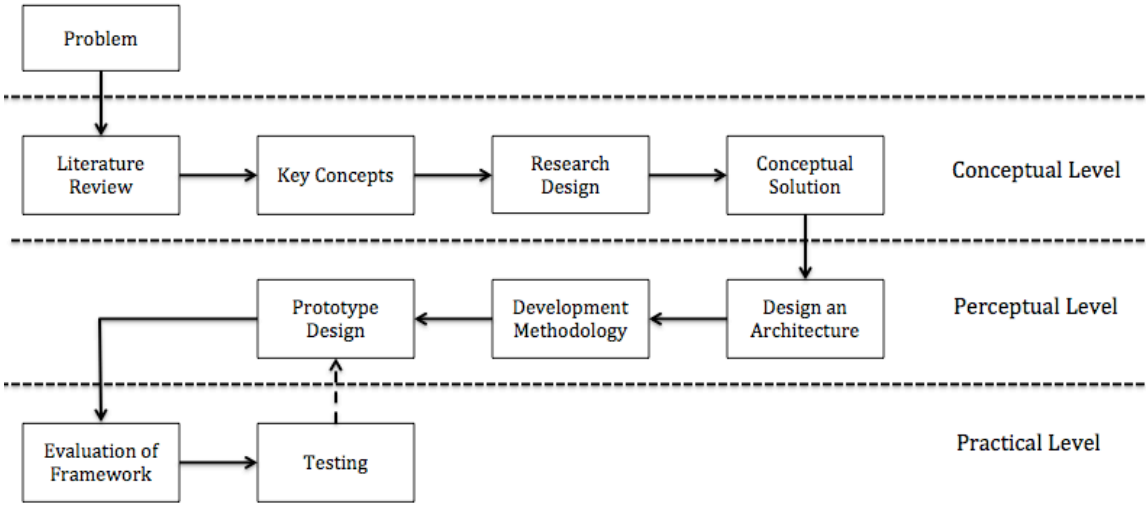


Figure 5. Science and engineering-based research approach.

Our first approach is to identify the research problems. This is followed by an extensive review of the literature in our domain of study, which in this case is the lipoprotein research domain. From the literature review, we extract the key concepts, and integrate these key concepts in the research design, from which we derive a conceptual solution. In the next phase, we proceed to design the

system architecture based on the conceptual solution. Subsequently, we incorporate aspects of existing methodologies towards the design of a new prototype. Once the prototype design has been developed, we evaluate this framework with case studies, followed by further testing and verification by domain experts. Based on the results of the case studies, we can then refine our prototype design to include the modifications, entailing an iterative model.

3.8 Conclusion

The main research issue identified in this thesis is the lack of a formal conceptualisation framework for the description, organisation and classification of lipoprotein-related information. We initially defined the key concepts in the lipoprotein research domain and discussed specific challenges within the domain which will serve as the motivation for the research. The underlying research issues were identified, including the issues of information explosion and unstructured domain knowledge, the heterogeneity of biomedical information, and difficulty in information retrieval. In response to these issues, we developed some key research questions which this thesis will aim to solve, in the context of lipoprotein research domain. With these key research questions in mind, we have chosen to develop Lipoprotein Ontology, a formal framework for the description, organisation and classification of lipoprotein-related information.

References

Bays, H. E., & Dujovne, C. A. (1998). Drug interactions of lipid-altering drugs. *Drug Safety, 19*(5), 355–371.

Bero, L., Oostvogel, F., Bacchetti, P., & Lee, K. (2007). Factors associated with findings of published trials of drug–drug comparisons: why some statins appear more efficacious than others. *PloS Med, 4*(6), e184.

Cuenca Grau, B., Horrocks, I., Kazakov, Y., & Sattler, U. (2008). Modular reuse of ontologies: Theory and practice. *Journal of Artificial Intelligence Research, 31*, 273-318.

Eisenberg, S. (1983). Lipoproteins and lipoprotein metabolism – a dynamic evaluation of the plasma fat transport system. *Journal of Molecular Medicine, 61*(3), 119-132.

Frayn, K. N. (2003). *Metabolic regulation: a human perspective*: Blackwell Publishing.

Frick, M. H., Elo, O., Haapa, K., et al. (1987). Helsinki Heart Study: primary prevention trial with gemfibrozil in middle-aged men with dyslipidemia. Safety of treatment, changes in risk factors, and incidence of coronary heart disease. *New England Journal of Medicine, 317*, 1237-1245.

Galliers, R. D. (1992). *Information Systems Research: Issues, Methods and Practical Guidelines*. Oxford: Blackwell Scientific Publications.

IDF, International Diabetes Federation. (2005). The IDF consensus worldwide definition of the metabolic syndrome.

Kannel, W. B., Castelli, W. P., & Gordon, T. (1979). Cholesterol in the prediction of atherosclerotic disease. New perspectives based on the Framingham study. *Annals of Internal Medicine, 90*(1), 85-91.

Kashyap, V., & Sheth, A. P. (1996). Semantic heterogeneity in global information systems: The role of metadata, context and ontologies. In M. Papazoglou & G. Schlageter (Eds.), *Cooperative Information Systems: Current Trends and Directions.*

NCEP, Expert Panel on Detection and Treatment of High Blood Cholesterol in Adults. (2001). Executive Summary of the Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III). *JAMA, 285*, 2486-2497.

Wazana, A. (2000). Physicians and the pharmaceutical industry. Is a gift ever just a gift? *JAMA, 283*(3), 373-380.

WHO, World Health Organization. (1998). Definitions, Diagnosis and Classification of Diabetes Mellitus and its Complications. Report of a WHO Consultation, Department of Noncommunicable Disease Surveillance. Geneva, Switzerland.

Chapter 4. Overview of the Solution

4.1 Introduction

In order to address the research issues elucidated in the previous chapter, we propose Lipoprotein Ontology, a formal framework for the description, organisation and classification of lipoprotein-related information. We begin this chapter with a brief review on ontology and its definitions. Subsequently, we revisit the underlying research issues associated with the representation of lipoprotein knowledge. We then discuss the role of ontology in addressing the research issues identified in this thesis, by providing examples of current applications of ontologies in the biomedical domain. Finally, this chapter provides an overview of Lipoprotein Ontology and introduces the key concepts and relations in the ontology.

4.2 Ontology Definition and Characteristics

The term ontology has its roots in the branch of philosophy known as metaphysics, and can be defined as the science of what exists, dealing with all categories and structures of objects, processes, events and their associations in every field of reality (Smith et al., 2005). The scheme of metaphysical ontology has inspired researchers of information sciences and knowledge engineering to develop shared and consistent representation framework for different domains of knowledge. In computer science, ontologies are engineering artefacts, which consist of sets of logical axioms that form a certain subset of reality or domain of discourse (Maedche, 2003). These rich conceptual schemas contain classes (concepts), properties (attributes) and relationships between the classes. By explicating the relationships between terms, ontologies represent a particular knowledge domain in a systematic and unambiguous way.

Since its incorporation into the knowledge engineering community, various groups have attempted to define ontology. As such, the definition of ontology has undergone much iteration over time. It is not the purpose of this section to provide an exhaustive list of ontological definitions, but to highlight some of the most significant and most commonly used definitions, in order to provide some guidelines towards ontology development.

One of the first definitions of an ontology was given as “the basic terms and relations comprising the vocabulary of a topic area as well as the rules for combining terms and relations to define extensions to the vocabulary” (Neches et al., 1991). This definition identifies components of an ontology such as terms (classes), relations (properties) and rules (restrictions), and implies that knowledge can be inferred from the combination of explicitly defined terms and relations.

Ontology is most commonly defined in the literature as “the explicit specification of a conceptualisation” (Gruber, 1993). Several versions have been proposed based on this definition, but by far the most thorough is elucidated by Studer et al.: ‘‘An ontology is a formal, explicit specification of a shared conceptualisation. Conceptualisation refers to an abstract model of some phenomenon in the world by having identified the relevant concepts of that phenomenon. Explicit means that the type of concepts used, and the constraints on their use are explicitly defined. Formal refers to the fact that the ontology should be machine-readable. Shared reflects the notion that an ontology captures consensual knowledge, that is, it is not private of some individual, but accepted by a group’’ (Studer et al., 1998).

Another core definition of ontology is stated as follows: “An ontology refers to an engineering artefact, constituted by a specific vocabulary used to describe a certain reality, plus a set of explicit assumptions regarding the intended meaning of the vocabulary. Usually a form of first-order-logic theory is used to represent these assumptions, vocabulary appears as unary and binary predicates, called concepts and relations, respectively.” (Maedche, 2003).

All the definitions above describe ontology as a technology used to represent and share knowledge about a domain by modelling concepts in that domain and

the relationships between those concepts (Gruber, 1991). For the purpose of this thesis, the following definition of ontology will be used.

Ontology

Definition: Ontology in this thesis is defined as a formal, explicit specification of a shared conceptualisation of a domain.

Formal

Definition: Formal refers to the representation of knowledge, composed of domain concepts and the relationships between them, in a formal or specification language in order to be machine-readable. Formalisation of the ontology in this thesis will be described through graphical visualisation.

Explicit

Definition: Explicit refers to the correctness and clarity of the defined concepts, that the type of concepts and relations used, and the constraints on their use are explicitly defined.

Shared

Definition: Shared refers to the notion that the ontology reflects consensual knowledge between a community of domain experts.

Conceptualisation

Definition: Conceptualisation is defined as materialisation of an abstract model or portion of reality. This can be achieved by describing the entities within a given domain in terms of concepts, and the relationships between them reflected as properties.

4.3 Ontology as the Proposed Solution

In section 3.3, we discussed several challenges in the lipoprotein domain due to the lack of formal framework for lipoprotein knowledge. Consequently, we identified the underlying research issues as follows:

- Information Explosion and Unstructured Domain Knowledge
- Heterogeneity of Biomedical Information
- Information Integration
- Inference of Knowledge

The use of ontologies has proven to be effective in the process of data organisation and information management, especially in biomedical domains (Bodenreider, 2008; Rubin et al., 2007). Therefore, we attempt to address the underlying research issues with Lipoprotein Ontology, a formal framework of lipoprotein concepts and their relationships. In this section, we elaborate on how such an ontological framework can resolve the above-mentioned problems. To justify our approach, we provide examples of current applications of ontologies in the biomedical domain to the corresponding issues.

4.3.1 Solution for Information Explosion and Unstructured Domain Knowledge

Lipoprotein Ontology addresses the issues of information explosion and unstructured domain knowledge by formally representing lipoprotein knowledge in terms of concepts (lipoprotein components) and relations (interactions between components) in a hierarchical and associative manner. As we have defined in the previous section, ontologies can be seen as a structured framework of concepts and relationships used to describe certain aspects of reality, and a set of explicit assumptions regarding the intended meaning of the vocabulary (Gruber, 1991; Guarino, 1998). Compared to various other classification schemes and structures, including thesauri and taxonomies, ontologies represent domain models in a more complete and precise manner (Huhns & Singh, 1997). Due to their machine-readability feature, ontological frameworks of concepts and relationships can be used to support the intelligent management of information by human users or computer systems. The formal

structure of ontologies also facilitates the process for classification and description of domain concepts and the relationships between them, thus providing some structure to domain knowledge. Based on these features, ontologies can be used as a reference for applications to obtain knowledge. These ontologies are generally referred to as reference ontologies and are applicable to many different functions. An example of such ontologies is the Foundational Model of Anatomy (FMA) (Rosse & Mejino Jr., 2003).

The FMA is a comprehensive reference ontology for human anatomy. Developed in collaboration with anatomists and knowledge engineers for the purpose of providing an electronically-accessible encyclopaedic reference for anatomic knowledge, FMA presents knowledge in the domain of anatomy as a set of entities with rich declarative relationships between them (Figure 6). Software applications which require anatomic knowledge about particular organs can then navigate to the corresponding entities and relations in FMA, in order to retrieve detailed information about anatomic structures needed by the application, such as adjacency, orientation and location.

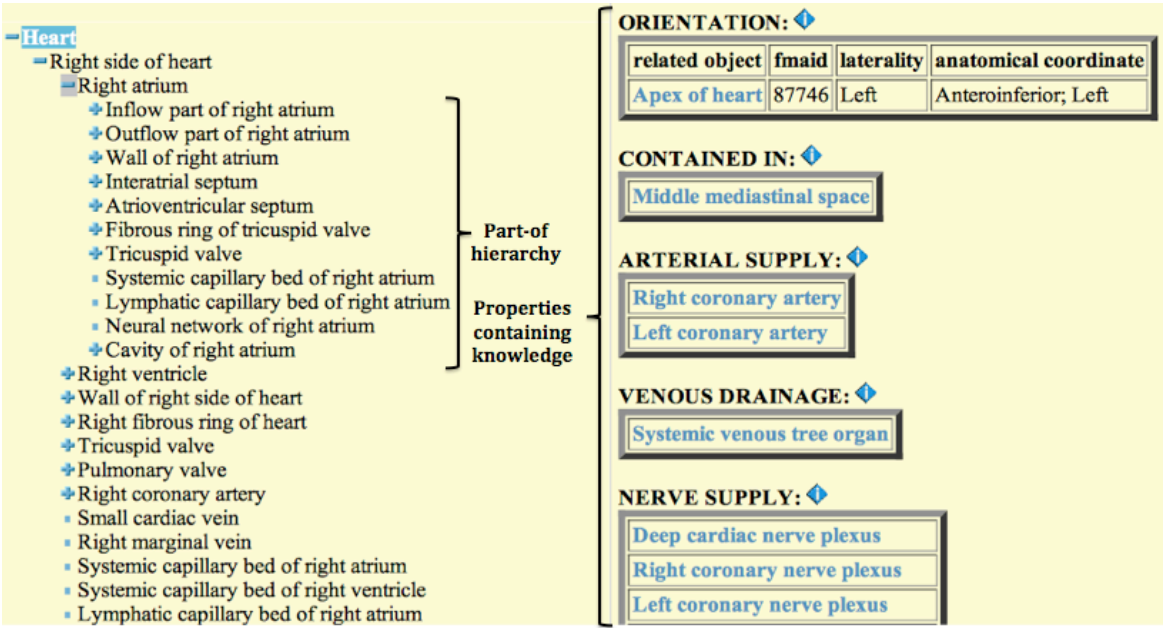


Figure 6. Foundational Model of Anatomy (FMA) is a reference ontology representing detailed knowledge in the anatomy domain. This screenshot, derived from the FMA Foundational Model Explorer page (http://fme.biostr.washington.edu/FME/index.html), shows that anatomic knowledge is modelled by specifying a large set of relations among the anatomic concepts. For example, the Heart is shown to have many relationships to other entities in the FMA, such as

Orientation, Containment, Vascular Supply (Arterial and Venous), as well as Nerve Supply. In addition, the ontology also shows hierarchical/subsumption relations between lower concepts and upper concepts via the part-Of hierarchy relations. For example, Wall of Right Atrium is part-Of Right Atrium, which is part-Of Right Side of Heart, and therefore part-Of Heart.

As an anatomic reference, FMA is particularly useful where anatomy might not be completely visualised in imaging procedures due to limited spatial resolution or to individual patient characteristics. Such detailed anatomic knowledge can serve as the basis to build various applications such as helping radiologists to interpret images and identify abnormalities in adjacent anatomic structures, informing health practitioners about diagnostic possibilities that might have been overlooked, or even predicting the anatomic consequences of penetrating injury (Rubin et al., 2006). In addition, FMA can also be used as the basis for providing anatomic context to other biomedical information (Brinkley, 1991).

4.3.2 Solution for Heterogeneity of Biomedical Information

By providing a controlled vocabulary of lipoprotein concepts, Lipoprotein Ontology alleviates heterogeneity issues in the lipoprotein domain. Ontologies clarify scientific discussions by serving as controlled terminologies for researchers to communicate their results consistently and effectively. As biomedical literature often contains many synonymous terms, acronyms and abbreviations which refer to the same concept, controlled terminologies establish a set of standardised terms for labelling concepts within a given domain, in order to ensure the consistent representation of the same concept among different agents. To further elaborate, ontologies provide a single identifier known as class or concept, to describe information about each entity. Ontologies can thus be used as a controlled terminology to describe biomedical entities in terms of their functions, disease involvement, etc., in a consistent way. In addition, ontologies can be augmented with terminological knowledge such as synonyms, abbreviations and acronyms. Therefore, ontologies enable the community to integrate resources by providing the ability to reliably identify a particular entity or a group of entities based on their biological relevance. These controlled sets of terms and relations are crucial for organising the information generated by different research groups in a consistent way, thereby facilitating communication and enabling interoperability between people as well as computer systems. There are a number of ontologies that were developed for

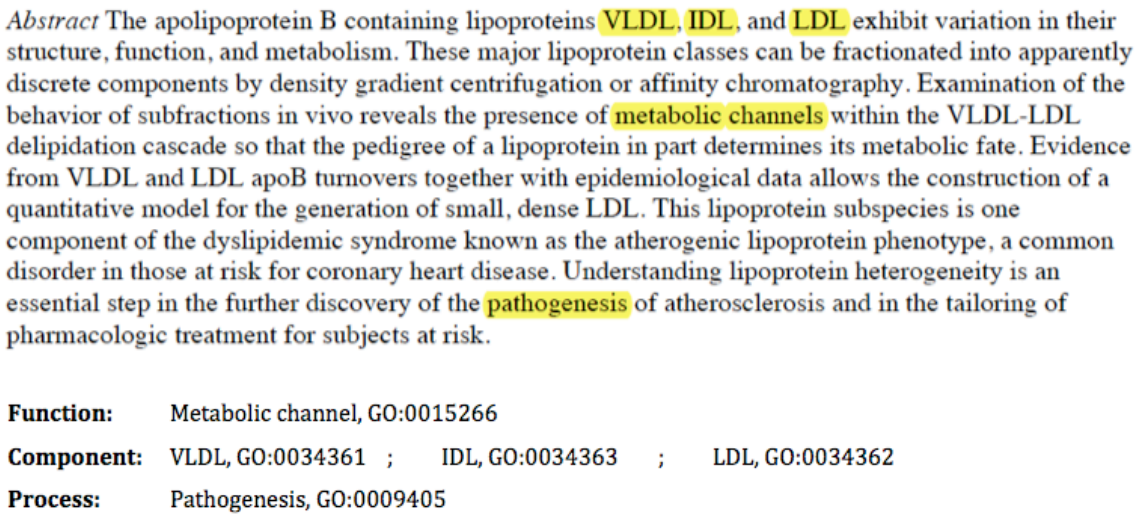
the purpose of resolving issues of heterogeneity. We provide some examples of biomedical ontologies which were developed towards specific purposes which correspond to heterogeneity issues defined in Chapter 3: inconsistent terminologies, differences in scope and granularity, differences in research output.

4.3.2.1 Inconsistent Terminologies

The Gene Ontology (GO) (GO Consortium, 2000) was designed for the standardised representation of various types of genes and gene products contained in different databases. Developed by the GO Consortium in 1998, the initial structure incorporated a collaborative amalgamation of three model organism databases of the Sacchamyces Genome Database, the Mouse Genome Database, and FlyBase. The GO now includes many other annotation groups.

The primary motivation behind GO was the observation that different databases describe the same biological processes, functions and cell components of gene products using different terms. Thus, GO serves to eliminate ambiguity across databases by providing a controlled vocabulary of terms, categorised under 3 non-overlapping ontologies: biological processes, molecular functions, and cellular components of gene products (GO Consortium, 2000). The GO terms have is-a, part-of and the “regulates” relations to other entities, which represents the relationship between biological concepts. These relations allow computer reasoning applications to infer subsumption or composition by tracing these relations respectively.

As whole genomes became readily available through high throughput studies, it was discovered that similar genes often have conserved functions in different organisms. Nucleic acid and polypeptide sequence data allowed easy comparative studies; however, while sequence comparison was easy, comparing functional annotation of those data was difficult. The GO facilitates the annotation process by associating genes and gene products to GO terms (Figure 7), thus making it possible to integrate and query knowledge from different databases in order to infer the functionality of newly discovered genes (Blake & Bult, 2006).



Abstract The apolipoprotein B containing lipoproteins VLDL, IDL, and LDL exhibit variation in their structure, function, and metabolism. These major lipoprotein classes can be fractionated into apparently discrete components by density gradient centrifugation or affinity chromatography. Examination of the behavior of subfractions in vivo reveals the presence of metabolic channels within the VLDL-LDL delipidation cascade so that the pedigree of a lipoprotein in part determines its metabolic fate. Evidence from VLDL and LDL apoB turnovers together with epidemiological data allows the construction of a quantitative model for the generation of small, dense LDL. This lipoprotein subspecies is one component of the dyslipidemic syndrome known as the atherogenic lipoprotein phenotype, a common disorder in those at risk for coronary heart disease. Understanding lipoprotein heterogeneity is an essential step in the further discovery of the pathogenesis of atherosclerosis and in the tailoring of pharmacologic treatment for subjects at risk.

Function: Metabolic channel, GO:0015266

Component: VLDL, GO:0034361 ; IDL, GO:0034363 ; LDL, GO:0034362

Process: Pathogenesis, GO:0009405

Figure 7. Gene Ontology (GO) used to create annotations based on biomedical text (Packard & Shepherd, 1997). In the excerpt from biomedical text shown, the molecular function (metabolic channel), cellular component (VLDL, IDL and LDL), and biological process (pathogenesis) of atherosclerosis, are annotated using the appropriate GO terms from each of the three GO ontologies.

In addition, GO annotations can be used to analyse the results of high throughput experiments. Each GO annotation has a source, allocated as evidence codes, attributed to it (du Plessis et al., 2011). These evidence codes promote further analysis of data by highlighting a cluster of genes with similar experimental results, which might infer common characteristics or functionality shared by that group of genes (Khatri & Draghici, 2005). There are various methods by which gene function can be measured quantitatively (du Plessis et al., 2011). One method is to measure the similarity as a function of the distance between the terms in an ontology graph (Pandey et al., 2008) or the number of common parents (Pekar & Staab, 2002). The most common approach is to calculate the distance between genes, taking into account all possible pairs of GO terms that are associated with both genes (du Plessis et al., 2011). Although slightly different in their approach, these measures ultimately enable the grouping of functionally related genes and identify biological themes shared between them.

4.3.2.2 Differences in Scope and Granularity

Originally developed as a classification system, the International Classification of Diseases (ICD) is the oldest and most important classification still in use (WHO, 2004). It is currently in its 10th revision (ICD-10), and has recently been formalised in OWL (Möller et al., 2013). Developed and maintained by the World Health Organisation, the ICD-10 are detailed classifications of known diseases and injuries. The ICD-10 is strongly goal-oriented, has a hierarchical structure and focuses upon major aetiological factors, diseases of worldwide importance, and causes of serious morbidity and mortality. The value of the hierarchical structure is in its support of abstraction, which is the process whereby lower level, more specific concepts can be generalised into broader, high level concepts to allow reasoning. The goal-oriented characteristic of the ICD-10 provides context for the abstraction, and the purpose of abstraction within the ICD-10 is to enable statistical analysis at the appropriate level, which can be used to support causal hypothesising or to identify potential health problems within a community. ICD-10 is also used internationally for morbidity and mortality statistics (Quan et al., 2005), clinical billing systems (Alexander et al., 2003) and automated decision support in medicine (Jao & Hier, 2010).

4.3.2.3 Differences in Research Output

By sharing the standard terms used by large databases, controlled terminologies facilitate the integration of data from different resources. An example is the National Cancer Institute (NCI) Metathesaurus, developed by the National Cancer Institute, which was designed to integrate molecular and clinical data in cancer research (Hartela et al., 2005). The NCI Metathesaurus provides a common vocabulary for cancer terms and relates them to each other in a DL framework, thus extending the inferential power of the ontology. By doing so, researchers are able to label experimental results in a systematic manner, as well as link their research findings to disease and other patterns (Sioutos & al., 2007). Annotating research data with ontology terms also enables efficient search and retrieval of data.

4.3.3 Solution for Information Integration

Lipoprotein Ontology addresses information integration issues in a number of ways. As we have previously discussed in the problem definition, the amount of biomedical information available is massive. In the setting of the current information explosion, the knowledge contained within a domain is practically beyond the capability of a single domain expert to manually process. Yet, biomedical discovery commonly occurs by integrating related, yet diverse information from different sources. Ontologies can streamline the process of integrating and accessing information across diverse resources. As described earlier, ontologies provide a means to make the semantics of a domain explicit by providing rich relations among its entities. Specifying the semantics of data in a variety of databases can enable researchers to integrate heterogeneous data across different databases. Ontologies can play several roles in information integration applications: they can provide a formally defined vocabulary for semantic annotations, they can be used to describe the structure of existing sources and the information that they store, and they can provide a comprehensive domain model by which information can be retrieved or integrated (Horrocks, 2008). Queries can be performed by DL reasoners through semantic annotations and formalised knowledge to retrieve and combine information from multiple sources (Stevens et al., 2000). In addition, it is possible for a single user or application to utilise different ontologies for different purposes. An example is the Open Biological and Biomedical Ontologies (OBO) Foundry, which is a library of ontologies designed to facilitate information sharing and integration in the biomedical domain.

With an increasing number of biologists recognising the value of ontologies, the proliferation of domain ontologies is becoming just as overwhelming as the original problem of information explosion. In an attempt to integrate these ontologies under one framework and facilitate collaboration within the bio-ontology community, the OBO Foundry was established. The OBO Foundry (Smith et al., 2007) aims to promote semantic interoperability between a group of ontologies covering different domains of biomedical reality, on the basis of an evolving set of common principles for ontology development.

Upper ontologies represent concepts that are universal and provide a standardised approach to KR at the topmost level of organisation in order to allow the integration of knowledge. They are not specific to a particular domain, and are often generic in order to deal with high-level abstraction and broad requirements of different domains, such as the theories of part and whole, dependence and boundaries. An example of upper ontologies in the biomedical domain is the Basic Formal Ontology (BFO). Of these upper ontologies, BFO was specifically developed to serve as an upper ontology for the integration of biomedical ontologies under the Open Biomedical Ontologies (OBO) consortium. The BFO operations on a distinction between continuants and occurrents; continuants are classified as entities in reality that continue to exist through time even while undergoing changes, whereas occurrents incorporate temporal dimension or time. The former typically encompass entities such as organisms which exist even as they undergo changes throughout metabolic processes. On the other hand, the latter usually refer to processes which occur through a period of time such that they can be divided into temporal parts or phases depending on the state that they are in (Smith et al., 2007).

With respect to this basic distinction, the BFO framework is divided into two types of ontologies: SNAP ontologies, which are a snapshot of continuant entities or all entities existing at a time, and SPAN ontologies, which describe processes which unfold through a given interval of time (Grenon & Smith, 2004). The interrelations between the two types of ontologies are defined such that BFO has the capability to deal with both static/spatial and dynamic/temporal features of reality. Both types of ontologies serve as the basis for other sub-ontologies, each of which contribute a certain portion of reality at a given level of granularity.

Such a framework provides guidance for the development of new ontologies based on simple guidelines and best practice models, and is particularly beneficial in avoiding common mistakes, especially for inexperienced ontology developers. Moreover, an upper ontology promotes coherence between neighbouring ontologies by providing the meta-level structures shared by the biomedical domain. Ultimately, the purpose of the OBO Foundry is to develop “an expanding family of ontologies designed to be interoperable and logically

well formed and to incorporate accurate representations of biological reality" (Smith & Brochhausen, 2010). Shown in Figure 8, OBO ontologies are roughly arranged from genotype to phenotype (Bodenreider, 2006).

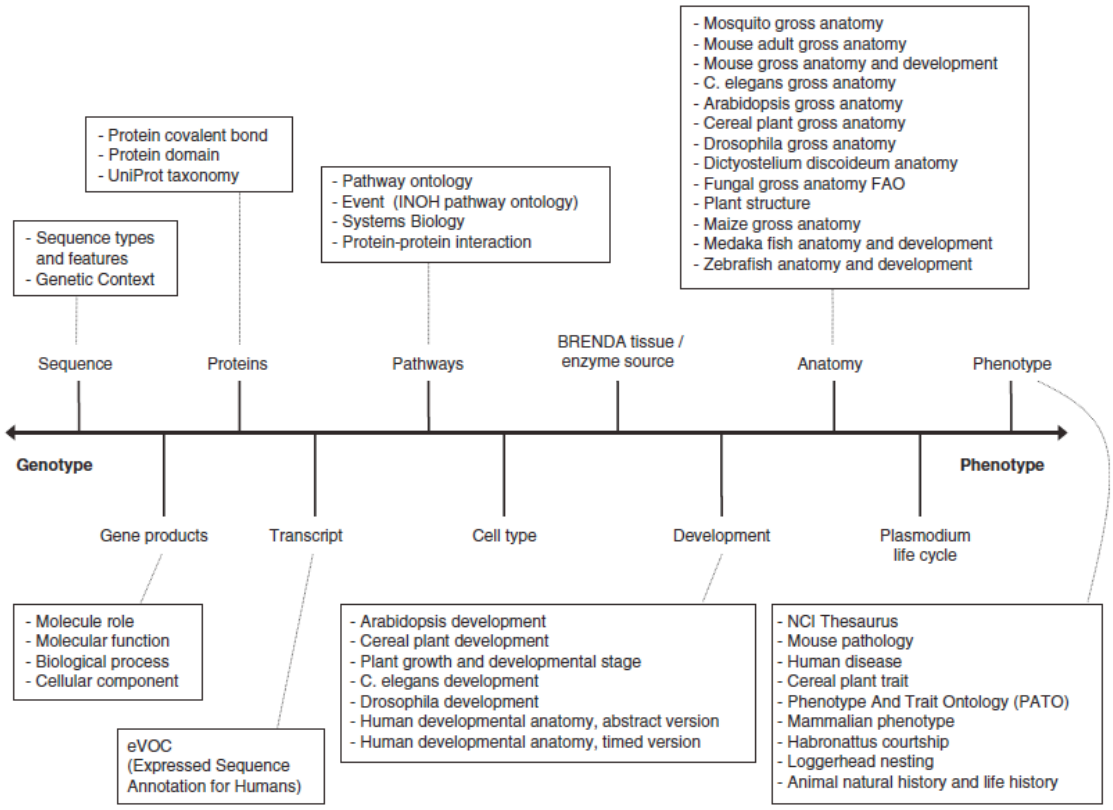


Figure 8. Open Biological and Biomedical Ontologies (OBO) arranged along a spectrum of genotype to phenotype, according to their main topic (Bodenreider, 2006).

In this way, OBO provides context to any specific ontology being queried, in relation to the genotype, anatomical location, stage of development, among others. Thus, by providing a common language across many resources, these non-overlapping and interoperable ontologies can have a potentially significant impact not only towards the holistic integration of biological knowledge, but also enable computational analysis of biological data across different levels of granularity.

4.3.4 Solution for Inference of Knowledge

Lipoprotein Ontology also supports the inference of knowledge and can provide the underlying basis for a wide variety of computational applications. Ontologies are particularly effective for the management of scientific knowledge as they offer several advantages. As ontologies operate on the basis of an open world assumption, relations which are not explicitly stated within the ontology can be inferred from the relations between explicitly defined terms. This way, ontologies extend the functionality of information models and terminological systems by supporting automated reasoning. Their formal structure, coupled with the properties of the relations between concepts, as well as rules of inference provided by DL, provide a justification towards the inference of knowledge that cannot be achieved by information modelling approaches or controlled vocabularies. Most importantly, information models which are represented by ontologies in the OWL format can be applied to Semantic Web search engines, thus increasing the processing and inferential capabilities of domain knowledge (Horrocks, 2002). In other words, biomedical ontologies enable machines to derive meaning from information and complete tasks with reduced human intervention. In this section, we discuss several examples of how ontologies can be used as the basis for various computational applications and support the inference of knowledge.

4.3.4.1 Data Mining

Ontologies can be used to infer new knowledge through data mining. The machine-readability feature of ontologies facilitates the process by which applications or software agents can access the information contained within the ontology, such as data mining tools. Data mining involves the extraction of information, and detection of patterns from a large body of information for further use, such as finding hidden patterns in the data and establishing predictive models based on this data (Fayyad et al., 1996). Some work has been done in the biomedical domain on integrating data across various bioinformatics databases and enable mining across various conceptual levels of biological information, culminating in a complex set of networks from which information can be derived (Gopalacharyulu et al., 2008). A more straightforward example of the application of data mining techniques on ontologies is the International

Classification of Diseases (ICD) discussed in Section 4.3.2. The availability of a large range of disease concepts and relations between diseases represented in the ICD enables data mining tools to detect rare events, such as adverse reaction to drugs, using ICD-9 (Jinjuvadia et al., 2007).

4.3.4.2 Decision-Based Support Systems

Computer-based information systems are increasingly being used by clinicians to aid in the diagnoses of diseases. Clinical systems are qualitatively different from the more general information systems used in health care administration or population studies, as they must cope with issues of scalability, detail and complexity of the information required for clinical decision. At the same time, the system needs to be intuitive and offer ease of use across a range of clinical settings. To address these requirements, the behaviour of such “intelligent” information systems needs to be influenced by semantic context or meaning of the information it is manipulating. Through the representation of concepts and their relations in formal languages, ontologies therefore serve as the semantic basis to enable easy access, retrieval and integration of information by clinical systems. In addition, DL reasoners are used in ontology applications due to their precision and reliability; this is particularly crucial in medical information management, where incorrect reasoning can have an adverse impact on patient care (Horrocks, 2008). Having ontologies as a core greatly increases the inferential power of any application. For example, if several different websites which contain medical information or provide medical services share the same underlying ontology, then software agents are able to extract and integrate information from these different sites. These agents can then use the aggregated information to answer complex user queries through reasoning. By making comparisons between the logical definitions of each concept and their relations, DL reasoners applied to ontologies are capable of inferring relations between concepts. Thus, ontologies can facilitate complex query processing in clinical systems (Huff, 2007).

4.4 Critical Review of Biomedical Ontologies

In the previous section, we have established how biomedical ontologies can be used to solve a number of research issues. This section presents a critical

review of some of these ontologies, as well as several others, to further validate the need to creating a new ontology rather than building on existing ontologies, and to justify the ontology design we have chosen in the lipoprotein research domain. We discuss this justification from two aspects: content and structure.

4.4.1 Content

Most of the biomedical ontologies discussed in the previous section do not contain any lipoprotein concepts. Of the ontologies that do, none provide adequate representations of lipoprotein concepts and their relations. For example, the Gene Ontology has a small section on Lipoproteins and Lipoprotein Metabolism (GO Consortium, 2000); however these concepts do not contain associative relations between them other than the hierarchical structure that they are organised in.

The Unified Medical Language System (UMLS) (http://umlsks.nlm.nih.gov), developed and maintained by the US National Library of Medicine, consists of a multitude of medical concepts from various categorisation systems which makes up the Metathesaurus base. Building on top of the Metathesaurus is the UMLS Semantic Network which draws both hierarchical and associative relationship types between each Metathesaurus concept. However, its associative relationship types are limited to five key categories: “physically related to”, “spatially related to”, “temporally related to”, “functionally related to” and “conceptually related to”. Such relationships do not represent the richness of the theories and pathways involved in lipoprotein metabolism. Therefore, the UMLS can only be used as a reference resource.

The two ontologies that are most closely associated with Lipoprotein Ontology are Lipid Ontology (Baker et al., 2008) and Protein Ontology (Natale et al., 2011). Although lipoproteins are essentially a lipid and protein complex, and both Lipid Ontology and Protein Ontology constitute a repository of lipid and protein concepts respectively, merging the two ontologies to form Lipoprotein Ontology is not a consideration, as they have not been designed for the organisation of lipoprotein concepts in particular. As a result, they lack many relevant concepts related to lipoproteins. However, it is possible to merge these

two ontologies under Lipoprotein Ontology in the future, in order to utilise the information they contain as the fine-grain details of Lipoprotein Ontology.
One of the main motivations for this thesis is to formalise concepts of lipoprotein dysregulation. As lipoprotein dysregulation is classified as a form of disease or disorder, we explored the ICD-10 (WHO, 2004) further and found a match for a subclass for lipoprotein disorders. Again, we face the same issue that these hierarchically structured concepts do not contain the rich associative relations that Lipoprotein Ontology offers. We will, however, map these concepts to Lipoprotein Ontology to facilitate the ease of ontology matching and alignment.

4.4.2 Structure

Ontologies used in classification systems are mostly hierarchical. The value of the hierarchical structure is in its support of abstraction, which is the process whereby lower level, more specific concepts can be generalised into broader, high level concepts to allow reasoning. For example, the ICD is a classification system for diseases which is strongly goal-oriented in its hierarchy and primarily serves as an epidemiological tool, although work to formalise the system in OWL is underway (Möller et al., 2013). The ICD focuses upon major aetiological factors, diseases of worldwide importance, and causes of serious morbidity and mortality. Views on disease models and prevalence are embedded in the system and cannot be changed in any way by the users. The purpose of abstraction within ICD is to enable statistical analysis at the appropriate level, which can be used to support causal hypothesis or to identify potential health problems within a community. The goal-oriented characteristic of the ICD provides context for the abstraction, which, as mentioned previously, cannot be modified. This can serve as a problem when the ICD utilised for different purposes, the classification will only be satisfactory when used in a setting where its aims are compatible with the embedded assumptions. Because of this reason, it has been considered inappropriate for use in general practice (GMSC-RCGP, 1988). If a classification does not represent an appropriate choice of concepts or embody a relevant abstraction, the appropriate response is to construct another classification. However, this can result in a serious fragmentation of medical terminology with a lack of standards. In an attempt to reconcile the numerous classifications currently in use, the Unified Medical Language System (UMLS) was developed. Another major shortcoming of classification systems is that they are often

enumerative and lack compositional features. Therefore, adding extensions to the content usually results in a combinatorial explosion of terms. This is not a serious issue when the classification systems are used for epidemiological and statistical purposes. However, when qualifiers and modifiers are included in a clinical system to increase the expressive power such as to represent the severity or progress of a disease, the total number of terms exponentially increase with respect to the modifier. For example, consider a classification system in which there are 100 diseases. If we add a modififer concept such as *severity* of a disease and limit this to three degrees of severity (mild, moderate and severe), then the total number of terms representing diseases becomes: 100 diseases x 3 severities + 100 original terms = 400 terms. If we introduce another modifier concept such as *progress* of a disease (better, same and worse), the result is now 1,600 terms.

As we have established, the ICD-10 is a traditional medical classification based on a single hierarchical structure, which represents a basic model of the domain knowledge (WHO, 2004). However, as new concepts are added to the ontology, it becomes increasingly difficult to reconcile the new perspective with respect to the purpose of the classification. The strong purpose specification of an ontology are greatly beneficial when used in an appropriate setting. In fact, one of the methodologies that we will discuss in the next section advocates the early establishment of purpose and scope in ontology development. However, at the same time, a deeply embedded goal can be detrimental to adaptation and management of the ontology. Therefore, in an attempt to resolve this problem, a semantic model can be built using a multi-axial approach to separate the basic concepts within the domain from the more complex ideas embedded in the hierarchical structure (Wingert et al., 1989). An example of multi-axial approach is the SNOMED CT (Systematized Nomenclature Of Medicine Clinical Terms), a systematically organised computer processable collection of medical terms (Wingert et al., 1989). The SNOMED CT comprises of six principal axes, which are based on a broad view of biomedical science such as topography, morphology, function, aetiology, disease, procedure.

4.5 Methodologies for Ontology Development

Numerous methodologies have been proposed for ontology development. We have reviewed these methodologies in detail, but due to keeping this section concise, we will be presenting a comparison of these methodologies. For the full review of the methodologies, please refer to Appendix H. Although these methodologies vary in their approach and purpose, they have basic commonalities with respect to different stages of purpose identification, domain conceptualisation, formalisation and evaluation of ontology. Some methodologies offer supplementary stages in order to fulfil certain requirements and be effective in different applications. It must also be mentioned that the process of ontology development is subjective and can be comparable to a design activity (Noy & McGuinness, 2001). Given the same set of requirements within the same domain, two ontology developers are most likely to produce ontologies that are quite different to one another. However, in spite of these differences, an ontology building methodology can be described to have a three-layered structure:

1. Top layer: An outline of the whole ontology development process that is similar to a software development process
2. Middle layer: Generic constraints and guidelines which specify the major steps
3. Bottom layer: Identification of concept and attributes

Of all the methodologies reviewed, the Knowledge Engineering Methodology (Noy & McGuinness, 2001) and the methodology proposed by Uschold & King (Uschold & King, 1995) are the only two methodologies which describe the steps of ontology development at the bottom and middle layers. Both methodologies are very similar in nature in that they both specify the purpose and scope of the ontology at the initial stage and include similar processes such as domain conceptualisation and formalisation of concepts. The only difference between the two is that the Uschold & King methodology offers additional steps, which are the evaluation and documentation processes.

Other methodologies are aimed mainly towards the middle and top layers by developing the ontology within the context of a management process. Therefore,

they are generally more focused towards the evaluation and maintenance of the ontology. For example, the TOVE methodology includes the incorporation of formal competency questions as an evaluation criterion for the effectiveness and completeness of the designed ontology (Gruninger & Fox, 1995). METHONTOLOGY focuses on the maintenance and quality of the ontology by incorporating project management activities (Fernández-López & Gomez-Perez, 2002) while the OnToKnowledge methodology works especially well for knowledge management applications (Staab et al., 2001). Respectively, the DILIGENT methodology focuses on user centrality and highlights the need to adapt the established ontology to its applicants' requirements (Pinto et al., 2004). The DOGMA methodology separates the domain axiomatisation (ontology base) from the application axiomatisation (commitment layer) in order to provide space for both ontology specific applications and their reusability (Spyns et al., 2008).

