8-1-2018 2:00 PM

Improving Prediction of Systemic Statin Exposure Using Concomitant Medications, Non-Linear Modelling and Novel SNP Discovery

Rhiannon Rose, *The University of Western Ontario*

Supervisor: Lizotte, Daniel J., *The University of Western Ontario*

A thesis submitted in partial fulfillment of the requirements for the Doctor of Philosophy degree in Epidemiology and Biostatistics

Abstract

Introduction: Statin drugs are a highly efficacious treatment for hypercholesterolemia and adherent treatment with statins reduces the risk of cardiovascular disease. Although statins are generally well tolerated, myalgia (muscle pain) is a common side effect and can lead to non-compliance with treatment. Increased systemic exposure may contribute to the development of myalgia. Drug-drug interactions inhibiting statin metabolism and impaired drug transporter function may lead to decreased statin clearance. Establishing accurate predictive models is an important step towards preventing adverse drug events by titrating statin dosing to limit systemic exposure.

Objectives: 1) To develop an algorithm to select concomitant medications for incorporation into the existing systemic exposure model and assess their predictive impact; 2) to apply nonlinear techniques to model systemic statin exposure, and assess their effectiveness and feasibility; 3) to identify novel genes and corresponding single nucleotide polymorphisms using next generation sequencing (NGS) in patients whose statin plasma concentrations were under-predicted using the original systemic exposure model to guide future biological research.

Methods: Data from a previously-collected prospective cohort of 130 patients prescribed rosuvastatin and 128 patients prescribed atorvastatin were used in this analysis. Concomitant medications were selected using penalized regression. Stable feature selection was achieved by repeated cross validation. Generalized additive models (GAMs) and support vector regression (SVR) were assessed as candidate nonlinear models. Candidate patients were chosen for NGS sequencing based on the proportional difference between their true and predicted values from the original systemic exposure model. Variant prioritization used the Sequence Kernel Association Test.

Results: Atorvastatin linear model fit was statistically significantly improved by incorporating the selected concomitant medications, but rosuvastatin model fit was not. Predictive performance was not improved using GAMs or SVR compared to linear regression, likely due to small sample size. Three candidate genes and corresponding observed variants were identified and discussed as potential predictors of systemic rosuvastatin exposure.

Conclusion: Linear modelling of systemic atorvastatin exposure can be improved by incorporating concomitant medications. The feasibility of using nonlinear predictive models is limited by small sample size. Future research on newly identified interacting medication and genetic variants may provide new insights regarding underlying molecular mechanisms affecting systemic statin exposure.

Keywords: Statins, hypercholesterolemia, nonlinear modelling, clinical pharmacology, plasma concentration, next generation sequencing

Acknowledgements

The completion of this thesis would not have been possible without the overwhelming support of so many people throughout my studies. First and foremost, I thank my supervisor Dr. Daniel Lizotte for being the most amazing mentor and supervisor a student could possibly wish for. Thank you so much for your support and guidance that has allowed me to grow as researcher and a person. I also wish to sincerely thank the members of my thesis committee, Drs. Ute Shwarz and Neil Klar for providing such excellent feedback and help throughout the course of my studies. Thanks very much to Drs. Stephanie Frisbee, GY Zou, and Kamran Sedig for providing excellent feedback during my thesis proposal examination. Thanks also to my final thesis examiners Drs. Ava Jean-Baptiste, Yun-Hee Choi, Cecelia Cotton and Steven Gryn for their insightful comments. Huge thanks to the National Sciences and Engineering Research Council (NSERC) for their financial support that made this degree a lot more tractable from a monetary perspective.

Many of the most valuable experiences I have had as a doctoral student were afforded to me by Dr. Richard Kim and the members of the personalized medicine lab. A huge thank you to Dr. Wendy Teft, Dr. Rommel Tirona, and all of the other people involved with the lab that I have had the pleasure of working with over the past few years - without you I wouldn't know how much I love working in the field of clinical pharmacology (and also what clinical pharmacology even is)! I've had so much fun being a part of the lab, and am extremely grateful that I had the opportunity to work on so many cool projects with so many cool people. Also, a massive thanks to Dr. Marianne DeGorter for laying the groundwork for the project that I was given the opportunity to participate in, and for collecting such a great dataset and allowing me to work with it.

The Department of Epidemiology and Biostatistics at Western University is full of people that are extremely helpful, kind and supportive. Thank you to Drs. Greta Bauer, Karen Campbell, Kelly Anderson, Saverio Stranges, Yun-Hee Choi, GY Zou, and the other members of the department for imparting so much information and guidance over the past four years. I wouldn't have been able to finish this degree without the department being so supportive and making accommodations for health problems I have dealt with on and off throughout my tenure here.

My entire family has been amazingly supportive over the course of my academic career thus far, and I thank all of you for all of the love and support you have shown me. I especially thank my dad, John Molenaar, and my stepmum Sheri Wiggins for helping me get through all of the ups and downs that come with adjusting to life as an adult. A big thank you to my marma, Jennifer Rose, my sister Sam, and my brothers Connor and Maarten. Thank you to Paul, Monique and Annie Johnston for welcoming me into your family with open arms; a

million thanks to my best friend Sarah “Facey” Johnston for always being there for me ever since we made friends in the sixth grade. Massive thanks to Adam Hartfiel for the many amazing years we spent together. Thanks also to Allie Engelhardt and Jenn Garrett for being super great friends.