Some methods also differ in the conceptualisation process. As they are focused towards the bottom and middle layers of ontology development, the Knowledge Engineering Methodology and Uschold & King methodology provide some guidelines towards the top down, bottom up, middle out approaches, providing the justification for each approach. Generally, methodologies commit to the middle or top layers commit to a specific approach, whether top down, bottom up or middle out approaches. For example, comparing the KACTUS (Bernaras et al., 1996) and SENSUS (Swartout et al., 1997) methods, the former involves developing the ontology by means of an abstraction process from an initial knowledge base, whereas in the latter, domain-specific ontologies are generated from a broad ontology.

Having reviewed a number of ontology building methodologies, it can be concluded that none of the approaches proposed are fully mature compared to software engineering methodologies (Corcho, 2003; Fernández-López & Gomez-Perez, 2002). In addition, these approaches are ad-hoc and do not necessarily serve as a standard methodology for ontology development (Noy & McGuinness, 2001). Rather, it is upon the ontology developer to consider the different approaches and adopt the features of these methodologies selectively which best suits their purpose.

We have chosen to develop our Lipoprotein Ontology based on the Knowledge Engineering Methodology due to its focus on conceptual design. However, we will also be incorporating various aspects of other methodologies to suit the purpose of our ontology, which will be discussed in the next chapter.

4.6.1 Overview of Lipoprotein Ontology

The main objective of this thesis is to address KR issues in the lipoprotein research domain. In this section, we provide an overview of Lipoprotein Ontology and highlight the important upper concepts of the ontology. The ontology will be elucidated in detail in Chapter 6.

Lipoprotein Domain Concept

Definition: Lipoprotein domain concept is defined as a set of lipoprotein concepts and relations. In order to represent the lipoprotein domain knowledge in a coherent way, these concepts and relations are organised in a hierarchical structure, under five distinct superclasses: *Classification*, *Metabolism*, *Pathophysiology*, *Aetiology* and *Treatment*. These superclasses are further defined as follows.

4.6.1 Classification

Definition: Classification is defined as the classification of Lipoprotein Entity. Lipoprotein Entity is defined as a soluble complex of lipids and proteins. Depending on the lipid and protein content, lipoprotein entities are classified accordingly into various subclasses. Examples of Lipoprotein Entity are: Chylomicron, High Density Lipoprotein.

4.6.2 Metabolism

Definition: Lipoprotein Metabolism is defined as the chemical processes which involve the synthesis and/or catabolism of Lipoprotein Entity. In this thesis, we classify “things” associated lipoprotein metabolism into three concepts: Physical Entity, Occurring Entity, and Participant Role.

4.6.2.1 Physical Entity

Definition: Physical Entity is defined as the representation of an entity that may participate in an interaction, a process or a relationship of significance. Physical Entity can be further categorised into two classes: Functional Entity and Structural Entity.

Functional Entity is a real thing involved in lipoprotein metabolism, defined by its properties or the function it performs. Some examples of Functional Entity are: Lipid, Protein.

Structural Entity is defined as a real thing, defined by its physico-chemical structure, and can be further divided into two classes: Macro Organic Structure (systems) and Micro Organic Structure. Examples of Macro Organic Structure are: Body Part, Organ. Examples of Micro Organic Structure are: Cell, Tissue.

4.6.2.2 Occurring Entity

Definition: Occurring Entity is defined as the representation of an entity that manifests, unfolds or develops through time, such as a discrete event, or a mutual or reciprocal action or influence that happens between participating physical entities, and/or other occurring entities. Occurring Entity can be further categorised into two classes: Process and Pathway.

Process is defined as the representation of the action which brings about a change from one entity to another entity. Examples of Process are: Hydrolysis, Synthesis.

Pathway is defined as the representation of a sequence of reactions which occurs whereby the products of one reaction are the substrates for subsequent reactions. Examples of Pathway are: Exogenous Pathway, Endogenous Pathway

4.6.2.3 Participant Role

Definition: Participant Role is defined as the function of a physical or conceptual entity, that is its role, in the execution of an event or process. Examples of Participant Role are: Action Role (Drug Action Role, Enzyme Action Role), Modifier (Activator, Inhibitor).

4.6.3 Pathophysiology

Definition: Pathophysiology is defined as the pathophysiology related to the dysregulation in lipoprotein disorders. Pathophysiology can be further categorised into two classes: Disorder and Symptom.

4.6.3.1 Disorder

Definition: Disorder is defined as disease or abnormality of function. Examples of disorder are: Cardiovascular Disease, Diabetes.

4.6.3.2 Symptom

Definition: Symptom is defined as the subjective indication of a disorder or disease. Examples of symptom are: Stroke, Dyslipidaemia.

4.6.4 Aetiology

Definition: Aetiology is defined as the causes of lipoprotein disorder. Aetiology can be further categorised into four classes: Lifestyle Cause, Genetic Cause, Disease States Cause and Drug Interaction.

4.6.4.1 Lifestyle Cause

Definition: Lifestyle Cause is defined as the causes of lipoprotein disorder due to lifestyle factors. Examples of these lifestyle factors are: Diet, Alcohol.

4.6.4.2 Genetic Cause

Definition: Genetic Cause is defined as the causes of lipoprotein disorder due to hereditary factors. Genetic causes are further categorised into two types: Defect (LPL Activity, LDL Clearance) and Deficiency (LDLr Deficiency, LPL Deficiency).

4.6.4.3 Disease State Cause

Definition: Disease State Cause is defined as the causes of lipoprotein disorder due to the disease states. Examples of these drugs are: Diabetes, Renal Failure.

4.6.4.5 Drug Interaction

Definition: Drug Interaction is defined as the causes of lipoprotein disorder due to the action of other drugs. Examples of these drugs are: Beta Blocker, Thiazide.

4.6.5 Treatment

Definition: Treatment is defined as the treatment of lipoprotein disorder. Treatment can be further categorised into three classes: Lifestyle Change, Drug and Combination Therapy.

4.6.5.1 Lifestyle Change

Definition: Lifestyle Change is defined as the treatment of lipoprotein disorder with changes in lifestyle. Examples of these lifestyle changes are: Dietary Supplements, PhysicalExercise.

4.6.5.2 Pharmacotherapy

Definition: Pharmacotherapy is defined as treatment of lipoprotein disorder with pharmacotherapy. Examples of Drug are: Statin, Fibrate.

4.6.5.3 Combination Therapy

Definition: Combination Therapy is defined as the treatment of lipoprotein disorder with a combination of treatment options.

4.6.6 Diagnostic Parameter

Definition: Diagnostic Parameter is defined as the requirement of diagnosis specific to the characteristics of an individual.

4.7 Conclusion

In this chapter, we introduced Lipoprotein Ontology, a framework for the conceptualisation and formal representation of lipoprotein knowledge, as the solution to the underlying research issues that were identified in Chapter 3. We initially provided some definitions of ontology and presented a clear argument on the role of ontologies in addressing the corresponding research issues. This is followed by a critical review of other biomedical ontologies, where we present validation that current ontologies in health and biomedical domains do not fulfil the requirements for a specific ontology for lipoproteins. We also reviewed existing methodologies on ontology development. This chapter concluded with an overview of Lipoprotein Ontology and defined the key concepts and relations in the ontology.

References

Alexander, S., Conner, T., & Slaughter, T. (2003). Overview of inpatient coding. *American Journal of Health-System Pharmacy, 60*(6), S11-14.

Baker, C.J., Kanagasabai R., Ang W.T., et al. (2008). Towards ontology-driven navigation of the lipid bibliosphere. *BMC Bioinformatics*, 9(1), S1-5.

Bernaras, A., Laresgoiti, I., & Corera, J. M. (1996). *Building and reusing ontologies for electrical network applications.* Paper presented at the Proceedings of the European Conference on Artificial Intelligence (ECAI 1996), Budapest, Hungary.

Blake, J. A., & Bult, C. J. (2006). Beyond the data deluge: data integration and bio-ontologies. *Journal of Biomedical Informatics, 39*(3), 314-320.

Bodenreider, O. (2006). Bio-ontologies: current trends and future directions. *Briefings in Bioinformatics, 7*(3), 256-274.

Bodenreider, O. (2008). Biomedical ontologies in action: role in knowledge management, data integration and decision support. *IMIA Yearbook 2008: Access to Health Information, 2008*(1), 67-79.

Brinkley, J. F. (1991). Structural informatics and its applications in medicine and biology. *Acad Med, 66*(10), 589-591.

Corcho, O. (2003). Methodologies, tools and languages for building ontologies. Where is their meeting point? *Data & Knowledge Engineering, 46*(1), 41-64.

du Plessis, L., Skunca, N., & Dessimoz, C. (2011). The what, where, how and why of gene ontology: A primer for bioinformaticians. *Briefings in Bioinformatics, 12*(6), 723-735.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth. *AI Magazine, 17,* 37-54.

Fernández-López, M., & Gomez-Perez, A. (2002). Overview of methodologies for building ontologies. *The Knowledge Engineering Review, 17*(2), 1-13.

GMSC-RCGP. (1988). The classification of general practice data. Final report of the GMSCRCGP Joint Computing Group Technical Working Party *BMA*. London.

GO Consortium. (2000). Gene ontology: tool for the unification of biology. *Nature Genetics, 25*, 25-29.

Gopalacharyulu, P. V., Lindfors, E., Miettinen, J., et al. (2008). An integrative approach for biological data mining and visualisation. *International Journal of Data Mining and Bioinformatics, 2*(1).

Grenon, P., & Smith, B. (2004). SNAP and SPAN: Towards dynamic spatial ontology. *Spatial Cognition and Computation, 4*, 99-104.

Gruber, T. R. (1991). *The role of common ontology in achieving sharable, reusable knowledge bases.pdf.* Paper presented at the KR'1991: Principles of Knowledge Representation and Reasoning, California, USA.

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition 5*(2), 1-23.

Gruninger, M., & Fox, M. S. (1995). *Methodology for the design and evaluation of ontologies.* Paper presented at the Methodology for the design and evaluation of ontologies, Montreal, Canada.

Guarino, N. (1998). *Formal ontology and information systems.* Paper presented at the Formal Ontology in Information Systems.

Hartela, F. W., de Coronadoa, S., Dionneb, R., Fragosoa, G., & Golbeckc, J. (2005). Modeling a description logic vocabulary for cancer research. *Journal of Biomedical Informatics, 38*(2), 114–129.

Horrocks, I. (2002). An ontology language for the semantic Web. Ieee Intelligent Systems. *IEEE Intelligent Systems, 17*(2), 74-75.

Horrocks, I. (2008). Ontologies and the semantic web. *Communications of the ACM 51*.

Huff, S. M. (2007). Ontologies, vocabularies, and data models. In R. A. Greenes (Ed.), *Clinical protege support: The road ahead* (pp. 307–324). Amsterdam: Academic Press.

Huhns, M. N., & Singh, M. P. (1997). Ontologies for Agents. *IEEE Internet Computing, 1*(6), 81-83.

Jao, C. S., & Hier, D. B. (2010). Clinical decision support systems: An effective pathway to reduce medical errors and improve patient safety. In C. S. Jao (Ed.), *Decision Support Systems*: INTECH.

Jinjuvadia K., Kwan, W., & Fontana, R. J. (2007). Searching for a needle in a haystack: use of ICD-9-CM codes in drug-induced liver injury. *American Journal of Gastroenterology, 102*(11), 2437-2443.

Khatri, P., & Draghici, S. (2005). Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics, 21*(18), 3587-3595.

Maedche, A. (2003). Ontology Learning for the Semantic Web. *Kluwer International Series in Engineering and Computer Science, 665*.

Möller, M., Sonntag, D., & Ernst, P. (2013). Modeling the International Classification of Diseases (ICD-10) in OWL. *Knowledge Discovery, Knowledge Engineering and Knowledge Management, 272*, 226-240.

Natale, D.A., Arighi, C. N., Barker, W. C., et al. (2011). The Protein Ontology: a structured representation of protein forms and complexes. *Nucleic Acids Research*, 39, D539-545.

Neches, R., Fikes, R. E., Finin, T., Gruber, T. R., Senator, T., & Swartout, W. R. (1991). Enabling technology for knowledge sharing. *AI Magazine, 12*(3), 36–56.

Noy, N. F., & McGuinness, D. L. (2001). Ontology development 101: a guide to creating your first ontology, from http://protege.stanford.edu/publications/ontology_development/ ontology101-noy-mcguinness.html

Packard, C. J., & Shepherd, J. (1997). Lipoprotein heterogeneity and apolipoprotein B metabolism *Arteriosclerosis, Thrombosis, and Vascular Biology, 17*(12), 3542-3556.

Pandey, J., Koyuturk, M., & Subramaniam, S. (2008). Functional coherence in domain interaction networks. *Bioinformatics, 24*(16), i28–34.

Pekar, V., & Staab, S. (2002). Taxonomy learning: factoring the structure of a taxonomy into a semantic classification decision. *COLING 2*, 786-792.

Pinto, H. S. a., Staab, S., & Tempich, C. (2004). *DILIGENT: Towards a fine-grained methodology for Distributed, Loosely-controlled and evolvInG*

Rosse, C., & Mejino Jr, J. L. V. (2003). A reference ontology for biomedical informatics: the foundational model of anatomy. *Journal of Biomedical Informatics, 36*, 478–500.

Quan, H., Sundararajan, V., Halfon, P., Fong, A., Burnand, B., Luthi, J.-C., Ghali, W. A. (2005). Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. *Medical Care, 43*(11), 1130-1139.

Rubin, D. L. (2006). Using ontologies linked with geometric models to reason about penetrating injuries. *Artificial Intelligence Medicine, 37*(3), 167-176.

Rubin, D. L., Shah, N. H., & Noy, N. F. (2007). Biomedical ontologies: a functional perspective. *Briefings in Bioinformatics, 9*(1), 75-90.

Sioutos, N. (2007). NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information. *Journal of Biomedical Information, 40*(1), 30-43.

Smith, B., Ceusters, W., Klagges, B., et al. (2005). Relations in biomedical ontologies. *Genome Biology, 6*(5), R46.

Smith, B., Ashburner, M., Rosse, C., et al. (2007). The OBO foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology, 25*, 1251-1255.
Smith, B., & Brochhausen, M. (2010). Putting biomedical ontologies to work. *Methods of Information in Medicine, 49*(2), 135-140.
Spyns, P., Tang, Y., & Meersman, R. (2008). An ontology engineering methodology for DOGMA *Journal of Applied Ontology, 3*(1-2), 13-39.
Staab, S. H., Schunurr, P., R., S., & Sure, Y. (2001). Knowledge processes and ontologies. *IEEE Intelligent Systems: Special Issue on Knowledge Management, 16*(1), 26-34.
Stevens R., Goble, C. A., & Bechhofer, S. (2000). Ontology-based knowledge representation for bioinformatics. *Briefings in Bioinformatics, 1*(4), 398-414.
Studer, R., Benjamines, V., & Fensel, D. (1998). Knowledge Engineering: Principles and Methods. *Data & Knowledge Engineering, 25*, 161-197.
Swartout, B., Patil, R., Knight, K., & Russ, T. (1997). *Toward distributed use of large-scale ontologies.* Paper presented at the Symposium on Ontological Engineering of AAAI. Stanford, California.
Uschold, M., & King, M. (1995). *Towards a methodology for building ontologies.* Paper presented at the Workshop on Basic Ontological Issues in Knowledge Sharing, United Kingdom.
Wingert, F., Rothwell, D., & Cote, R. (1989). Automated indexing into SNOMED and ICD. In J. R. Scherrer, R. A. Cote & S. H. Mandil (Eds.), *Computerised natural medical language processing for knowledge engineering* (pp. 5-17). North Holland: Elsevier Science Publishers.
WHO, World Health Organization. (2004). International Statistical Classification of Diseases and Health Related Problems.

Chapter 5. Research Methodology

5.1 Introduction

In the previous chapter, we provided justifications for Lipoprotein Ontology as the solution to the corresponding research issues that were identified in Chapter 3. We presented a critical review of biomedical ontologies, as well as a comparison of existing methodologies on ontology development. Subsequently, we arrived at the conclusion that ontology building is a subjective process, and that features from different ontology development methods can be adopted to suit the purpose of the ontology. This chapter details the steps we have undertaken in developing Lipoprotein Ontology. The methodology used to build Lipoprotein Ontology is based on the Knowledge Engineering Methodology; however, we will also incorporate various features of other methodologies in our ontology development process.

5.2 Overview of the Methodology

In developing Lipoprotein Ontology, our methodology covers four broad processes: specification, conceptualisation, formalisation and evaluation (Chen & Hadzic, 2010). Knowledge acquisition occurs throughout the four stages. Detailed steps for these four processes are shown in Figure 9 and elaborated in the subsequent sections.

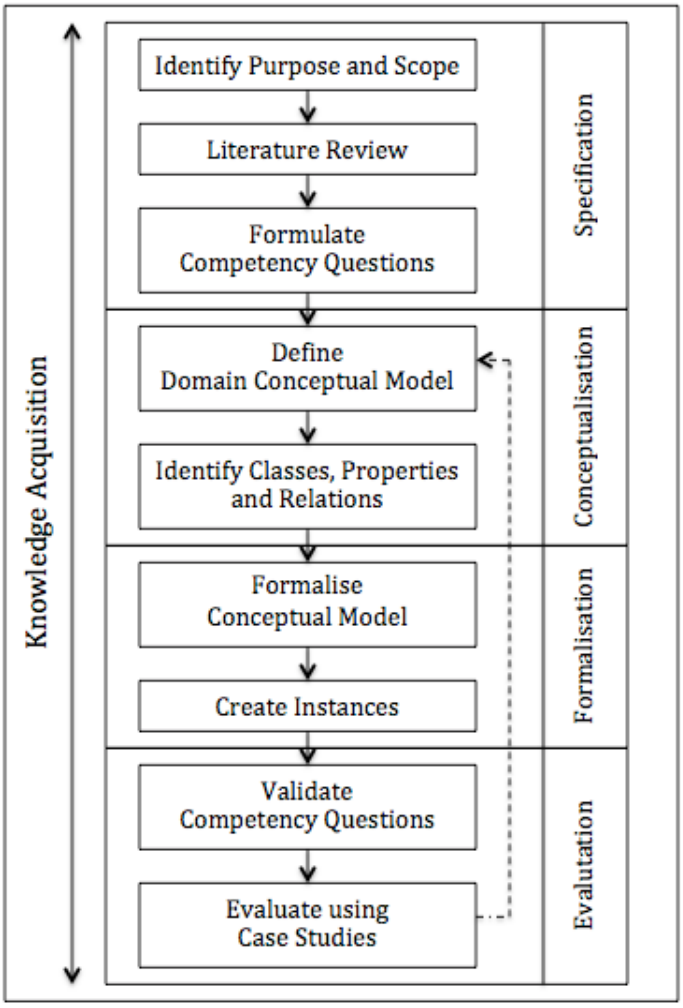


Figure 9. A schematic view of the methodology for Lipoprotein Ontology.

5.3 Specification

The specification phase involves 3 sub-phases:

- Identification of purpose and scope
- Literature review
- Formulation of competency questions

5.3.1 Identification of Purpose and Scope

The first step to developing Lipoprotein Ontology was to identify the key objective, followed by the purpose of the ontology. The aim of this thesis is to develop an ontology for the lipoprotein domain which formally represents

knowledge about lipoprotein-related concepts and their relationships. This ontology will serve the following purposes which correspond to the underlying research issues defined in Chapter 3:

- Formal representation of lipoprotein research domain
- Controlled vocabulary for lipoprotein concepts to alleviate issues of information heterogeneity
- Lipoprotein knowledge integration and management
- Inference of lipoprotein knowledge

Subsequently, we established the scope of Lipoprotein Ontology. This is to ensure that the ontology created is purpose-driven and contain the right level of granularity for knowledge-based queries (Noy & McGuinness, 2001). Here, we defined the domain and range of our work, as well as some of the features of the ontology.

The scope of Lipoprotein Ontology is focused mainly on the organisation of lipoprotein concepts and coverage of the concepts and theories in the lipoprotein research domain. However, further refinement and evolution of Lipoprotein Ontology could consider other issues such as the adaptation of the ontology to various communities of users.

The domain of this work is the lipoprotein research domain, which was extensively reviewed in Chapter 2. The range of the project starts from the specification phase, followed by the conceptualisation of lipoprotein knowledge, ontology design, formal representation of the conceptual model, as well as evaluation of Lipoprotein Ontology. We specified the areas of lipoprotein knowledge that the ontology aims to represent, and identified the resources and databases from which we extracted knowledge for the development of our ontology (Spyns et al., 2008). For the lipoprotein domain, we used PubMed as the literature database from which we extracted lipoprotein concepts from peer-reviewed journal articles and conference proceedings, as well as other biomedical textbooks and articles concerning lipoproteins. The selected resources must be representative of the domain-related concepts and theories which are used by domain experts and researchers. Lipoprotein Ontology was formalised in OWL and visualised in Protégé 4.2.

Some key features of Lipoprotein Ontology include:

- Evidence-based: The concepts in Lipoprotein Ontology were manually extracted from peer-reviewed journal articles and conference proceedings from PubMed as well as biomedical textbooks.
- Reusability: The concepts and relations in Lipoprotein Ontology should be reusable for purposes beyond those anticipated in the ontology model.
- Extensibility: Lipoprotein Ontology also has to have the potential to evolve to accommodate new concepts that extend well beyond the current concepts. It is imperative to structure the ontology in such a way that allows for legacy use to consider ontology evolution.
- Completeness: Although ontologies operate on the basis of an open world assumption and can therefore never be classified as entirely "complete", Lipoprotein Ontology will attempt to model the lipoprotein domain as described in the literature review.
- Compatibility: Some of the concepts in Lipoprotein Ontology will be adopted from neighbouring ontologies such as Lipid Ontology and Protein Ontology. However, as ontology alignment is a significant research area in itself, it is beyond the scope of this thesis to align entire ontologies with Lipoprotein Ontology. As a compromise, we have reused concepts in our ontology from its neighbouring ontologies whenever possible. This gives us the potential to merge Lipoprotein Ontology and other ontologies for future work.

5.3.2 Literature Review

In the next part of the specification phase, a broad literature survey on lipoproteins was conducted to define the most important lipoprotein concepts. First, we identified the resources and databases from which we can extract knowledge for the development of our ontology. The selected resources must be representative of the domain-related concepts which are used by domain experts and researchers. Therefore, we chose to curate our resources from PubMed, the largest online repository for biomedical literature from MEDLINE, biomedical science journals, and other online books (PubMed, 2013). PubMed is developed and maintained by the National Center for Biotechnology Information (NCBI), at the US National Library of Medicine (NLM).

For our work, we manually extracted lipoprotein concepts from peer-reviewed journal articles and conference proceedings from the PubMed database, as well as other biomedical textbooks and articles concerning lipoproteins. We initially searched for review articles for a general overview of the lipoprotein research domain, using the search term “lipoproteins” in PubMed. The relevant texts and statements from various knowledge resources were selected and recorded, to define the basic concepts for Lipoprotein Ontology. From our review of the literature, we identified the core motivations behind lipoprotein research and subsequently categorised them according into five sub-concepts: *Classification*, *Metabolism*, *Pathophysiology*, *Aetiology* and *Treatment*. These concepts provided the structure for the formulation of competency questions in the next step.

5.3.3 Formulation of Competency Questions

The formulation of competency questions supports the iterative process of knowledge acquisition and also serves as a validation technique for the correctness and consistency of the ontology. These competency questions cover the six sub-ontologies of Lipoprotein Ontology: *Classification*, *Metabolism*, *Pathophysiology*, *Aetiology*, *Treatment* and *Diagnostic Parameter*.

5.4 Conceptualisation

The conceptualisation phase involves 2 sub-phases:

- Defining the domain conceptual model
- Identification of classes, properties and relations

5.4.1 Defining the Domain Conceptual Model

The primary aim of the conceptualisation phase is to create an abstract model of the lipoprotein domain knowledge. In the literature review step of the specification phase, we identified five important concepts as the sub-ontologies under which the corresponding lipoprotein concepts can be categorised: *classification of lipoproteins*, *lipoprotein metabolism pathways*, *pathophysiology of lipoprotein dysregulation (dyslipidaemia)*, *aetiology (causes) of dyslipidaemia*

and *treatment of dyslipidaemia*. These concepts constitute the initial framework for the lipoprotein conceptual model.

We developed the initial lipoprotein conceptual model using the multi-axial approach (Wingert et al., 1989) to represent Lipoprotein Ontology based on five overarching sub-ontologies: *Classification*, *Metabolism*, *Pathophysiology*, *Aetiology* and *Treatment*, under which the associated concepts were identified and categorised in a hierarchical structure. We presented this work at the 22nd IEEE International Symposium on Computer-Based Medical Systems in August, 2009 (Chen & Hadzic, 2009). During this time, we evaluated our ontology and have since added one more sub-ontology *Diagnostic Parameter* to incorporate physiological measures for diagnostic purposes.

We extended the initial conceptual model by extracting the concepts associated with the five key lipoprotein concepts. The search terms that were used to extract these concepts were: “classification of lipoproteins”, “lipoprotein metabolism”, “lipoprotein disorders”, “dyslipidaemia”, “causes”, “treatment”. This was carried out via the same literature review process described in the specification phase: by manually extracting these concepts from peer-reviewed journal articles and conference proceedings from the PubMed database, as well as other biomedical textbooks and articles.

Simultaneously, we also reviewed the literature for various ontology building methodologies, as well as other established ontologies to structure our conceptual model in the most appropriate manner. At the early stages of concept building, it is imperative to develop the basic ontological framework towards modularity, for the purposes of reusability, maintenance and evolution (Rector, 2003). This can be done through the process of normalisation, which is a common procedure in databases for similar reasons. In developing our conceptual model, we adopted some features of the normalisation process according to (Rector, 2003). The key requirements of this process include:

- The concepts to be reused can be identified and separated from the whole taxonomy.
- Maintenance can be split amongst authors who can work independently.

- Concepts can evolve independently and new concepts can be added with minimal side effects.
- The differences between various categories of information are represented explicitly for both human understanding and formal machine inference.

With respect to these requirements, we define the Lipoprotein Domain Concept to include two general upper concepts:

- Self-standing concepts: Broadly, self-standing concepts are “things” in the physical and conceptual world. By the same token, self-standing concepts of the Lipoprotein Domain Concept are “things” in the lipoprotein knowledge domain. Hence, we categorise the six lipoprotein sub-ontologies discussed in the beginning of this section under the general category of Self Standing Concept.
- Refining concepts: Value types and values which partition conceptual spaces e.g. “small, medium, large”, “mild, moderate, severe”, etc. For refining concepts:

  a) There should be a taxonomy of primitive “value types” which is disjoint;

  b) The primitive children of each value type should form a disjoint exhaustive partition, i.e. the values should “cover” the “value type”.

At this stage, we also considered the use of these ontologies for other purposes such as enrichment of our ontology and ontology merging for future work (Noy & McGuinness, 2001; Spyns et al., 2008).

We elaborate on this conceptual model and provide justifications for our choice of ontology structure in Chapter 6.

5.4.2 Identification of Classes, Properties and Relations

In the next step of the conceptualisation phase, we identified and described hundreds of concepts, their definitions and binary relationships between them in the form of theoretical axioms, using a Word document. From this, we defined

lipoprotein concepts as classes, arranged in a hierarchical structure under their respective sub-ontologies. The class hierarchy was developed by means of a combination of top-down and bottom-up classification strategies (Noy & McGuinness, 2001). Using a top-down approach, we initially began with the creation and description of the six sub-ontologies defined in the previous stage. Some of the relevant middle-level concepts then fall under their respective top-level concepts via a top-down method. Others were formed using a bottom-up approach through which we first identified several specific concepts and then abstracted a more general and representative concept for them.

Having established lipoprotein classes in a hierarchical structure, we subsequently assigned properties (relationships) to the corresponding lipoprotein classes (Noy & McGuinness, 2001; Spyns et al., 2008). The resultant model of classes and their relationships appear in the form of theory statements such as “A partOf B” or “X hasProperty Y”. As a result, relations between these classes are described in an unambiguous way and the possible values are filled. It must be reiterated that the process of knowledge acquisition occurs throughout the four stages, hence as new classes are identified, the associated properties and relations must be filled accordingly. We elaborate on this process in Chapter 6.

5.5 Formalisation

The formalisation phase involves 2 sub-phases:

- Formalisation of the conceptual model
- Creation of instances

5.5.1 Formalisation of the Conceptual Model

To formalise the lipoprotein conceptual model from the previous phase, it is necessary to implement the ontology using an ontology language and its corresponding tool. As we have previously defined in Chapter 4.2, “formalisation” refers to the translation of the conceptualised knowledge into a machine-readable and formal language. Lipoprotein Ontology is represented by OWL as the ontology language due to its superiority compared to its preceding

languages (Grau et al., 2008; Horrocks et al., 2003). As an extension of RDF and DL, OWL balances expressivity with reasoning, which makes it the most suitable language for the purpose of Lipoprotein Ontology. OWL is the language used by prominent ontology editors, including Protégé. Most importantly, OWL is the standardised language of the Semantic Web.

Ontology tools are capable of translating the concept definitions and descriptions into a predefined ontology language in an automatic way. We opted to develop Lipoprotein Ontology using the most recent version of the Protégé ontology editor at the writing of this thesis, Protégé 4.2. This tool allows us to store the specified concepts in a class hierarchy and facilitates the description and definition of their properties, constraints and describes the relationships with other concepts. The Protégé tool automatically translates concept definitions and descriptions into the formal OWL language. For example, the formal representation of "**Chylomicron** is a subclass of LipoproteinEntity and has component of at least one **Cholesterol**" appears as:

```
<owl:Class rdf:about="http://www.owl-ontologies.com/lipoprotein_ontology.owl
#Chylomicron">
  <rdfs:subClassOf rdf:resource="http://www.owl-
ontologies.com/lipoprotein_ontology.owl#LipoproteinEntity"/>
    <rdfs:subClassOf
        <owl:Restriction>
            <owl:onProperty
rdf:resource="http://www.owl-
ontologies.com/lipoprotein_ontology.owl#hasComponent">
            <owl:someValuesFrom
rdf:resource="http://www.owl-
ontologies.com/lipoprotein_ontology.owl#Cholesterol"/>
        </owl:Restriction>
    </rdfs:subClassOf>
</owl:Class>
```

Further details about the formalisation process of the conceptual framework are presented in Chapter 7.

5.5.2 Creation of Instances

In this step, we populate the ontology by creating instances for their given classes. Due to the extensive nature of Lipoprotein Ontology, we will be creating instances for the purpose of evaluation. We refer to our list of competency questions as a guide in this process.

5.6 Evaluation

After the ontology has been developed, it must be evaluated for:

- **Completeness**: The concepts and relationships are explicitly stated and each definition is complete.
- **Consistency**: The definitions are consistent and do not include contradictory information. In addition, inferences need to be consistent with existing definitions and axioms, and should be clear and logical.
- **Conciseness**: The ontology does not store any unnecessary definitions.
- **Extensibility**: Users can add new definitions to the ontology and more knowledge to its definitions without altering the set of well-defined properties.
- **Minimal encoding bias**: Conceptualisations need to be specified at knowledge-level and not at a symbol or notation level. As such, the proper use of relations is necessary in order to maintain the integrity of the ontology. For example, for the superclass **LPMetabolism**, although it may appear to be more convenient to use the subclass relation “isA” instead of “partOf” expression, the “isA” relation is not the correct relation for its respective subclasses **PhysicalEntity**, **OccurringEntity** and **ParticipantRole**. Rather, we used the “partOf” relation in order to maintain the correct formal definition for their corresponding axioms.

With respect to these criteria, the evaluation phase involves 3 sub-phases:

- Concept coverage
- Validation of competency questions
- Evaluation using case studies

5.6.1 Concept Coverage

A test set, which consists of randomly selected abstracts of published papers from publication databases, can be used to evaluate the conceptual coverage of the designed ontology. The developed ontology is used to encode knowledge from this test set. Conceptual coverage evaluation can then be carried out by calculating the percentage of domain concepts, which are covered or represented by the ontology.

5.6.2 Validation of Competency Questions

The first step of the evaluation phase involved testing the complete Lipoprotein Ontology against the competency questions that were defined at the initial specification stage of ontology development. This was carried out in order to check that the model has successfully represented relationships present in the initial documents or definitions. How well does the model perform when it is faced with information that is not explicitly in the scope of its design? I.e. What inferences can we draw from it?

5.6.3 Evaluation using Case Studies

The established Lipoprotein Ontology was then evaluated using the concept coverage criterion (Hartmann et al., 2005); we developed two case studies or scenarios to evaluate whether the ontology represents the majority of lipoprotein-related concepts used in the literature. The lipoprotein-related concepts abstracted from these scenarios were given to the ontology tool. Then we calculated the percentage of concepts within the scenarios that had equal or similar concepts in Lipoprotein Ontology. In the later stages, new concepts will be added and created axioms will be further refined to ensure that the ontology meets the reusability, consistency, clarity coherence, minimal encoding bias, minimal ontological commitment, simplicity and correctness criteria (Brank et al., 2005). The process of conceptual coverage method and our evaluation results are described in Chapter 8.

5.7 Conclusion

In this chapter, we described the methodology used to develop Lipoprotein Ontology. This methodology incorporated different stages of other existing methodologies, and included the following processes: specification, conceptualisation, formalisation and evaluation. In the next chapter, we present the conceptual framework of Lipoprotein Ontology.

References

Brank J., Grobelnik M., & Mladenić, D. (2005). *A survey of ontology evaluation techniques.* Paper presented at the Proceedings of the Conference on Data Mining and Data Warehouses (SiKDD 2005).

Chen, M., & Hadzic, M. (2009). *Lipoprotein ontology as a functional knowledge base.* Paper presented at the Proceedings of the 22nd IEEE International Symposium on Computer-Based Medical Systems (CBMS 2009), New Mexico, USA.

Grau, B. C., Horrocks, I., Motik, B., et al. (2008). OWL 2: the next step for OWL. *Web Semantics: Science, Services and Agents on the World Wide Web, 6*(4), 309–322.

Hartmann, J., Sure, Y., Giboin, A., et al. (2005). Methods for ontology evaluation *Knowledge web project deliverable*

Horrocks, I., Patel-Schneider, P. F., & van Harmelen, F. (2003). From SHIQ and RDF to OWL: the making of a web ontology language. *Web Semantics: Science, Services and Agents on the World Wide Web, 1*(1), 7–26.

Noy, N. F., & McGuinness, D. L. (2001). Ontology development 101: a guide to creating your first ontology, from http://protege.stanford.edu/publications/ontology_development/ ontology101-noy-mcguinness.html

PubMed. (2013), from http://www.ncbi.nlm.nih.gov/pubmed

Rector, A. L. (2003). *Modularisation of domain ontologies implemented in description logics and related formalisms including OWL.* Paper presented at the 2nd International Conference on Knowledge Capture, Florida, USA.

Spyns, P., Tang, Y., & Meersman, R. (2008). An ontology engineering methodology for DOGMA *Journal of Applied Ontology, 3*(1-2), 13-39.

Chapter 6. Conceptual Framework of Lipoprotein Ontology

6.1 Introduction

This chapter presents the conceptual framework for the representation of lipoprotein domain knowledge, and elaborates on some of the upper level concepts which are represented in Lipoprotein Ontology. Building on the literature review on lipoproteins we have thoroughly conducted in Chapter 2, we hereby present lipoprotein-related concepts in a structured hierarchy according to ontology principles defined in Chapter 5.4.1. The corresponding properties and restrictions of these concepts will be elucidated in Chapter 7 during the process of formalisation. Lipoprotein Ontology was developed and then implemented with Protégé 4.2. The visualisation of the lipoprotein conceptual framework and concept hierarchies will be presented in this chapter accordingly as figures.

6.2 Overview of of Lipoprotein Ontology

In this section, we describe the structure of Lipoprotein Ontology and outline the upper concepts that will be discussed in the remainder of the chapter. We refer to Chapter 4.6 for the definitions of these concepts in Lipoprotein Ontology. We will include the membership criteria to each class in an informal manner as part of the conceptualisation process, in order to organise lipoprotein concepts into their corresponding upper concepts. The attributes and relations of these concepts will be represented as a formal framework in the next chapter.

First and foremost, we reiterate the definition of ontology as “the formal, explicit specification of a shared conceptualisation of a domain.” Therefore, it is important that we clearly establish our domain of interest at the initial stage of

ontology development. We have done this by creating two major subclasses under the main ontology concept **Thing**, which represents the class of all things (Figure 10):

- **LipoproteinDomainConcept**: A set of lipoprotein concepts and relations that makes up the lipoprotein domain knowledge
- **MetaConcept***:* All other concepts that do not belong to the lipoprotein domain concept.

Having defined **LipoproteinDomainConcept** as the class in which a set of lipoprotein concepts and relations are structured under, our next step was to organise the domain conceptual framework into two subclasses for the purpose of modularity, as discussed in Chapter 5.4.1. In order to preserve the integrity of the two upper concepts to the **LipoproteinDomainConcept**, we attached the prefix "LP" to these concepts to denote their unique associations with the lipoprotein domain. Thus, the two subclasses of **LipoproteinDomainConcept** are represented as follows (Figure 10):

- **LPSelfStandingConcept**: "Things" or concepts in the lipoprotein domain domain.
- **LPRefiningConcept**: Value types and values which partition conceptual spaces e.g. "small, medium, large", "mild, moderate, severe", etc.

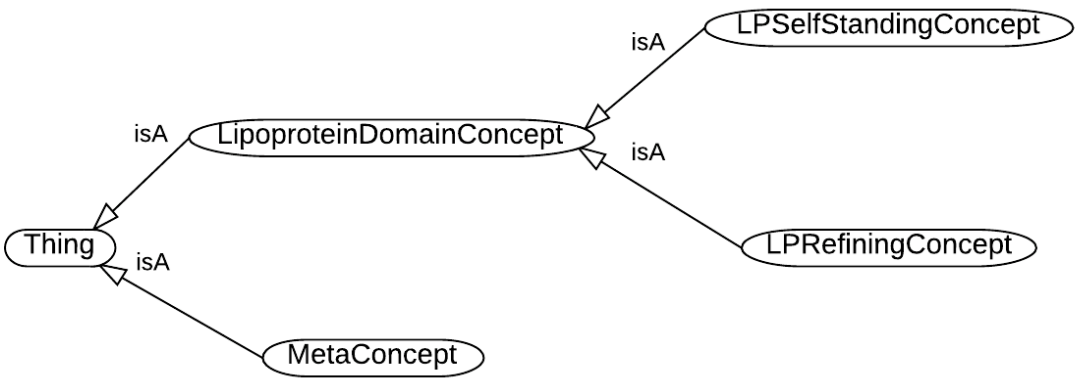


Figure 10. Upper concepts of Lipoprotein Ontology, *LipoproteinDomainConcept* and *MetaConcept* with parent *Thing*. *LipoproteinDomainConcept* contains two subclasses *LPSelfStandingConcept* and *LPRefiningConcept*.

Since **LPSelfStandingConcept** is the class in which concepts in the lipoprotein knowledge domain are organised under, it serves as the main conceptualisation of the lipoprotein domain knowledge, which can be represented by six sub-ontologies (Figure 11). Similarly, we attached the prefix "LP" to these concepts to denote their unique associations with the lipoprotein domain.

- **LPClassification**
- **LPMetabolism**
- **LPPathophysiology**
- **LPAetiology**
- **LPTreatment**
- **LPDiagnosticParameter**

Lipoprotein concepts and relations are organised under the corresponding sub-ontologies, which we will discuss in the subsequent sections. These concepts are not mutually exclusive, and can be reused by more than one sub-ontology. This notion is particularly in alignment with the nature and role of lipoprotein components in the lipoprotein metabolism pathway. For example, the concept **Cholesterol** is used in two sub-ontologies: **Metabolism** and **Classification**. **Cholesterol** is a **Sterol** component classified under **PhysicalEntity**, which is a concept with partOf relation to **Metabolism**. At the same time, **Cholesterol** is also categorised in the **Classification** sub-ontology as one of the critical components of all classes of lipoproteins, collectively known as **LipoproteinEntity**, i.e. **Chylomicron**, **VLDL**, **LDL**, **IDL** and **HDL**.

LPRefiningConcept is the top level concept which classifies value types and values which partition conceptual spaces. **LPRefiningConcept** for Lipoprotein Ontology covers the following concepts (Figure 11):

- **LPValuePartition**
- **LPQualitativeMeasure**
- **LPInheritanceType**

Each of these concepts can be categorised further into their respective subclasses, and related to other concepts in the ontology, which we will discuss in the next chapter.

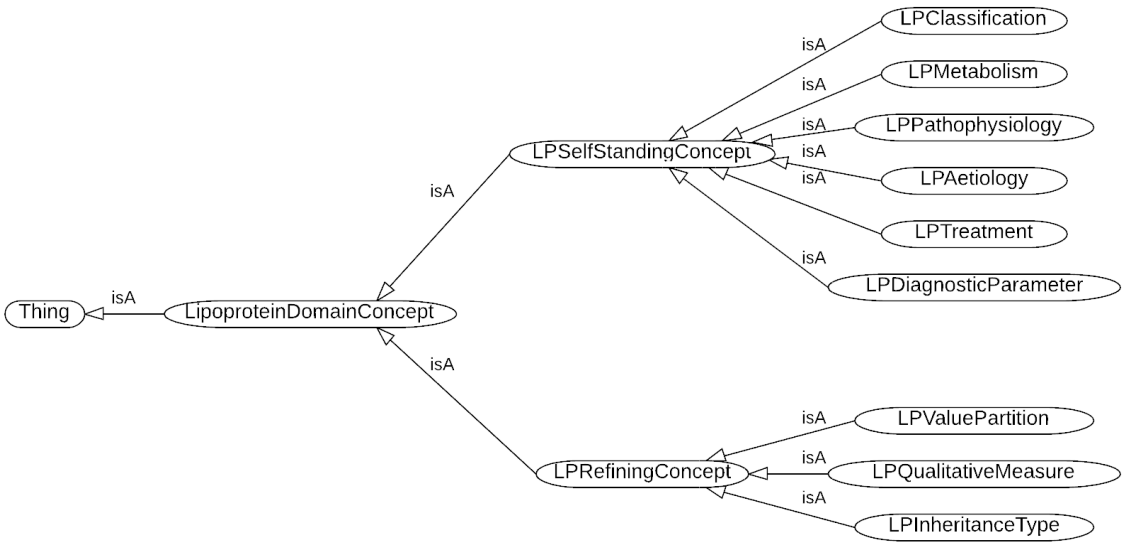


Figure 11. *LipoproteinDomainConcept* represented by two major subclasses, *LPSelfStandingConcept* and *LPRefiningConcept*, with their respective subclasses.

6.3 Classification

"Definition: Classification is defined as the classification of Lipoprotein Entity. Lipoprotein Entity is defined as a soluble complex of lipids and proteins. Depending on the lipid and protein content, lipoprotein entity are classified accordingly into various subclasses."

From the above, the sub-ontology **LPClassification** is defined as the classification of **LipoproteinEntity**. Respectively, **LipoproteinEntity** is defined as a soluble complex of lipids and proteins. **LipoproteinEntity** defines lipoprotein classes as follows (Figure 12):

- **Chylomicron**
- **ChylomicronRemnant**
- **VeryLowDensityLipoprotein**
- **IntermediateDensityLipoprotein**
- **LowDensityLipoprotein**
- **HighDensityLipoprotein**
- **Lipoprotein(a)**

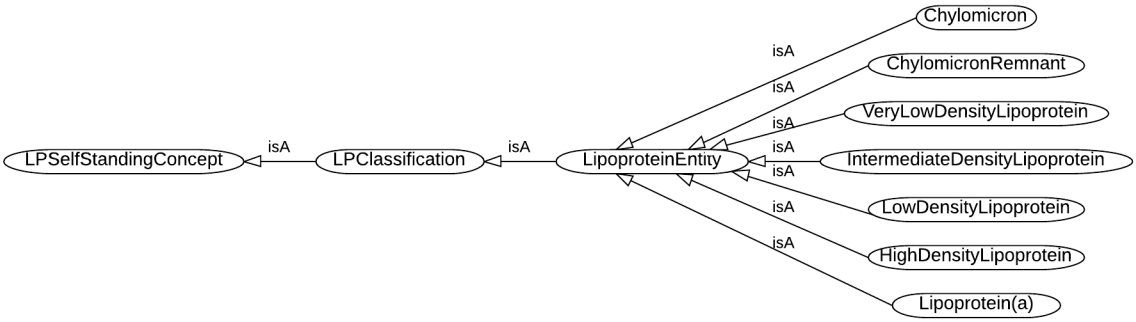


Figure 12. Sub-ontology of *LPClassification* with parent *LPSelfStandingConcept* and subclass *LipoproteinEntity*, with its respective subclasses.

Lipoprotein classes are categorised accordingly under **LipoproteinEntity** based on several criteria that confer membership into the class: content, electrophoretic mobility and function. These attributes will be discussed further in Chapter 7 in formal language.

Content

As defined in Chapter 2.2.1, **LipoproteinEntity** is a soluble complex of lipids and proteins. Within the circulation, these components are in a state of constant flux, changing in composition and structure as lipoproteins are metabolised via various lipoprotein metabolism pathways. As lipoprotein contents are constantly being passed from one to another during metabolism, the distinction between classes is somewhat variable. Depending on the lipid and protein content, lipoprotein entities are classified accordingly into various types. Table 1 from Chapter 2.2.1 (Appendix A) summarises the lipid content and composition that constitute the major lipoprotein classes, **Chylomicron**, **ChylomicronRemnant**, **VeryLowDensityLipoprotein**, **IntermediateDensity Lipoprotein**, **LowDensityLipoprotein**, **HighDensity Lipoprotein**, **Lipoprotein(a)**.

Electrophoretic Mobility

Because the lipoprotein divisions by density are arbitrary and vary somewhat within their own classes, they can also be classified according to their electrophoretic mobility on agarose gels into α, pre-β and β lipoproteins, corresponding to HDL, VLDL and LDL lipoprotein fractions respectively. Chylomicrons remain at the electrophoretic origin (Snyder, 1977).

Function

The functions of the different lipoprotein classes are determined by their lipid and apolipoprotein components. Chylomicrons are synthesised in the intestines for the transport of dietary triglycerides to various tissues. VLDL are synthesised in the liver for the export of endogenous triglycerides, while IDL and LDL are derived from the metabolism of VLDL in circulation. The function of LDL is to deliver cholesteryl esters to the liver and peripheral tissues. HDL are synthesised and assembled in the liver and intestines or formed from the metabolism of other lipoproteins in circulation, and from cellular lipids at the cell membranes. HDL remove excess cholesterol from cells and transport it to liver and other tissues for metabolism and excretion (Chan et al., 2004).

6.4 Metabolism

"Definition: Lipoprotein Metabolism is defined as the chemical processes which involve the synthesis and/or catabolism of Lipoprotein Entity."

Up to this point, concept hierarchies were represented as hierararchical *is-a* relationships. For the concept of **LPMetabolism**, we represent the associative relationship between **LPMetabolism** and its corresponding concepts using the *partOf* relation. In Figure 13 below, we show the representation of the *isA* and *partOf* relations using different arrows. We will discuss the difference between these relations in the next chapter.

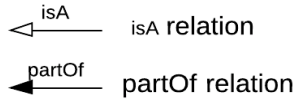


Figure 13. Representation of hierarchical *isA* and associative *partOf* relations using empty and solid arrows, respectively.

Understanding the metabolism of lipoproteins involves a clear elucidation of the components that participate in the lipoprotein metabolic pathways. This includes separating components from pathways, processes from location. The conceptualisation of the sub-ontology **LPMetabolism** can be categorised into three classes (Figure 14):

- **PhysicalEntity**
- **OccurringEntity**
- **ParticipantRole**

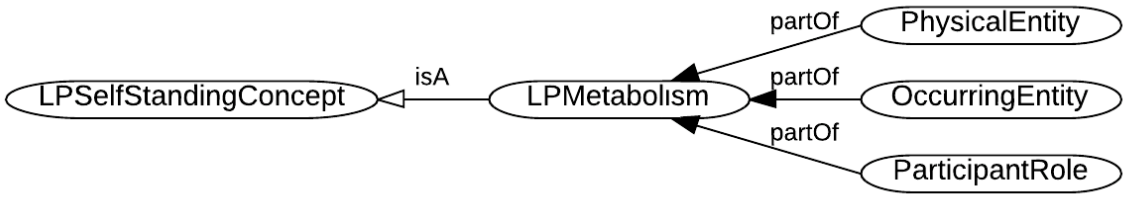


Figure 14. Sub-ontology of *LPMetabolism* with parent *LPSelfStandingConcept* and subclasses *PhysicalEntity, OccurringEntity* and *ParticipantRole.*

The upper concepts which belong to the **LPMetabolism** sub-ontology are derived from Systems Biology Ontology (Courtot et al., 2011), with some of the concept definitions referenced from MeSH (Lowe & Barnett, 1994). These concepts will be discussed in detail in the following sections.

6.4.1 Physical Entity

"Definition: Physical Entity is defined as the representation of an entity that may participate in an interaction, a process or a relationship of significance."

PhysicalEntity can be further categorised into two classes (Figure 15):

- **FunctionalEntity**
- **StructuralEntity**

FunctionalEntity is a real thing involved in lipoprotein metabolism, defined by its properties or the function it performs. The list below is not fully exhaustive of all existing functional entities; for the purpose of this thesis, **FunctionalEntity** includes the following subclasses (Figure 15):

- **Protein**
- **Lipid**
- **CoEnzyme**
- **Enzyme**
- **Neurotransmitter**

- **FreeRadical**
- **Receptor**
- **Metabolite**
- **Hormone**
- **UnitOfGeneticInformation**

StructuralEntity is a real thing, defined by its physico-chemical structure, and can be further divided into two classes (Figure 15):

- **MacroOrganicStructure**
- **MicroOrganicStructure**

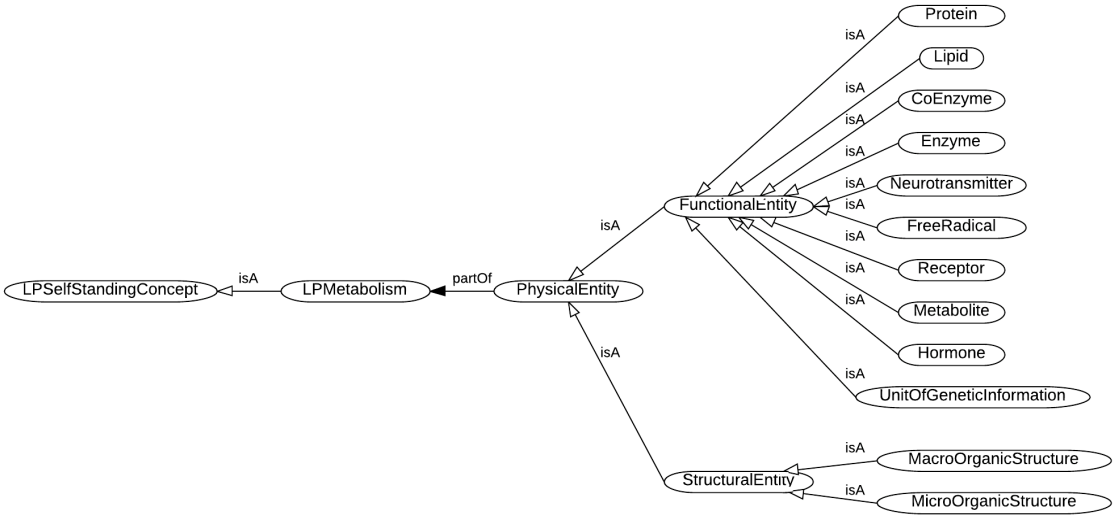


Figure 15. Concept of *PhysicalEntity* with parent *LPMetabolism* and subclasses *FunctionalEntity* and *StructuralEntity*, with their respective subclasses.

6.4.2 Occurring Entity

"Definition: Occurring Entity is defined as the representation of an entity that manifests, unfolds or develops through time, such as a discrete event, or a mutual or reciprocal action or influence that happens between participating physical entities, and/or other occurring entities."

OccurringEntity can be further categorised into two classes:

- **Process**
- **Pathway**

Process is defined as the representation of the action which brings about a change from one entity to another entity. In lipoprotein metabolism, these processes include but are not exclusive to (Figure 16):

- **Absorption**
- **Acylation**
- **Assembly**
- **Catabolism**
- **Clearance**
- **Esterification**
- **GeneTranscription**
- **Hydrolysis**
- **Inhibition**
- **Oxidation**
- **Phagocytosis**
- **Secretion**
- **Synthesis**
- **Transfer**
- **Transport**
- **Uptake**

Pathway is defined as the representation of a sequence of reactions which occurs whereby the products of one reaction are the substrates for subsequent reactions. In Chapter 2.2.2, we have identified three major pathways in lipoprotein metabolism as: 1. **ExogenousPathway**, dealing with the transport and metabolism of dietary lipids; 2. **EndogenousPathway**, dealing with the lipids that are synthesised in the liver, and; 3. **ReverseCholesterolTransport**, dealing with the synthesis and metabolism of HDL. However, in our ontology, we also include other pathways which are significant in the metabolism of various lipoprotein components. For example, we have included the **HMG-CoAReductasePathway** concept as it is one of the target treatment pathways for dyslipidaemia. Therefore, the **Pathway** class includes the following concepts, among others (Figure 16):

- **ExogenousPathway**
- **EndogenousPathway**

- **ReverseCholesterolTransport**
- **Gluconeogenesis**
- **Glycerol-3-PhosphatePathway**
- **Glyceroneogenesis**
- **Glycolysis**
- **HMG-CoAReductasePathway**
- **MonoacylglycerolPathway**

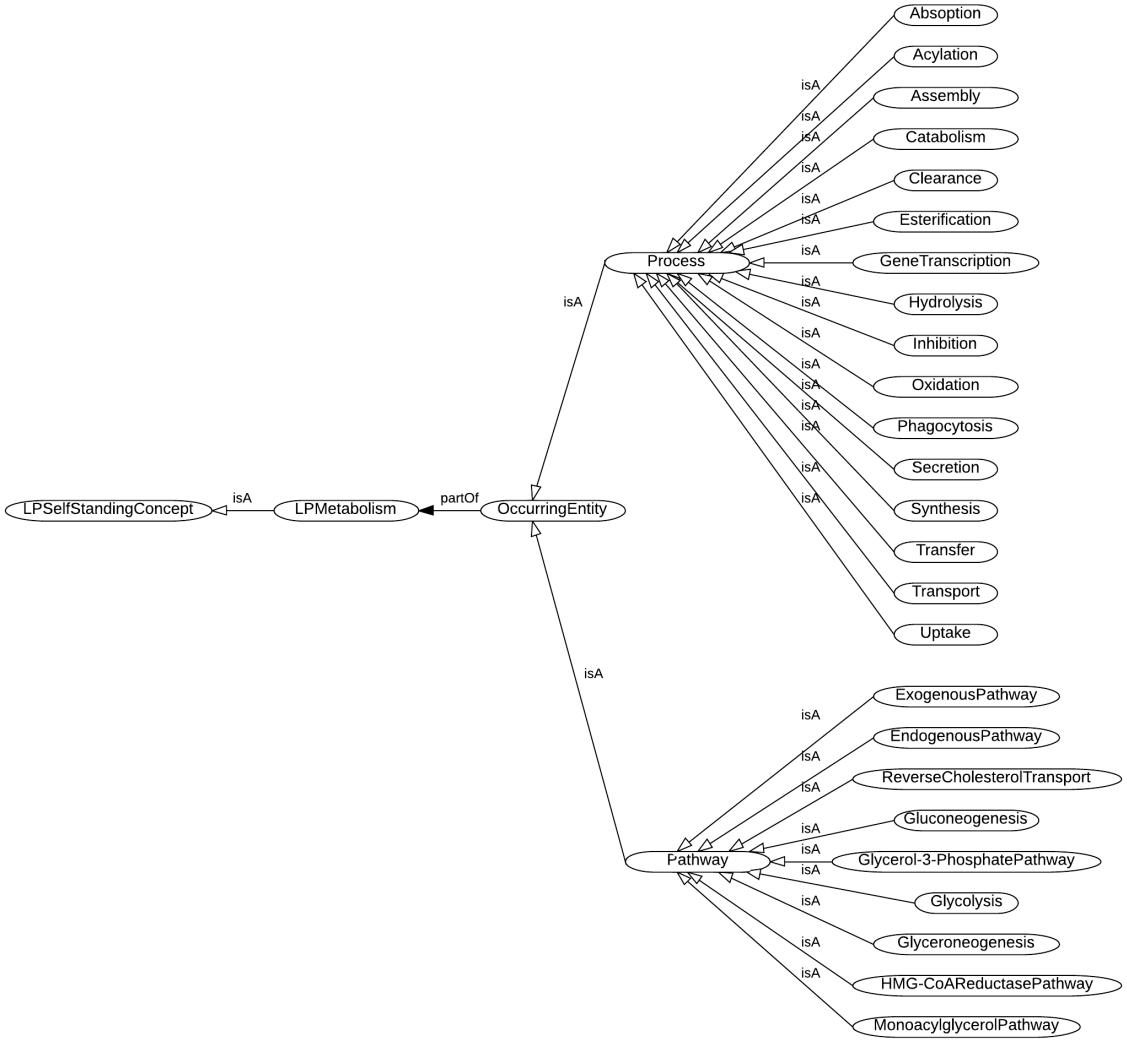


Figure 16. Concept of *OccurringEntity* with parent *LPMetabolism* and subclasses *Process* and *Pathway*, with their respective subclasses.

6.4.3 Participant Role

"Definition: Participant Role is defined as the function of a physical or conceptual entity, that is its role, in the execution of an event or process."

In order to assign roles to lipoprotein components, it is necessary to build a concept for this purpose. Therefore the **ParticipantRole** describes the function of various lipoprotein components and can be further categorised into six classes (Figure 17):

- **Reactant**
- **Product**
- **Modifier**
- **FunctionalCompartment**
- **ActionRole**
- **ActionResponse**

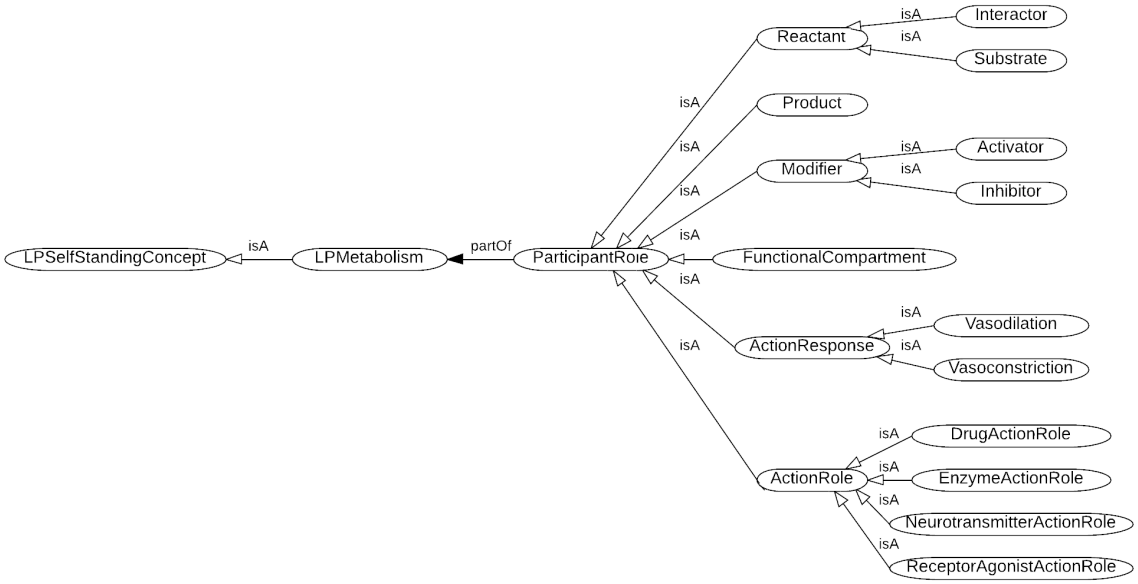


Figure 17. Concept of *ParticipantRole* with parent *LPMetabolism* and subclasses *Reactant, Product, Modifier*, *FunctionalCompartment, ActionRole* and *ActionResponse* with their respective subclasses.

6.5 Pathophysiology

"Definition: Pathophysiology is defined as the pathophysiology related to the dysregulation in lipoprotein disorders."

Dysregulation in lipoprotein metabolism occurs as a consequence of alterations in the kinetics of lipoproteins. To classify the sub-ontology **LPPathophysiology**, we separate concepts of the actual disorder from their physical indication into two classes using the associative *partOf* relationship (Figure 18).

- **Disorder**
- **Symptom**

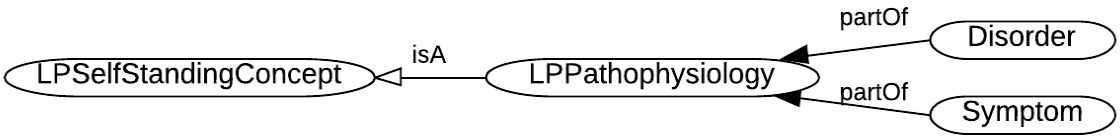


Figure 18. Sub-ontology of *LPPathophysiology* with parent *LPSelfStandingConcept* and subclasses *Disorder* and *Symptom.*

6.5.1 Disorder

"Definition: Disorder is defined as disease or abnormality of function. Examples of lipoprotein disorder are: Cardiovascular Disease, Diabetes."

Disorder associated with lipoprotein dysregulation can be further categorised into eight classes (Figure 19). Again, it should be noted that this list is not fully exhaustive of all disorders associated with lipoprotein dysregulation and can be extended to include other concepts.

- **MetabolicSyndrome**
- **InsulinResistance**
- **LiverDisorder**
- **RenalDisorder**
- **LipidDisorder**
- **ThyroidDisease**
- **Diabetes**
- **CardiovascularDisease**

As a number of definitions have been proposed to diagnose **MetabolicSyndrome**, we have included the most common definitions that are currently used in practice.

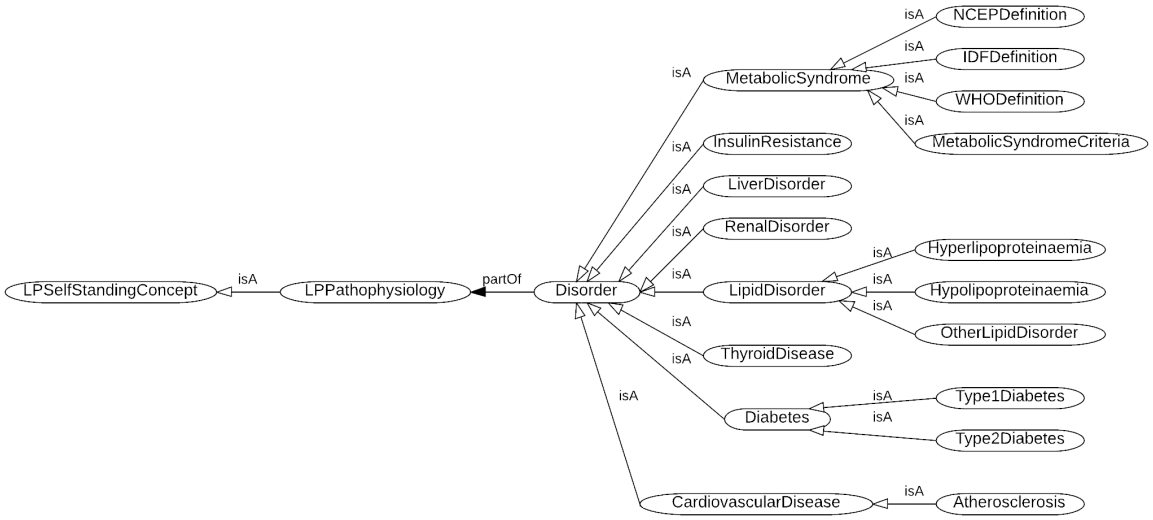


Figure 19. Concept of *Disorder* with parent *LPPathophysiology* and subclasses *MetabolicSyndrome*, *InsulinResistance*, *LiverDisorder*, *RenalDisorder*, *LipidDisorder*, *ThyroidDisease*, *Diabetes* and *CardiovascularDisease*, and their respective subclasses.

6.5.2 Symptom

"Definition: Symptom is defined as the subjective indication of a disorder or disease."

Symptom associated with lipoprotein dysregulation can be further categorised into following classes, among others (Figure 20):

- **Dyslipidaemia**
- **CentralObesity**
- **Hyperglycaemia**
- **Hypertension**
- **Hyperinsulinaemia**
- **Hyperuricaemia**
- **Microalbuminuria**

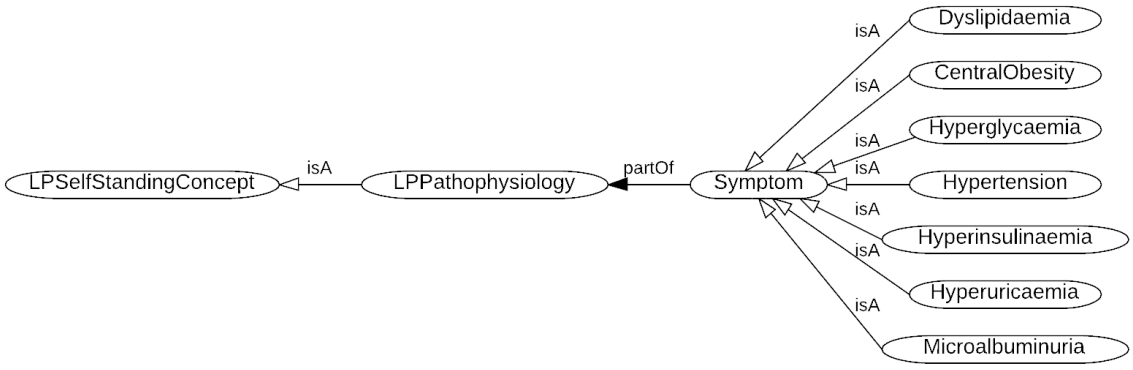


Figure 20. Concept of *Symptom* with parent *LPPathophysiology*, with subclasses *Dyslipidaemia*, *CentralObesity*, *Hyperglycaemia*, *Hypertension*, *Hyperinsulinaemia*, *Hyperuricaemia* and *Microalbuminuria*.

6.6 Aetiology

"Definition: Aetiology is defined as the causes of lipoprotein disorder."

There are several contributing factors to dyslipidaemia. Dyslipidaemia can be the result of lifestyle, the interaction between genetic predisposition and environmental factors, or other diseases, or a combination of both. Drug interactions may also contribute to lipoprotein dysregulation. Therefore, **LPAetiology** can be categorised into three classes (Figure 21):

- **LifestyleCause**
- **GeneticCause**
- **DiseaseStateCause**
- **DrugInteraction**

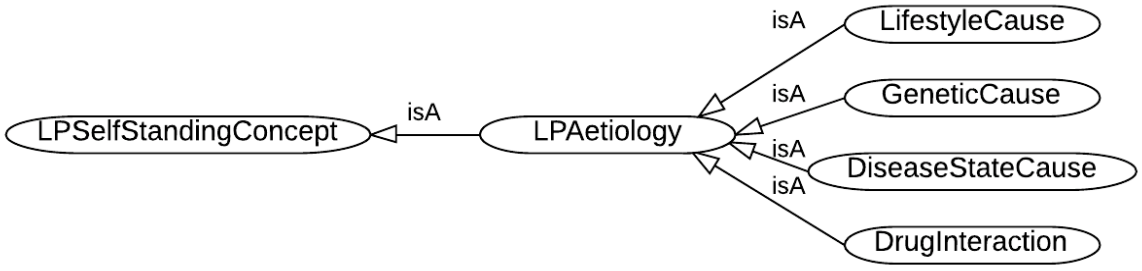


Figure 21. Sub-ontology of *LPAetiology* with parent *LPSelfStandingConcept* and subclasses *LifestyleCause*, *Genetic*, *DiseaseStateCause* and *Drug Interaction.*

6.6.1 Lifestyle Cause

"Definition: Lifestyle is defined as the causes of lipoprotein disorder due to lifestyle factors."

LifestyleCause of lipoprotein dysregulation can be further categorised into five classes(Figure 22). This list is not fully exhaustive of all lifestyle causes associated with dyslipidaemia and can be extended to include other concepts.

- **Smoking**
- **Alcohol**
- **Diet**
- **Exercise**
- **EmotionalWellbeing**

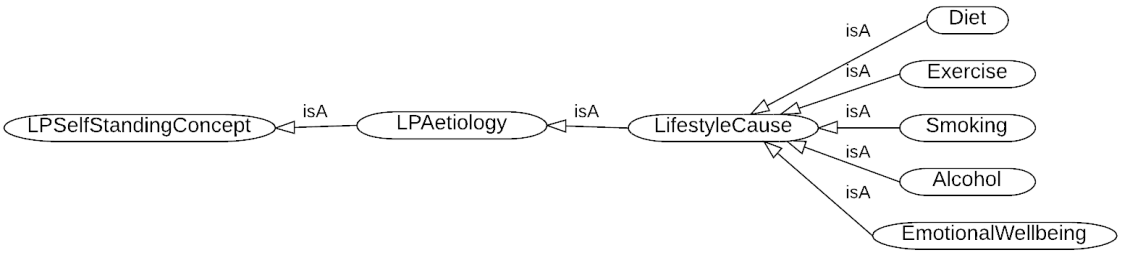


Figure 22. Concept of *LifestyleCause* with parent *LPAetiology* and subclasses *Diet*, *Exercise*, *Smoking*, *Alcohol* and *EmotionalWellbeing*.

6.6.2 Genetic Cause

"Definition: Genetic Cause is defined as the causes of lipoprotein disorder due to hereditary factors."

GeneticCause is further categorised into two types (Figure 23):

- **Defect**
- **Deficiency**

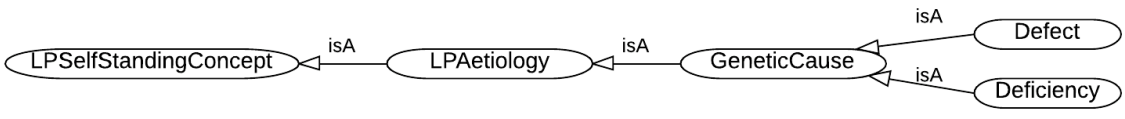


Figure 23. Concept of *Genetic* with parent *LPAetiology* and subclasses *Defect* and *Deficiency.*

6.6.3 Disease State Cause

"Definition: Disease State Cause is defined as the causes of lipoprotein disorder due to the disease states."

DiseaseStateCause is the concept given for disease states which cause lipoprotein disorders and will be explained further in the next chapter (Figure 24).

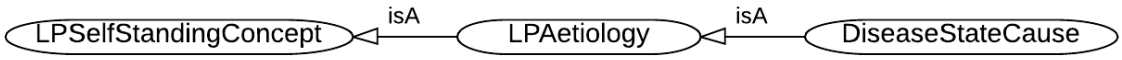


Figure 24. Concept of *DiseaseStateCause* with parent *LPAetiology*.

6.6.4 Drug Interaction

"Definition: Drug Interaction is defined as the causes of lipoprotein disorder due to the action of other drugs"

DrugInteraction is the concept given for the interaction of drug classes which causes lipoprotein disorders and will be explained further in the next chapter (Figure 25).

Figure 25. Concept of *DrugInteraction* with parent *LPAetiology*.

6.7 Treatment

"Definition: Treatment is defined as the treatment of lipoprotein disorder."

Treatment options to correct dyslipidaemia include drugs such as statins, fibrates, bile acid sequestrants and cholesterol absorption inhibitors, as well as lifestyle changes, or a combination of both. Therefore, **LPTreatment** can be further categorised into three classes, among others (Figure 26):

- **LifestyleChange**
- **Pharmacotherapy**
- **CombinationTherapy**

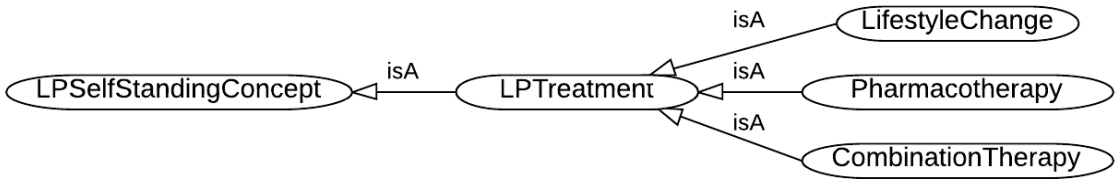


Figure 26. Sub-ontology of *LPTreatment* with parent *LPSelfStandingConcept* and subclasses *LifestyleChange*, *Pharmacotherapy* and *CombinationTherapy.*

6.7.1 Lifestyle Change

"Definition: Lifestyle Change is defined as the treatment of lipoprotein disorder with changes in lifestyle."

Examples of these **LifestyleChange** concepts include but are not exclusive to (Figure 27):

- **DietarySupplement**
- **PhysicalExercise**
- **WeightLoss**

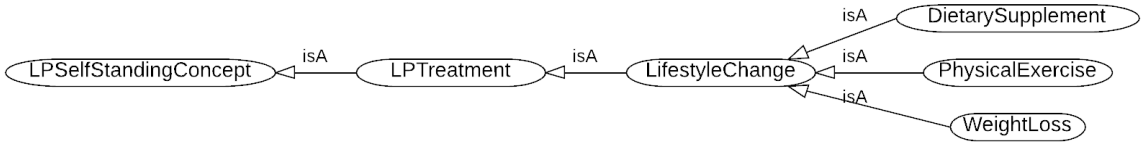


Figure 27. Concept of *LifestyleChange* with parent *LPTreatment* and subclasses *DietarySupplement*, *PhysicalExercise* and *WeightLoss.*

6.7.2 Pharmacotherapy

"Definition: Pharmacotherapy is defined as treatment of lipoprotein disorder with pharmacotherapy."

Pharmacotherapy of lipoprotein dysregulation include but are not exclusive to (Figure 28):

- **Statin**
- **Fibrate**
- **NicotinicAcid**
- **BileAcidSequestrant**

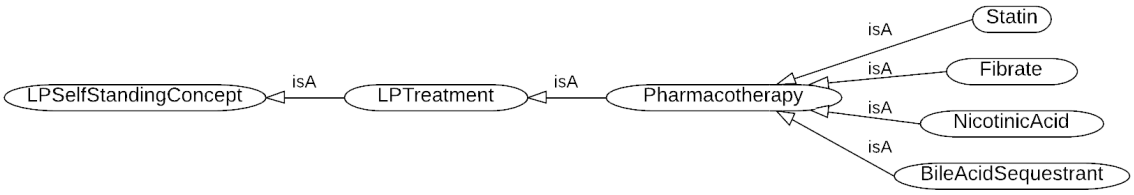


Figure 28. Concept of *Pharmacotherapy* with parent *LPTreatment* and subclasses *Statin*, *Fibrate*, *NicotinicAcid* and *BileAcidSequestrant*.

6.7.3 Combination Therapy

"Definition: Combination Therapy is defined as the treatment of lipoprotein disorder with a combination of treatment options."

CombinationTherapy is the concept given for the combination of treatment options for lipoprotein dysregulation (Figure 29).

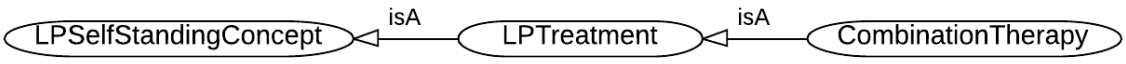


Figure 29. Concept of *CombinationTherapy* with parent *LPTreatment*.

6.8 Diagnostic Parameters

"Definition: Diagnostic Parameter is defined as the requirement of diagnosis specific to the characteristics of an individual."

In clinical practice, dyslipidaemia can be diagnosed by measuring lipoprotein components, and comparing these levels to established guidelines according to the gender and geographic ancestry of an individual. Since these concepts have been classified respective to their categories as discussed previously, the concept **LPDiagnosticParameter** includes the requirement of diagnosis specific to the characteristics of an individual. These are (Figure 30):

- **Gender**
- **Ethnicity**
- **GeographicAncestry**
- **Patient**

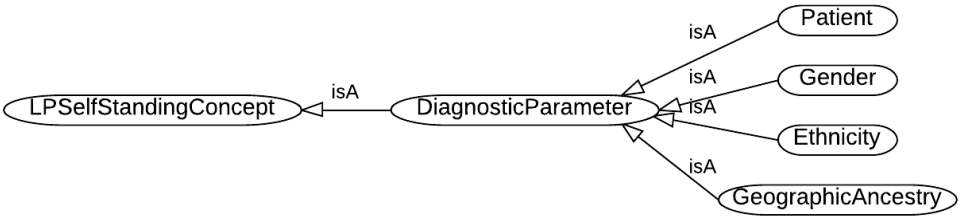


Figure 30. Sub-ontology of *LPDiagnosticParameter* with parent *LPSelfStandingConcept* and subclasses *Patient*, *Gender*, *Ethnicity* and *GeographicAncestry*.

6.8.1 Gender

For this thesis, **Gender** will be classified into two classes (Figure 31):

- **Male**
- **Female**

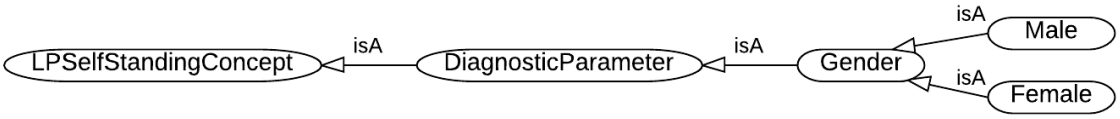


Figure 31. Concept of *Gender* with parent *LPDiagnosticParameter*.

6.8.2 Ethnicity

For this thesis, **Ethnicity** will be classified into three classes (Figure 32):

- **Asian**
- **Caucasian**
- **OtherNonspecifiedEthnicity**

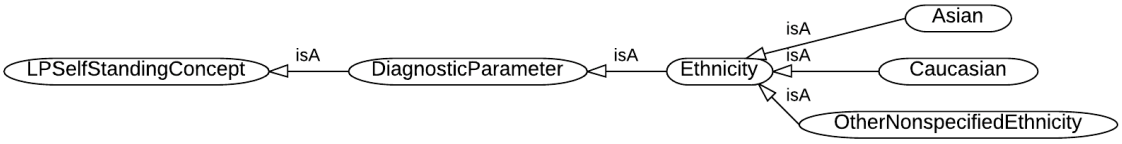


Figure 32. Concept of *Ethnicity* with parent *LPDiagnosticParameter*.

6.8.3 Geographic Ancestry

GeographicAncestry can be categorised into the following classes (Figure 33). This list is not fully exhaustive of all geographical locations and can be extended to include other countries.

- **Australia**
- **Europe**
- **NorthAmerica**
- **SouthAmerica**
- **Asia**
- **MiddleEast**
- **Africa**
- **MixedGeographicAncestry**
- **OtherNonSpecifiedCountry**

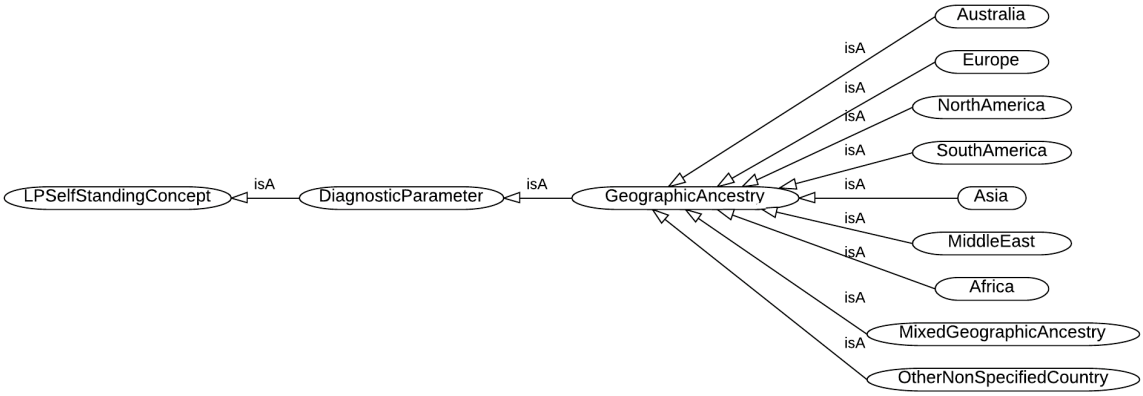


Figure 33. Concept of *GeographicAncestry* with parent *LPDiagnosticParameter* and subclasses *Australia*, *Europe*, *NorthAmerica*, *SouthAmerica*, *Asia*, *MiddleEast*, *Africa*, *MixedGeographicAncestry* and *OtherNonSpecifiedCountry*.

6.8.4 Patient

Patient is the concept given for the instances of patients that are classified or inferred by the ontology (Figure 34).

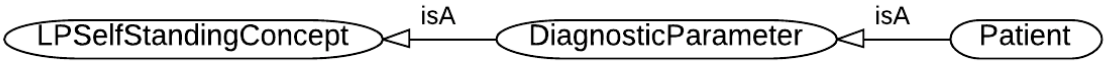


Figure 34. Concept of *Patient* with parent *LPDiagnosticParameter.*

6.9 Conclusion

In this chapter, we presented the conceptual framework for Lipoprotein Ontology using the ontology editor Protégé 4.2. First, we provided an outline of Lipoprotein Ontology and introduced the upper concepts of the ontology. We explained how these concepts are structured according to ontology principles defined in Chapter 5.4.1. Subsequently, we elaborated on the conceptualisation process of each of the six sub-ontologies of the lipoprotein domain knowledge, namely: *Classification*, *Metabolism*, *Pathophysiology*, *Aetiology*, *Treatment* and *Diagnostic Parameter*. These concept hierarchies were presented as figures. The corresponding properties and restrictions of these concepts will be discussed further in Chapter 7 during the process of formalisation.

References

Chan, D. C., Barrett, P. H. R., & Watts, G. F. (2004). Lipoprotein transport in the metabolic syndrome: methodological aspects of stable isotope kinetic studies. *Clinical Science, 107*(221-232).

Courtot, M., Juty, N., Knüpfer, C., et al. (2011). Model storage, exchange and integration. *Molecular Systems Biology, 7*, 543.

Lowe, H. J., & Barnett, G. O. (1994). Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches. *JAMA, 271*(14), 1103-1108.

Snyder, F. e. (1977). Lipid metabolism in mammals. In F. Snyder (Ed.), *Lipid metabolism in mammals* (Vol. 1). New York: Plenum Press.

Chapter 7. Formalisation of Lipoprotein Concepts

7.1 Introduction

In the previous chapter, we presented the conceptual framework of Lipoprotein Ontology. We have opted to use Protégé 4.2 as our ontology development tool, and OWL as the ontology language for the implementation of Lipoprotein Ontology. This chapter explains how the conceptualisation process in the previous chapter was formalised using OWL and its operators. First, we provide an overview of KR languages and justify the rationale behind our choice of implementation tool and language. This is followed by a description of OWL components and the structure of OWL ontologies. Finally, visualisation of Lipoprotein Ontology is presented as screenshots from Protégé to represent various functions of concept classification, definition and description, using various OWL constructs.