I have made some amazing friends while at Western. Thank you Gina Bhullar, Jordan Edwards, Jaky Kueper, Jason Black, Kathryn Nicholson, Ayden Scheim and all of the many others who helped brighten my days while doing research and coursework during my degree. My graduation cohort partners in crime Markus Gulilat, Aze Wilson and Adrienne Borrie made getting to the end of my degree way more enjoyable and motivating, and their research projects were a ton of fun to be involved with. A million thanks to Brent Davis for providing a beacon of hope for success during the trials of NGS data processing - many many more buckets of tears would have been shed over that awful analysis without your help. Extra special thanks to Adrienne Borrie for being an amazing, genuine and caring person. Working with you on your research was a great experience, but more importantly, your emotional support and friendship for the past few years has gotten me through some incredibly tough times and I wouldn’t be the person I am today without you.

Finally, I would like to thank the hordes of musicians who have made such inspiring music to listen to while writing code and papers, and the authors who have gifted the world with such entertaining stories to read while procrastinating on getting work done. Thank you to the entire internet, and especially those people on it who have the expertise and time to post helpful things on StackOverflow. I would like to thank the universe for providing everyone with such cool opportunities for self-improvement, and finally, my cat Bells for being perpetually adorable.

Dedication

This thesis is dedicated to my fellow survivors of mental illness -
we may have to play the game of life on hard mode, but we can still win

❤ ❤ ❤

Table of Contents

List of Figures

List of Tables

List of Appendices

Chapter 1

Background

1.1 Use of Statins for Hypercholesterolemia

Cardiovascular disease (CVD) is a leading cause of mortality around the globe; it has been estimated that cardiovascular diseases are responsible for approximately 30 percent of global mortality[1]. Hypercholesterolemia (elevated blood levels of low-density lipoprotein (LDL) cholesterol) is a major risk factor for CVD. When high levels of LDL cholesterol are in the circulatory system, these particles penetrate the innermost layer of cells in the walls of arteries and accumulate, causing inflammation. The immune system recognizes this as damage, and capped plaques full of fat-engorged leukocytes form in the area. Stroke and myocardial infarction can then result from these plaques being disrupted and forming clots in the bloodstream[1]. It has also been determined that cardioprotective cholesterol exists in the form of high-density lipoprotein (HDL), which removes lipids from the blood more effectively than LDL as the lipids are packed more densely, and also stops the formation and oxidization of LDL plaques[1]. Prior to the discovery of statin drugs, blood cholesterol management strategies were limited to

diet modification, and the use of niacin, fibrates, probucol and bile-acid sequestrants, most of which had limited success or undesirable side effects[2]. The development of statin drugs was a major breakthrough for treating CVD as statins provide a reliable and highly effective means of reducing LDL cholesterol levels within the body[3]. Statins achieve lipid-lowering by competitively inhibiting 3-hydroxy-3-methylglutaryl-CoA (HMG-CoA), which is a key enzyme in the cholesterol synthesis pathway[1]. Atorvastatin (trade name Lipitor) and rosuvastatin (trade name Crestor) are very popular prescription options for statin therapy, and have been found to be generally well tolerated. Atorvastatin was found to have greater cholesterol-lowering capabilities than previously developed statin drugs including pravastatin, lovastatin, simvastatin and fluvastatin in a randomized open-label clinical trial (the CURVES study)[4]. A benefit of rosuvastatin is that it has a high potency to lower cholesterol, and has a lower potential for interactions with other medications compared to atorvastatin. There is less of a potential for interactions with other drugs since rosuvastatin is minimally metabolized before performing its intended action, unlike other statins[5]; interactions may happen when different chemicals associated same metabolic enzymes or drug transporters, and are competitvely inhibiting or inducing one another. Additionally, like all statin drugs it is primarily active in the liver rather than the surrounding tissues (in which it can produce muscle pain), and it has a good duration of action[6]. A substantial body of research exists comparing the effectiveness and safety of atorvastatin versus rosuvastatin[7;8;9;10]. One meta-analysis on the risks and benefits of these two drugs comparatively found that rosuvastatin is more effective at reducing LDL cholesterol than atorvastatin at 1:1 and 1:2 dose ratios of rosuvastatin to atorvastatin; there was no significant difference found at the dose ratio of 1:4 for rosuvastatin to atorvastatin[10]. Dosing for atorvastatin and rosuvastatin is regulated according to prescribing information; within these

regulations, atorvastatin dosing typically ranges from 10-80 mg per day with oral administration, while rosuvastatin doses generally range from 5-40 mg per day[11].

1.2 Statin-Induced Myopathy

Major adverse effects of treating hyperlipidemia using statins are myopathy and myalgia; these are problems that affect the skeletal muscles[12]. In extreme cases statins can cause rhabdomyolysis, in which striated muscles experience damage, disintegration and necrosis. Muscle cells contain creatine kinase (CK), which leaks into plasma as a result of the muscle damage[13]. Elevated CK plasma level is used as a metric of muscle damage, and the diagnostic threshold for myopathy is a concentration that is >10 times greater than the normal range in plasma[14;15;16]. In severe cases rhabdomyolysis can be fatal[17]; however, muscle damage to this extent is very rare (0.003 - 0.1%)[18]. More often patients experience myalgia, a less severe disorder characterized by muscle pain and weakness, but not usually elevated levels of CK[19]. Even so, less severe statin-associated myalgia has the potential to severely decrease the quality of life for patients taking this medication; additionally it may decrease compliance with the medication[20], resulting in a lowered dose of the medication, or pursuit of an alternative therapy.