7.2 Knowledge Representation Languages

In order to enable the sharing of information by software agents and facilitate the automatic processing of information, conceptual models have to be implemented in a formal language with well-defined syntax. As discussed in the Chapter 6, we have chosen to develop Lipoprotein Ontology using the ontology editor Protégé 4.2. At the time of writing of this thesis, Protégé is the leading ontology editor across various disciplines, with a community of approximately 220,000 users internationally (http://protege.stanford.edu).

Although various other KR languages exist, in this section we focus mainly on the formal language that is supported by Protégé, the Web Ontology Language 2 (OWL 2). OWL 2 was developed by the World Wide Web Consortium (W3C)

as the standard to define ontologies on the Semantic Web (W3C, 2012), and is therefore a suitable language for the formalisation of Lipoprotein Ontology.

This section introduces various KR languages associated with ontology building, such as Resource Descriptive Framework (RDF), Resource Descriptive Framework Schema (RDFS) and the Web Ontology Language (OWL).

7.2.1 RDF and RDFS

The Resource Descriptive Framework (RDF) (http://w3.org/RDF) is a language for representing information about resources on the World Wide Web (Berners-Lee et al., 2001). RDF serves as a framework for metadata description and is represented as a machine-readable format of subject-predicate-object triplet from the English language. In contrast to object-oriented languages which are resource-centric, RDF is a property-centric language with an XML-based syntax (Baader et al., 2003). A resource-centric language requires the definition of concepts and their properties in a centralised method, which does not allow a property to be defined independently of its associated class. Thus, a new property cannot be defined unless it is represented by a predefined class or the creation of a new class for it. On the other hand, RDF is a semantic web language which allows the representation of information in a free manner, by providing the facility to define properties independently of the classes. Therefore, RDF has the capability to describe any class with any existing property. However, RDF has no vocabulary associated with it; as such, anything can be expressed in it as long as it follows a subject-predicate-object structure.

The lack of vocabulary in RDF poses a problem for machine processing due to the insufficient context or meaning associated with the data model. Thus, RDF Schema (RDFS) (http://w3.org/TR/rdf-schema) was developed to extend RDF vocabulary to allow the description of concepts and attributes. This allows the creation of meaningful statements and facilitates the exchange of information between computer systems more efficiently due to the additional context. RDFS introduces the concept of classes, which allows classes to be specialised into subclasses that can inherit from multiple upper classes. Similarly, properties can be arranged in a hierarchical format from general to specific. RDF is then used to make actual statements using the structure described previously. In addition

to the class hiearchy, RDFS has other features such as such as providing the facility to set the domain and range of properties (Baader et al., 2003). RDFS also allows the creation of unions and intersections of classes, thereby effectively creating new classes. However, ontologies require greater vocabulary than one that RDF and RDFS offer, in addition to formal semantics. For this reason, the Web Ontology Language (OWL) was created on top of RDF and RDFS, and is our language of choice to represent Lipoprotein Ontology.

7.2.2 OWL and OWL 2

The modelling capabilities of RDF and RDFS allow vocabularies to be organised in hierarchies of class/subclass and property/subproperty relationships, operate domain and range restrictions and define instance memberships. However, they still lack other significant features of an ontology, such as cardinality restrictions, describing disjoint classes, among others. These shortcomings can be addressed by an ontology language.

The Web Ontology Language (OWL) (http://w3.org/TR/owl-features) is a family of KR languages for authoring ontologies (W3C, 2012). Known in previous revisions as DAML (Stein, 2002), OIL (Horrocks, 2000), and their successor DAML+OIL (Horrocks, 2001), OWL extends the basic fact-stating function of RDF and the structuring capabilities of RDFS in various ways. In addition to declaring classes and properties, and organising them in a subsumption hierarchy accordingly, OWL has greater expressivity with the capability to provide restrictions on how properties behave that are local to a class. Furthermore, OWL also allows Boolean combinations such as intersectionOf, unionOf and complementOf, to combine both owl:Classes and class expressions (i.e. anonymous classes), or disjoint with other classes (W3C, 2009).

To satisfy different requirements, OWL comes in three different forms: OWL Lite, OWL DL and OWL Full. The difference between these types is their increasing expressiveness from simple constraints via computational completeness to syntactic freedom of RDF with no computational guarantees (Horrocks et al., 2003).

OWL Lite is used in situations where the features of RDFS are required with simple constraints (W3C, 2009). As OWL Lite offers the lowest formal complexity, such as setting cardinality values of 0 or 1, it allows a more tractable inference as well as provides a quick migration path from simple classification hierarchies such as taxonomies (W3C, 2009). Due to its low complexity, OWL Lite is considered to be too limited for our requirements.

OWL DL is an extension of OWL Lite that offers maximum semantic expressiveness while retaining computational function which guarantees complete computation in finite time (W3C, 2009). OWL DL is based on DL, a family of logic-based KR languages that have formal semantics based on first order logic and allows automatic reasoning (Baader et al., 2003). As automatic reasoning is one of the requirements for Lipoprotein Ontology, OWL DL is suitable for developing our ontology.

Out of the three OWL versions, OWL Full provides the greatest semantic expressiveness. However, as it lacks computational guarantees, it is not suitable for our requirements, as it is necessary that Lipoprotein Ontology be computable in finite time in order to be useful in its application.

In 2009, OWL was extended with new features, thereby known as OWL 2 (http://w3.org/TR/owl2-overview). OWL 2 is currently the standard ontology language recommended by W3C (W3C, 2012). Some of the new features of OWL 2 include the more user-friendly Manchester syntax, and a functional-style syntax which specifies the language structure and allows OWL 2 ontologies to be written in a compact form (W3C, 2012).

OWL 2 is the ontology language supported by our chosen ontology editor, Protégé 4.2, and retains all the characteristics of OWL 1 in terms of its expressivity. It has two corresponding dialects to OWL 1: OWL 2 DL, which is used to refer to ontologies interpreted using Direct Semantics, and OWL 2 Full, which can only be interpreted under RDF-Based Semantics (W3C, 2012). Therefore, based on the purpose of our ontology and the need for reasoning capabilities, we have chosen to use OWL 2 DL to develop Lipoprotein Ontology.

In addition to automatic reasoning function, OWL 2 DL utilises Manchester syntax that can be used to make working with OWL easier. In the Semantic Web context, where users with a wide range of expertise might be expected to create or modify ontologies, readability and general ease of use are important considerations for an ontology language. This provides further justification towards our choice to use OWL 2 DL to develop our ontology.

7.3 Components of OWL Ontologies

OWL ontologies are composed of individuals, classes and properties. As discussed previously, we have opted to use OWL 2 to develop our ontology, which corresponds to SROIQ(D) logic. DL allow concepts to be defined and described using a rich set of operators such as intersection, union and negation, in addition to other property restrictions. Thus, complex concepts can be built up in definitions out of simpler concepts using the OWL components **Classes**, **Properties** and **Individuals**, described below.

7.3.1 Individuals

Individuals represent the objects in our domain of interest and can be referred as being instances of classes according to the fulfilment of specific conditions.

7.3.2 Classes

Classes are concrete representation of concepts composed of formal descriptions that specify the conditions that must be satisfied by an individual for it to be a member of the class. For example, in Lipoprotein Ontology, the class **Chylomicron** contains all instances of chylomicron molecules. There are typically two types of classes in an ontology: primitive classes which are defined in terms of their necessary properties or attributes, and defined concepts which are defined in terms of properties which are both necessary and sufficient. We will discuss the difference between the two types of concepts in the subsequent sections.

7.3.3 Properties

Properties represent the relations between two classes or individuals, and can also be placed in a subsumption (property/subproperty) hierarchy for the purpose of abstraction. There are three types of properties:

- **Object property** links instances of one class to another class. E.g. The property hasComponent can be used to link the instances of Chylomicron and Cholesterol as "**Chylomicron** *hasComponent* **Cholesterol**".
- **Datatype property** specifies datatype values for individuals by linking them to an rdf literal or an XML Schema Datatype such as integer, float, string, Boolean etc. E.g. The datatype property hasSize can be used to describe the size range (between 75nm and 1200nm) that Chylomicron particles can fall under as "**Chylomicron** *hasSize* some (integer[>= 75] and integer[<= 1200])". As lipoprotein particles are classified into different classes according to the size range in diameter which is unique to that specific class, any particle that falls between the range specified in our example above is automatically classified as a chylomicron.
- **Annotation property** is used for metadata description such as comments, references to resources, etc. E.g. The annotation property of the class Chylomicron is defined as "Comment: A class of lipoproteins that carry dietary cholesterol and triglycerides from the small intestines to the tissues."

7.3.3.1 Property Characteristics

OWL allows the meaning of object properties to be enriched through the use of property chacteristics. In this section, we describe some of the property characteristics used in Lipoprotein Ontology and provide examples to illustrate how they are applied in our ontology.

Inverse properties

Each object property may have a corresponding inverse property. For example, if the inverse property of *hasComponent* is *isComponentOf*, we can infer from

"Chylomicron *hasComponent* Cholesterol" that "Cholesterol *isComponentOf* Chylomicron".

Functional properties

If a property is functional, there can be at most one unique value for each individual. For example, lipoproteins can be classified according to their electrophoretic mobility on agarose gels into α, pre-β and β lipoproteins, corresponding to HDL, VLDL and LDL lipoprotein fractions respectively. Chylomicrons remain at the electrophoretic origin. Therefore, for each subclass of **LipoproteinEntity** which has the corresponding functional property *hasElectrophoreticMobilityValue*, we can automatically infer the class that an individual belongs to, given the electrophoretic mobility value.

Transitive properties

If a property P is transitive and associates object X to object Y, and object Y to object Z, then object X is simultaneously associated with object Z via property P. For example, for the transitive property *hasMetabolismProduct*, if "**VeryLowDensityLipoprotein** *hasMetabolismProduct* **IntermediateDensityLipoprotein**" and "**IntermediateDensity Lipoprotein** *hasMetabolismProduct* **LowDensityLipoprotein**", then we can infer that **VeryLowDensityLipoprotein** *hasMetabolismProduct* **LowDensityLipoprotein**.

Symmetric properties

If object X is associated with object Y via property P, then object Y is associated with object X along the same property P. For example, the symmetric property *isAssociatedWith* is its own inverse property.

7.3.3.2 Property Restrictions

In addition to designating property characteristics, OWL also provides the facility to further constrain the range of a property through property restrictions. By assuming a restriction type as an abstract or anonymous class, we may describe a given class of individuals as a subclass of that anonymous class

which satisfies the restriction criteria. We have applied the following property restrictions to Lipoprotein Ontology and provide examples for each type of restriction.

Quantifier restrictions

Quantifier restrictions can be further categorised into existential restriction and universal restrictions.

Existential restrictions, written in OWL as "someValuesFrom", describe classes whose individuals are related, via a given property, to individuals of some other class. For example, the statement "**Chylomicron** *hasComponent* some **Cholesterol**" describes instances which belong to the class **Chylomicron** that have at least one (some) *hasComponent* relationship to members of the class **Cholesterol**. Here, the restriction "*hasComponent* some **Cholesterol**" can be abstracted as an anonymous class whose members are those individuals from the class **Chylomicron** that satisfy the restriction, that is, have at least one *hasComponent* relationship with **Cholesterol**.

Universal restrictions, written in OWL as "allValuesFrom", define classes of individuals that are associated only with individuals of another specified class along a certain property. For example, the statement "apoB-48 *isAssociatedWith* only **Chylomicron**" describes instances which belong to the anonymous class of individuals that only have *isAssociatedWith* relationship to the class **Chylomicron**. The universal restriction is rarely used in Lipoprotein Ontology since lipoprotein concepts are highly interrelated.

Cardinality restrictions

Cardinality restrictions allow us to describe the class of individuals that have at least, at most or exactly a specified number of relationships with other individuals or datatype values. For a given property P, a *minimum cardinality restriction* specifies the minimum number of P relationships that an individual must participate in. A *maximum cardinality restriction* specifies the maximum number of P relationships that an individual can participate in. A *cardinality restriction* specifies the exact number of P relationships that an individual must

participate in. *Qualified cardinality restrictions* are more specific than *cardinality restrictions* in that they state the class of the objects within the restriction.

For example, the statement "**MetabolicSyndrome** *hasSymptom* min 3 Thing" describes the set of individuals that are members of the class **MetabolicSyndrome** and that have at least three *hasSymptom* relationships with other (distinct) individuals. This can also be applied to the various cardinality restriction types accordingly.

hasValue restrictions

A *hasValue restriction* describes the set of individuals that have at least one relationship along a specified property to a specified individual. For example, the *hasValue restriction* **Patient** *hasGeographicAncestry* Australia (where Australia is an individual) describes the set of individuals (the anonymous class of individuals) that belongs to the class **Patient** and has at least one relationship along the *hasGeographicAncestry* property to the specific instance Australia.

7.3.4 Necessary and Sufficient Conditions

Having described OWL properties, including property characteristics and restrictions, we will now discuss their usage in OWL classes to build up complex definitions from simpler definitions. Two types of OWL classes were implemented in Lipoprotein Ontology.

Primitive class is a class that only has necessary conditions. For example, for the primitive class **LipoproteinEntity**, if an individual is a member of the class **LipoproteinEntity** then it must satisfy the condition. However, even if any (random) individual satisfies these necessary conditions, it cannot be automatically inferred as a member of the class **LipoproteinEntity**, as the conditions are not sufficient to be able to say this.

Defined class is a class with at least one set of necessary and sufficient conditions, and can also be referred to as an equivalent class. For example, for the defined class **Chylomicron**, if an individual is a member of the class **Chylomicron** then it must satisfy the condition. On top of that, if any (random)

individual satisfies the conditions, then the individual must be a member of the class **Chylomicron**. Therefore, the conditions are not only necessary for membership of the class **Chylomicron** but also sufficient to determine that something which satisfies these conditions is also a member of the class.

7.4 Structure of OWL Ontologies

The relations between concepts or individuals (and their properties) can be categorised into two types: hierarchical and associative relationships.

Hierarchical relationships organise concepts into “isA” or “partOf” hierarchies. The finite concepts in most of the ontologies are linked together using this type of relationship. Ontology-based representations are typically hierarchical models, following the natural tendency of the human mind to organise mental models through the formation of hierarchies. The value of the hierarchical structure is in its support of abstraction, which is the process whereby lower level, more specific concepts can be generalised into broader, upper level concepts to allow the automatic reasoning of subsumption (superclass/subclass) relationships (Rosch, 1978). Ontologies can also be represented as graph structures or trees or network models in order to meet the more complex requirements of an expert system (Helsper & van der Gaag, 2002).

Associative relationships describe the dynamic interactions between entities and the influences that different components have on each other. Such influences can take various forms including causation, alteration, regulation, etc. In this regard, associative relationships in scientific domains should represent facts by identifying the existing connections between different concepts and describing the influences that those concepts have on each other through scientific evidence from the literature, rather than the consensus of domain experts (Smith et al., 2005). For example, an ontological statement such as “cell cycle checkpoint regulates cell cycle” in Gene Ontology (GO Consortium, 2000) is grounded in the scientific fact that the cell cycle is constantly and necessarily regulated by the cell cycle checkpoint (GO Consortium, 2000). Therefore, rather than defining ontology relationships through the agreement of domain experts, we have developed relations in Lipoprotein Ontology based on scientific evidence obtained from the literature.

7.5 Notation and Syntax

In this thesis, the following conventions are used:

- English text versions of definitions or concepts to be represented are written between double quotes “protein and lipid complex”.
- Things that appear on the screen are given in a bold sans-serif font, e.g. **LipoproteinOntology**.
- Classes (concepts) and property names (attributes, slots or roles) are written in camel back notation in bold and italics respectively, e.g. **LipoproteinEntity**, *hasComponent*.
- Class names always begin with an uppercase letter, and with singular noun. Property names always begin with a lower case letter.

Given these conventions, the representation of lipoproteins, defined as “protein and lipid complex”, is **LipoproteinEntity**. The relationship between chylomicrons, a kind of lipoprotein, and lipoprotein entity, is hierarchical and is represented by the “isA” property relation. Therefore, in Lipoprotein Ontology, the previous statement is written as: **Chylomicron** *isA* **LipoproteinEntity**.

Protégé 4.2 follows the Manchester syntax, which is used when editing class expressions. These notations are shown in the table below.

Restrictions

OWL	DL Symbol	Manchester OWL Syntax	Example
someValuesFrom	∃	some	*hasCause* some **X**
allValuesFrom	∀	only	*hasCause* only **X**
hasValue	∋	value	*hasCause* value **X**
minCardinality	≥	min	*hasCause* min 3
cardinality	=	exactly	*hasCause* exactly 3
maxCardinality	≤	max	*hasCause* max 3

Boolean Class Constructors

OWL	DL Symbol	Manchester OWL Syntax	Meaning
intersectionOf	⊓	and	**Female** and **Patient**
unionOf	⊔	or	**Male** or **Female**
complemenOf	¬	not	not

7.6 Formalisation of Lipoprotein Concepts

In the previous section, we discussed the various components of OWL ontologies and the types of relations that OWL ontologies hold. Here, we describe the process of formalising Lipoprotein Ontology with OWL components of Classes, Properties and Individuals. The relationships between Lipoprotein Ontology concepts will be shown as figures throughout this chapter. Due to space constraints, we cannot link every single concept that has been defined in our ontology on these diagrams; however, wherever possible, we have shown how concepts from different sub-ontologies were reused, in order to view them within context. In addition, for some concepts, we include screenshots of the properties frame in Protégé to illustrate the properties and restrictions associated with the corresponding classes.

The underlying DL formalism of our chosen implementation language OWL DL allows the use of a reasoner to check that the statements and definitions in the ontology are mutually consistent, ensuring the integrity of the ontology throughout the development process. Therefore, we have chosen to present the inferred model of our ontology, as it can be seen as a formal validation method to check for inconsistencies in the ontology implementation. In addition, the inferred model also provides a greater level of detail and relationships in some cases.

The differences between various reasoners were reviewed in (Dentler et al., 2011). In this thesis, we have used the Protégé built-in HermiT 1.3.7 reasoner, as it has the capability to determine the consistency of an ontology, identify subsumption relationships between concepts, among other features (Dentler et

al., 2011). In addition, it provides support for Semantic Web Rule Language (SWRL), which can be applied to the ontology at a much later stage to provide additional rules on top of the OWL knowledge base (Dentler et al., 2011; O'Connor, 2006).

The upper concepts and class hierarchy of **LipoproteinDomainConcept** are represented in Figure 35 below.

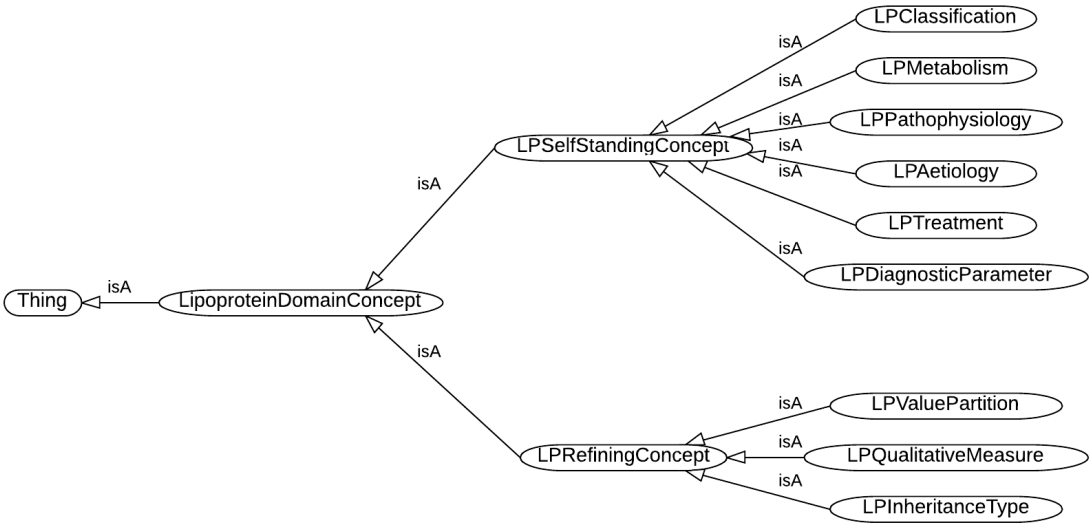


Figure 35. *LipoproteinDomainConcept* represented by two major subclasses, *LPSelfStandingConcept* and *LPRefiningConcept*, with their respective subclasses.

In the subsequent sections, we discuss the upper classes of **LPSelfStandingConcept**, which we refer to as sub-ontologies of Lipoprotein Ontology, as they contain the majority of lipoprotein domain knowledge. These include **LPClassification**, **LPMetabolism**, **LPPathophysiology**, **LPAetiology**, **LPTreatment** and **LPDiagnosticParameter**.

7.6.1 LPClassification

In this section, we formalised the various subtypes of lipoproteins according to several criteria which confer membership into the class such as content, electrophoretic mobility and function. These criteria were defined by various object and datatype properties which link concepts between **LPClassification** and **LPMetabolism** sub-ontologies. The sub-ontology **LPClassification**

contains the subclass **LipoproteinEntity**, which defines different lipoprotein classes as follows:

- **Chylomicron**
- **ChylomicronRemnant**
- **VeryLowDensityLipoprotein**
- **IntermediateDensityLipoprotein**
- **LowDensityLipoprotein**
- **HighDensityLipoprotein**
- **Lipoprotein(a)**

LipoproteinEntity is linked to different subclasses of **FunctionalEntity**, which is asserted to be *partOf* **LPMetabolism** via the *hasComponent* property. In other words, for an individual to belong to the class **LipoproteinEntity**, it has to fulfil the following conditions: *hasComponent* some **Triglyceride**, **Cholesterol**, **CholesterylEster**, **Phospholipid** and **Apolipoprotein**.

Due to the subsumption rule, all subclasses of **LipoproteinEntity** inherit the properties of the concept accordingly; however, it is not sufficient to classify different subtypes of **LipoproteinEntity**. Hence, we defined each individual subclass of **LipoproteinEntity** according to their respective characteristics. Figure 36 shows the concept **Chylomicron**. As we have discussed extensively in the literature review, the lipoprotein metabolism process involves a highly dynamic and constant remodelling of particles where lipid and apolipoprotein components are exchanged and/or catabolised, which results in the transformation of one lipoprotein particle to another. For example, in the endogenous pathway, the triglycerides in VLDL are hydrolysed as VLDL is metabolised into IDL and subsequently LDL. Because each lipoprotein subclass can be defined according to the content of their lipid and apolipoprotein components, referred to as percentage weight (that is, a VLDL cannot have the same lipid and apolipoprotein content as an LDL), the most appropriate definition of the different lipoprotein subclasses would be according to their lipid/apolipoprotein content. Hence, for a lipoprotein particle to be classified as a certain type of lipoprotein, it has to meet the necessary and sufficient conditions using the datatype properties:

- *hasTriglyceridePercentageWeight*
- *hasCholesterolPercentageWeight*
- *hasCholesterylEsterPercentageWeight*
- *hasPhospholipidPercentageWeight*
- *hasApolipoproteinPercentageWeight*

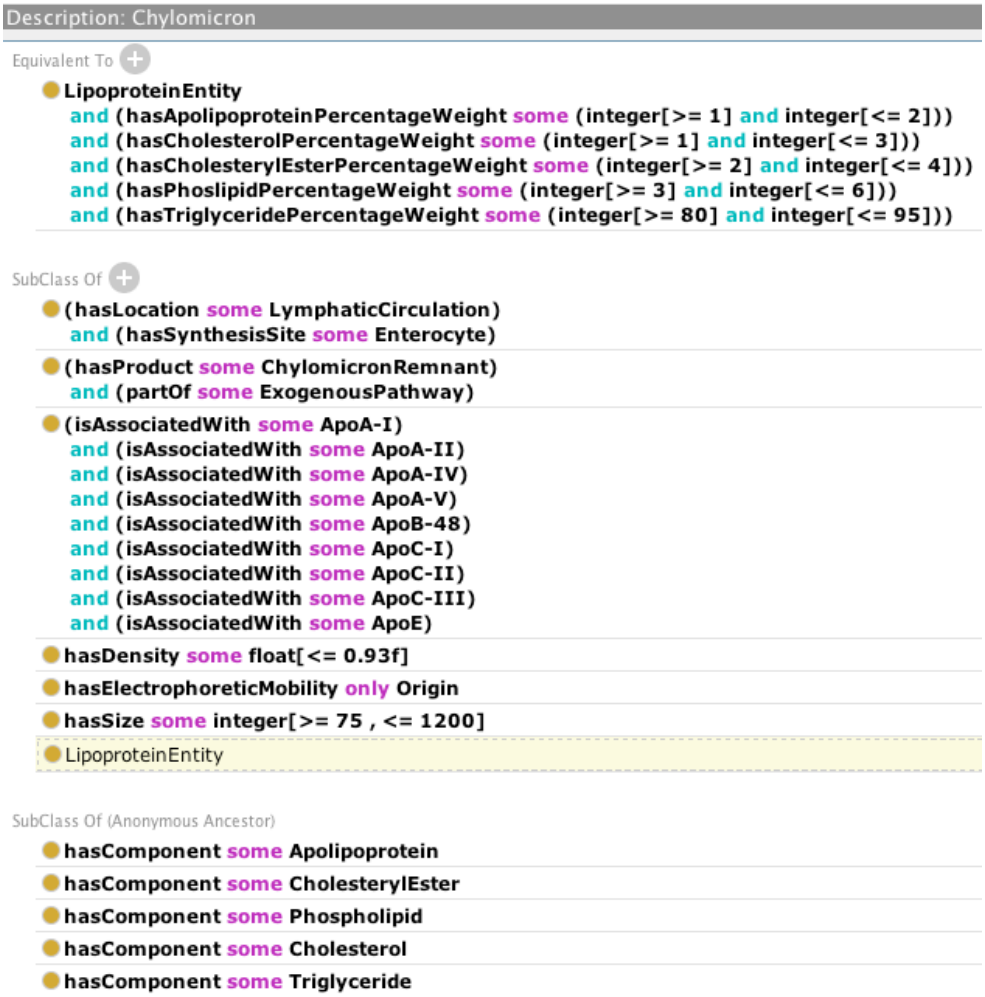


Figure 36. Formalisation of the concept *Chylomicron.*

In addition, the class **Chylomicron** is linked to other concepts in the sub-ontology **LPMetabolism** via properties such as:

- *isAssociatedWith* (symmetric property). For example, if **Chylomicron** *isAssociatedWith* **ApoA-I**, the reasoner will infer that **ApoA-I** is *AssociatedWith* **Chylomicron**

- *hasProduct* (transitive property and inverse of the property isProductOf). Given **Chylomicron** *hasProduct* **ChylomicronRemnant**, this assertion implies that **ChylomicronRemnant** *isProductOf* **Chylomicron**
- *PartOf*
- *hasSynthesisSite*
- *hasLocation*
- *hasElectrophoreticMobility*
- *hasDensity* (datatype property)
- *hasSize* (datatype property)

The Figure above also shows the inherited properties from the superclass **LipoproteinEntity**, under *SubClassOf* (Anonymous Ancestor).

Similarly, we have defined necessary and sufficient conditions for the other subclasses of **LipoproteinEntity**: **ChylomicronRemnant**, **VeryLowDensityLipoprotein**, **IntermediateDensityLipoprotein**, **LowDensityLipoprotein**, **HighDensityLipoprotein** and **Lipoprotein(a)**.

7.6.2 LPMetabolism

LPMetabolism concepts are commonly reused across the different Lipoprotein Ontology sub-ontologies, such as **LPClassification,** discussed in the previous section. In fact, **LPMetabolism** can be seen as the core of Lipoprotein Ontology, as it contains and describes the various atomic components of lipoprotein physiology, and relates these components to other aspects of lipoprotein research. In order to model lipoprotein metabolism in context of the biological domain as a whole, we categorised important components of **LipoproteinEntity** according to their appropriate classes. We refer to this process as normalisation, where we categorised concepts according to a structured hierarchy which allows for the easy extension and/or modification to a particular concept. For this reason, we separated the following concepts, defined to be *partOf* **LPMetabolism** as:

- **PhysicalEntity**: An entity which participates in a process or relationship
- **OccurringEntity**: An event or entity which develops through time
- **ParticipantRole**: The role of an entity

We note that the concept names above are derived from Systems Biology Ontology (Courtot et al., 2011), with some of the concept definitions referenced from MeSH (Lowe & Barnett, 1994}. This gives us the potential to merge Lipoprotein Ontology and Systems Biology Ontology for future work.

7.6.2.1 PhysicalEntity

As its name suggests, **PhysicalEntity** contains the various components associated with **LPClassification**, categorised into the following classes:

- **FunctionalEntity**: Defined by its properties or functions
- **StructuralEntity**: Defined by its structure

In Chapter 6.4.1, we presented all the subclasses of **FunctionalEntity** and **StructuralEntity** respectively. Figure 37 illustrates how some of these concepts are related to Chylomicron in **LPClassification** as follows:

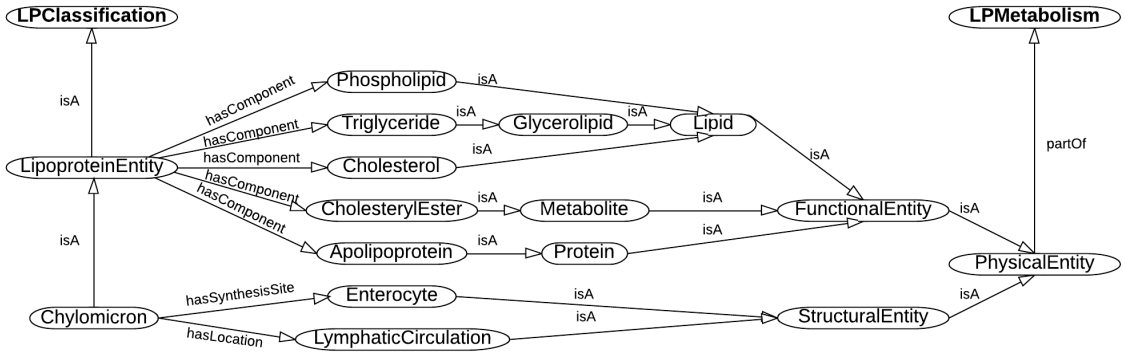


Figure 37. Relations between *LPClassification* concepts and *LPMetabolism* concepts.

In the Figure above, class names are represented in rounded boxes, and their attributes or relationships to other classes are described along a line and arrowhead pointing towards the direction of the property that is being described.

The relation *isA* naturally confers a hierarchical or subsumption relationship between two concepts. This implies that if "**Cholesterol** *isA* **Lipid** *isA* **FunctionalEntity** *isA* **PhysicalEntity**", then "**Cholesterol** *isA* **PhysicalEntity**". The subsumption relationship also leads to the inheritance of the properties of the superclass. Therefore, if "**LipoproteinEntity** *hasComponent* **Cholesterol**" and "**Chylomicron** *isA* **LipoproteinEntity**", then the equivalent is true: "**Chylomicron** *hasComponent* **Cholesterol**".

Although the diagram illustrates a clear representation of the binary relationships between any two concepts, we are unable to show every single relation they contain, due to space constraints. For example, we have defined the property *hasComponent* to have an inverse property *isComponentOf*. Hence, based on the statement: "**LipoproteinEntity** *hasComponent* **Cholesterol**", the reasoner can infer an equivalent statement to be: "**Cholesterol** *isComponentOf* **LipoproteinEntity**".

7.6.2.2 OccurringEntity

OccurringEntity is categorised into the following classes:

- **Pathway**: Defined as the representation of a sequence of reactions
- **Process**: Defined as the action that causes a change from one entity to another

As metabolic pathways are extremely complex, highly dynamic and involve many different components from the concept **PhysicalEntity**, it is practically impossible to formalise these two concepts in a stand-alone manner. Hence, rather than presenting the formal definition of concepts of **OccurringEntity** individually, we illustrate the formalisation process of the three lipoprotein metabolism pathways:

- **Exogenous pathway**: The transport and metabolism of dietary lipids (Figure 38)
- **Endogenous pathway**: The metabolism of hepatically synthesised lipids (Figure 39)
- **Reverse cholesterol transport**: The synthesis and metabolism of HDL (Figure 40)

Class names are represented in rounded boxes, and their attributes or relationships to other classes are described along a line and arrowhead pointing towards the direction of the property that is being described. Dotted boxes represent **Process** or **Pathway** concepts which are used to describe the relationship between one entity to another. Enlarged versions of the figures below are shown in Appendix I, J and K.

Due to the complexity of these pathways and space constraints, we are unable to show every single concept, but have represented the most interesting and relevant concepts involved in lipoprotein metabolism. Although the diagrams demonstrate a clear representation of the binary relations between two concepts in the context of their respective metabolism pathways, we are unable to show every single property they contain.

7.6.2.2.1 ExogenousPathway

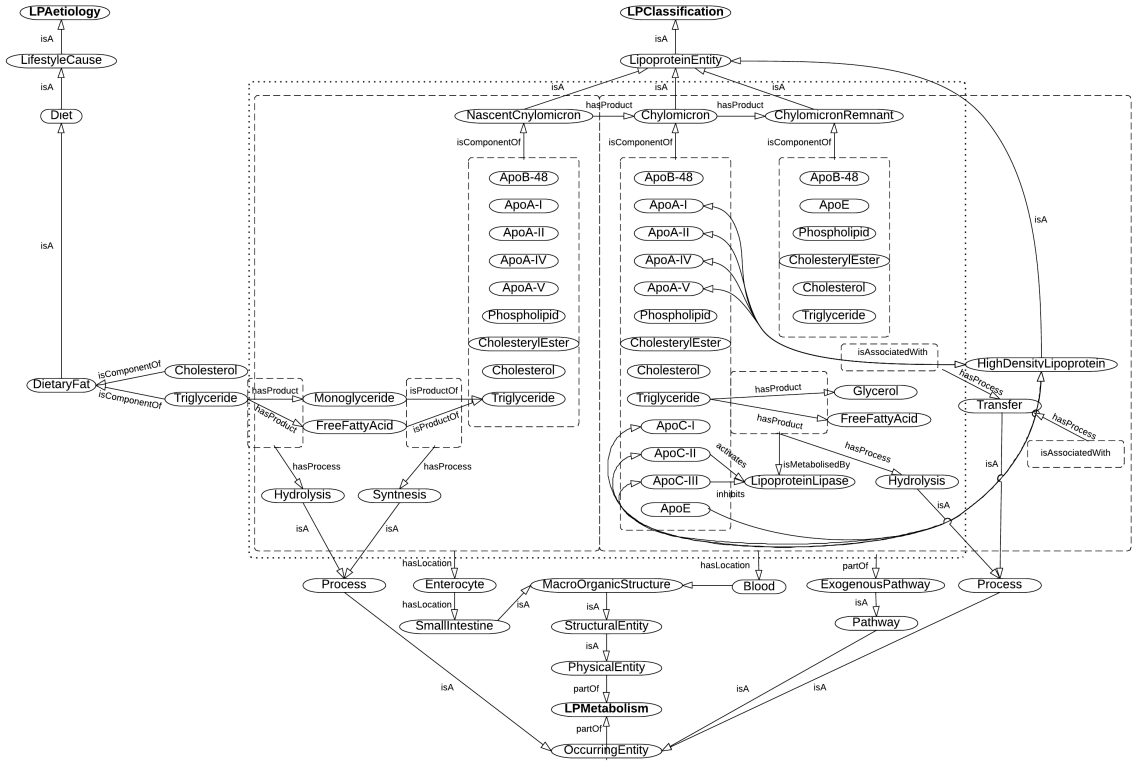


Figure 38. Formalisation of the concept *ExogenousPathway*.

The concept **ExogenousPathway** describes the transport of dietary lipids, defined to be "**Cholesterol** *isA* **Lipid** and *isComponentof* **DietaryFat**" and "**Triglyceride** *isA* **Lipid** and *isComponentof* **DietaryFat**". After a meal, the hydrolysis of triglycerides into monoglycerides and free fatty acids is described as "**Triglyceride** *hasProduct* **Monoglyceride** and **FreeFattyAcid**" via "*hasProcess* **Hydrolysis**". The resynthesis of monoglycerides and free fatty acids is described as "**Monoglyceride** and **FreeFattyAcid** *isProductOf* **Triglyceride**" via "*hasProcess* **Synthesis**". Here, we note that the property *hasProduct* has an inverse property *isProductOf*; this is to keep the properties of the concepts consistent, which allows it to be used in different context.

Triglycerides are subsequently combined with cholesterol, cholesteryl ester, triglycerides, apo B-48 and apo As to form nascent chylomicrons in the enterocytes: "**Triglyceride**, **Phospholipid**, **Cholesterol** and **CholesterylEster** *isComponentOf* **LipoproteinEntity**" and "*hasLocation* **Enterocyte**". As "**NascentChylomicron** *isA* **LipoproteinEntity**", this confers exactly the same properties. This is logically appropriate as every instance of **LipoproteinEntity** has to have the components **Triglyceride**, **Phospholipid**, **Cholesterol** and **CholesterylEster**. However, apolipoproteins are slightly different. With the exception of the principal apolipoproteins (ApoB-100, ApoB-48, ApoA-I And ApoA-II), apolipoproteins are exchanged and transferred between the different subtypes of lipoproteins. Hence, we used the relation "isAssociatedWith", defined to be a symmetric property (illustrated by double arrows), which allows the reasoner to infer the association between apolipoproteins to their respective classes. For example, "**ApoA-I** *isAssociatedWith* **NascentChylomicron**" infers the symmetric statement "**NascentChylomicron** *isAssociatedWith* **ApoA-I**".

Nascent chylomicrons are then secreted into the circulation where they acquire apoCs and apoE from HDL, and referred to as chylomicrons. This is described as "**NascentChylomicron** *hasProduct* **Chylomicron**", "**Chylomicron** *isAssociatedWith* **ApoB-48**, **ApoA-I**, **ApoA-II**, **ApoA-IV**, **ApoA-V**, **ApoC-I**, **ApoC-II**, **ApoC-III** and **ApoE**" and "*hasLocation* **Circulation**". We defined the process of apolipoprotein transfer between chylomicrons and HDL as "**ApoC-I**, **ApoC-II** and **ApoC-III** *isAssociatedWith* **HighDensityLipoprotein**" and "hasProcess **Transfer**".

ApoC-II activates lipoprotein lipase located in adipose and muscle tissues, where triglycerides in chylomicrons are hydrolysed into free fatty acids and glycerol by the action of lipoprotein lipase. This process is inhibited by apoC-III. In formal language, we describe the statement as: "**ApoC-II** *activates* **LipoproteinLipase**", "**ApoC-III** inhibits **LipoproteinLipase**" and "**LipoproteinLipase** *hasLocation* **AdiposeTissue** and **MuscleTissue**". "**Triglyceride** *hasProduct* **Glycerol** and **FreeFattyAcid**" and "*isMetabolisedBy* **LipoproteinLipase**" and "*hasProcess* **Hydrolysis**". As triglycerides are hydrolysed, chylomicrons gradually shrink to form chylomicron remnants, during which ApoAs and ApoC-II are transferred to HDL. We describe this as "**Chylomicron** *hasProduct* **ChylomicronRemnant**".

7.6.2.2.2 EndogenousPathway

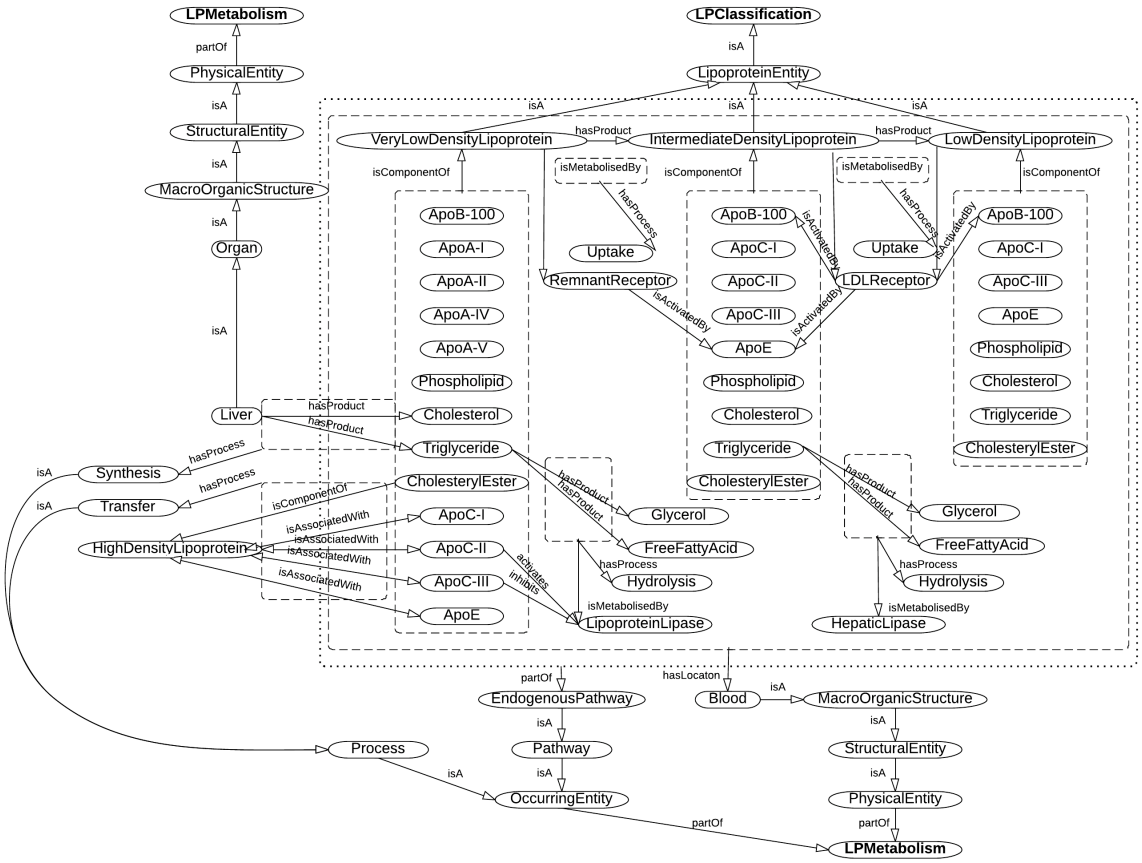


Figure 39. Formalisation of the concept *EndogenousPathway*.

The **EndogenousPathway** describes the transport of lipids synthesised in the liver as "**Liver** *hasProduct* **Cholesterol** and **Triglyceride**" and "*hasProcess* **Synthesis**". VLDL carry triglycerides, cholesterol, apoB-100, and acquire cholesteryl esters, apoCs and apo E from HDL. This is described as "**ApoC-I**, **ApoC-II**, **ApoC-III** and **ApoC-III** *isAssociatedWith* **VeryLowDensityLipoprotein** and **HighDensityLipoprotein**" and "*hasProcess* **Transfer**".

The metabolism of VLDL is facilitated by apoC-II and inhibited by the apoC-III content of the lipoprotein particles, where triglycerides are hydrolysed into grlycerol and free fatty acids by lipoprotein lipase. This process has been defined previously as "**ApoC-II** *activates* **LipoproteinLipase**", "**ApoC-III** inhibits **LipoproteinLipase**" and "**LipoproteinLipase** *hasLocation* **AdiposeTissue** and **MuscleTissue**". "**Triglyceride** *hasProduct* **Glycerol** and **FreeFattyAcid**" and "*isMetabolisedBy* **LipoproteinLipase**" and "*hasProcess* **Hydrolysis**". As VLDL lose triglycerides, they form IDL, and subsequently LDL through the same

process. Therefore, "**VeryLowDensityLipoprotein** *hasProduct* **IntermediateDensityLipoprotein**" and "**IntermediateDensityLipoprotein** *hasProduct* **LowDensityLipoprotein**". Because essentially all "**VeryLowDensityLipoprotein** *hasProduct* **LowDensityLipoprotein**", we have defined the property *hasProduct* to be transitive to infer as such.

VLDL that are not metabolised to form IDL are taken up by the remnant receptor, while IDL that are not metabolised to form LDL are taken up by the LDL receptors, which is activated by apoB-100 and apoE. LDL also binds to LDL receptors, which is activated by apoB-100. Therefore, the uptake of VLDL, IDL and LDL are defined as follows: "**VeryLowDensityLipoprotein** *isMetabolisedBy* **RemnantReceptor**"; "**IntermediateDensityLipoprotein** *isMetabolisedBy* **LDLReceptor**", "isActivatedBy ApoB-100 and ApoE" and "*hasProcess* **Uptake**"; and "**LowDensityLipoprotein** *isMetabolisedBy* **LDLReceptor**", "isActivatedBy ApoB-100" and "*hasProcess* **Uptake**".

7.6.2.2.3 ReverseCholesterolTransport

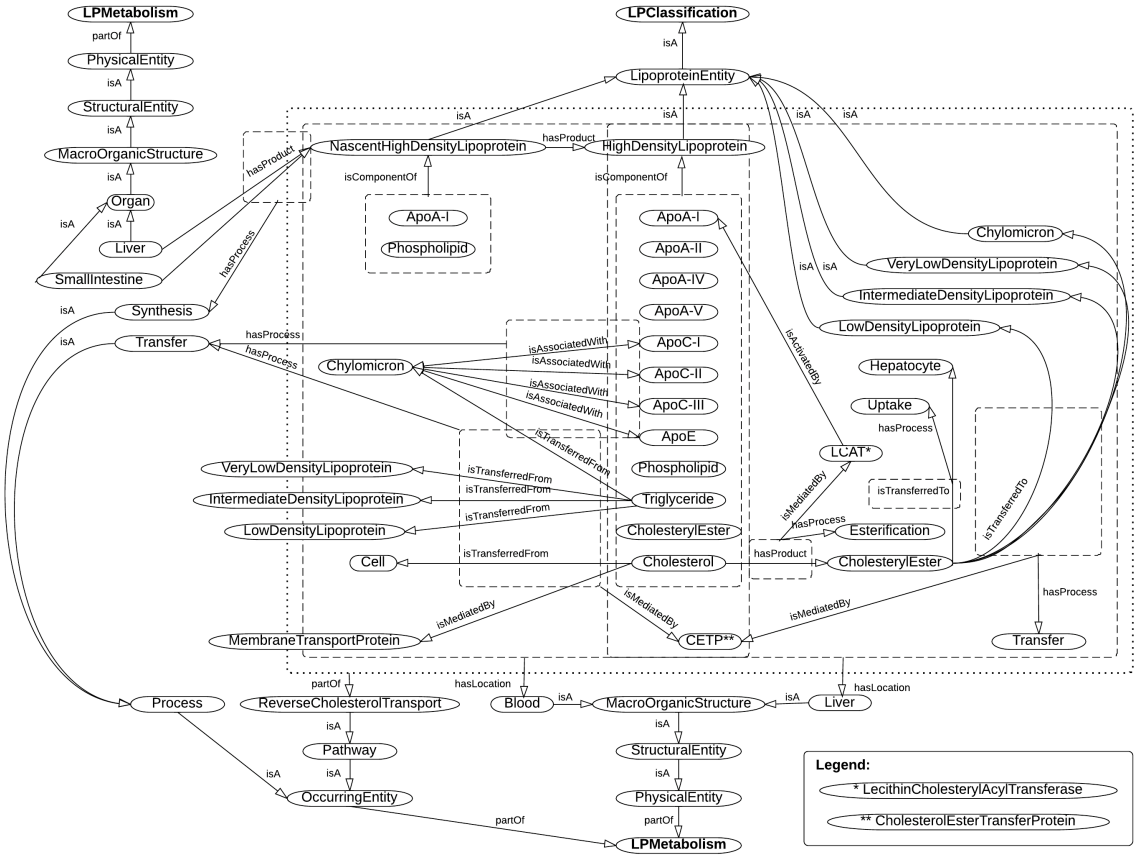


Figure 40. Formalisation of the concept *ReverseCholesterolTransport*.

The **ReverseCholesterolTransport** describes the removal of excess cholesterol from cells from the circulation back to the liver by HDL. HDL is synthesised in the liver and small intestines and released into the circulation as nascent HDL, which contains apoAs and phospholipid. This is defined as "**Liver** and **SmallIntestine** *hasProduct* **NascentHighDensityLipoprotein**" and "*hasProcess* **Synthesis**".

Nascent HDL then acquires cholesterol from cell membranes, which is mediated by membrane transport protein, described as "**Cholesterol** *isTransferredFrom* **Cell**" and "*isMediatedBy* **MembraneTransportProtein**" and "*hasProcess* **Transfer**", as well as triglycerides from other lipoproteins, described as "**Triglyceride** *isTransferredFrom* **Chylomicron**, **VeryLowDensityLipoprotein**, **IntermediateDensityLipoprotein** and **LowDensityLipoprotein**", "*isMediatedBy* **CETP**" and "*hasProcess* **Transfer**". The transfer of these components to nascent HDL leads to mature HDL: "**NascentHighDensityLipoprotein** *hasProduct* **HighDensityLipoprotein**".

Free cholesterol in mature HDL is esterified into cholesteryl esters by lecithin:cholesterol acyltransferase (LCAT) upon activation by apoA-I. This process is described as "**Cholesterol** *hasProduct* **CholesterylEster**", "*hasProcess* **Esterification**", "*isMediatedBy* **LCAT**" and "**LCAT** *isActivatedBy* **ApoA-I**".

Cholesteryl esters can then be taken up by hepatocytes or transferred to apoB-containing lipoproteins such as VLDL, IDL and LDL, by cholesterol ester transfer protein (CETP) in exchange for triglycerides. This process is defined as "**CholesterylEster** *isTransferredTo* **Chylomicron**, **VeryLowDensityLipoprotein**, **IntermediateDensityLipoprotein** and **LowDensityLipoprotein**", "*isMediatedBy* **CETP**" and "*hasProcess* **Transfer**". Earlier, we described the maturation process of nascent HDL to mature HDL to involve the transfer of triglycerides. Hence, we demonstrate the relationship between these two processes to be the same, except with inverse components and products.

7.6.2.1 ParticipantRole

Last but not least, for modularity purposes, we have created the concept **ParticipantRole** to describe the function or role of a **PhysicalEntity**. Figure 41 shows an example of how subclasses of **ParticipantRole** are used to relate concepts under **PhysicalEntity.** An enlarged version of the figure is shown in Appendix L.

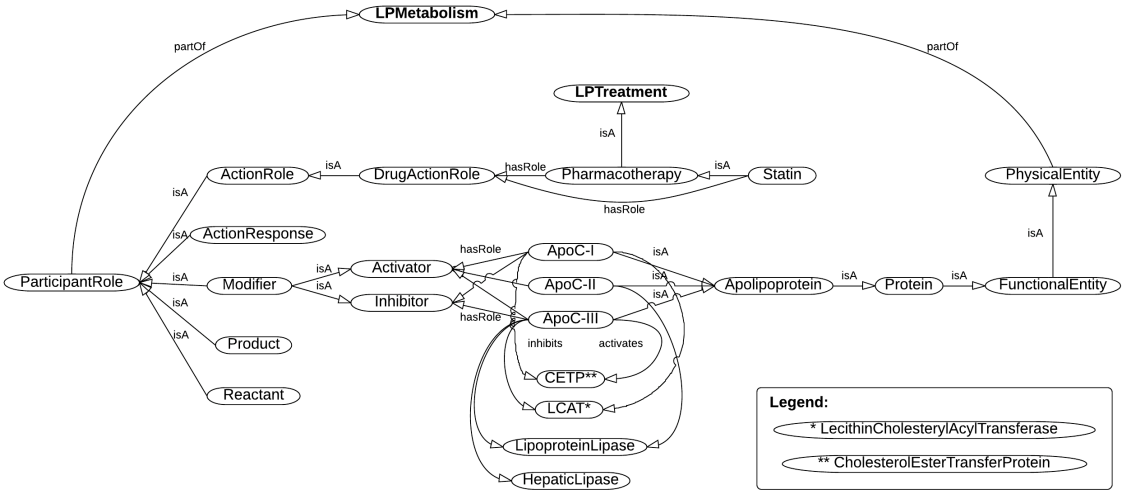


Figure 41. Relations between *ParticipantRole* concepts and *PhysicalEntity* concepts in *LPMetabolism* and *LPTreatment*.

Although classified under the same functional group, many physiological components have different and sometimes opposing functions, and interact differently to the same substance. A very good example of this is the different types of apolipoproteins. Although "Apolipoprotein isA Protein", they have completely different effects on enzymes. We define them as follows:

- **ApoC-I** *hasRole* **Activator** and *activates* **LCAT**
- **ApoC-I** *hasRole* **Inhibitor** and *inhibits* **CETP**
- **ApoC-II** *hasRole* **Activator** and *activates* **LipoproteinLipase**
- **ApoC-III** *hasRole* **Activator** and *activates* **CETP**
- **ApoC-III** *hasRole* **Inhibitor** and *inhibits* **LCAT**, **LipoproteinLipase** and **HepaticLipase**

In addition, the ParticipantRole concept is used to describe the role of other classes in general, such as "**Pharmacotherapy** *hasRole* **DrugActionRole**", "**Enzyme** *hasRole* **EnzymeActionRole**", and so on.

7.6.3 LPPathophysiology

The sub-ontology **LPPathophysiology** contains concepts that are used to define the dysregulation of lipoprotein metabolism. These concepts are primarily used to describe the diagnostic requirements in the class **Patient**. To define the sub-ontology **LPPathophysiology**, we separate concepts of the actual disorder from their physical indication into two classes using the assosiative *partOf* relationship:

- **Disorder**: Defined to be a disease
- **Symptom**: Defined to be the subjective indication of a disease

7.6.3.1 Disorder

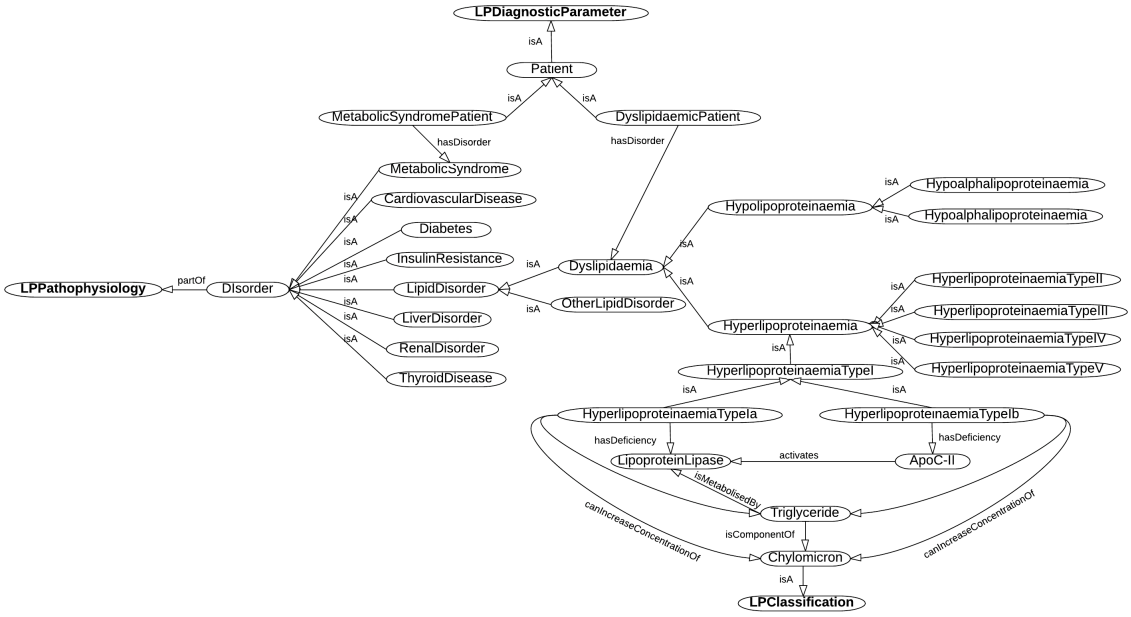


Figure 42. Relations between *Disorder* concepts in *LPMetabolism* and *LPTreatment*.

Figure 42 illustrates the link between **Disorder** concepts in **LPMetabolism** and **LPTreatment**. An enlarged version of the figure is shown in Appendix M. The class **Disorder** contains disease concepts that are related to the dysregulation of lipoprotein metabolism. Although this condition is usually referred to as "dyslipidaemia", we stress that the concept of lipoprotein dysregulation is not necessarily ONLY present in patients with dyslipidaemia. Rather, lipoprotein dysregulation can also (and commonly occur) in patients suffering from

cardiovascular disease, diabetes, and other disorders. Therefore, we defined the concept **Disorder** to include diseases such as:

- MetabolicSyndrome
- CardiovascularDisease
- Diabetes
- InsulinResistance
- LipidDisorder
- LiverDisorder
- RenalDisorder
- ThyroidDisease

For the purpose of this thesis, we will only discuss **Dyslipidaemia** in the context of lipoprotein dysregulation, although the concepts above have been properly defined in our full ontology. These concepts can also serve as linkage to their respective ontologies for fine-grained details of the corresponding concepts, in the case of ontology merging.

Dyslipidaemia is defined to be a **LipidDisorder**. Primary dyslipidaemia, which is caused by genetic factors, is defined to have the subclasses **Hypolipoproteinaemia** and **Hyperlipoproteinaemia**. Again, a these subclasses have been defined in our ontology, we will select an example to illustrate how the process was carried out, as well as the link between the different sub-ontologies, highlighted in bold. As shown in the figure above, **HyperlipoproteinaemiaTypeI** contains two subclasses:

- **HyperlipoproteinaemiaTypeIa**
- **HyperlipoproteinaemiaTypeIb**

We have defined in **LPMetabolism** (Chapter 7.6.2.2.2 EndogenousPathway) that "**ApoC-II** *activates* **LipoproteinLipase**", "**Triglyceride** *isMetabolisedBy* **LipoproteinLipase**" and "**Triglyceride** *isComponentOf* **Chylomicron**". Therefore, if we defined "**HyperlipoproteinaemiaTypeIa** *hasDeficiency* **LipoproteinLipase**" and "**HyperlipoproteinaemiaTypeIb** *hasDeficiency* **ApoC-II**", then "**HyperlipoproteinaemiaTypeIa** and **HyperlipoproteinaemiaTypeIb** *canIncreaseConcentrationOf* **Triglyceride** and **Chylomicron**".

Dyslipidaemia can also be caused by factors other than genetic causes, referred to as secondary dyslipidaemia. We distinguish the difference between primary dyslipidaemia (described above) and secondary dyslipidaemia by creating a separate statement for the latter as "**DyslipidaemicPatient** *hasDisorder* **Dyslipidaemia**". This concept will be discussed further in the **LPAetiology** section.

7.6.3.2 Symptom

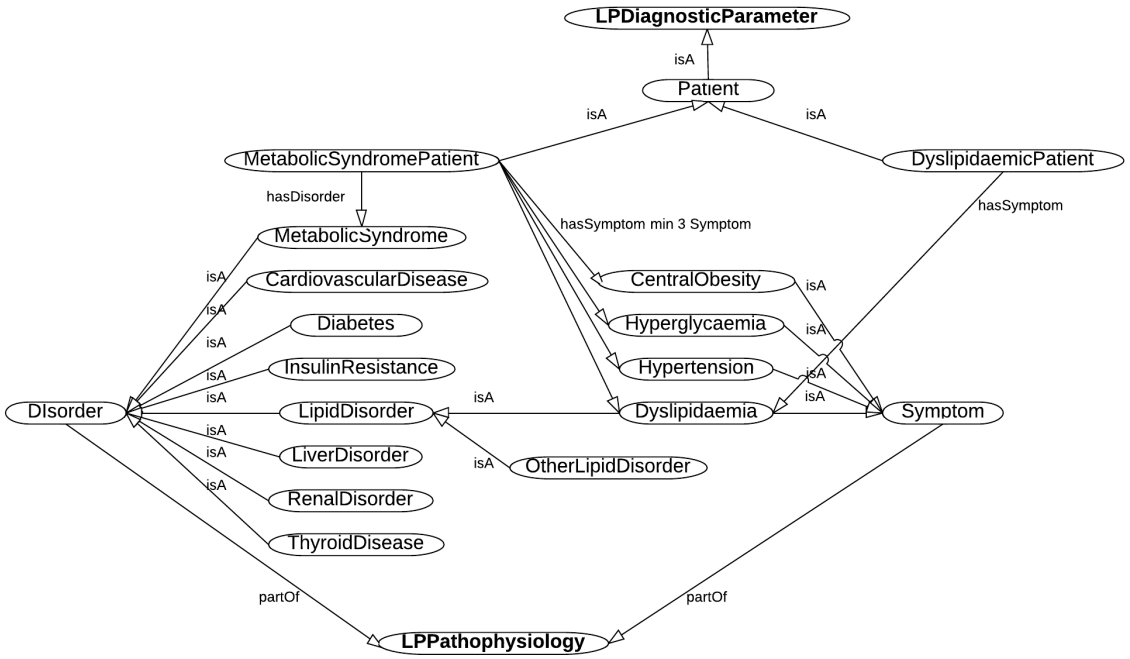


Figure 43. Relations between *Symptom* concepts and *Disorder* concepts in *LPPathophysiology* and *LPDiagnosticParameter*.

Figure 43 demonstrates the relationship between **Symptoms** and **Disorder** concepts in **LPMetabolism** and **LPTreatment**. An enlarged version of the figure is shown in Appendix N. For the purpose of this thesis, the concept **Symptom** includes the four criteria of **MetabolicSyndrome**:

- **Dyslipidaemia**
- **CentralObesity**
- **Hyperglycaemia**
- **Hypertension**

These concepts are linked to the type of **Patient** associated with a particular symptom. For example, "**DyslipidaemicPatient** *hasSymptom* **Dyslipidaemia**", "**ObesePatient** *hasSymptom* **CentralObesity**", "**HyperglycaemicPatient** *hasSymptom* **Hyperglycaemia**" and "**HypertensivePatient** *hasSymptom* **Hypertension**". The properties of **DyslipidaemicPatient**, **ObesePatient**, **HyperglycaemicPatient** and **HypertensivePatient**, i.e. values that define the classification of a symptom according to individual characteristics according to the NCEP/ATP III guideline are discussed in detail in **LPDiagnosticParameter**.

Furthermore, since the metabolic syndrome is defined to have three out of the four criteria above, we have applied a cardinality restriction on the concept **MetabolicSyndromePatient** *hasSymptom* min 3 **Symptom**. Therefore, if an instance of a "**Patient** *hasSymptom* **Dyslipidaemia**, **CentralObesity** and **Hypertension**" AND has fulfilled the necessary and suffficient conditions for their corresponding **Patient** types, then the instance can be inferred to be a **MetabolicSyndromePatient**.

7.6.4 LPAetiology

The sub-ontology **LPAetiology** describes the different causes of lipoprotein dysregulation and contains the following subclasses:

- **LifestyleCause**
- **GeneticCause**
- **DiseaseStateCause**
- **DrugInteraction**

Figure 44 illustrates some of the relations that these concepts have with concepts in other sub-ontologies. An enlarged version of the figure is shown in Appendix O. Here, we demonstrate the link between **LPAetiology**, **LPClassification**, **LPPathophysiology** and **LPTreatment** concepts.

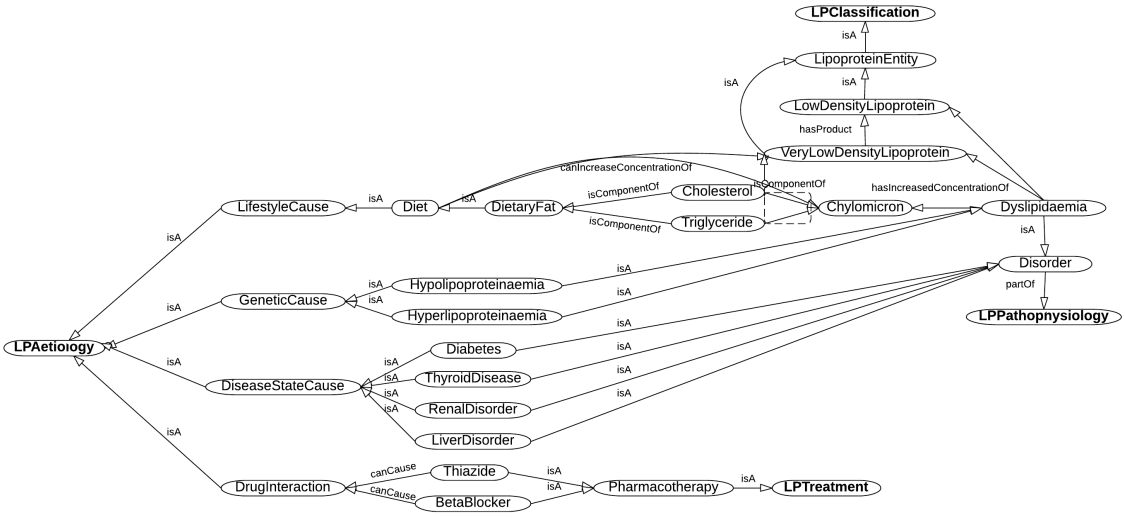


Figure 44. Relations between *LPAetiology*, *LPClassification*, *LPPathophysiology* and *LPTreatment* concepts.

One of the most common causes of lipoprotein dysregulation is **LifestyleCause** due to a sedentary lifestyle with lack of physical exercise and excessive dietary intake of saturated fat and cholesterol. Although **LifestyleCause** contains other subclasses which we have discussed in Chapter 6.6.1, we will be discussing the effect of diet on lipoprotein dysregulation. First, we define "**DietaryFat** *isA* **Diet** *isA* **LifestyleCause**". We have defined previously that "**Chylomicron** and **Triglyceride** *isComponentOf* **DietaryFat**". In LPAetiology, we state that "**Diet** *canIncreaseConcentrationOf* **Chylomicron** and **VeryLowDensityLipoprotein**" Inversely, "**Dyslipidaemia** *hasIncreasedConcentrationOf* **Chylomicron** and **VeryLowDensityLipoprotein**".

In Chapter 7.6.3.1, we defined "**Hypolipoproteinaemia** and **Hyperlipoproteinaemia** *isA* **Dyslipidaemia**". Due to their genetic origins, we link these lipid disorders to the concept "*isA* **GeneticCause**" in **LPAetiology**. Similarly, we stated that "**Diabetes**, **ThyroidDisease**, **RenalDisorder** and **LiverDisorder** *isA* **Disorder**". Hence, in this section, we define the same concepts to be "*isA* **DiseaseStateCause**". **DrugInteraction** is a rather ambitious class, which has the potential to offer much inferential capability in future work. For the purpose of this thesis, we have associated the concepts "**Thiazide** and **BetaBlocker** *hasCause* **DrugInteraction**" and "*isA* **Pharmacotherapy** *isA* **LPTreatment**". For the time being, we have only selected six types of pharmacotherapy medication to treat lipoprotein

dysregulation. However, there are many different drugs that exist that can be included in the superclass **Pharmacotherapy** for the treatment of different **Disorder** that have been or can be added to our ontology. With a much larger subset of treatment options defined in the ontology, it will be possible to create restrictions such that, if a certain drug is mixed with another type, then a drug-drug interaction can occur which may have negative implications on the patient, and can be classified under the class **DrugInteraction**.

7.6.5 LPTreatment

The sub-ontology **LPTreatment** describes treatment options to manage lipoprotein dysregulation and contains the subclasses:

- **LifestyleChange**
- **Pharmacotherapy**
- **CombinationTherapy**

Figure 45 illustrates some of the links that these concepts have with concepts in other sub-ontologies. An enlarged version of the figure is shown in Appendix P. Here, we demonstrate the link between **LPTreatment**, **LPClassification**, **LPMetabolism**, **LPPathophysiology** and **LPAetiology** concepts.

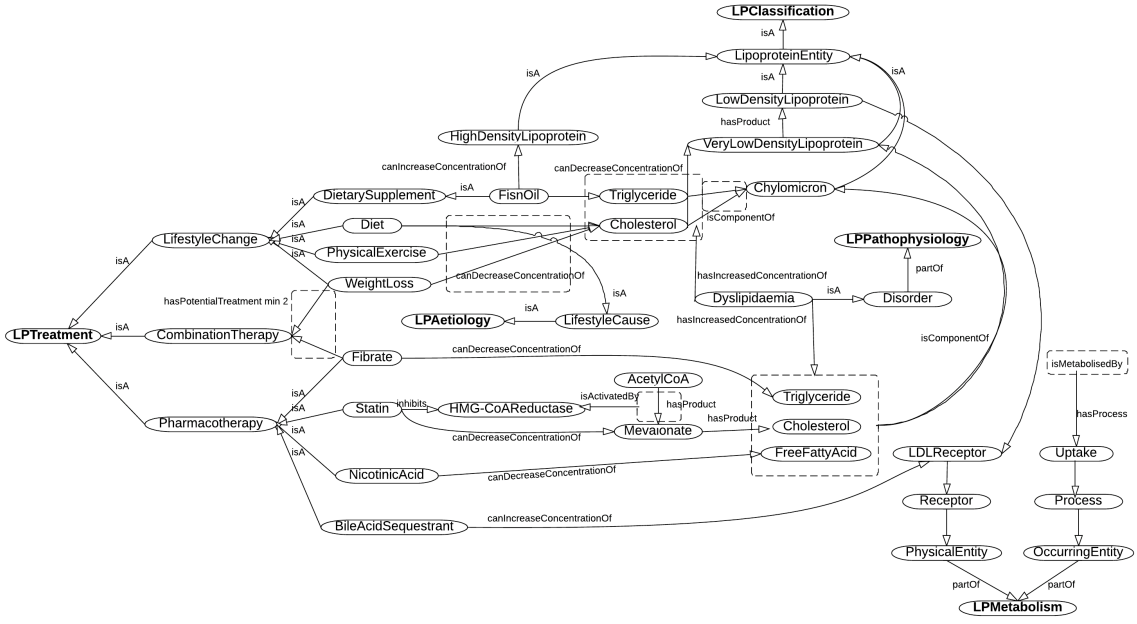


Figure 45. Relations between *LPTreatment*, *LPClassification*, *LPMetabolism*, *LPPathophysiology* and *LPAetiology* concepts.

The first line of approach for treatment of dyslipidaemia is **LifestyleChange**, which is very closely related to **LifestyleCause** in **LPAetiology**. Although these concepts seem very similar at a glance, we created two separate concepts (**LifestyleCause** and **LifestyleChange**) in order to describe them within the context of aetiology and treatment. **LifestyleChange** is defined as the treatment of lipoprotein disorder with changes in lifestyle. Studies have found that fish oil consumption was shown to raise HDL cholesterol and reduce triglycerides, which would in turn decrease the concentration of chylomicron and VLDL. We descibe this as "**FishOil** *isA* **DietarySupplement** *isA* **LifestyleCause**", "**FishOil** *canDecreaseConcentrationOf* **HighDensityLipoprotein**" and "**FishOil** *canDecreaseConcentrationOf* **Triglyceride**". Similarly, "**Diet**, **PhysicalExercise** and **WeightLoss** *canDecreaseConcentrationOf* **Cholesterol**". As we have defined "**Cholesterol** and **Triglyceride** *isComponentOf* **Chylomicon** and **VeryLowDensityLipoprotein**", a decreased concentration of cholesterol and triglyceride implies the decrease in chylomicron and VLDL.

The concept **Pharmacotherapy** contains four drug classes that have been shown to have different lipid lowering properties due to different actions on different receptors. For example, we make the following statement: "**Statin** *inhibits* **HMG-CoAReductase**". Given "**AcetylCoA** *hasProduct* **Mevalonate**", "*isActivatedBy* **HMG-CoAReductase**" and "*hasProduct* **Cholesterol**", the inhibition of HMG-CoAReductase would mean that "**Statin** *canDecreaseConcentrationOf* **Mevalonate**". Since the property *hasProduct* is defined to be transitive, it would imply that "**Statin** *canDecreaseConcentrationOf* **Cholesterol**". We also state "**Fibrate** *canDecreaseConcentrationOf* **Triglyceride**" and "**NicotinicAcid** *canDecreaseConcentrationOf* **FreeFattyAcid**". "**BileAcidSequestrant** *canIncreaseConcentrationOf* **LDLReceptor**". As "**LowDensityLipoprotein** *isMetabolisedBy* **LDLReceptor**" and "*hasProcess* **Uptake**", and increase in LDL receptor would therefore imply faster metabolism of LDL.

For the purpose of this thesis, we have applied a cardinality restriction on **CombinationTherapy**, defined as "*hasPotentialTreatment* min 2". Therefore, any instance with at least two relations containing *hasPotentialTreatment*, then it is inferred to be an instance of **CombinationTherapy**.

7.6.6 LPDiagnosticParameter

The sub-ontology **LPDiagnosticParameter** contains the subclasses **Gender**, **Ethnicity**, **GeographicAncestry** and **Patient**.

7.6.6.1 Gender

For the class **Gender**, we created two subclasses **Female** and **Male**, with equivalent classes *hasGender* **Female** and *hasGender* **Male**, respectively. We will be applying these concepts to the class **Patient** to distinguish different lipid parameters according to different genders. In addition, it is imperative that we defined the necessary and sufficient conditions as such, in order for the reasoner to infer the gender of the **Patient** accordingly.

7.6.6.2 Ethnicity

For the purpose of our ontology, we created three subclasses of **Ethnicity**: **Asian**, **Caucasian** and **Other**, with equivalent classes *hasEthnicity* **Asian**, *hasEthnicity* **Caucasian** and *hasEthnicity* **Other**. This is with respect to the lipoprotein guideline we have opted to use for Lipoprotein Ontology, which lists different lipid measurements for two ethnicities of **Patient**: **Asian** and **Caucasian**. Therefore, we created the concept **Other** for all other ethnicities, which can be can easily be extended to include other ethnic groups. Again, the necessary and sufficient conditions allow us to infer the ethnicity of the **Patient** accordingly.

7.6.6.3 Geographic Ancestry

The class **GeographicAncestry** expands the concept **Ethnicity** a little further and serves as a more realistic view of the world. We created a list of geographical regions which can be linked with the class **Patient** and **Ethnicity** via the object property *hasEthnicity* to define the origins of the individual. This list is in no way complete, but can be extended further to refine the concept **GeographicAncestry**. At this stage, it is beyond the scope of this thesis, but this concept can potentially be linked to the Gene Ontology to infer the possible genetic profile of an individual based on their ethnicity and/or geographic

ancestry. This has implications in genetic testing, and determining different racial/environmental influences on individual lipid profiles.

7.6.6.4 Patient

For the class **Patient**, we created various subclasses which link their respective instances to the subclasses of **Symptom** (under the sub-ontology **LPPathophysiology**) via the object property *hasSymptom*. We focused specifically on the definition of the **MetabolicSyndrome** as well as the formalisation process of the different risk value indicators for dyslipidaemia, classified as **RiskValuePartition**.

7.6.6.4.1 Classification of Metabolic Syndrome Patients

As we have elaborated in Chapter 2.3.3, several guidelines exist towards the identification of the metabolic syndrome, such as the NCEP/ATP III (NCEP, 2001), WHO (WHO, 1998) and IDF (IDF, 2005) definitions. For the purpose of this thesis, we have used the NCEP/ATP III guideline (Table 8), as it is the most commonly used definition for the metabolic syndrome (Huang, 2009). Furthermore, as the next update (NCEP/ATP IV) is anticipated to be released shortly, it remains to be the most current guideline for diagnostic purposes in a clinical setting (Martin et al., 2012).

Table 8. NCEP/ATP III definition of the metabolic syndrome (NCEP, 2001).

Criteria	Three or more criteria
Central Obesity	Waist circumference: * *Caucasian*: ≥ 102 cm (men), ≥ 88 cm (women) * *Asian*: ≥ 90 cm (men), ≥ 80 cm (women)
Hyperglycaemia	* *Fasting plasma glucose level:* ≥ 110 mg/dL
Dyslipidaemia	* *Triglyceride levels:* ≥ 150 mg/dL * *HDL-cholesterol level:* < 40 mg/dL (men), < 50 mg/dL (women)
Hypertension	* *Systolic BP:* ≥ 130 mmHg * *Diastolic BP:* ≥ 85 mmHg

Based on the NCEP/ATP III guideline, a person is classified as a metabolic syndrome patient if s/he suffers from any three of the following symptoms: central obesity, hyperglycaemia, dyslipidaemia and hypertension. Hence, we had to clearly identify patients which fall under each criteria as **ObesePatient**, **HyperglycaemicPatient**, **DyslipidaemicPatient** and **HypertensivePatient** respectively, prior to defining the concept **MetabolicSyndromePatient**. As an example, Figure 46 shows the concept **ObesePatient**, for which we defined to have the necessary and sufficient conditions to infer that instances of **ObesePatient** *hasSymptom* **CentralObesity**. In addition, we have also defined **ObesePatient** to have datatype property *hasWaistCircumference* for instances which are restricted along the object properties *hasEthnicity* as well as *hasGender* in order to correctly classify the gender and ethnicity of patients according to the NCEP/ATP III criteria.

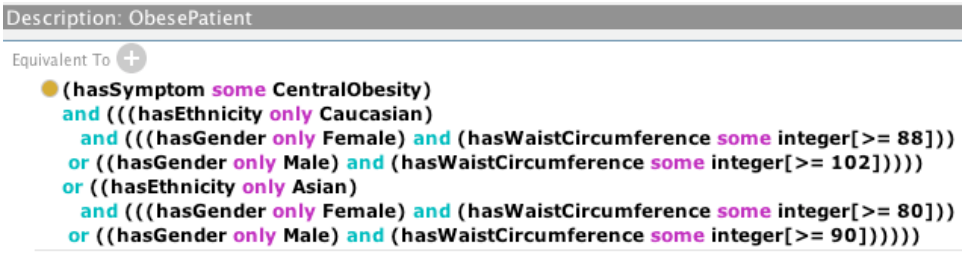


Figure 46. Formalisation of the concept *ObesePatient* in Protégé.

Similarly, we have defined necessary and sufficient conditions for the concept **HyperglycaemicPatient**, for which they are linked to the property *hasSymptom* to **Hyperglycaemia**, as well as the datatype property *hasFastingPlasmaConcentration*.

For the concept **DyslipidaemicPatient**, we have defined necessary and sufficient conditions which include the property *hasSymptom* **Dyslipidaemia**, as well as datatype properties *hasTriglycerideConcentration* and *hasHDLConcentration*. The datatype property *hasHDLConcentration* is restricted along the object properties *hasEthnicity* as well as *hasGender* in order to correctly classify the gender and ethnicity of patients according to the NCEP/ATP III criteria.

Correspondingly, we have defined necessary and sufficient conditions for the concept **HypertensivePatient** to include the property *hasSymptom*

Hypertension, as well as datatype properties *hasBloodPressureSystolicValue* and *hasBloodPressureDiatolicValue*.

As the class **MetabolicSyndromePatient** contains individuals who suffer from three or more of the criteria defined above, we applied the cardinality restriction *hasSymptom* min 3 Thing to specify the minimum number (3) of *hasSymptom* relationships that an instance must have order to be inferred as a **MetabolicSyndromePatient**. In other words, **MetabolicSyndromePatient** has the equivalent class **Patient** and (*hasSymptom* min 3 **Thing**).

7.6.6.4.2 Risk Value Partition

In Chapter 2.4.1, we presented various guidelines which are used for the diagnosis of dyslipidaemia, such as the NCEP/ATP III (NCEP, 2001), AHA (Miller et al., 2011) and AACE (Jellinger et al., 2012). The guidelines provide an indication of risk (normal, borderline risk, high risk, very high risk) for various plasma levels of total triglycerides, cholesterol, LDL cholesterol and HDL cholesterol. Again, as the NCEP/ATP III guideline is the most frequently used guideline for dyslipidaemia management, we have incorporated this classification in our ontology (Table 9).

Table 9. NCEP/ATP III lipid profile classification of risk factors (NCEP, 2001).

LDL-Cholesterol (mg/dL)	Risk Value Indicator
< 100	Optimal
100 - 129	Near optimal
130 - 159	Borderline high
160 - 189	High
≥ 190	Very high
Total Cholesterol (mg/dL)	**Risk Value Indicator**
< 200	Optimal
200 - 239	Borderline high
240 -249	High
≥ 250	Very High
Total Triglyceride (mg/dL)	**Risk Value Indicator**
< 150	Normal
150 - 199	Borderline high
200 - 499	High
≥ 500	Very high
HDL-Cholesterol (mg/dL)	**Risk Value Indicator**
< 40	High
> 60	Optimal

To formalise and define the concept of risk value indicators associated with different plasma levels of total triglycerides, cholesterol, LDL cholesterol and

HDL cholesterol, we created the class **RiskValuePartition**, which contains the subclasses **Normal**, **BorderlineRisk**, **HighRisk** and **VeryHighRisk**, disjoint to one another.

We defined concepts for **Normal** (Figure 47), **BorderlineRisk**, **HighRisk** and **VeryHighRisk** based on the guideline above using necesssary and sufficient conditions: object property **hasRiskValue**, as well as datatype properties **hasTotalTriglycerideConcentration**, **hasTotalCholesterolConcentration**, **hasLDLCholesterolConcentration** and **hasHDLCholesterolConcentration**.

Figure 47. Formalisation of the concept *Normal* in Protégé.

Having defined the subclasses of **RiskValuePartition**, we subsequently created a covering axiom for **RiskValuePartition** such that **RiskValuePartition** is covered by the classes **Normal**, **BorderlineRisk**, **HighRisk** and **VeryHighRisk**. This is necessary due to the open world assumption feature of ontologies; a covering axiom ensures that a member of **RiskValuePartition** must be a member of either **Normal** or **BorderlineRisk** or **HighRisk** or **VeryHighRisk**.

7.7 Conclusion

This chapter discussed the formalisation process of the lipoprotein conceptual framework described in Chapter 6. First, we provided an overview of KR languages associated with ontology building, and the justification behind our choice of implementation tool and language. This is followed by a brief introduction to the components of OWL ontologies and how they were utilised in Lipoprotein Ontology. We then discussed the structure of OWL ontologies, as well as the notation and syntax used throughout this chapter. We concluded this chapter by presenting a visualisation of Lipoprotein Ontology in figures and screenshots from Protege to represent various functions of concept classification, definition and relations, using various OWL constructs.

References

Baader, F., Calvanese, D., McGuinness, D., et al. (2003). *The Description Logic Handbook: Theory, Implementation and Applications. Cambridge Press*.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web, from http://www.scientificamerican.com/article.cfm?id=the-semantic-web

Courtot, M., Juty, N., Knüpfer, C., Waltemath, D., et al. (2011). Model storage, exchange and integration. *Molecular Systems Biology, 7*, 543.

GO Consortium. (2000). Gene ontology: tool for the unification of biology. *Nature Genetics, 25*, 25-29.

Helsper, E. M., & van der Gaag, L. C. (2002). *Building bayesian networks through ontologies.* Paper presented at the 15th European Conference on Artificial Intelligence.

Horrocks, I. (2000). The Ontology Interference Layer (OIL).

Horrocks, I. (2001). DAML+OIL.

Horrocks, I. (2008). Ontologies and the semantic web. *Communications of the ACM 51*.