1.2.1 Myalgia Definition, Incidence and Risk Factors

The exact definition of what constitutes clinically relevant myalgia differs between different studies, but in general encompasses the experience of muscle pain, soreness, weakness, cramping, tenderness, stiffness and/or heaviness[20]. The muscle pain or weakness caused as a result of statin exposure can appear or worsen during exercise, when general muscle injury can oc-

cur; however, many patients also experience myalgia at rest[20]. The lack of a general consensus on an exact definition of myalgia between studies on adverse drug events resulting from statin exposure makes estimating the prevalence difficult. In clinical trials of statins, the incidence of myalgia is reported to be around 1.5-3%, comparable to that of patients taking a placebo[20;18]. The incidence of myalgia observed in these trials is probably lower than would be seen in the general clinical population however, because often patients particularly at risk for myotoxicity are excluded from study populations, as are patients who have extensive comorbidities and concomitant medication use[12]. Unfortunately, no clinical trials have been performed using a consensus definition of myalgia and non CK-elevated myopathy, which makes it difficult to accurately compare the risk of these adverse drug events between different HMG-CoA inhibiting drugs, although atorvastatin is thought to have a higher risk than other statin drugs[12]. Various community-based studies have estimated the incidence of myalgia to be between 5-20%[19;18;21], affecting a significant portion of patients treated using statins.

Specific risk factors for statin-induced myalgia are predominantly those that affect the concentration of statins in the plasma and surrounding muscle tissue, interacting medications, and factors for independent muscle injury. These include statin dose, polymorphisms decreasing the function of drug transporters responsible for hepatic statin uptake or efflux, impairment of liver and kidney function, age (especially persons over 80 years), frail body condition, female sex, conditions like diabetes and hypothyroidism, and acute factors such as addictive drug use, excessive alcohol consumption, heavy exercise and muscle inflammation, and extensive surgery[12;15]. Ethnicity has also been shown to be strongly associated with statin plasma level, with Asian patients having a rosuvastatin plasma concentration approximately double that of Caucasian patients, despite taking the same dose of medication[22]. Importantly, the risk of

myalgia among patients taking statin drugs is dose dependent; systemic exposure to statins outside of the liver where their primary mechanism of action lies, particularly in the plasma and surrounding muscle tissue[11], is thought to play a causal role in the development of these adverse drug events[19]. The risk of myalgia based on systemic exposure has been found to be independent of the level of lipid-lowering achieved by statin therapy[12], although this effect is debated in the literature.

1.2.2 Statin Plasma Level and Drug Transporters

Drug transporters are important determinants of statin concentration in the liver, bloodstream, and muscle tissue[23]. Uptake transporters of the organic anion transporting polypeptides (OATP) family are members of the solute carrier transport *SLCO* superfamily of genes[24], and are largely responsible for delivering statins into the liver. For example, OATP1B1 is particularly important when considering the uptake and distribution of statins in the body, as it is thought to be the main carrier of statins into the liver; indeed, OATP1B1 is expressed only on the basolateral membrane of the liver[23]. Polymorphisms decreasing the function of this transporter are associated with higher area under the curve (AUC) of plasma exposure for many types of statins and may correspond to higher risk for patients of experiencing muscle pain or muscle damage as a result of toxicity[23]. Importantly, OATP1B1 c.521T>C, an impaired-function SNP resulting in a change from thymine to cytosine, has been shown to strongly associate with myopathy in patients prescribed simvastatin and atorvastatin, as reported in genome wide association studies (GWAS)[11]. The change from thymine to cytosine results in an amino acid change from valine to alanine. Patients with decreased hepatic expression of this transporter also have a higher risk

of myopathy because when less of the drug is entering the liver, the bioavailability of the drug in the plasma increases[12].

Efflux transporters such as ATP-binding cassette (ABC) transporters are also important for the distribution and removal (excretion) of statin drugs from the body[23]. Polymorphisms reducing the function of ABCG2, also known as breast cancer resistance protein (BCRP), have been shown to increase the concentration of rosuvastatin in the plasma[25]. With certain SNPs such as ABCG2 c.421T>C, the transporters are unable to export as much rosuvastatin from the liver, and the resultant higher concentration in the liver causes a greater lipid-lowering effect[23]. Other efflux transporters that can affect statin plasma concentration are ABCB1 (P-glycoprotein /P-gp) and ABCC2 (multidrug resistance-associated protein 2/ MRP2)[26].

A number of potential causal mechanisms for muscle damage caused by statin use have been hypothesized. It has been observed that drug transporters responsible for cellular statin uptake are also expressed in human muscle tissue, such as OATP2B1[19;23]. Because statin drugs lower cholesterol by distrupting HMG CoA Reductase within the melavonate pathway, it is possible that the reduction of intermediaries on this pathway within muscle tissue could be a potential cause. Ubiquinone or Co-Q10 has been observed to be depeleted in muscle tissue as a result of high-dose statin exposure. It is a component of the mitochondrial electron transport chain and antioxidant, and thus is important for many cellular functions. However, a supplementation study failed to observe a role of Co-Q10 in reversing symptoms[21]. It is also possible that a decreased mitochondrial volume could be implicated in the development of myalgia[27]. Another hypothesis that has been supported by experimental evidence is that myalgia could be mediated by statin-induced upregulation of the phospholipase C pathway which causes a large increase in the influx of calcium (Ca2+) into cells, disrupting calcium

metabolism[28]. Other mechanisms that might play a role in the development of myopathy and myalgia are increases in muscle cell apoptosis via pathways upregulated by statin exposure, and a depletion of enzymes produced by cholesterol synthesis that in turn destabilizes muscle cell membranes[15;29;21].

1.3 Statin-Drug or Herb-Drug Interactions

1.3.1 Statin-Drug Interactions Involving Cytochrome P450 (CYP) Enzymes

Hypercholesterolemia and cardiovascular disease tend to cluster with other conditions such as diabetes, obesity and metabolic syndrome[15]; because of the high level of comorbidity in these populations and the effectiveness of statins for treating high cholesterol, polypharmacy and concomitant medication use are very likely among patients who are prescribed statin drugs[11]. This is problematic because some statins (i.e. atorvastatin) are significantly metabolized by the cytochrome P450 (CYP) 3A family of enzymes, namely CYP3A4 and CYP3A5[5]. Atorvastatin clearance (around 90%) occurs primarily through CYP3A4, which also metabolizes a significant portion of other prescription drugs[11]. When drugs share a metabolic pathway in this manner, the level of statins circulating in plasma increases as the statin molecules are inhibited from binding with the CYP3A4 enzymes by other drugs, in turn increasing the systemic exposure of the drug[12].

Some examples of drugs that may pose a problem with statins because they are known inhibitors of CYP3A enzymes are cyclosporine, erythromycin, itraconazole and HIV protease

inhibitors[30]. Adverse drug events become much more likely in the context of concomitant medication use; while rhabdomyolysis is rare the number needed to treat in order to see one case increases from 1 in 22717 for patients on statin monotherapy to 1 in 1672 for patients concurrently taking statins and fibrates[14]. An estimated about 60% of rhabdomyolysis cases in patients using statin therapy for lipid-lowering are thought to be caused by concomitant drug use[12].

Because atorvastatin is metabolized by CYP3A4, patients taking it are more likely to experience drug-drug interactions (DDIs) than patients taking rosuvastatin, which is not metabolized before uptake into the liver. In general, the following types of medication tend to affect plasma concentrations of atorvastatin via CYP3A4 inhibition: cyclosporine, an immunosuppressant[31]; antibiotic medications such as erythromycin and clarithromycin; antifungal drugs such as itraconazole, ketoconazole and fluconazole; HIV protease inhibitors such as ritonavir, nelfinavir and indinavir; calcium channel antagonists such as diltiazem and verapamil used for treating hypertension and arrhythmias; grapefruit juice; and others drugs inhibiting CYP3A4.

From the standpoint of metabolism, rosuvastatin is less likely than atorvastatins to have DDIs, because it is not metabolized by CYP3A4. However, since it is minimally metabolized by CYP2C9, there is a chance of interactions with drugs that are also metabolized by that enzyme due to competitve inhibition by other metabolites that interact with this enzyme. Additionally, drugs that tend to independently cause muscular concerns can interact synergistically with statins to worsen myalgias[15;20;32;12;33].

1.3.2 Statin-Drug Interactions Involving Drug Transporters

Another avenue for potential DDIs with statin drugs involve competitive inhibition of OATP1B1, as it is a transporter of many substances besides statin drugs. Substances that are also transported by OATP1B1 include most bile acids, protease inhibitors, some anti-diabetic medications, among others[34]. To this date, approximately 65 substances have been identified as inhibitors of this drug transporter[34]. Similar to the mechanism behind CYP3A4-related DDIs with atorvastatin, when other substances transported by OATP1B1 are present in the body along with statins, the capacity of the transporter to deliver statins to the liver is decreased, as OATP1B1 is occupied by the transport of other substances. A detailed list of drugs with the potential to interact with statin medications can be found in Tables 1.1, 1.2, 1.3 and 1.4.

1.4 Predictive Modelling of Systemic Statin Exposure

Two multiple linear regression models were developed at London Health Sciences Center (LHSC) by DeGorter et al.[23] for the purpose of studying the relationship between patient factors and statin plasma concentration. Atorvastatin and rosuvastatin were modelled separately to better account for differences in drug metabolism and transport. Both models controlled for age, body mass index (BMI), sex, self-reported ethnicity, statin dose, and time since last dose taken. The atorvastatin predictive model additionally controlled for the plasma concentration of 4β-hydroxycholesterol.

The covariates of interest found to be significant for predicting the log-transformed atorvastatin plasma level included 4β-Hydroxycholesterol (a marker for CYP3A activity), and the OATP1B1 polymorphisms c.521T>C, and c.388A>G. The variables age, dose, and time since

last dose were also found to be statistically significant. The covariates of interest found to be statistically significant for predicting the log-transformed rosuvastatin plasma level included the polymorphisms *SLCO1B1* c.521T>C, and *ABCG2* c.421C>A. As in the atorvastatin model, the variables age, dose, and time since last dose were also found to be statistically significant. The prediction tool based off of these regression models[23] was designed to keep the resultant statin plasma level below the 90th percentile (as it is expected that 1 in 10 individuals will experience statin-associated myopathy).