Horrocks, I., Patel-Schneider, P. F., & van Harmelen, F. (2003). From SHIQ and RDF to OWL: the making of a web ontology language. *Web Semantics: Science, Services and Agents on the World Wide Web, 1*(1), 7–26.

Huang, P. L. (2009). A comprehensive definition for metabolic syndrome. *Disease Models and Mechanisms, 2*(5-6), 231–237.

IDF, International Diabetes Federation. (2005). The IDF consensus worldwide definition of the metabolic syndrome.

Jellinger, P. S., Smith, D., Mehta, A., et al. (2012). American Association of Clinical Endocrinologists' guidelines for management of dyslipidemia and prevention of atherosclerosis. *Endocrine Practice, 18*(1), 1-78.

Lowe, H. J., & Barnett, G. O. (1994). Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches. *JAMA, 271*(14), 1103-1108.

Miller, M., Stone, N. J., Ballantyne, C., Bittner, V., Criqui, M. H., Ginsberg, H. N., Pennathur, S. (2011). Triglycerides and Cardiovascular Disease: A Scientific Statement From the American Heart Association. *Circulation, 123*, 2292-2333.

NCEP, Expert Panel on Detection and Treatment of High Blood Cholesterol in Adults. (2001). Executive Summary of the Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III). *JAMA, 285*, 2486-2497.

O'Connor, M. J. (2006). Semantic Web Rule Language, from http://protege.cim3.net/file/pub/files/SWRL/SWRLTalkProtegeShortCourse2006.pdf

Smith, B., Ceusters, W., Klagges, B., et al. (2005). Relations in biomedical ontologies. *Genome Biology, 6*(5), R46.

W3C. (2012). OWL 2 Web Ontology Language. http://w3.org/TR/owl2-overview.

Chapter 8. Evaluation of Lipoprotein Ontology

8.1 Introduction

In this chapter, we evaluate Lipoprotein Ontology against several criteria. There are a number of evaluation methods for ontologies which include comparison to a "golden standard", functional performance, conceptual coverage, among others (Brank et al., 2005; Gangemi et al., 2005; Gruninger & Fox, 1995; Hartmann et al., 2005). It is important to note that the validity of an ontology might vary between different users or different domains. In this thesis, we have used three strategies to evaluate Lipoprotein Ontology as follows:

- Evaluation of the syntactic quality of Lipoprotein Ontology
- Evaluation of the conceptual coverage of Lipoprotein Ontology
- Evaluation of the practical application of Lipoprotein Ontology through the validation of competency questions and case studies

8.2 Ontology Evaluation Aspects

Different evaluation methods are used for different purposes and applications. As ontologies are complex structures, it is often more practical to evaluate different aspects of the ontologies separately rather than focusing on the ontologies as a whole (Brank et al., 2005). Broadly, these aspects include:

- **Syntactic level**: An ontology which was manually constructed can be evaluated at the syntactic level by using syntactic considerations such as the presence of natural-language documentation, as well as the adherence towards several design criteria, which we will discuss further in the next section.

- **Conceptual or lexical level**: The vocabulary used to represent classes, instances and axioms in the ontology is evaluated through comparisons with various sources of data concerning the domain of interest such as domain-specific text corpora. This method is also referred to as the concept coverage method.

- **Contextual or application level**: The context of the ontology can be evaluated through the use of competency questions that were set out in at the initial stages of ontology development. The ontology can also be evaluated at the application level through the use of case studies and the capability of the ontology to answer scenarios.

8.3 Evaluation of the Syntactic Quality of Lipoprotein Ontology

The basis of Lipoprotein Ontology was syntactically derived from rigorous literature review of peer-reviewed scientific journals and biomedical textbooks in the lipoprotein domain, which was documented in Chapter 2 of this thesis. General aspects of lipoprotein research were extracted and identified as upper self-standing concepts of Lipoprotein Ontology such as *Classification*, *Metabolism*, *Pathophysiology*, *Aetiology*, *Treatment* and *Diagnostic Parameter*. The quality of Lipoprotein Ontology was evaluated at the syntactic level using the five design criteria as follows (Brank et al., 2005)

- **Completeness**: Concepts in Lipoprotein Ontology along with their relationships with other concepts are explicitly stated, and their properties restricted accordingly. Each definition is documented in natural language but described as formal axiom in the ontology. Most of the definitions in Lipoprotein Ontology are defined with necessary and sufficient conditions with respect to their scientific definitions, but some of the simpler concepts are defined with only necessary conditions, in order to build complex definitions out of atomic concepts.

- **Consistency**: The definitions in Lipoprotein Ontology are consistent and do not include contradictory information. In order to ensure the consistency of the concepts in the ontology, a single term is given to every lipoprotein concept; where acronyms, synonyms and abbreviations

exist which refer to the same concept, we have included these terms in the annotation property of the concept respectively. In addition, inferences are found to be satisfactorily consistent with existing definitions and axioms, described in Chapter 8.5.1.

- **Conciseness**: Lipoprotein Ontology does not store any unnecessary definitions. Every concept is linked to at least one other concept in Lipoprotein Ontology and contains the respective properties and restrictions.

- **Extensibility**: Despite the conciseness of Lipoprotein Ontology, users can add new definitions to the ontology and more knowledge to the definitions based on the existing vocabulary, without altering the set of well-defined properties and/or definitions. Also, due to the hierarchical structure and normalised design of Lipoprotein Ontology, new concepts can be added under existing concepts without affecting neighbouring concepts and their properties. From these new concepts, we can then build even more complex definitions in the ontology.

- **Minimal encoding bias**: Conceptualisations are specified at knowledge-level and not at a symbol or notation level. As such, the proper use of relations is necessary in order to maintain the integrity of the ontology. For example, for the superclass **LPMetabolism**, although it may appear to be more convenient to use the subclass relation “isA” instead of “partOf” expression, the “isA” relation is not the correct relation for its respective subclasses **PhysicalEntity**, **OccurringEntity** and **ParticipantRole**. Rather, we used the “partOf” relation in order to maintain the correct formal definition for their corresponding axioms.

8.4 Evaluation of the Conceptual Coverage of Lipoprotein Ontology

One of the most common methods of ontology evaluation is through its concept coverage. This evaluates the ontology at the lexical or vocabulary level as described in the previous section, and can be done by determining the percentage of domain concepts which are covered or represented by the ontology.

Our evaluation method involves the comparison of concepts contained in Lipoprotein Ontology against a selected data source from the lipoprotein research domain. Articles were extracted from PubMed, the largest online repository for biomedical literature, selected and analysed for evaluation of Lipoprotein Ontology. On 9th March 2013 we searched for the keyword "lipoprotein" in the PubMed database. The search engine originally returned a total of 157,964 articles; we narrowed down our search to "humans" by applying a filter on the database to only include articles on human subjects, which reduced the result to 115,815 articles. From this set of articles, 90 article abstracts were randomly selected according to the core lipoprotein concepts in Lipoprotein Ontology, namely *Classification*, *Metabolism*, *Pathophysiology*, *Aetiology*, *Treatment* and *Diagnostic Parameter*. Please refer to Appendix Q for the complete list of the abstract titles.

In the next phase, we examined these 90 article abstracts and extracted their lipoprotein-related concepts. Lipoprotein Ontology is then used to encode the knowledge from this test set. Determining the conceptual coverage includes the calculation of the percentage of sentences that can be fully represented by Lipoprotein Ontology.

For example, we analysed the sentence "VLDL, the main secretory lipoprotein of the liver, contains cholesterol, phospholipid, triglyceride, newly synthesized B-100 and small amounts of E and C apopeptides" (Olson, 1998) from one of the randomly selected abstracts, in terms of its lipoprotein-related concepts. From this statement, we identified a number of lipoprotein concepts such as **VeryLowDensityLipoprotein**, **Liver**, **Cholesterol**, **Phospholipid**, **Triglyceride**, **ApoB-100**, **ApoE** and **ApoC**. These concepts were categorised under the various sub-ontologies of Lipoprotein Ontology accordingly. For example, **VeryLowDensityLipoprotein** was classified under **LPClassification**, and the following were classified under **LPMetabolism**: **Liver**, **Cholesterol**, **Phospholipid**, **Triglyceride**, **ApoB-100**, **ApoE** and **ApoC**. Figure 48 illustrates the process of manual concept mapping.

ABSTRACT The idea of a fat transport system in the plasma of mammals evolved slowly over three centuries. At the turn of this century, it was discovered that plasma globulins contained lecithin and that the digestion of plasma proteins with pepsin liberated small amounts of fat and cholesterol. The high density lipoprotein (HDL) was first isolated from horse serum in 1929 and the low density lipoprotein (LDL) in 1950. It was then shown that flotation of plasma in the ultracentrifuge revealed an array of lipoproteins that included VLDL, LDL and HDL and permitted quantitation. Subsequently, it was discovered that the free fatty acids (FFA) in plasma were bound to albumin and varied with feeding and fasting. From further studies, it was concluded that lipoprotein-bound triglycerides were delivered to adipose cells for uptake after meals; during fasting, the fat cells secreted FFA, which provided fuel for many tissues. The protein components of the lipoproteins (apopeptides) were characterized in the period from 1960 to 1970 and the LDL-receptor was identified in 1974. Fat transport was then established as a receptor-mediated delivery system of lipoproteins to targeted tissues. Defects in this system due to genetically altered or absent receptors explained dyslipidemias, which promoted atherosclerosis, xanthomatosis and Alzheimer's disease.

LPClassification: HDL, LDL, VLDL
LPMetabolism: Lecithin, protein, pepsin, fat, cholesterol, free fatty acids, albumin, triglyceride
LPPathophysiology: Dyslipidemia, atherosclerosis, xanthomatosis, Alzheimer's disease
LPAetiology: Genetic, absent receptor

Figure 48. The mapping of lipoprotein concepts in an article abstract to Lipoprotein Ontology (Olson, 1998).

Table 10 shows the results of concept mapping for the abstract that was randomly selected.

Table 10. Concept mapping of lipoprotein concepts in the article abstract shown in Figure 48.

Sub-Ontology Category	Extracted Concept from the Article Abstract	Equivalent or Similar Lipoprotein Ontology Concept
LPClassification	High density lipoprotein	HighDensityLipoprotein
	Low density lipoprotein	LowDensityLipoprotein
	VLDL	VeryLowDensityLipoprotein
	LDL	LowDensityLipoprotein
	HDL	HighDensityLipoprotein
LPMetabolism	Lecithin	Lecithin
	Protein	Protein
	Pepsin	-
	Fat	Lipid
	Cholesterol	Cholesterol
	Free fatty acid	FreeFattyAcid
	Albumin	Albumin
	Triglyceride	Triglyceride
LPPathophysiology	Dyslipidaemia	Dyslipidaemia
	Atherosclerosis	Atherosclerosis
	Xanthomatosis	-
	Alzheimer's disease	-
LPAetiology	Genetic	GeneticCause
	Absent receptor	ReceptorDeficiency

To calculate the concept coverage of Lipoprotein ontology, we used the following formula (Olson, 1998). Coverage(S) is defined as the measure of the level of coverage of a set of concepts S in a given ontology. Specifically, Coverage({C1,…,Cn}) was computed using the following formula (with |O| being the size of the ontology O given as the number of concepts included in O):

$$\text{Coverage}(\{C1,\ldots,Cn\}) = \frac{|\{\text{Covered}(C1) \cup \ldots \cup \text{Covered}(Cn)\}|}{|O|}$$

Table 11 presents a summary of our evaluation results from the mapping process of the 90 randomly selected article abstracts in the lipoprotein domain.

Table 11. Evaluation results of the concept mapping process.

Lipoprotein Ontology Sub-Ontology Category	% of Lipoprotein Concepts with Equal or Similar Concepts
LPClassification	100
LPMetabolism	83.3
LPPathophysiology	79.1
LPAetiology	62.9
LPTreatment	61.2
LPDiagnosticParameter	76.5

Our results indicate that we were able to match almost all of the extracted lipoprotein concepts under our lipoprotein sub-ontologies, namely LPClassification, LPMetabolism, LPPathophysiology, LPAetiology, LPTreatment and LPDiagnosticParameter. Therefore, these sub-ontologies can serve as upper-level categories for lipoprotein-related concepts from other biomedical ontologies.

8.5 Evaluation of the Practical Application of Lipoprotein Ontology

The practical application of Lipoprotein Ontology was evaluated at the context level using two methods:

- Validation of competency questions
- Evaluation of Lipoprotein Ontology through case studies

8.5.1 Validation of Competency Questions

We evaluated Lipoprotein Ontology against the list of competency questions that were defined at the specification stage of ontology development. The competency questions encompass the six sub-ontologies of Lipoprotein Ontology: Classification, Metabolism, Pathophysiology, Aetiology, Treatment and Diagnostic Parameter. Validating the competency questions was carried out in order to check that the ontology has successfully represented relationships present in the initial documents or definitions, and that it has satisfied its purpose.

8.5.1.1 Classification

Can the ontology identify lipoprotein particles based on their properties?

To address this competency question, we created two unknown instances LipoproteinX and LipoproteinY, which belong to the general class **Thing**, and not to any **LipoproteinEntity** subclass. Without asserting the lipoprotein class these unknown instances belong to, we assigned properties from which the reasoner can automatically classify these instances to belong to their respective classes. Based on this, we initially verified that the ontology is capable of identifying all instances of the class **LipoproteinEntity**. As we have described in the previous chapter, each **LipoproteinEntity** subclass contains instances that were asserted to be members of their respective classes. This serves as a baseline measure for us to verify the correct identity of the unknown instances.

DL Query: LipoproteinEntity

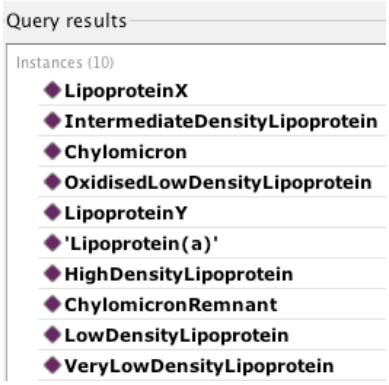


Figure 49. Query result for LipoproteinEntity.

In the first example, shown in Figures 50 A, B and C below, we queried the ontology from the most general, narrowing down to more specific properties, such as the size of the lipoprotein particle, apolipoprotein content and electrophoretic mobility value.

DL Query (A): **LipoproteinEntity** and (*hasSize* some integer[>=75, <=1200])
DL Query (B): **LipoproteinEntity** and (*hasSize* some integer[>=75, <=1200]) and (isAssociatedWith some **ApoB-48**)
DL Query (C): **LipoproteinEntity** and (*hasSize* some integer[>=75, <=1200]) and (*isAssociatedWith* some **ApoB-48**) and (*hasElectrophoreticMobility* some **Origin**)

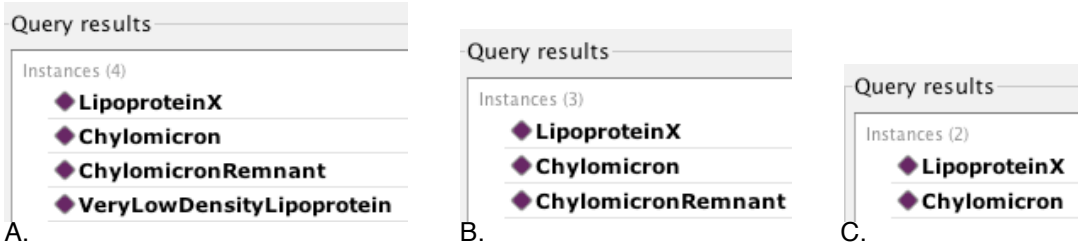


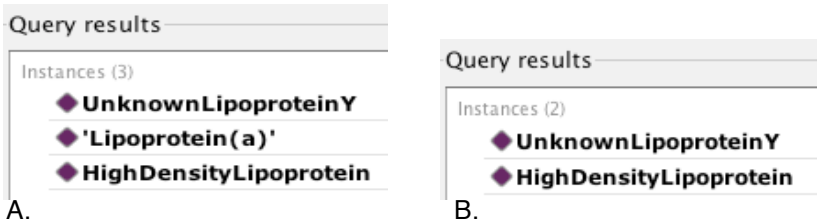
Figure 50. Query result for competency question Classification – Size.

Referring to our table of major lipoprotein classes in Appendix A, the corresponding sizes of the three lipoprotein classes are as follows: **Chylomicron** (75mm-1200nm), **ChylomicronRemnant** (30-80nm) and **VeryLowDensityLipoprotein** (30-80nm). Based on the example above, Query A returned instances which have the datatype property *hasSize* between 75nm and 1200nm: Chylomicron (78nm), ChylomicronRemnant (80nm), VeryLowDensityLipoprotein (77nm) and LipoproteinX (79nm). This narrowed down the identity of LipoproteinX to be either of these four lipoprotein classes. If another query parameter was added to Query B, such as the object property *isAssociatedWith* a certain type of apolipoprotein, e.g. **ApoB-48**, the query results automatically returned the instances Chylomicron, ChylomicronRemnant and LipoproteinX, thus excluding VeryLowDensityLipoprotein, as it does not contain ApoB-48. To narrow down our search further, we applied an additional query parameter to Query C, the object property *hasElectrophoreticMobility* only **Origin**. As we have restricted the property to have the universal restriction "only" along the class/value **Origin**, we can satisfactorily infer that LipoproteinX is a

Chylomicron. This was verified by the query results, which also returned the instance Chylomicron.

The second example, shown in Figures 51 A and B below, is similar to the first, where we queried the ontology from the most general, narrowing down to more specific properties, such as the density of the lipoprotein particle and apolipoprotein content.

DL Query (A): **LipoproteinEntity** and (*hasDensity* some float[>=1.063f, <=1.21f])
DL Query (B): **LipoproteinEntity** and (*hasDensity* some float[>=1.063f, <=1.21f]) and (*isAssociatedWith* some **ApoA-I**)



A. B.

Figure 51. Query result for competency question Classification – Density.

Referring to our table of major lipoprotein classes in Appendix A, the corresponding densities of the two lipoprotein classes are as follows: **Lipoprotein(a**) (1.050-1.120g/ml) and **HighDensityLipoprotein** (1.063-1.210g/ml). Based on the example above, Query A returned instances which have the datatype property *hasDensity* between 1.063 and 1.210g/ml: HighDensityLipoprotein, Lipoprotein(a) and LipoproteinY. This narrowed down the identity of LipoproteinY to be either of these three lipoprotein classes. If another query parameter was added to Query B, such as the object property *isAssociatedWith* a certain type of apolipoprotein, e.g. ApoA-I, the query results automatically returned the instances HighDensityLipoprotein and LipoproteinX, thus excluding Lipoprotein(a) as it does not contain ApoA-I. Based on this, we can satisfactorily infer that LipoproteinY is a **HighDensityLipoprotein**. This is verified by the query results, which also returns the instance HighDensityLipoprotein.

8.5.1.2 Metabolism

Can the ontology identify the apolipoprotein classes that are associated with their corresponding lipoprotein entities?

First, we verified that the reasoner is capable of identifying all instances of the class **Apolipoprotein**.

DL Query: Apolipoprotein

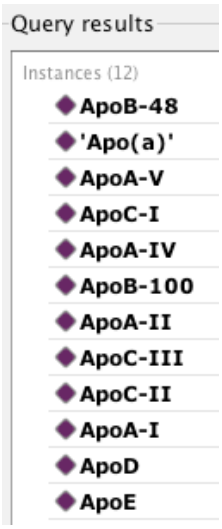


Figure 52. Query result for Apolipoprotein.

The relationship between the subclasses of **LipoproteinEntity** and the subclasses of **Apolipoprotein** is described using the object property *isAssociatedWith* as follows: **Chylomicron** *isAssociatedWith* some **ApoA-I**, **ApoA-II**, **ApoA-IV**, **ApoA-V**, **ApoB-48**, **ApoC-I**, **ApoC-II**, **ApoC-III**, **ApoE**. It is important to note here that the object property *isAssociatedWith* was defined to be symmetric; thus, the subclasses of Apolipoprotein contain no additional properties. In Figures 53 A and B below, we illustrate the inferencing capability of the symmetric property, *isAssociatedWith*.

DL Query (A): **Apolipoprotein** and (*isAssociatedWith* some **Chylomicron**])
DL Query (A): **Apolipoprotein** and (*isAssociatedWith* some **ChylomicronRemnant**])

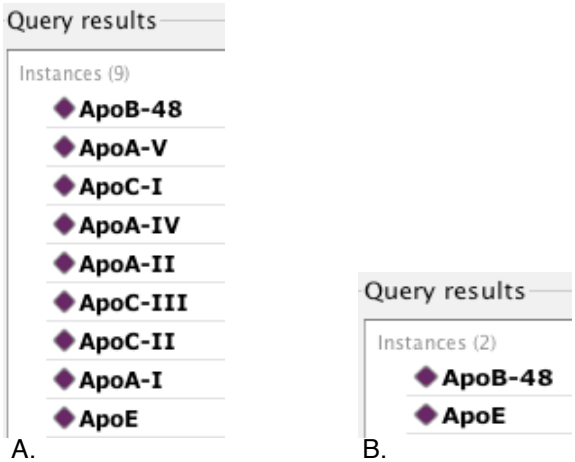


Figure 53. Query result for competency question Metabolism.

As the figure above illustrates, due to the symmetric property *isAssociatedWith* that was asserted to describe the relationship between the class **Chylomicron** and its corresponding **Apolipoprotein** subclasses, the reasoner is able to infer the **Apolipoprotein** subclasses that are associated with the lipoprotein class **Chylomicron** along the symmetric property *isAssociatedWith*. Respectively, we show this relation with the lipoprotein class **ChylomicronRemnant** and demonstrate the exact same inference result.

8.5.1.3 Pathophysiology

Can the ontology identify metabolic sydrome based on a patient's symptoms?

We used two unknown instances DisorderX and DisorderY, which belong to the general class **Thing**, and not to any **Disorder** subclass. Without asserting the classes of disorders these unknown instances belong to, we created properties from which the reasoner can automatically classify these instances to belong to their respective classes. We initially verified that the ontology is capable of identifying all instances of the class **Disorder**. This serves as a baseline measure for us to verify the correct identity of the unknown instances.

DL Query: **Disorder**

Figure 54. Query result for Disorder.

DisorderX contains the properties *hasSymptom* some **Dyslipidaemia** and **Hypertension**, whereas DisorderY contains the properties *hasSymptom* some **Dyslipidaemia**, **Hypertension** and **Hyperglycaemia**. Figure 55 shows the result of our query with the following statement.

DL Query: **Patient** and (hasDisorder some **MetabolicSyndrome**)

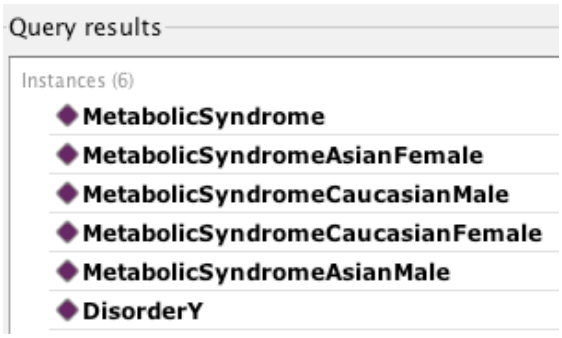


Figure 55. Query result for competency question Pathophysiology.

As we have defined metabolic syndrome to have the property *hasSymptom* min 3 **Thing**, the reasoner has identified DisorderY to be an instance of metabolic syndrome based on the properties (i.e. three *hasSymptom* relations) it contains.

8.5.1.4 Aetiology

Can the ontology identify patients who are at risk of developing dyslipidaemia based on their lifestyle and/or genetic predisposition?

To address this competency question, we presented an example using two unknown individuals PatientX and PatientY, which belong to the general class **Thing**. Without indicating any symptoms and disorders, we created properties from which the reasoner can infer these individuals who are at risk of developing dyslipidaemia, which belong to the class **PatientAtRisk**. First, we verified that the ontology is capable of identifying all instances of the class **PatientAtRisk**, shown in Figure 56 below.

DL Query: **PatientAtRisk**

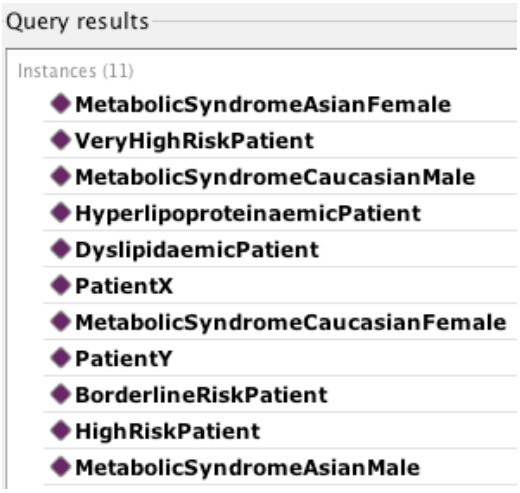
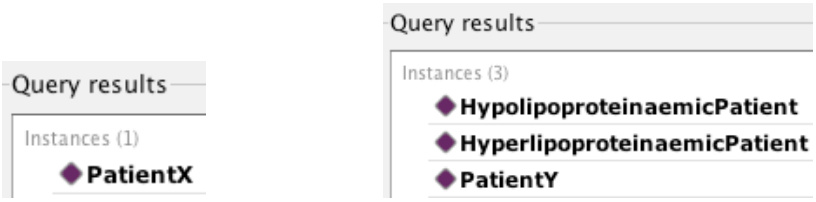


Figure 56. Query result for PatientAtRisk.

DL Query (A): **PatientAtRisk** and (*hasLifestyleCause* some **Smoking**)
DL Query (B): **PatientAtRisk** and (*hasCause* some **GeneticCause**)



A. B.

Figure 57. Query result for competency question Aetiology – LifestyleCause.

In Figure 57 A, the individual PatientX has been identified to be an instance of **PatientAtRisk** due to a smoking habit. Figure 57 B shows instances of the anonymous class which contain the intersection of **PatientAtRisk** due to genetic factors: PatientY, HyperlipoproteinaemicPatient and HypolipoproteinaemicPatient. In order to identify the disorder that PatientY is suffering from, we query the ontology using a property that PatientY is associated with, shown in Figure 58 below.

DL Query: **PatientAtRisk** and (*hasCause* some **GeneticCause**) and (*hasDeficiency* some **ApoB-100**)

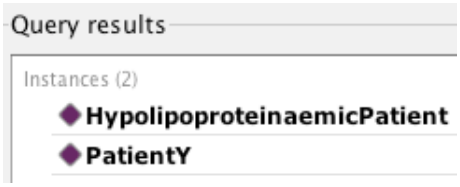


Figure 58. Query result for competency question Aetiology – GeneticCause.

Therefore, based on the ontology query, PatientY is inferred to be a member of **PatientAtRisk**, and suffering from hypolipoproteinaemia

8.5.1.5 Treatment

Can the ontology identify potential treatments for patients with high cholesterol, high triglyceride, and both?

Figures 59 A, B and C illustrate the types of treatments that can be prescribed for patients with high cholesterol, high triglyceride and both, respectively. We present the evaluation results as classes instead of their instances in order to keep this section concise.

DL Query A: **LPTreatment** and (*canDecreaseConcentrationOf* some **Cholesterol**)
DL Query B: **LPTreatment** and (*canDecreaseConcentrationOf* some **Triglyceride**)
DL Query C: **LPTreatment** and ((*canDecreaseConcentrationOf* some **Cholesterol**) or (*canDecreaseConcentrationOf* some **Triglyceride**))

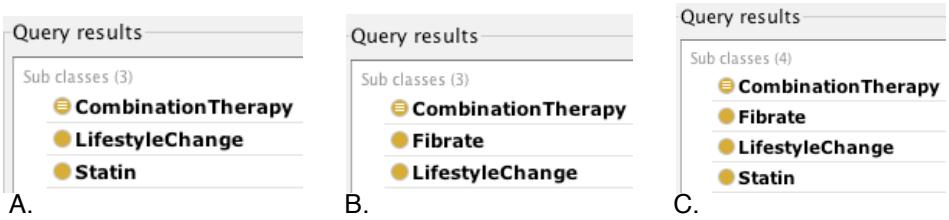


Figure 59. Query result for competency question Treatment.

According to the ontology reasoner, the ideal treatments for lowering cholesterol and triglyceride are **CombinationTherapy**, **LifestyleChange**, **Statin** and **Fibrate**.

8.5.1.6 Diagnostic Parameter

Can the ontology identify patients who are at risk of developing dyslipidaemia?

The **DiagnosticParameter** sub-ontology was evaluated against several criteria. The first step in our evaluation is to determine if the reasoner is able to classify the different subclasses of the class **Patient** based on the necessary and sufficient conditions we have defined for certain classes. Figures 60 and 61 show a comparison between the asserted and inferred subclasses of **Patient**.

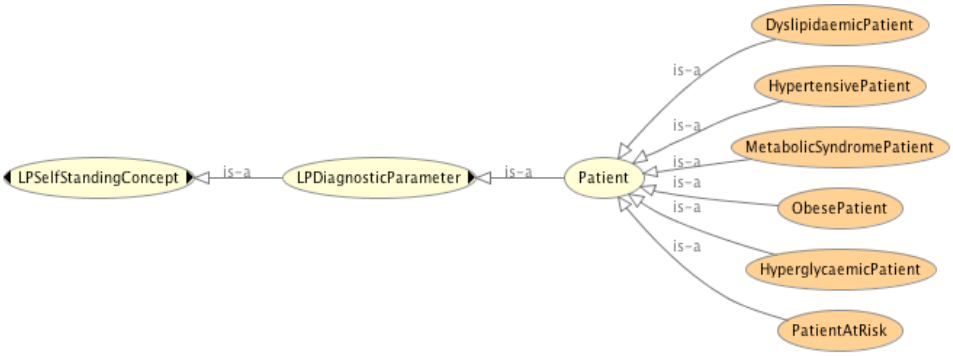


Figure 60. Asserted model of the subclasses of the concept *Patient* in Protégé.

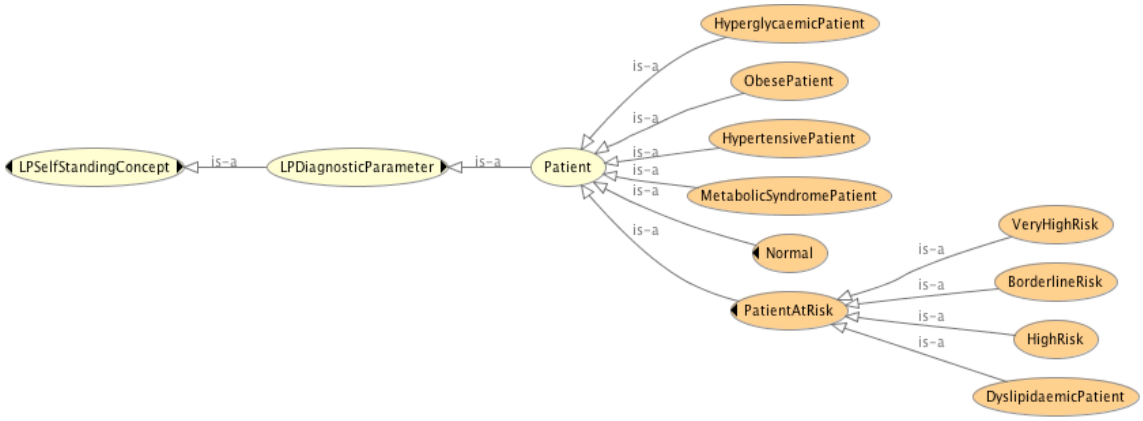


Figure 61. Inferred model of the subclasses of the concept *Patient* in Protégé.

In addition to the subclasses of **Patient** discussed in Chapter 7.6.6.4.1, we created another subclass of **Patient**, named **PatientAtRisk**, to classify the subset of patients which fall under the risk category (normal, borderline risk, high risk, very high risk) for various levels of total triglycerides, cholesterol, LDL cholesterol and HDL cholesterol. We also created the class **RiskValuePartition**, described in Chapter 7.6.6.4.2, which contains the subclasses **Normal**, **BorderlineRisk**, **HighRisk** and **VeryHighRisk**, to formalise and define the concept of risk value indicators associated with different plasma levels of total triglycerides, cholesterol, LDL cholesterol and HDL cholesterol.

Based on the necessary and sufficient conditions that we have defined for each concept, the reasoner was then able to classify the ontology according to the respective definitions. Figure B above demonstrates that the concepts **BorderlineRisk**, **HighRisk**, **VeryHighRisk**, along with **DyslipidaemicPatient** are now classified under the concept **PatientAtRisk**, whereas the concept **Normal** is classified under the concept **Patient**.

8.5.2 Evaluation of Lipoprotein Ontology Through Case Studies

Another method of evaluating the practical usefulness of Lipoprotein Ontology is through the use of case studies or scenarios. For this purpose, we have developed a case study for evaluating the application of Lipoprotein Ontology specifically in clinical diagnosis and management of dyslipidaemia.

Consider the disease "metabolic syndrome", a clinical syndrome characterised by three or more criteria such as central obesity, hyperglycaemia, dyslipidaemia and hypertension (NCEP, 2001), as shown in Table 8 in Chapter 7.6.6.4.1 (Page 185). For convenience, we included the table below.

Table 8. NCEP/ATP III definition of the metabolic syndrome (NCEP, 2001).

Criteria	Three or more criteria
Central Obesity	Waist circumference: * *Caucasian*: ≥ 102 cm (men), ≥ 88 cm (women) * *Asian*: ≥ 90 cm (men), ≥ 80 cm (women)
Hyperglycaemia	* *Fasting plasma glucose level:* ≥ 110 mg/dL
Dyslipidaemia	* *Triglyceride levels:* ≥ 150 mg/dL * *HDL-cholesterol level:* < 40 mg/dL (men), < 50 mg/dL (women)
Hypertension	* *Systolic BP:* ≥ 130 mmHg * *Diastolic BP:* ≥ 85 mmHg

As the table illustrates, the criteria for central obesity differs between genders and ethnicities. The criteria for HDL cholesterol level is also different between men and women. Given that this information has been defined in Lipoprotein Ontology, we develop the following scenario.

6 patients visit a clinic for a regular check up. Table 12 summarises each of the patient profile and the measurements taken:

- Waist circumference
- Total triglyceride concentration
- Total cholesterol concentration
- LDL cholesterol concentration
- HDL cholesterol concentration
- Fasting plasma glucose level
- Blood pressure

Table 12. Patient profile according to ethnicity, waist circumference, total triglyceride concentration, total cholesterol concentration, LDL cholesterol concentration, HDL cholesterol concentration, fasting plasma glucose level and blood pressure

Name	Ethnicity	Disorder	Waist Circumference (cm)	Total Triglyceride Concentration (mg/dL)	Total Cholesterol Concentration (mg/dL)	LDL Cholesterol Concentration (mg/dL)	HDL Cholesterol Concentration (mg/dL)	Fasting Plasma Glucose Level (mg/dL)	Blood Pressure (mmHg)
John	Asian	-	85	145	100	95	65	105	120/80
Jane	Caucasian	-	80	140	170	90	70	100	110/80
Mark	Asian	Central Obesity Dyslipidaemia Hypertension Hyperglycaemia	110	510	260	200	30	170	150/90
Mary	Caucasian	Central Obesity Dyslipidaemia Hypertension Hyperglycaemia	95	500	250	190	35	160	135/95
Anthony	Asian	Hypertension Hyperglycaemia	100	150	220	140	50	150	160/95
Amanda	Caucasian	Central Obesity Hypertension	120	160	210	135	60	100	140/90

Based on the measurements above with respect to their gender and ethnicity, they are shown to suffer from the associated conditions listed under the column "disorder". Given this scenario, we propose the questions:

- Can the ontology identify the respective disorders associated with each patient, and identify the patient(s) suffering from the metabolic syndrome?
- Can the ontology identify the risks associated with each individual?

First, we created several asserted instances of **Patient** shown in Figure 62, which serve as baseline measures for us to verify the correct classification of the inferred instances of the class **Patient**. For example, asserted instances of **MetabolicSyndrome** include MetabolicSyndromeAsianMale, MetabolicSyndromeAsianFemale, MetabolicSyndromeCaucasianMale and MetabolicSyndromeCaucasianFemale. These instances were defined according to their respective object and datatype properties, and serve as baseline measures for us to verify the correct identity of the inferred instances of **MetabolicSyndrome**. Based on the necessary and sufficient conditions that we have defined for **MetabolicSyndrome**, any instance of **Patient** that has met the conditions for **MetabolicSyndrome** will be inferred to be an instance of **MetabolicSyndrome**.

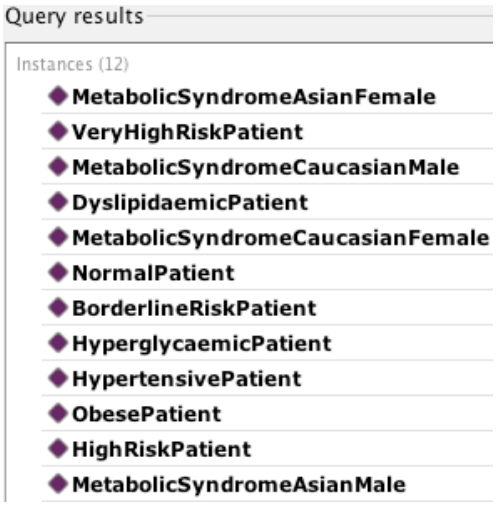
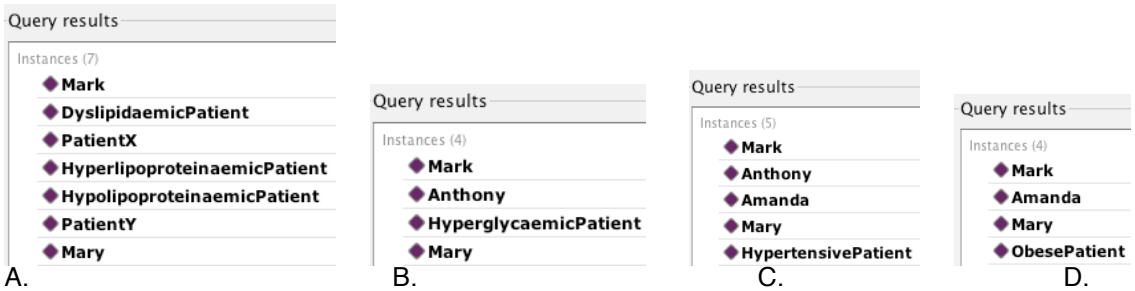


Figure 62. Query result to Patient.

Without asserting the disorders they are suffering from, we created instances of **Patient** (John, Jane, Mark, Mary, Anthony and Amanda) with various object properties such as *hasGender* and *hasEthnicity*. In addition, for every instance we entered various datatype property values such as *hasTriglycerideConcentration*, *hasHDLConcentration*, *hasWaistCircumference* and *hasBloodPressureValue* (Table 9). In the first example, shown in Figures 63 A, B, C and D below, we queried the ontology for instances of patients who suffer from a particular disorder, using different query parameters.

DL Query (A): **DyslipidaemicPatient**
DL Query (B): **HyperglycaemicPatient**
DL Query (C): **HypertensivePatient**
DL Query (D): **ObesePatient**



A. B. C. D.

Figure 63. Query result for patients suffering from different disorders.

Figures 63 A, B, C and D show the inference results to the corresponding queries, which match to the values in the "disorder" column that were not asserted for each individual. Figure 64 shows that based on the datatype property values that were entered for the instances, the ontology is capable of classifying individuals who belong to the class **MetabolicSyndromePatient**.

DL Query: **MetabolicSyndromePatient**

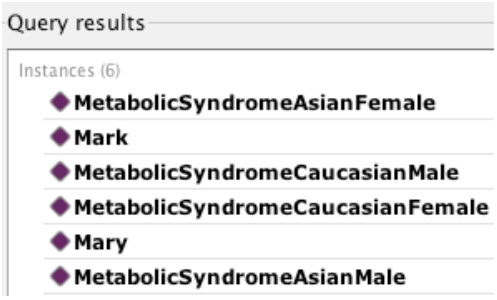


Figure 64. Query result to MetabolicSyndromePatient.

Based on the query results, we can then satisfy that Mark and Mary are both individuals suffering from the metabolic syndrome disorder.

In addition to the inference of disorders according to the corresponding properties, the reasoner was also able to classify individuals who are at risk of developing dyslipidaemia according to different values (Normal, BorderlineRisk, HighRisk and VeryHighRisk) shown in Figure 65 A, B, C and D.