1.4.1 Patient Population

Patient-recruitment and data collection for the project were conducted by DeGorter et al. at the University of Western Ontario . A total of 299 adult dyslipidemic patients (over 18 years of age) taking a stable dose of atorvastatin or rosuvastatin were recruited and followed prospectively at the LHSC between August 2009 and May 2011[23]. Of these, 9 patients (3 on rosuvastatin and 6 on atorvastatin) had statin plasma levels that were undetectable, and so were excluded from analysis[23]. Inclusion criteria were that the study subject must have taken their medication within 24 hours of their last clinic visit and availability of a blood sample, and must not have been on an alternate-day dosing regime for their statin medications. After providing written informed consent following approval by the Research Ethics Board of Western University, the following information was collected: detailed medical history, time of last oral statin dose, and self-reported ethnicity. Information on plasma statin concentration, LDL-C response, determination of total cholesterol, and genotyping were subsequently obtained through analysis of a blood sample provided by each patient. For full details on recruitment, sample size, and

number of patients excluded, refer to DeGorter et al.[23].

Table 1.1: Medications potentially impacting statin pharmacokinetics (A-A)

Drug/Class	Type	Hypothesized DDI Mechanism	Source
Amiodarone	Antiarrhythmic	CYP3A4 Inhibitor	35;36;29;37;15;21
Angiotensin II Receptor Blocker			
Candesartan	Antihypertensive	ABCG2 Inhibitor	38
Fimasartan			39
Losartan		OATP1B1 Inhibitor	38
Telmisartan			40;41
Olmesartan		OATP1B3 Inhibitor	42
Valsartan			43
Antibiotics			
Azithromycin	Antibiotic	CYP3A4 Inhibitor	36;21
Ciprofloxacin			36
Clarithromycin		ABCB1 Inhibitor	35;12;36;11;33;41
Erythromycin		OATP1B1 Inhibitor	35;12;36;38;33;41
Telithromycin			36;11;33
Troleandomycin			36
Anticoagulants	Anticoagulant	CYP3A4 Inhibitor	
Acenocoumarol		CYP2C9 Inhibitor	44
Warfarin		OATP16A1 Inhibitor	35;12;36;45;21
Antiepileptics	Anticonvulsant	CYP3A4 Inducer	
Carbamazepine			35
Oxycarbazepine			44
Antidepressants	Antidepressant	CYP3A4 Inhibitor	
Fluvoxamine			35;44
Fluoxetine			35
Sertraline			35;44
Venlafaxine			35
Antidiabetics	Antidiabetic	CYP3A Inducer	
Troglitazone		CYP2C9 Inducer	46;35;47
Pioglitazone			38
Azole Antifungals	Antifungal	CYP3A4 Inhibitor	29
Fluconazole		CYP2C9 Inhibitor	35;12;36;11;21
Itraconazole			35;12;36;11;38;33
Ketoconazole			35;12;36;38;33
Posaconazole			11

Table 1.2: Medications potentially impacting statin pharmacokinetics (B-E)

Drug/Class	Type	Hypothesized DDI Mechanism	Source
Barbituates	Anticonvulsant	CYP3A4 Inducer	
Phenobarbital		CYP2C9 Inducer	[35]
Bile Acids			
Glycodeoxycholate	Bile Acid	OATP1B1 Inhibitor	[34]
Glycochenodeoxycholate			
Taurolithocholate			
Calcium Channel Antagonists	Antihypertensive	CYP3A4 Inhibitor	[48] [29]
Amlodipine			[38;41]
Azelnidipine			[36]
Benidipine			[36]
Diltiazem			[12;35;36;33;37;21]
Mibefradil			[35;36;33]
Nicardipine			[38]
Nifidipine			[36]
Verapamil			[12;35;36;38;15;33]
Clopidigrel	Anti-platelet	CYP3A4 Inhibitor	[49]
		CYP2C19 Inhibitor	[36]
Colchicine	Anti-gout	CYP3A4 Inhibitor	[21]
Cyclophosphamide	Antineoplastic	CYP3A4 Inducer	[35]
Cyclosporine	Immunosuppressant	CYP3A4 Inhibitor	
		OATP1B1 Inhibitor	[12;15;33]
		OATP2B1 Inhibitor	[36;21;35;29]
		OATP1B3 Inhibitor	[11;38;33;34]
		NTCP Inhibitor	
Corticosteroids	Corticosteroid	CYP3A4 Inducer	[35;50;47]
Dexamethasone		ABCB1 Inducer	[50]
Danazol		CYP3A4 Inhibitor	[33;51]
Digoxin	Antiarrhythmic	ABCB1 Inhibitor	[35;36;43;52]
Erlotinib	Antineoplastic	CYP3A4 Inhibitor	[36]
Estrone 3-Sulfate	Hormone	OATP1B1 Inhibitor	[34]

Table 1.3: Medications potentially impacting statin pharmacokinetics (F-N)

Drug/Class	Type	Hypothesized DDI Mechanism	Source
Fibrates	Lipid-Lowering Agent		
Bexafibrate		OATP1B1 Inhibitor	32
Clofibrate		CYP2C9 Inhibitor	12
Fenofibrate		CYP2C8 Inhibitor	12
Gemfibrozil			12;36;15;29;19
			12;36;15;20;29
		Glucoronidation Inhibitor	11;21
Fexofenadine	Antihistamine	OATP1B1 Inhibitor	43
Flavenoids			
Baicalin	Flavenoid	OATP1B1 Inducer	53
Silymarin		OATP1B1 Inhibitor	34
Grapefruit /Citrus juice	Nutrient	CYP3A4 Inhibitor	35;36;5;29
		OATP1B1 Inhibitor	21;12;38
Histamine H_2 Receptor Antagonists	Antacid	CYP2C9 Inhibitor	
Cimetidine			12
Ranitidine			12
HIV Protease Inhibitors	Antiviral		
Amprenavir		OATP1B1 Inhibitor	35
Atazanavir		ABCG2 Inhibitor	21
Indinavir		CYP3A4 Inhibitor	11;34
Lopinavir			12;41;21
Nelfinavir			36;11;15;54
Ritonavir			12;36;41;21
Saquinavir			12;36;11;15;54;33
Tipranavir			36;21
			36
Metformin	Antidiabetic	ABCC2 Inhibitor	55
Midazolam	Sedative	CYP3A4 Inhibitor	35
Nefazodone	Antidepressant	CYP3A4 Inhibitor	36;35;12;33;29;21
		CYP2C9 Inhibitor	

Table 1.4: Medications potentially impacting statin pharmacokinetics (P-Z)

Drug/Class	Type	Hypothesized DDI Mechanism	Source
Phenytoin	Anticonvulsant	CYP3A4 Inducer CYP2C9 Inducer	35;36;44
Proton Pump Inhibitor Esomeprazole Omeprazole	Antacid	 ABCB1 Inhibitor CYP3A4 Inducer	 56 12;35;38;56
Repaglinide	Antidiabetic	OATP1B1 Inhibitor	43;57
Rifamycins Rifampicin Rifamycin SV	Antibiotic	CYP3A4 Inducer General CYP Inducer OATP1B1 Inhibitor	 33;58;35;36;59 34
Sitagliptin	Antidiabetic	CYP3A4 Inhibitor	60;61
Sirolimus	Immuno-suppressant	CYP3A4 Inhibitor	44
Sulfonamides Sulfaphenazole Sulphamethoxazole	Antibiotic	CYP2C9 Inhibitor	 35 44
Tacrolimus	Immunosuppressant	CYP3A4 Inihibitor	35 36
Tamoxifen	Antineoplastic	CYP3A4 Inhibitor	35
Vitamin B3 (Niacin)	Nutrient	OATP16A1 Inhibitor	12;29;62;5
Vitamin D	Nutrient	CYP Inducer	62;63
Warfarin	Anti-coagulant	CYP2C9 Substrate CYP 3A4 Substrate OATP16A1 Inhibitor	35;12;36;45;21
Zafirlukast	Anti-asthmatic	CYP3A4 Inhibitor	44

References

[1] Philip A Rea. Statins: From fungus to pharma: The curiosity of biochemists, mixed with some obvious economic incentives, created a family of powerful cardiovascular drugs. American Scientist, 96(5):408–415, 2008.

[2] Jonathan A Tobert. Lovastatin and beyond: the history of the HMG-CoA reductase inhibitors. Nature reviews Drug discovery, 2(7):517–526, 2003.

[3] Akira Endo. The discovery and development of HMG-CoA reductase inhibitors. Journal of Lipid Research, 33(11):1569–1582, 1992.

[4] Peter Jones, Stephanie Kafonek, Irene Laurora, Donald Hunninghake, et al. Comparative dose efficacy study of atorvastatin versus simvastatin, pravastatin, lovastatin, and fluvastatin in patients with hypercholesterolemia (the CURVES study). The American Journal of Cardiology, 81(5):582–587, 1998.

[5] Russell A Wilke, Jason H Moore, and James K Burmester. Relative impact of CYP3A genotype and concomitant medication on the severity of atorvastatin-induced muscle damage. Pharmacogenetics and Genomics, 15(6):415–421, 2005.

[6] Fergus McTaggart, Linda Buckett, Robert Davidson, Geoffry Holdgate, Alex McCormick, Dennis Schneck, Graham Smith, and Michael Warwick. Preclinical and clinical pharmacology of rosuvastatin, a new 3-hydroxy-3-methylglutaryl coenzyme A reductase inhibitor. The American Journal of Cardiology, 87(5):28–32, 2001.

[7] Peter H Jones, Michael H Davidson, Evan A Stein, Harold E Bays, James M McKen-

ney, Elinor Miller, Valerie A Cain, and James W Blasetto. Comparison of the efficacy and safety of rosuvastatin versus atorvastatin, simvastatin, and pravastatin across doses (STELLAR* Trial). The American Journal of Cardiology, 92(2):152–160, 2003.

[8] Anders G Olsson, Helge Istad, Olavi Luurila, Leiv Ose, Steen Stender, Jaakko Tuomilehto, Olov Wiklund, Harry Southworth, John Pears, JW Wilpshaar, et al. Effects of rosuvastatin and atorvastatin compared over 52 weeks of treatment in patients with hypercholesterolemia. American Heart Journal, 144(6):1044–1051, 2002.

[9] Dennis W Schneck, Robert H Knopp, Christie M Ballantyne, Ruth McPherson, Rohini R Chitra, and Steven G Simonson. Comparative effects of rosuvastatin and atorvastatin across their dose ranges in patients with hypercholesterolemia and without active arterial disease. The American Journal of Cardiology, 91(1):33–41, 2003.

[10] John Wlodarczyk, David Sullivan, and Michael Smith. Comparison of benefits and risks of rosuvastatin versus atorvastatin from a meta-analysis of head-to-head randomized controlled trials. The American Journal of Cardiology, 102(12):1654–1662, 2008.

[11] R Elsby, C Hilgendorf, and K Fenner. Understanding the critical disposition pathways of statins to assess drug–drug interaction risk during drug development: it's not just about OATP1B1. Clinical Pharmacology & Therapeutics, 92(5):584–598, 2012.