DL Query (A): **Normal**
DL Query (B): **BorderlineRisk**
DL Query (C): **HighRisk**
DL Query (D): **VeryHighRisk**

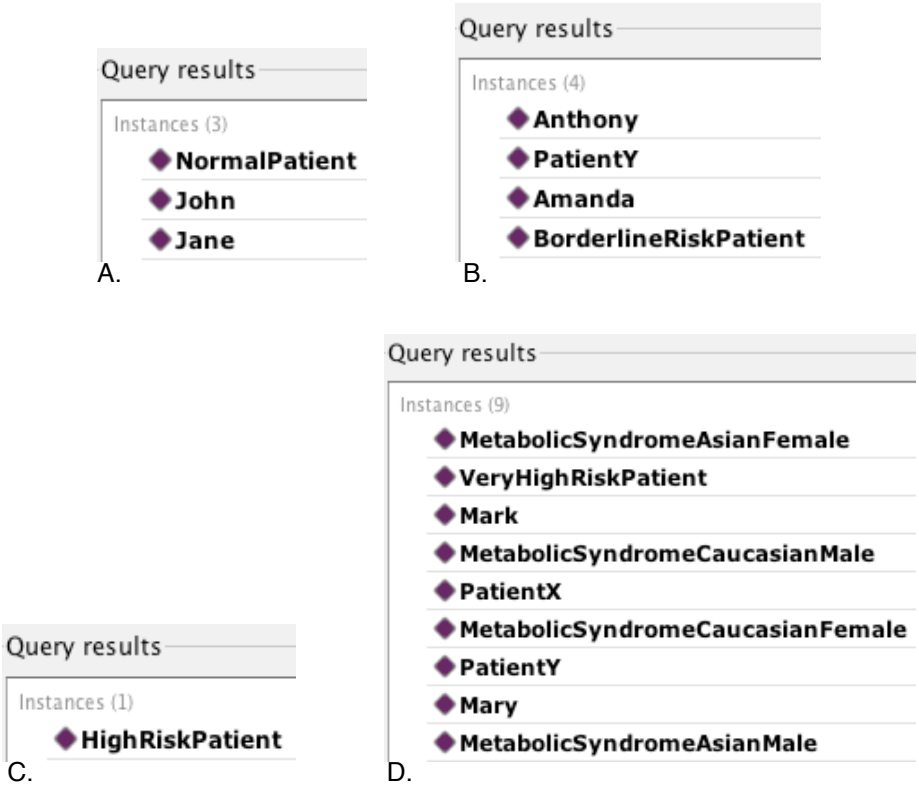


Figure 65. Query result for individual risk indicator

As the figures above illustrate, we were able to identify individuals according to the risk of developing dyslipidaemia based on their lipid profile.

8.2 Conclusion

This chapter presents the evaluation phase of Lipoprotein Ontology. Here, we evaluated our ontology using three measures. First, we evaluated the syntactic quality of our ontology using five design criteria which include: completeness, consistency, conciseness, extendibility and minimal encoding bias. Subsequently, the conceptual coverage of Lipoprotein Ontology was evaluated by mapping the concepts from 90 randomly selected abstracts from the PubMed database. Finaly, the practical usefulness of Lipoprotein Ontology was evaluated through the validation of competency questions, as well as a case study scenario which demonstrate the application of the ontology.

References

Brank, J., Grobelnik, M., & Mladenić, D. (2005). A survey of ontology evaluation techniques. Paper presented at the Proceedings of the Conference on Data Mining and Data Warehouses (SiKDD 2005).

Gangemi, A., Catenacci, C., Ciaramita, M., & Lehmann, J. (2005). A theoretical framework for ontology evaluation and validation. Paper presented at the Methodology for the design and evaluation of ontologies, Montreal, Canada.

Gruninger, M., & Fox, M. S. (1995). Methodology for the design and evaluation of ontologies. Paper presented at the Methodology for the design and evaluation of ontologies, Montreal, Canada.

Hartmann, J., Sure, Y., Giboin, A., et al. (2005). Methods for ontology evaluation Knowledge web project deliverable

NCEP, Expert Panel on Detection and Treatment of High Blood Cholesterol in Adults. (2001). Executive Summary of the Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III). *JAMA, 285*, 2486-2497.

Olson, R. E. (1998). Discovery of the lipoproteins, their role in fat transport and their significance as risk factors. The Journal of Nutrition, 128, 439S–443S.

Chapter 9. Summary and Future Work

9.1 Introduction

This chapter provides a recapitulation of the work that was carried out in this thesis, with an emphasis on how our work has addressed the research objectives that were presented in Chapter 1. This is followed by a brief discussion on the significance of our research. We also elaborate on the limitations of Lipoprotein Ontology, and identify potential areas which can be developed further as future work. Finally, this thesis concludes with a reflection on the wider biomedical challenge.

9.2 Summary of the Thesis

In Chapter 1, we provided an overview of the lipoprotein research domain, which covers the classification of lipoproteins, lipoprotein metabolism, lipoprotein disorders (dyslipidaemia), causes and treatment of dyslipidaemia. We then discussed how existing literature fail to present an interrelated model of lipoprotein concepts or demonstrate the links between them. Subsequently, we highlighted some issues within the lipoprotein research domain, which serve as the motivation of our research. These issues include the complexity of lipoprotein metabolism pathways, the impact of dyslipidaemia on human health, causes of dyslipidaemia, as well as the diagnosis and treatment of dyslipidaemia. We followed this discussion by presenting the scope of the problem and the objectives of this thesis as follows:

- Objective 1. To develop an overall methodology framework towards the development of Lipoprotein Knowledge Representation.

- Objective 2. To conceptualise lipoprotein concepts and structure the lipoprotein domain into six sub-domains as follows: describing the

classification of lipoproteins in healthy individuals (Classification), modelling the complex metabolism of lipoproteins through systematic relationships between lipoprotein concepts (Metabolism), transferring this approach to the classification of lipoproteins in patients with lipoprotein disorders (Pathophysiology), mapping the causes of dyslipidaemia (Aetiology), modelling treatment options for lipoprotein disorders (Treatment) and providing a consistent representation of diagnostic parameters for dyslipidaemia (Diagnostic Parameter).

- Objective 3. To formalise the conceptualisation of lipoprotein knowledge in OWL representation language in the following sub-ontologies: *Classification*, *Metabolism*, *Pathophysiology*, *Aetiology*, *Treatment* and *Diagnostic Parameter*.

- Objective 4. To evaluate Lipoprotein Ontology through the use of case studies, which have been developed to evaluate the robustness and consistency of the lipoprotein knowledge framework.

In Chapter 2, we presented the current state of research in the lipoprotein domain. First, we introduced various lipoprotein concepts, including the different lipoprotein classes, components and metabolism pathways. We then examined the implications of dyslipidaemia on health, by describing the key components in dyslipidaemia, clinical and metabolic features of dyslipidaemia, as well as the significance of dyslipidaemia in the metabolic syndrome. Subsequently, we elaborated on the management of dyslipidaemia in terms of the diagnostic parameters, causes and treatment of dyslipidaemia. We then described the nature of biomedical information and discussed KR techniques in the management of biomedical data. Accordingly, we introduced ontologies as the potential solution for knowledge management issues.

Chapter 3 stated the main research issue identified in this thesis as the lack of a formal conceptualisation framework for the description, organisation and classification of lipoprotein-related information. We first defined the key concepts in the lipoprotein research domain and discussed challenges unique within the domain, which served as the motivation of study. The underlying research issues were identified, including the issues of information explosion and

unstructured domain knowledge, the heterogeneity of biomedical information, and difficulty in information retrieval. Based on these issues, we derived the following key research questions:

1. How can we model lipoprotein concepts and relationships such that they can be used as a representation of the lipoprotein research domain? Specifically, how can we represent complex lipoprotein metabolism pathways in terms of concepts and relationships?
2. How can the model be used to aid knowledge integration and management in the lipoprotein domain? How can we use this framework to infer knowledge to aid diagnosis of lipoprotein dysregulation?
3. How can such a model faciliate collaboration between users? For instance, how can we develop this platform for easy transferrance of research output from different research groups?
4. How can we validate our framework such that it is consistent and easily extensible?

With these key research questions in mind, we discussed our choice of research approach towards the development of Lipoprotein Ontology, a formal framework for the description, organisation and classification of lipoprotein-related information.

In Chapter 4, we provided an overview of the solution to the problem definition in Chapter 3. We introduced some ontology definitions and characteristics, and discussed the role of ontologies in addressing the corresponding research issues, by providing examples of current applications of ontologies in the biomedical domain. This is followed by a critical review of other biomedical ontologies, which do not fulfil the requirements for a specific lipoprotein ontology. We also reviewed existing methodologies on ontology development which serve as the justification towards our choice of methodology. Finally, this chapter concluded with an overview of Lipoprotein Ontology and defined the key concepts and relations in the ontology.

In Chapter 5, we described the methodology that was used to develop Lipoprotein Ontology. This methodology was based on the Knowledge Engineering Methodology due to its focus on conceptual design. However, we

have also incorporated various aspects of other methodologies to suit the purpose of Lipoprotein Ontology, which were discussed accordingly. Our methodology incorporated different stages of other existing methodologies, and included the following processes: specification, conceptualisation, formalisation and evaluation. This chapter addressed Objective 1 presented in Chapter 1.

Chapter 6 elaborated on the conceptual framework for the representation of lipoprotein domain knowledge, and elaborated on some of the upper level concepts which are represented in Lipoprotein Ontology. Essentially, this chapter served as the conceptual backbone for our ontology. Here, we organised lipoprotein-related concepts in a structured hierarchy using the ontology editor Protege 4.2, according to ontology principles that we have previously defined in an earlier section. We described the conceptualisation process of each of the six sub-ontologies of the lipoprotein domain knowledge, namely: *Classification*, *Metabolism*. *Pathophysiology*, *Aetiology*, *Treatment* and *Diagnostic Parameter*. Collectively, these sub-ontologies make up the lipoprotein conceptual framework, which is in alignment with Objective 2, namely: the classification of lipoproteins in healthy individuals (Classification), modelling the complex metabolism of lipoproteins through systematic relationships between lipoprotein concepts (Metabolism), transferring this approach to the classification of lipoproteins in patients with lipoprotein disorders (Pathophysiology), mapping the causes of dyslipidaemia (Aetiology), modelling treatment options for lipoprotein disorders (Treatment) and providing a consistent representation of diagnostic parameters for dyslipidaemia (DiagnosticParameter).

In Chapter 7, we discussed the formalisation process of the lipoprotein conceptual framework described above. We have opted to use Protege 4.2 as our ontology development tool, OWL as the ontology language for the implementation of Lipoprotein Ontology and HermiT 1.3.8 as the reasoner. First, we provided an overview of knowledge representation languages associated with ontology building, and the justification behind our choice of implementation tool and language. This is followed by a brief introduction to the components of OWL ontologies and how they were utilised in Lipoprotein Ontology. We then discussed the structure of OWL ontologies, as well as the notation and syntax used throughout this chapter. This chapter concluded with a visualisation of Lipoprotein Ontology presented as screenshots from Protege to represent

various functions of concept classification, definition and description, using various OWL constructs. Therefore, Chapter 7 adequately addressed Objective 3. At the time of writing, Lipoprotein Ontology contained 469 classes, 66 object properties, 25 data properties and 237 instances. Figure 66 shows the complete ontology metrics including axiom and logical axiom counts.

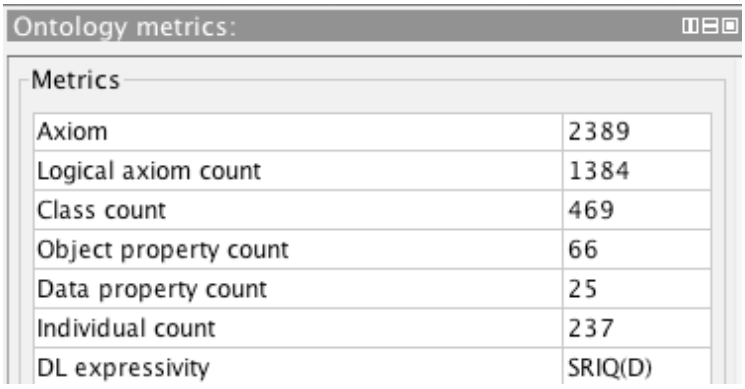


Figure 65. Ontrology metrics for Lipoprotein Ontology in Protégé.

Chapter 8 served as the evaluation phase of Lipoprotein Ontology. We evaluated our ontology using three methods. First, the syntactic quality of the ontology was evaluated using five design criteria: completeness, consistency, conciseness, extendibility and minimal encoding bias. Next, we mapped the concepts from 90 randomly selected article abstracts from the PubMed database to Lipoprotein Ontology in order to evaluate its conceptual coverage. Our evaluation results indicate that there was an excellent match between the concepts from the selected articles and the Lipoprotein Ontology concepts, the best match being 100% and the lowest being 61.2%. Finally, we evaluated the practical application of Lipoprotein Ontology through the validation of the competency questions that were developed during our methodology phase, as well as a case study scenario. The results demonstrate that the ontology contained the necessary concepts and relations in addressing the competency questions, thus meeting Objective 4.

9.3 Research Significance

The development of Lipoprotein Ontology has significance for both lipoprotein and ontology domains.

- At the time of writing this thesis, Lipoprotein Ontology is the first ontology framework designed specifically for the lipoprotein research domain. Lipoprotein Ontology provides a formal framework for lipoprotein concepts and relationships that can be used to integrate, link and represent the disorganised and dispersed knowledge of lipoproteins in a systematic and meaningful way.

- The development of a formal, explicit and common vocabulary framework for the lipoprotein domain enables interoperability between research groups, thereby allowing for more efficient retrieval and analysis of information, saving time and resources.

- Lipoprotein Ontology has the potential to provide an overview of various research subjects and empirical findings such that different concepts and empirical results can be categorised accordingly and viewed in relation to one another. Therefore, through the unified structure of Lipoprotein Ontology, some previously undiscovered relationships among different theories and concepts can be revealed, which can serve as a motivation for researchers to carry out studies to address gaps in the literature.

- Lipoprotein Ontology can serve as a basis for the design of ontology-driven tools and applications for various information retrieval, analysis and data mining purposes. For example, Lipoprotein Ontology can be applied to Semantic Web search engines through which information can be retrieved, managed and analysed in an intelligent way. This initiative is particularly in alignment with recent worldwide attempts to develop and replace the current form of web with Semantic Web.

- Lipoprotein Ontology can be used as the basis for the development of applications for the diagnosis and treatment of lipoprotein disorders. By incorporating specific aspects of lipoprotein research in Lipoprotein Ontology, not only in terms of the classification of lipoproteins, but also understanding the metabolic pathway, pathophysiology, causes of lipoprotein dysregulation and treatment options for lipoprotein disorders this impacts not only on identifying the risks, but also provides effective preventative measures.

9.4 Limitations and Future Work

Lipoprotein Ontology is an ambitious endeavour and much of its potential has been discussed in this thesis, however there are a number of limitations. Some of these issues are discussed below and pave the way for future work.

9.4.1 Limitations in Conceptual Representation

As discussed previously in this thesis, the challenge of maintaining a comprehensive repository of lipoprotein knowledge is beyond the capability of a single person, and the success of an ontology ultimately depends on the end users. Although we consider Lipoprotein Ontology to offer sufficient formal representation of lipoprotein concepts for the purpose of this thesis, ontology development is a subjective process. Therefore, characteristic of all ontologies, the work is not presented as a complete task. There is a need for collaboration with domain experts for further development and maintenance of the ontology for further use. In view of this limitation, we have designed our ontology to be easily extensible to accommodate new concepts without compromising the original framework with respect to future work. This includes the possibility of exporting the ontology to other forms, such as UML.

9.4.2 Lack of Collaboration

The metabolism of lipoproteins involves many different enzymes, receptors and other components interacting in various pathways, all occurring in tandem, and with one reaction frequently affecting the rate of a previous or subsequent reaction. However, delving further into these processes and components would require additional knowledge from domain experts in their respective fields. Deciding the granularity of the ontology was a challenge we faced initially, but the evaluation process has shown that the ontology contains all the necessary lipoprotein concepts to a satisfactory extent. Some of the advantages of an ontology include having an open world assumption as well as extensibility. Hence, if necessary, the ontology can be extended to include other finer entities (concepts) involved in lipoprotein metabolism. Having a fine-grained view of the overall sequence of events and the reaction rates of the various pathways can potentially extend the inferential power of the ontology in predicting the effect of

disease states or an agent such as drug on fat absorption. For this to occur, there needs to be collaboration between lipoprotein specialists in the field of physiology, biochemistry, molecular biology and pharmacology.

9.4.3 Parameters for Lipoprotein Kinetics

Lipoprotein kinetics involves breaking down lipoprotein metabolism pathways into separate compartments and applying mathematical formulas to determine the rates of reaction which occur in different compartments (Chan et al., 2004). Having lipoprotein kinetics mathematically annotated to the ontology will greatly improve the reasoning capabilities of the ontology; the end-user will be able to harness the computational power of the ontology in a relatively straightforward manner once the concepts and rules have been built in. Due to constraints on the scope of Lipoprotein Ontology, we have decided to exclude lipoprotein kinetics from the ontology, but this can certainly be included in our future work.

9.4.4 The Notion of Time

Despite the processing power of OWL, the language does not provide any support for the representation of temporal information (O'Connor & Das, 2011). The notion of time is particularly relevant in biomedical data and clinical applications of ontologies as the ability to infer patterns over time has tremendous potential. Some of these promising applications which come to mind include analysing longitudinal studies and mining for patterns and trends, extrapolating risk factors given the patient's history, current lifestyle habits and diagnostic parameters, as well as monitoring the success of a treatment course. Several systems have been developed which address the temporal dimension such as the Time Oriented Database (Wiederhold, 1981) and the Arden Syntax (Hripcsak et al., 1994), which allow the extraction of information according to certain temporal patterns by associating an instant timestamp with particular records. A lightweight yet expressive temporal model has also been developed to facilitate the consistent representation as well as querying of temporal information in OWL ontologies (O'Connor & Das, 2011). As the model was designed to be integrated with existing ontologies, applying this model to Lipoprotein Ontology could have potential benefits in the inference of patterns with respect to time, and will be considered in future versions of this work.

9.4.5 Sociological Limitations

As with any endeavour, the challenge to the adoption of Lipoprotein Ontology is largely sociological. Although the biomedical community has been moving steadily forward towards an age of bioinformatics and biomedical ontologies, the success of Lipoprotein Ontology is dependent on its acceptance by the community. As much as possible, we have engineered the development of Lipoprotein Ontology as a response towards issues within the lipoprotein domain; however societal limitations still exist to a certain extent.

9.5 Conclusion

We conclude this thesis by presenting a summary of our work, with an emphasis on how we have addressed our objectives. We have developed an ontological framework for the formal representation of lipoprotein concepts in terms of their *Classification*, *Metabolism*, *Pathophysiology*, *Aetiology* and *Treatment*. By providing formal specifications of lipoprotein concepts in a system of hierarchical and associative relations, Lipoprotein Ontology serves as the semantic framework for the computable modelling of lipoprotein domain knowledge. The integration of knowledge is particularly important in translating research outputs from experimental studies into clinical care. Lipoprotein Ontology is designed to bridge the gap between lipoprotein research and clinical practice by providing a controlled terminology which enables researchers and practitioners to integrate and relate heterogeneous information in lipoprotein research, as well as link research findings to disease patterns. Associating research data with ontology terms also enables effective retrieval and intelligent querying of information. Furthermore, successful collaboration between research groups or software agents would benefit from a common platform of shared and agreed knowledge. On a more ambitious outlook, Lipoprotein Ontology can serve as the basis for the design of intelligent applications such as semantic search engines and tools for the diagnosis and treatment of dyslipidaemia. It is not anticipated that such applications could perform better than human experts; however they could play an important role in filtering the flood of data to the point where human experts could apply their knowledge sensibly.

References

Chan, D. C., Barrett, P. H. R., & Watts, G. F. (2004). Lipoprotein kinetics in the metabolic syndrome: Pathophysiological and therapeutic lessons from stable isotope studies. *Clinical Biochemistry Review, 25*, 31-48.

Hripcsak, G., Ludemann, P., Allan Pryor, T., et al. (1994). Rationale for the Arden Syntax. *Computers and Biomedical Research, 27*, 291-324.

O'Connor, M. J., & Das, A. K. (2011). *A method for representing and querying temporal information in OWL.* Paper presented at the Biomedical Engineering Systems and Technologies.

Wiederhold, G. (1981). Databases for healthcare *Lecture Notes in Medical Informatics*. Heidelberg, Germany: Springer.

Appendix A. Major Lipoprotein Classes

A comparison of density, size, electrophoretic mobility, percentage of total mass that is triglyceride (TG), cholesterol (C), cholesteryl ester (CE), phospholipid (Ph), protein (P) and apolipoproteins (apo) content

Lipoprotein	Density (g/ml)	Size (nm)	Electrophoretic mobility	TG (%)	C (%)	CE (%)	Ph (%)	P (%)	Major apo	Other apo	Other constituents
Chylomicrons	<0.930	75-1200	Origin	80-95	1-3	2-4	3-6	1-2	B-48	A-I, A-II, A-IV, A-V, C-I, C-II, C-III, E	Vitamin A
Chylomicron remnants	0.930-1.006	30-80	Slow pre-β						B-48	E	Vitamin A
VLDL	0.930-1.006	30-80	Pre-β	45-65	4-8	16-22	15-20	6-10	B-100	A-I, A-II, A-IV, A-V, C-I, C-II, C-III, D, E	Vitamin A
IDL	1.006-1.019	25-35	Slow pre-β						B-100	E, C-I, C-II, C-III	Vitamin E
LDL	1.019-1.063	18-25	β	4-8	6-8	45-50	18-24	18-22	B-100	E, C-I, C-III, A-I	Vitamin E
HDL	1.063-1.210	5-12	α	2-7	3-5	15-20	26-32	45-55	A-I	A-II, A-IV, A-V, C-I, C-II, C-III, D, E	LCAT, CETP
Lp(a)	1.050-1.120	25	Pre-β						B-100	(a)	

Abbreviations: VLDL, very low density lipoprotein; IDL, intermediate density lipoprotein; LDL, low density lipoprotein; HDL, high density lipoprotein; Lp(a), lipoprotein (a)

Appendix B. Summary of Lipoprotein Metabolism Processes

Summary of lipoprotein metabolism processes, location and major components

Lipoprotein	Location	Process	Major components
Chylomicron	Intestine	Assembly, secretion	Apo A, apo B, phospholipid, cholesterol, cholesteryl ester, triglyceride
	Lymph	Transfer	Apo C, apo E
	Plasma	Transfer	Apo C, cholesterol
	Endothelial cell	Hydrolysis	Triglyceride, phospholipid
		Transfer	Fatty acid, cholesterol, phospholipid, apo C, apo A, apo E
VLDL	Liver	Assembly, secretion	Apo B, apo C, phospholipid, cholesterol, cholesteryl ester, triglyceride
	Plasma	Transfer	Apo C, apo E, cholesteryl ester
	Endothelial cell	Hydrolysis	Triglyceride, phospholipid
		Transfer	Fatty acid, cholesterol, phospholipid, apo C, apo E
LDL	Plasma	Formation	From VLDL and chylomicrons
		Exchange	Cholesterol, phospholipid
	Peripheral tissues, liver	Uptake, catabolism, regulation	Cholesterol, cholesteryl ester
HDL	Liver	Assembly, secretion	Apo A, apo E, phospholipid, cholesterol
	Plasma	Formation	From chylomicrons
		Acyltransfer	Phosphatidylcholine, cholesterol, cholesteryl ester, apo C, apo E
		Transfer	Cholesteryl ester, apo C, apo E
		Exchange	Apo C, cholesterol, phospholipid
	Liver, peripheral tissues	Uptake, catabolism	Cholesterol, cholesteryl ester

Abbreviations: VLDL, very low density lipoprotein; IDL, intermediate density lipoprotein; LDL, low density lipoprotein; HDL, high density lipoprotein; apo, apolipoprotein

Appendix C. Major Human Apolipoproteins

Major human apolipoproteins[a], lipoprotein classes, molecular weight, concentration in plasma, synthesis site and function

Apolipoprotein[a]	Lipoprotein class[b]	Molecular weight (Da)	Concentration in plasma (mg/dl)	Synthesis site and function
Apo A-I	**HDL**, *chylomicron*	28,100	130	Secreted by the intestine, binds ABCA1 on macrophages, critical antioxidant protein of HDL, activates LCAT
Apo A-II	**HDL,** *chylomicron*	17,400	40	Synthesised by the liver, inhibits HL
Apo A-IV	**Chylomicron,** *VLDL, HDL*	44,500	15	Synthesised exclusively by the small intestine and the hypothalamus, inhibits food intake, activates LCAT*
Apo B-48	**Chylomicron**	512,000	250	Exclusively found in chylomicrons, secreted by the intestine and liver, derived from apoB-100 gene by RNA truncation
Apo B-100	**VLDL, IDL, LDL**	242,000		Major protein of LDL, secreted exclusively by the liver, binds to LDLr
Apo C-I	**VLDL**, **Chylomicron**, *HDL*	6,600	3	Synthesised mainly by the liver and the brain*, inhibits receptor-mediated uptake of TRL, inhibits CEPT, inhibits LPL-mediated hydrolysis of TG, activates LCAT*
Apo C-II	**VLDL**, **Chylomicron**, *HDL*	8,800	12	Synthesised mainly by the liver, activates LPL
Apo C-III	**VLDL**, **Chylomicron**, *HDL*	9,000	12	Synthesised mainly by the liver and the intestine*, normally binds to HDL, but binds to TRL in hypertriglyceridaemic patients, inhibits LPL, inhibits HL, inhibits apoB and apoE mediated binding of lipoproteins to LDLr, inhibits TRL binding to LSR, inhibits LCAT, stimulates CETP
Apo D	**HDL**	22,000	12	Expressed widely in human tissue, closely associated with LCAT
Apo E	**VLDL**, **HDL,** **Chylomicron**, *IDL*	34,200	7	Synthesised mainly by the liver and the brain, binds to LDL receptor, activates LCAT*
Apo (a)	**Lp(a)**	300-000-800,000	0.1-40	Disulfide bonded to apoB-100, forms a complex with LDL identified as Lp(a); strongly resembles plasminogen; may deliver cholesterol to sites of vascular injury, high risk association with premature CVD

Abbreviations: apo, apolipoprotein, VLDL, very low density lipoprotein; IDL, intermediate density lipoprotein; LDL, low density lipoprotein; HDL, high density lipoprotein; Lp(a), lipoprotein (a); ABCA1, ATP-binding cassete transporter; TG, triglycerides; TRL, triglyceride-rich lipoprotein; LPL, lipoprotein lipase; HL, hepatic lipase; LDLr, LDL receptor; LCAT, lecithin-cholesterol acyltransferase; CETP, cholesteryl ester transfer protein; LSR, lipolysis-stimulated lipoprotein receptor ; * To a lesser extent

Appendix D. Clinical Definitions of The Metabolic Syndrome

Clinical definitions of metabolic syndrome from WHO, World Health Organization; NCEP/ATP III, National Cholesterol Education Program Adult Treatment Panel III; IDF, International Diabetes Foundation

	WHO 1998	NCEP/ATP III 2001; updated 2004	IDF 2005
Criteria	Hyperglycemia/insulin resistance plus two or more of four criteria	Three or more of five criteria	Central obesity plus two or more of four criteria
Central Obesity	* *Waist/hip ratio:* > 0.9 (men) > 0.85 (women) * *And/or body mass index:* > 30 kg/m2	Waist circumference: * *Caucasian*: ≥ 102 cm (men) ≥ 88 cm (women) * *Asian*: ≥ 90 cm (men) ≥ 80 cm (women) * Lower cut-offs (≥ 94 cm [men], ≥ 80 cm [women]) for some non-Asian adults with strong genetic predisposition to insulin resistance	Waist circumference: * *Europid, African, Mediterranean and Middle Eastern (Arab)*: ≥ 94 cm (men) ≥ 80 cm (women) * *South Asian, Chinese, South/Central American*: ≥ 90 cm (men) ≥ 80 cm (women) * *Japanese*: ≥ 85 cm (men) ≥ 90 cm (women)
Hyperglycemia	Insulin resistance, diabetes, impaired fasting glucose, impaired glucose tolerance	* *Fasting plasma glucose level:* ≥ 5.6 mmol/L * *Or current drug treatment for elevated glucose level*	* *Fasting plasma glucose level:* ≥ 5.6 mmol/L * *Or previous diagnosis of type 2 diabetes*
Dyslipidemia	* *Triglyceride levels:* ≥ 1.7 mmol/L **And/or HDL-cholesterol level:* < 0.9 mmol/L (men), < 1.0 mmol/L (women)	* *Triglyceride levels:* ≥ 1.7 mmol/L * *HDL-cholesterol level:* < 1.0 mmol/L (men), < 1.3 mmol/L (women) * *Or current drug treatment for hypertriglyceridemia/low HDL-cholesterol level*	* *Triglyceride levels:* ≥ 1.7 mmol/L * *HDL-cholesterol level:* < 0.9 mmol/L (men), < 1.1 mmol/L (women) * *Or current drug treatment for hypertriglyceridemia/low HDL-cholesterol level*
Hypertension	* *Blood pressure (BP):* ≥ 140/90 mmHg	* *Systolic BP:* ≥ 130 mmHg * *Diastolic BP:* ≥ 85 mmHg * *Or current drug therapy for known hypertension*	* *Systolic BP:* ≥ 130 mmHg * *Diastolic BP:* ≥ 85 mmHg * *Or current drug therapy for known hypertension*
Others	Microalbuminuria: * *Urinary albumin excretion rate > 20 µg/min* * *Or urinary albumin/creatinine ratio > 3.5 mg/mmol*		

Appendix E. Classification of Cardiovascular Risk Factors

Classification of risk factors of various lipids according to LDL-cholesterol, total plasma cholesterol, total plasma triglycerides, HDL cholesterol and plasma apolipoprotein B levels

	NCEP/ATP III 2001, updated 2004	AACE 2012		AHA 2001
LDL-Cholesterol (mg/dL)				
< 100	Optimal	Optimal		Optimal
100 - 129	Near optimal	Near optimal		Near optimal
130 - 159	Borderline high	Borderline high		Borderline high
160 - 189	High	High		High
≥ 190	Very high	Very high		Very high
Total Cholesterol (mg/dL)				
< 200	Optimal	Optimal		Optimal
200 - 239	Borderline high	Borderline high		Borderline high
≥ 240	High	High		High
Total Triglyceride (mg/dL)				
< 100				Optimal
< 150	Normal	Normal		Normal
150 - 199	Borderline high	Borderline high		Borderline high
200 - 499	High	High		High
≥ 500	Very high	Very high		Very high
HDL-Cholesterol (mg/dL)		**Men**	**Women**	
< 40	High	High	High	High
> 60	Optimal	Optimal	Optimal	Optimal
Apo B (ml/dL)				
< 80		Optimal (for patients with CVD or those with diabetes plus ≥ 1 CVD risk factor)		
< 90		Optimal (for patients at risk for CVD, including those with diabetes)		

Abbreviations: NCEP/ATP III, National Cholesterol Education Program Adult Treatment Panel III; AACE, American Association of Clinical Endocrinologists; AHA, American Heart Association

Appendix F. Fredrickson/WHO Classification of Primary Hyperlipidaemia

Fredrickson/WHO classification of primary hyperlipidaemia according to type with respect to its synonym, type of defect, electrophoresis patter, lipoprotein abnormality, concentrations of plasma cholesterol and triglycerides, clinical features as well as recommended treatment

Type	Synonym	Defect	Electrophoresis Pattern	Lipoprotein Abnormality	Plasma Cholesterol concentration	Plasma Triglyceride concentration	Clinical Features	Treatment
I	Familial hyperchylomicronaemia	↓ LDL Defective ApoC-II	↑ Chylomicron zone	↑ Chylomicron	Normal to ↑	↑↑↑↑	Pancreatitis, lipemia, skin eruption, xanthoma, hepatosplenomegaly	Diet
IIa	Familial hypercholesterolaemia	↓ LDL receptor	↑ Beta lipoprotein zone	↑ LDL	↑↑	Normal	Xanthelasma, arcus senilis, tendon xanthoma	Cholestyramine, cholestipol, statin, niacin
IIb	Familial combined hypercholesterolaemia	↓ LDL receptor ↑ apoB	↑ Pre-beta and beta-lipoprotein zone	↑ LDL ↑ VLDL	↑↑	↑↑		Statin, niacin, fibrate
III	Familial dysbetalipoproteinaemia	Defective ApoE2 synthesis	Beta band is broad in beta band zone	↑ IDL	↑↑	↑↑↑	Tubo-eruptive xanthoma, palmar xanthoma	Statin, fibrate
IV	Familial hyperlipaemia	↑ VLDL production ↓ catabolism	↑ pre-beta lipoprotein zone	↑ VLDL	Normal to ↑	↑↑		Statin, niacin, fibrate
V	Endogenous hypertriglyceridaemia	↑ VLDL production ↓ LPL	↑ pre-beta lipoprotein and chylomicron zone	↑ VLDL ↑ Chylomicron	↑ to ↑↑	↑↑↑↑		Niacin, fibrate

Abbreviations: VLDL, very low density lipoprotein; IDL, intermediate density lipoprotein; LDL, low density lipoprotein

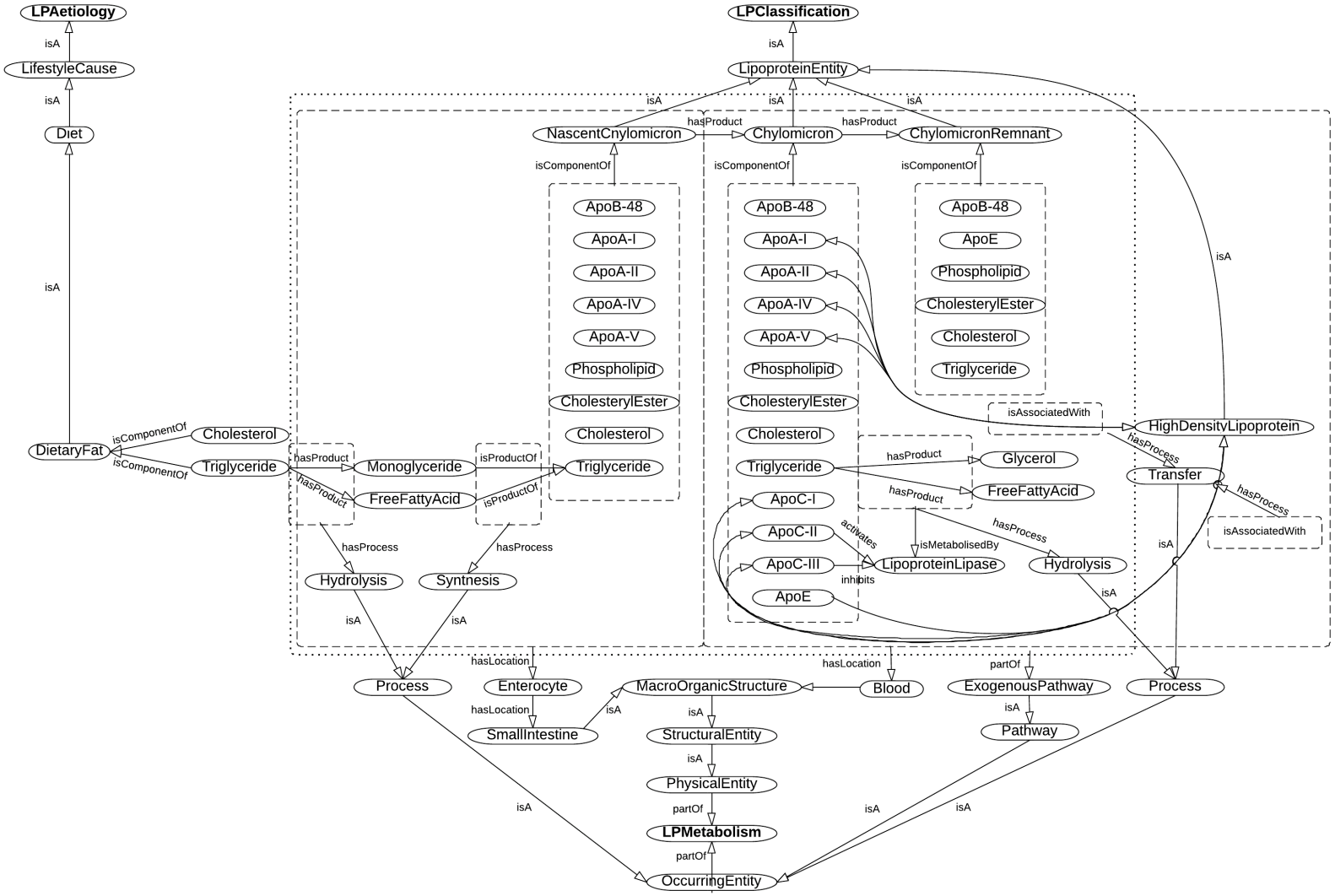
Appendix G. Pharmacotherapy Effects on Lipoprotein Metabolism

Drug classes and their effects on lipoprotein metabolism

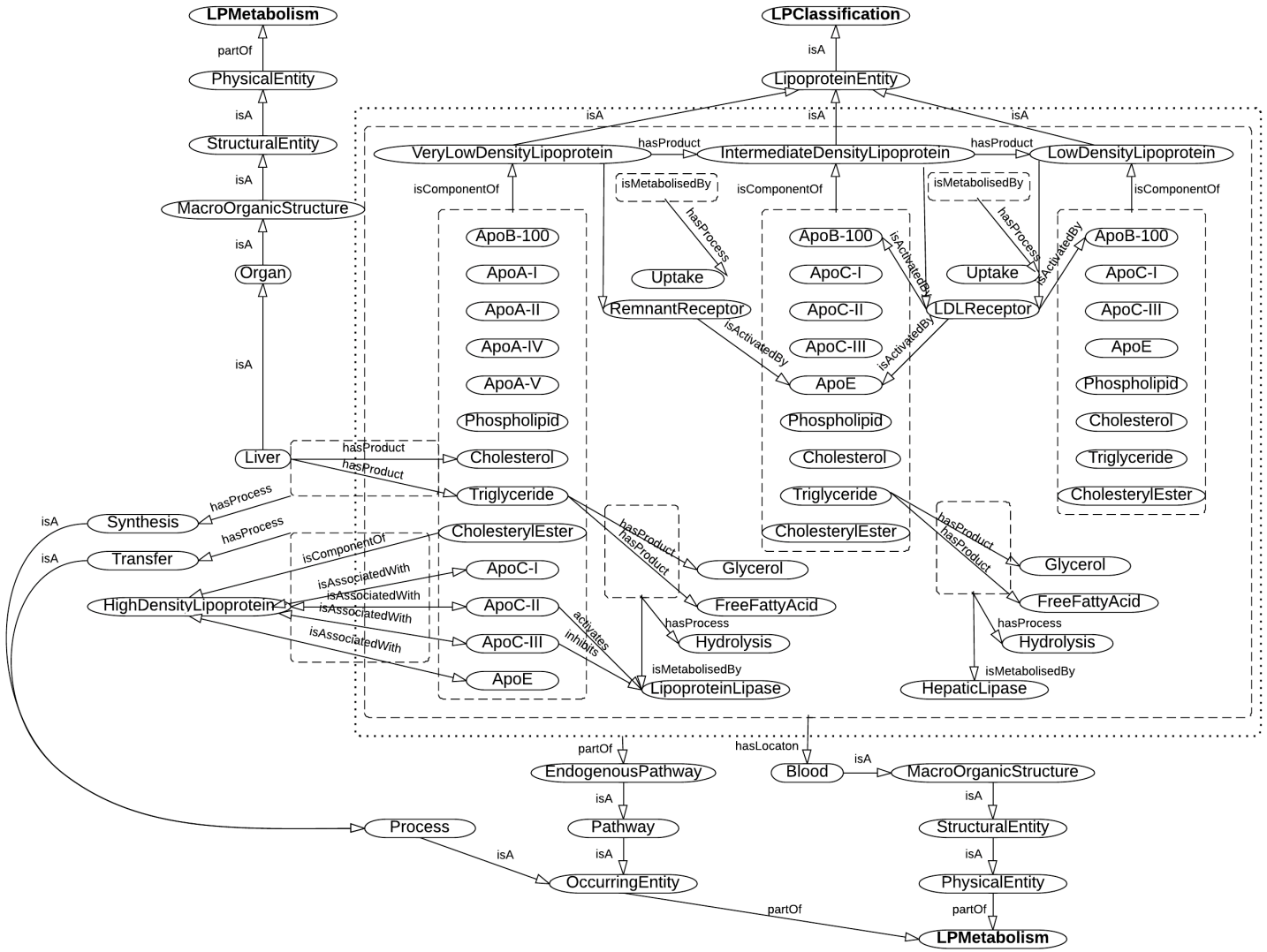
Drug Class	Agents and Daily Doses	Lipid/Lipoprotein Effects		Side Effects	Contraindications
Statins	Lovastatin (20-80 mg) Pravastatin (20-40 mg) Simvastatin (20-80 mg) Fluvastatin (20-80 mg) Atorvastatin (10-80 mg) Cerivastatin (0.4-0.8 mg)	LDL HDL TG	↓ 18-55% ↑ 5-15% ↓ 7-30%	Myopathy Increased liver enzymes	*Absolute:* • Active or chronic liver disease *Relative:* • Concomitant use of certain drugs*
Fibrates	Gemfibrozil (600 mg) Fenofibrate (200 mg) Clofibrate (1000 mg)	LDL HDL TG	↓ 5-20% ↑ 10-20% ↓ 20-50%	Dyspepsia Gallstones Myopathy	*Absolute:* • Severe renal disease • Severe hepatic disease
Bile acid sequestrants	Cholestyramine (4-16 g) Colestipol (5-20 g) Colesevelam (2.6-3.8 g)	LDL HDL TG	↓ 15-30% ↑ 3-5% No change	Gastrointestinal distress Constipation Decreased absorption of other drugs	*Absolute:* • Dysbetalipoproteinemia • TG >400 mg/dL *Relative:* • TG >200 mg/dL
Nicotinic acid	Immediate release (crystalline) (1.5-3 gm) Extended release (Niaspan®) (1-2 g) Sustained release (1-2 g)	LDL HDL TG	↓ 5-25% ↑ 15-35% ↓ 20-50%	Flushing Hyperglycaemia Hyperuricaemia (gout) Upper GI distress Hepatotoxicity	*Absolute:* • Chronic liver disease • Severe gout *Relative:* • Diabetes • Hyperuricemia • Peptic ulcer disease

Abbreviations: LDL, low density lipoprotein; HDL, high density lipoprotein; TG, triglycerides