[12] Yiannis S Chatzizisis, Konstantinos C Koskinas, Gesthimani Misirli, Chris Vaklavas, Apostolos Hatzitolios, and George D Giannoglou. Risk factors and drug interactions predisposing to statin-induced myopathy. Drug Safety, 33(3):171–187, 2010.

[13] Ana L Huerta-Alardín, Joseph Varon, and Paul E Marik. Bench-to-bedside review: Rhabdomyolysis–an overview for clinicians. Critical Care, 9(2):158–169, 2004.

[14] David J Graham, Judy A Staffa, Deborah Shatin, Susan E Andrade, Stephanie D Schech, Lois La Grenade, Jerry H Gurwitz, K Arnold Chan, Michael J Goodman, and Richard Platt. Incidence of hospitalized rhabdomyolysis in patients treated with lipid-lowering drugs. JAMA, 292(21):2585–2590, 2004.

[15] Grant T. Generaux, Fiorenza M. Bonomo, Marta Johnson, and Kelly M. Mahar Doan. Impact of SLCO1B1 (OATP1B1) and ABCG2 (BCRP) genetic polymorphisms and inhibition on LDL-C lowering and myopathy of statins. Xenobiotica, 41(8):639 – 651, 2011.

[16] Andrew L Mammen. Statin-associated autoimmune myopathy. New England Journal of Medicine, 374(7):664–669, 2016.

[17] Ana Alfirevic, Dermot Neely, Jane Armitage, Hector Chinoy, Robert G Cooper, Reijo Laaksonen, Daniel F Carr, Katarzyna M Bloch, Joe Fahy, Anita Hanson, et al. Phenotype standardization for statin-induced myotoxicity. Clinical Pharmacology & Therapeutics, 96(4):470–476, 2014.

[18] M Needham and FL Mastaglia. Statin myotoxicity: a review of genetic susceptibility factors. Neuromuscular Disorders, 24(1):4–15, 2014.

[19] Michael J Knauer, Bradley L Urquhart, Henriette E Meyer zu Schwabedissen, Ute I Schwarz, Christopher J Lemke, Brenda F Leake, Richard B Kim, and Rommel G Tirona.

Human skeletal muscle drug transporters determine local exposure and toxicity of statins. Circulation Research, 106(2):297–306, 2010.

[20] Terry A Jacobson. Toward pain-free statin prescribing: clinical algorithm for diagnosis and management of myalgia. In Mayo Clinic Proceedings, volume 83, pages 687–700. Elsevier, 2008.

[21] Loukianos S Rallidis, Katerina Fountoulaki, and Maria Anastasiou-Nana. Managing the underestimated risk of statin-associated myopathy. International Journal of Cardiology, 159(3):169–176, 2012.

[22] Y Tomita, K Maeda, and Y Sugiyama. Ethnic variability in the plasma exposures of OATP1B1 substrates such as HMG-CoA reductase inhibitors: A kinetic consideration of its mechanism. Clinical Pharmacology & Therapeutics, 94(1):37–51, 2013.

[23] Marianne K DeGorter. Statin Transport by Hepatic Organic Anion-Transporting Polypeptides (OATPs). PhD thesis, The University of Western Ontario, 2012.

[24] Megan Roth, Amanda Obaidat, and Bruno Hagenbuch. OATPs, OATs and OCTs: the organic anion and cation transporters of the SLCO and SLC22A gene superfamilies. British Journal of Pharmacology, 165(5):1260–1287, 2012.

[25] JE Keskitalo, O Zolk, MF Fromm, KJ Kurkinen, PJ Neuvonen, and M Niemi. ABCG2 polymorphism markedly affects the pharmacokinetics of atorvastatin and rosuvastatin. Clinical Pharmacology & Therapeutics, 86(2):197–203, 2009.

[26] YY Lau, Y Huang, L Frassetto, and LZ Benet. Effect of OATP1B trans-

porter inhibition on the pharmacokinetics of atorvastatin in healthy volunteers. Clinical Pharmacology & Therapeutics, 81(2):194–204, 2007.

[27] Hannu Päivä, Karin M Thelen, Rudy Van Coster, Joél Smet, Boel De Paepe, Kari M Mattila, Juha Laakso, Terho Lehtimäki, Klaus Bergmann, Dieter Lütjohann, et al. High-dose statins and skeletal muscle metabolism in humans: A randomized, controlled trial. Clinical Pharmacology & Therapeutics, 78(1):60–68, 2005.

[28] Reijo Laaksonen, Mikko Katajamaa, Hannu Päivä, Marko Sysi-Aho, Lilli Saarinen, Päivi Junni, Dieter Lütjohann, Joél Smet, Rudy Van Coster, Tuulikki Seppänen-Laakso, et al. A systems biology strategy reveals biological pathways and plasma biomarker candidates for potentially toxic statin-induced changes in muscle. PloS One, 1(1):e97 (1–9), 2006.

[29] Sivakumar Sathasivam. Statin induced myotoxicity. European Journal of Internal Medicine, 23(4):317–324, 2012.

[30] David Williams and John Feely. Pharmacokinetic-pharmacodynamic drug interactions with HMG-CoA reductase inhibitors. Clinical Pharmacokinetics, 41(5):343–370, 2002.

[31] Susan Tofte. Cyclosporine. Journal of the Dermatology Nurses' Association, 3(3):161–162, 2011.