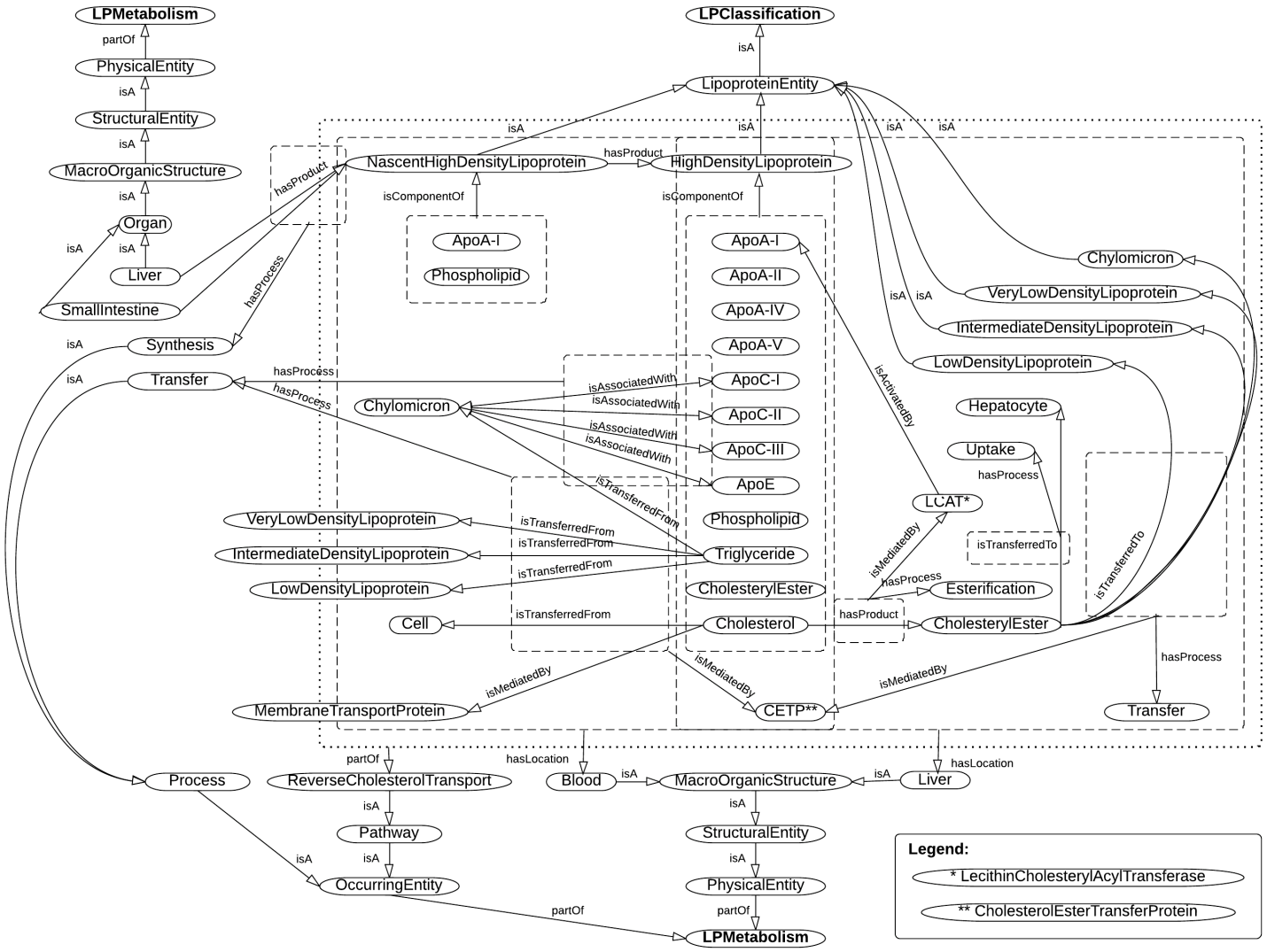
Appendix I. Formalisation of LPMetabolism – Exogenous Pathway



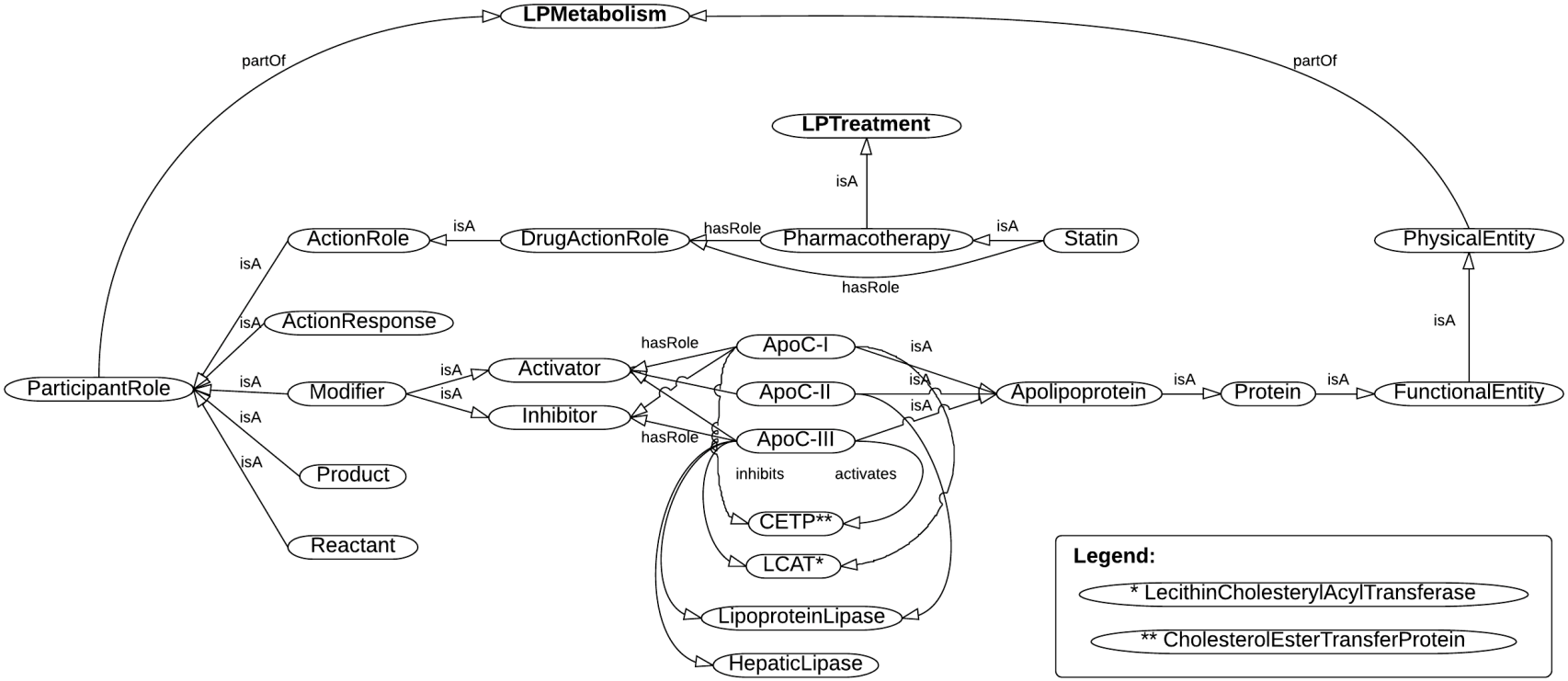
Appendix J. Formalisation of LPMetabolism – Endogenous Pathway



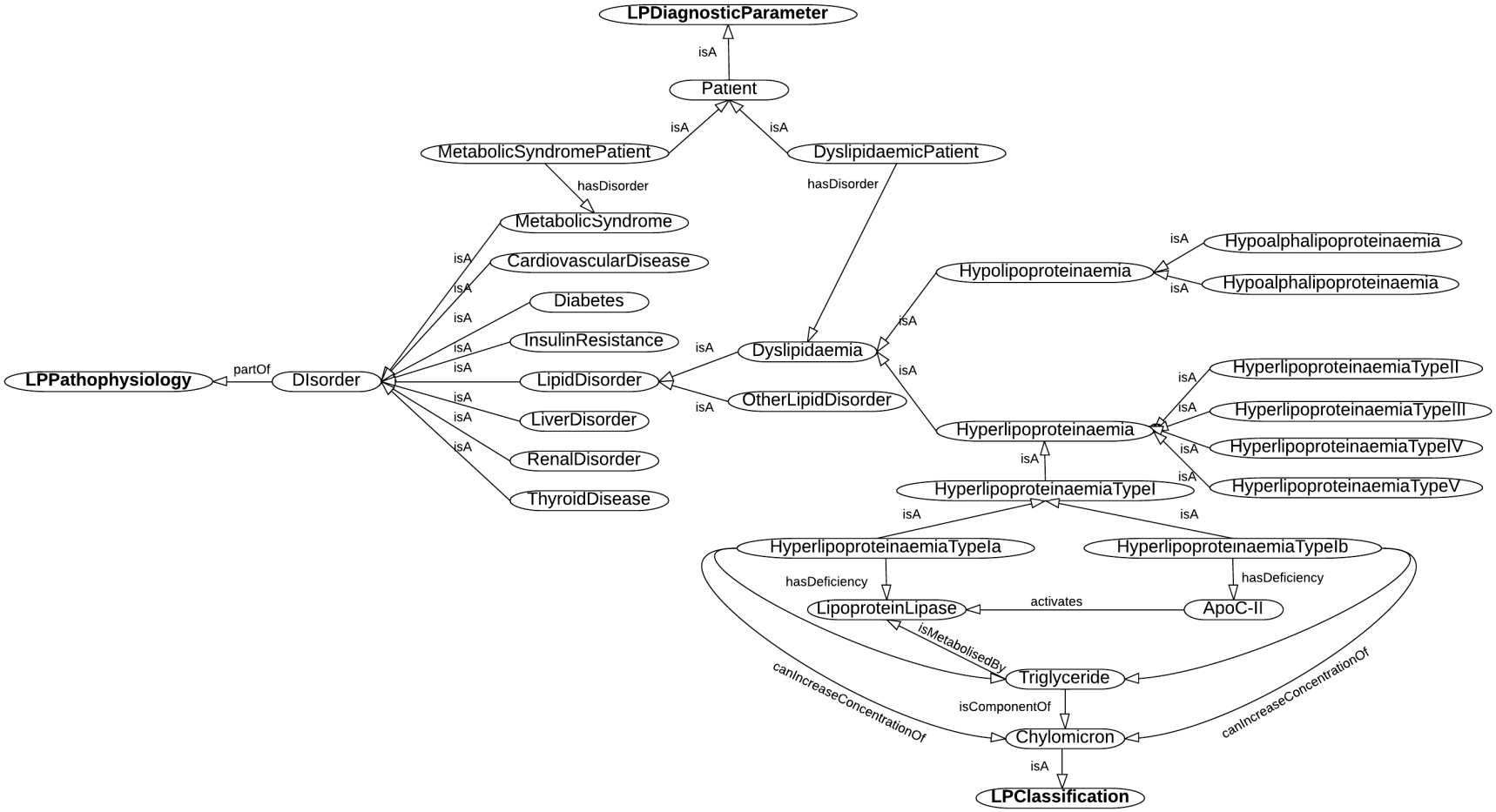
Appendix K. Formalisation of LPMetabolism – Reverse Cholesterol Transport



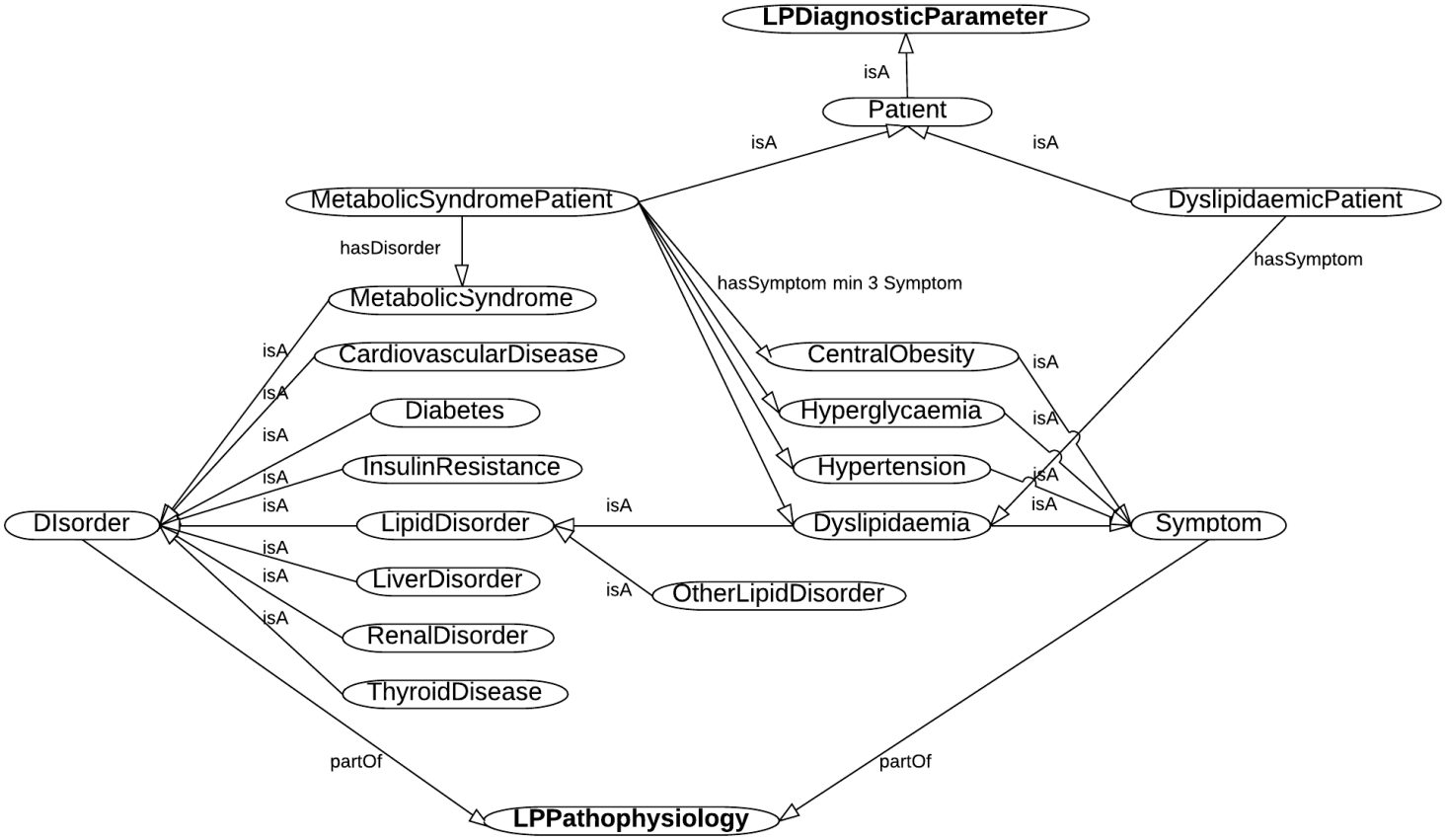
Appendix L. Formalisation of LPMetabolism – Participant Role



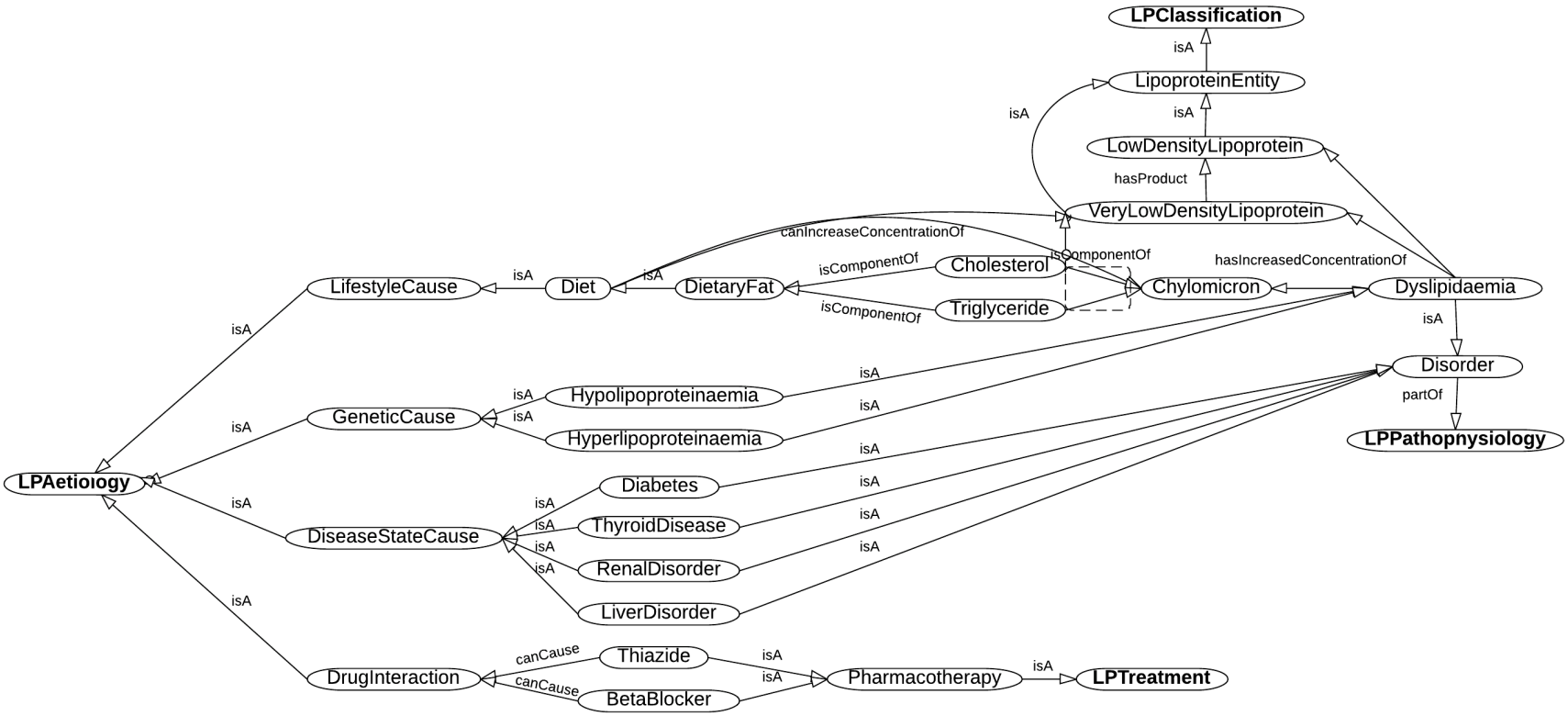
Appendix M. Formalisation of LPPathophysiology - Disorder



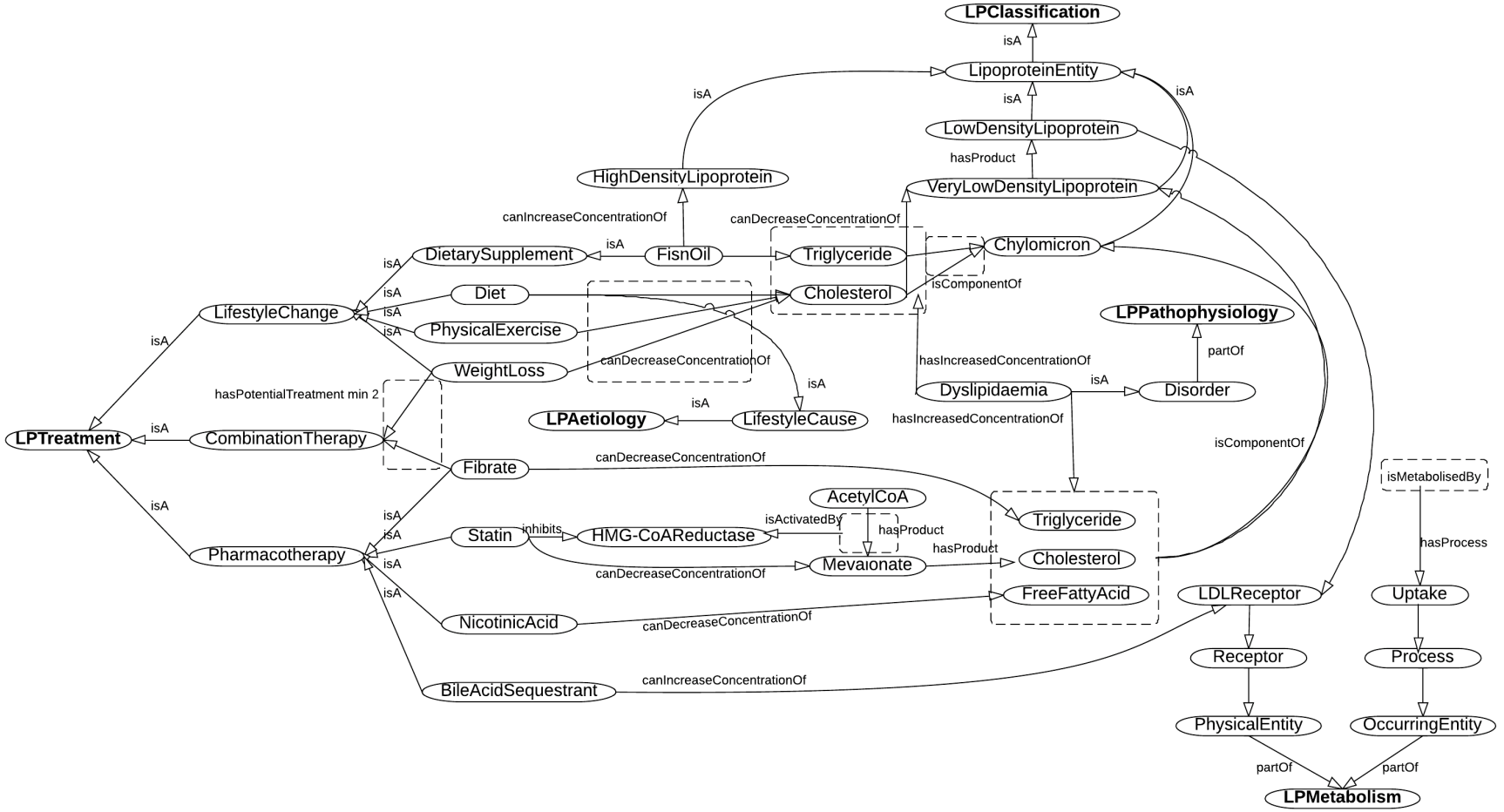
Appendix N. Formalisation of LPPathophysiology – Symptom



Appendix O. Formalisation of LPAetiology



Appendix P. Formalisation of LPTreatment



Appendix H. Article Abstracts for the Evaluation of Lipoprotein Ontology

No.	Title of Paper
1.	2009 Canadian Cardiovascular Society Canadian guidelines for the diagnosis and treatment of dyslipidemia and prevention of CVD in adults
2.	A receptor-mediated pathway for cholesterol homeostasis
3.	A simplified approach to lipoprotein kinetics and factors affecting serum cholesterol and triglyceride concentrations
4.	Additional cardiovascular risk factors associated with excess weight in children and adolescents. The Belo Horizonte Heart Study
5.	An ABC of apolipoprotein C-III: A clinically useful new cardiovascular risk factor?
6.	Analysis for cholesterol in all lipoprotein classes by single vertical ultracentrifugationof fingerstick blood and controlled-dispersion flow analysis
7.	Apolipoproteins and metabolism in atherosclerosis
8.	Characterization of electrophoretic lipoprotein fractions: Immunochemical and electron microscopic studies
9.	Characterization of plasma lipoproteins separated and purified by agarose-column chromatography
10.	Cholesterol metabolism, LDL, and the LDL receptor
11.	Cholesterol metabolism
12.	Chylomicron metabolism
13.	Classification of hyperlipidaemias and hyperlipoproteinaemias
14.	Classification of lipoproteins and lipoprotein disorders
15.	Comparison of the efficacy and safety of rosuvastatin versus atorvastatin, simvastatin, and pravastatin across doses (STELLAR Trial)
16.	Detection, evaluation and treatment of high blood cholesterol in adults (ATP III)
17.	Diet and triglyceride metabolism
18.	Dietary fiber and its interaction with drugs
19.	Differences in apolipoprotein and lipid composition between human chylomicron remnants and very low density lipoproteins isolated from fasting and postprandial plasma
20.	Discovery of the lipoproteins, their role in fat transport and their significance as risk factors
21.	Disorders of lipid metabolism
22.	Disorders of lipoprotein metabolism
23.	Drug interactions of lipid-altering drugs
24.	Drug interactions with lipid-lowering drugs: Mechanisms and clinical relevance

25.	Dynamics of lipoprotein transport in the human circulatory system
26.	Effects of alcohol on lipoprotein lipase, hepatic lipase, cholesteryl ester transfer protein, and lecithin-cholesterol acyltransferase in high-density lipoprotein cholesterol elevation
27.	Effects of alcohol on plasma lipoproteins and cholesterol and triglyceride metabolism in man
28.	Effects of atorvastatin treatment on the oxidatively modified low density lipoprotein in hyperlipidemic patients
29.	Efficacy and safety of controlled-release niacin in dyslipoproteinemic veterans
30.	ESC/EAS Guidelines for the management of dyslipidaemias
31.	Exploring the lipoprotein composition using Bayesian regression on serum lipidomic profiles
32.	Familial hypobetalipoproteinemia: A review
33.	Functional overlap between “chylomicra” and “very low density lipoproteins” of human plasma during alimentary lipaemia
34.	Functions and interrelationships of different classes of plasma lipoproteins
35.	Genetic basis of lipoprotein disorders
36.	Genetic factors affecting blood lipoproteins: The candidate gene approach
37.	Genetic influences on susceptibility to atherosclerosis in the young
38.	Genetic markers in atherosclerosis: A review
39.	Heterogeneity of human plasma very low density lipoproteins. Separation of species differing in protein components
40.	Hypertriglyceridemia: Its etiology, effects and treatment
41.	Hypolipoproteinaemia
42.	Influence of diet and physical exercise on plasma lipid
43.	Joint distribution of lipoprotein cholesterol classes. The Framingham study
44.	Lipid absorption and metabolism
45.	Lipid composition of human serum lipoproteins
46.	Lipid composition of lipoproteins of normal human plasma
47.	Lipid depletion in atheromatous coronary arteries in rhesus monkeys after regression diets
48.	Lipid-lowering drugs
49.	Lipid metabolism
50.	Lipoprotein heterogeneity and apolipoprotein B metabolism
51.	Lipoprotein kinetics in the metabolic syndrome: Pathophysiological and therapeutic lessons from stable isotope studies
52.	Lipoprotein management in patients with cardiometabolic risk
53.	Lipoprotein structure and metabolism
54.	Lipoproteins and cardiovascular reactivity

55.	Lipoproteins and lipoprotein metabolism: A dynamic evaluation of the plasma fat transport system
56.	Low-density lipoprotein, non-high-density lipoprotein, and apolipoprotein B as targets of lipid-lowering therapy
57.	Low-density lipoprotein subclasses: Mechanisms of formation and modulation
58.	Metabolic relationships among the plasma lipoproteins
59.	Metabolic syndrome: The danger signal in atherosclerosis
60.	Molecular genetics of lipoprotein disorders
61.	Pathophysiology of dyslipidaemia in the metabolic syndrome
62.	Plasma lipid concentrations: The concept of normality and its implications for detection of high cardiovascular risk
63.	Plasma lipids and lipoproteins in liver disease
64.	Plasma lipoproteins: Genetic influences and clinical implications
65.	Plasma lipoproteins, lipid transport, and atherosclerosis - recent developments.
65.	Plasma triglyceride determines structure-composition in low and high density lipoproteins
66.	Plasma or patient, paper electrophoresis or physician? The four-P problem in classification of hyperlipidaemia
67.	Postheparin plasma lipoprotein and hepatic lipase are determinants of hypo- and hyperalphalipoproteinemia
68.	Prediction of coronary heart disease using risk factor categories
69.	Prevention and treatment: A tale of two strategies
70.	Recommendations for the management of dyslipidemia 2003 update
71.	Risk for myopathy with statin therapy in high-risk patients
72.	Separation of the main lipoprotein density classes from human plasma by rate-zonal ultracentrifugation
73.	Secondary hyperlipidaemia
74.	Serum lipoproteins in four European communities: A quantitative comparison
75.	Simple scoring scheme for calculating the risk of acute coronary events based on the 10-year follow-up of the prospective cardiovascular Münster (PROCAM) study
76.	Standardization of measurements for cholesterol, triglycerides, and major lipoproteins
77.	Statin safety and drug interactions: Clinical implications
78.	Structural heterogeneity of apoB-containing serum lipoproteins visualized using cryo-electron microscopy
79.	Studies of the mechanisms of carbohydrate-induced lipaemia in normal man
80.	Studies on the composition and structure of plasma lipoproteins: Distribution of lipoprotein families in major density classes of normal human plasma lipoproteins
81.	The anti-coronary club: A dietary approach to the prevention of coronary heart disease - a 7 year report
82.	The distribution and chemical composition of ultracentrifugally separated lipoproteins in human serum

83.	The enterohepatic circulation of bile acids as they relate to lipid disordersx
84.	The metabolic heterogeneity of human very low density lipoprotein triglyceride
85.	The metabolic pathways of high-density lipoprotein, low-density lipoprotein, and triglycerides: A current review
86.	The plasma lipoproteins: Structure and metabolism
87.	Therapy of lipid disorders
88.	Translating molecular discoveries into new therapies for atherosclerosis
89.	Triglycerides and cardiovascular disease: A scientific statement from the American Heart Association
90.	Waiting for the National Cholesterol Education Program Adult Treatment Panel IV Guidelines, and in the meantime, some challenges and recommendations

Appendix H. Methodologies for Ontology Development

Introduction

Numerous methodologies have been proposed for ontology development. The following describes some of these methodologies, of which we have provided a comparison and critical review in Chapter 4.5.

1.1 Knowledge Engineering Methodology

A simple Knowledge Engineering Methodology was proposed by Noy & McGuinness to develop an ontology (Noy & McGuinness, 2001). Three fundamental rules serve as the basis for the methodology: 1. Modelling a domain depends on the application the ontology is built for; 2. Ontology development is an iterative process; 3. Concepts in the ontology should be as close to objects (nouns) and relationships (verbs) in the domain of interest. The methodology covers seven steps: purpose and scope identification, reusing existing ontologies if applicable, highlighting important terms in the ontology, defining the classes and class hierarchy, defining the properties of classes, defining restrictions, and creating instances. The first step defines the purpose and scope of the ontology by developing competency questions to which an ontology should be able to answer. These questions will also be used to validate the ontology later on. The second step involves refining and extending existing ontologies towards the domain of interest and intended purpose. This is particularly useful if the system is required to interact with other applications which have already committed to particular ontologies. The third step is the informal conceptualisation phase which consists of obtaining a comprehensive list of important terms and relations. In the fourth step, the concepts obtained from the previous stage are arranged in a class hierarchy. There are several possible approaches, which includes a *top-down* approach of defining the most general concepts in the domain and subsequent specialisation of the concepts, *bottom-up* approach of starting with the most specific concepts and subsequent grouping of these concepts into more general concepts, and a *middle-out* approach that combines the top-down and bottom-up approaches (Uschold & Gruninger, 1996). The fifth step involves defining the properties of the concepts that have been established in the previous step. The sixth step consists of

describing the value type (e.g. string or integer), allowed values, the number of values (cardinality) a property might have. The final step involves creating individual instances of the classes in the hierarchy. Defining an individual instance of a class requires allocating an individual instance to a class and describing its property.

1.2 Uschold & King Methodology

Uschold & King proposed a general methodology towards ontology development, based on the experience of creating Enterprise Ontology (Uschold & King, 1995). The methodology incorporates four main steps: purpose identification, ontology building, evaluation and documentation. In the initial stage, the ontology builder defines the purpose of the ontology and its intended applications. Competency questions can be drawn to assist in the clarification of the purpose, as well as serve as a reference document for validating the completed ontology. The ontology building stage includes three sub-phases: ontology capture, ontology coding and integration of available ontologies. Ontology capture involves identifying domain concepts and their relationships via a middle-out approach: by identifying the most general concepts, which are then used to branch out to upper and lower concepts by generalisation or specialisation respectively. In the coding phase, the concepts and relationships specified in the previous stage are formalised using a representation language. The integration task considers the possible incorporation of other existing ontologies into the ontology. In the third stage, the developed ontology is evaluated. The evaluation can be made with respect to reference criteria such as competency questions which were established in the initial phase, requirement specifications as well as compliance with the real world. The documentation stage prescribes the documentation of all underlying assumptions about the concepts in a given ontology in order to facilitate the process of effective knowledge sharing and reuse (Uschold & King, 1995). Enterprise Ontology was developed in Ontolingua, which will be described in the next section.

1.3 TOVE Methodology

The TOVE methodology was originally developed to aid enterprise process modelling at the Toronto University (Gruninger & Fox, 1995). The process includes the use motivating scenarios as well as a set of competency questions to determine the scope of the ontology to be modelled, and aims to generate an enterprise model which is capable of deducing answers to users' implicit queries. It places a special emphasis on the formulation of informal competency questions to which the ontology must be an answer. TOVE incorporates six stages: creating motivating scenarios, formulating informal competency questions, formalising terms extracted from the competency questions, developing formal competency questions, establishing first-order logic axioms, and validation using completeness theorems. In the motivating scenario stage, the ontology developer identifies and describes situations and applications to which ontology is expected to offer solutions. The proposed motivating scenario also incorporates a number of intuitively potential solutions to the identified problems as well as a rationale for including certain objects in the ontology. In the next stage, the requirements or questions to which ontology should provide answers are specified and described in an informal way. Subsequently, a set of terms are extracted from the informal competency questions and formally expressed in first-order logic or KIF language. The terminology specification relies on the identification of the objects, their attributes and relations in the given domain. In the next stage, the established ontology terminology is used to develop the formal competency questions. Through an iterative process, first-order logic axioms are defined in order to provide semantics for the ontology terms and concepts. Axioms provide terms with appropriate definitions and impose restrictions on their interpretations. The process of axiom specification is basically directed by the predefined formal competency questions. Axioms are necessary and must adequately express the competency questions and their potential solutions. If an insufficient number of axioms have been proposed, then they must be refined and extended, and if necessary, other axioms added until there are adequate axioms for representing questions and solutions. In the final stage, called the completeness theorem, the expert will define the conditions under which the ontology has offered complete solutions to the competency questions. The formal competency questions, in this phase, is used to prove the completeness theorems of the established ontology (Gruninger & Fox, 1995).

1.4 DOGMA Methodology

The DOGMA (Developing Ontology-Grounded Methods and Applications) methodology (Spyns et al., 2008) separates the conceptualisation of a domain (ontology base) from their application (commitment layer), which allows for reusability and scalability in reasoning about formal semantics. The methodology offers a special paradigm for the separation of the domain axiomatisation (the ontology base) from the application axiomatisation (the commitment layer) with the purpose of finding a solution for the trade-off problem which often exists between an ontology's usability and its reusability. The advantage of DOGMA allows domain experts and users to have multiple views and requirements for different applications while using the same stored, meaning-independent conceptualisation (Spyns et al., 2008). Moreover, the DOGMA proposes the notion of the context which can be considered as an identifier to restrict the interpretation of each term to the specified concepts which exist within the context of that term (Jarrar & Meersman, 2008).

1.5 METHONTOLOGY

The METHONTOLOGY framework was developed by by Polytechnic University of Madrid, based on IEEE standards for Developing Software Life Cycle Processes, 1074-1995 (Fernández-López, M., & Gomez-Perez, A., 2002) The methodology basically enables the construction of ontologies at the knowledge level. METHONTOLOGY proposes three separate, yet overlapping set of activities to develop ontologies based on an evolving life cycle and prototype refinement: ontology development process, support activities and project management process. The developmental process in METHONTOLOGY incorporates the stages of requirement specification, domain conceptualisation, formalisation of the conceptual model in a formal language, implementation of the formal model and maintenance of the implemented ontology. The support actions may include knowledge acquisition, documentation, evaluation and integration of other ontologies. Finally, the project management activities concern the tasks of planning, control and quality assurance (Fernández-López, M., & Gomez-Perez, 2002). WebODE, which will be discussed in the next section, is a support tool for this methodology.

1.6 OnToKnowledge

Developed at the Karlsruhe University, OnToKnowledge is a process-oriented methodology that focuses on knowledge management and maintenance in enterprises with respect to an analysis of usage scenarios (Staab et al., 2001). OnToKnowledge is based on a two-loop architecture: knowledge process and knowledge meta process. Knowledge process includes knowledge acquisition and evolution process. Knowledge meta process is a methodology of ontology development and consists of five steps: feasibility study, kick off, which includes specifying ontology requirements, identifying competency questions and considering the use of other ontologies; refinement, where a mature ontology is produced; evaluation, where the ontology is validated according to the initial requirements and competency questions; maintenance (Fense et al., 1999).

1.7 DILIGENT Methodology

The DILIGENT methodology aims to support DIstributed, Loosely controlled and evolvInG Engineering of oNTologies (Pinto et al., 2004). It can be considered as an extension of ontology engineering methodologies such as OnToKnowledge or METHONTOLOGY with an emphasis on user centrality. It plans to integrate automatic agents in the ontology evolution process, allowing the ontology engineer to adapt to the unremitting change of domain knowledge. The DILIGENT methodology involves five major steps: building, local adaptation, analysis, revision and local update. In the building stage, the initial ontology is built by a small number of domain experts, users, ontology engineers and knowledge engineers. The ontology model at this stage need not be complete. After this preliminary ontology has been established, the users start working with it, developing their local ontologies by adapting it to their local requirements. The users are also able to modify the core ontology through a control board which records all modifications. At the analysis stage, local ontologies and the requests for change are analysed by the control board so that similarities among them can be discovered. After that, decisions will be made as to which modifications need to be applied to the core ontology to meet various users' requirements. However, the new adaptations and localisation need to be revised in order to ensure that the core ontology has not lost its sharable quality. Hence,

the revision stage aims to adapt the ontology to various applicants' requirements, enhancing its acceptance, consensuality and sharedness (Pinto et al., 2004). Experts from different areas take responsibility for the revision of the ontology. For example, users evaluate the usability and advantages of the ontology, providing feedback to ontology engineers through their requests and requirements. Respectively, the existence of factual mistakes and the degree to which the ontology represents the intended knowledge domain are assessed by domain experts. Correspondingly, knowledge engineers and ontology engineers evaluate the technical dimensions of the ontology such as its efficiency, logical properties or standard conformance, trying to update as well as hold the balance of different applied ontology modifications. Finally, at the stage of local update, applicants update their local ontologies to cope with the revisions which have been introduced to the modified core ontology (Vrandecic et al., 2005). The DILIGENT methodology is particularly suitable for de-centralised knowledge management systems. It also offers flexibility in the use of ontology language or formalisms (Vrandecic et al., 2005).

1.8 KACTUS Methodology

The KACTUS methodology involves building the ontology on the basis of a knowledge base through the process of abstraction, following a bottom-up strategy (Bernaras et al., 1996). Initially, a knowledge base is built towards a specific application, but as more requirements are needed, the knowledge base is subsequently generalised into an ontology and adapted to the new applications. The more applications are built, the more general the ontology becomes. By applying this method recursively, the ontology gradually expands and represents consensual knowledge needed in all applications.

1.9 SENSUS Methodology

The SENSUS methodology derives domain-specific ontologies from large ontologies and enables reusability of knowledge since they have a common underlying structure (Swartout et al., 1997). The process involves the identification of "seed" terms which are relevant to a given domain, and linked manually to an upper ontology. All the concepts from the seed terms to the root

of the ontology are included; any missing term is added manually and the previous step is performed again to ensure completeness. In addition, nodes which have a large number of associations through them would have the entire subtree under the node added on the basis of relevance.

References

Bernaras, A., Laresgoiti, I., & Corera, J. M. (1996). *Building and reusing ontologies for electrical network applications.* Proceedings of the European Conference on Artificial Intelligence (ECAI 1996), Budapest, Hungary.

Fernández-López, M., & Gomez-Perez, A. (2002). Overview of methodologies for building ontologies. *The Knowledge Engineering Review, 17*(2), 1-13.

Gruninger, M., & Fox, M. S. (1995). *Methodology for the design and evaluation of ontologies.* Methodology for the design and evaluation of ontologies, Montreal, Canada.

Jarrar, M., & Meersman, R. (2008). Ontology Engineering -The DOGMA Approach *Advances in Web Semantics I* (Vol. LNCS 4891): Springer.

Noy, N. F., & McGuinness, D. L. (2001). Ontology development 101: a guide to creating your first ontology, from http://protege.stanford.edu/publications/ontology_development/ ontology101-noy-mcguinness.html

Pinto, H. S. a., Staab, S., & Tempich, C. (2004). *DILIGENT: Towards a fine-grained methodology for DIstributed, Loosely-controlled and evolvInG*
Engineering of oNTologies. Proceedings of the 16th European Conference on Artificial Intelligence (ECAI 2004).

Spyns, P., Tang, Y., & Meersman, R. (2008). An ontology engineering methodology for DOGMA *Journal of Applied Ontology, 3*(1-2), 13-39.

Swartout, B., Patil, R., Knight, K., & Russ, T. (1997). *Toward distributed use of large-scale ontologies.* Symposium on Ontological Engineering of AAAI, Stanford, California.

Uschold, M., & Gruninger, M. (1996). Ontologies: Principles, Methods and Applications. *Knowledge Engineering Review, 11*(2).

Uschold, M., & King, M. (1995). *Towards a methodology for building ontologies.* Workshop on Basic Ontological Issues in Knowledge Sharing, UK.

Vrandecic, D., Pinto, S., Sure, Y., & Tempich, C. (2005). The DILIGENT Knowledge Processes. *Journal of Knowledge Management, 9*(5), 85-96.