[32] David B Miller and J David Spence. Clinical pharmacokinetics of fibric acid derivatives (fibrates). Clinical Pharmacokinetics, 34(2):155–162, 1998.

[33] Pertti J Neuvonen, Mikko Niemi, and Janne T Backman. Drug interactions with lipid-lowering drugs: mechanisms and clinical relevance. Clinical Pharmacology & Therapeutics, 80(6):565–581, 2006.

[34] Maria Karlgren, Gustav Ahlin, Christel AS Bergström, Richard Svensson, Johan Palm, and Per Artursson. In vitro and in silico strategies to identify oatp1b1 inhibitors and predict clinical drug–drug interactions. Pharmaceutical Research, 29(2):411–426, 2012.

[35] Debasish Maji, Shehla Shaikh, Dharmesh Solanki, and Kumar Gaurav. Safety of statins. Indian Journal of Endocrinology and Metabolism, 17(4):636–646, 2013.

[36] Anida Čaušević Ramosevac and Sabina Semiz. Drug interactions with statins. Acta Pharmaceutica, 63(3):277–293, 2013.

[37] Michele Thai, Emily Reeve, Sarah N Hilmer, Katie Qi, Sallie-Anne Pearson, and Danijela Gnjidic. Prevalence of statin-drug interactions in older people: a systematic review. European Journal of Clinical Pharmacology, 72(5):513–521, 2016.

[38] Hideki Fujino, Tsuyoshi Saito, Yoshihiko Tsunenari, and Junji Kojima. Interaction between several medicines and statins. Arzneimittelforschung, 53(03):145–153, 2003.

[39] Kwang-Hee Shin, Tae-Eun Kim, Sung Eun Kim, Min Goo Lee, Im-Sook Song, Seo Hyun Yoon, Joo-Youn Cho, In-Jin Jang, Sang-Goo Shin, and Kyung-Sang Yu. The effect of the newly developed angiotensin receptor II antagonist fimasartan on the pharmacokinetics of atorvastatin in relation to OATP1B1 in healthy male volunteers. Journal of Cardiovascular Pharmacology, 58(5):492–499, 2011.

[40] Miao Hu, Hon-Kit Lee, Kenneth KW To, Benny SP Fok, Siu-Kwan Wo, Chung-Shun Ho, Chun-Kwok Wong, Zhong Zuo, Thomas YK Chan, Juliana CN Chan, et al. Telmisartan increases systemic exposure to rosuvastatin after single and multiple doses, and

in vitro studies show telmisartan inhibits ABCG2-mediated transport of rosuvastatin. European Journal of Clinical Pharmacology, 72(12):1471–1478, 2016.

[41] Mijeong Son, Jinju Guk, Yukyung Kim, Dong Woo Chae, Young-A Heo, Dongjun Soh, and Kyungsoo Park. Pharmacokinetic interaction between rosuvastatin, telmisartan, and amlodipine in healthy male korean subjects: A randomized, open-label, multiple-dose, 2-period crossover study. Clinical Therapeutics, 38(8):1845–1857, 2016.

[42] Hyerang Roh, Hankil Son, Donghwan Lee, HeeChul Chang, Chohee Yun, and Kyungsoo Park. Pharmacokinetic interaction between rosuvastatin and olmesartan: a randomized, open-label, 3-period, multiple-dose crossover study in healthy Korean male subjects. Clinical Therapeutics, 36(8):1159–1170, 2014.

[43] Kenneth Kellick. Organic ion transporters and statin drug interactions. Current Atherosclerosis Reports, 19(65), 2017.

[44] Maria Zhelyazkova-Savova, Silvia Gancheva, and Vera Sirakova. Potential statin-drug interactions: prevalence and clinical significance. Springerplus, 3(168):1–8, 2014.

[45] Abdul Naveed Shaik, Tonika Bohnert, David A Williams, Lawrence L Gan, and Barbara W LeDuc. Mechanism of drug-drug interactions between warfarin and statins. Journal of Pharmaceutical Sciences, 105(6):1976–1986, 2016.

[46] Alawi A Alsheikh-Ali, Heather M Abourjaily, and Richard H Karas. Risk of adverse events with concomitant use of atorvastatin or simvastatin and glucose-lowering drugs (thiazolidinediones, metformin, sulfonylurea, insulin, and acarbose). The American Journal of Cardiology, 89(11):1308–1310, 2002.

[47] Vinod Ramachandran, Vsevolod E Kostrubsky, Bernard J Komoroski, Shimin Zhang, Kenneth Dorko, James E Esplen, Stephen C Strom, and Raman Venkataramanan. Troglitazone increases cytochrome P-450 3A protein and activity in primary cultures of human hepatocytes. Drug Metabolism and Disposition, 27(10):1194–1199, 1999.

[48] V Garaliene, V Barsys, S Giedraitis, R Benetis, and A Krauze. The role of external CA2+ in the action of CA2+-channel agonists and antagonists on isolated human thoracic arteries. Journal of Physiology and Pharmacology, 65(1):25–31, 2014.

[49] Luiz F Pinheiro, Carolina N França, Maria C Izar, Simone P Barbosa, Henrique T Bianco, Soraia H Kasmas, Gustavo D Mendes, Rui M Povoa, and Francisco AH Fonseca. Pharmacokinetic interactions between clopidogrel and rosuvastatin: effects on vascular protection in subjects with coronary heart disease. International Journal of Cardiology, 158(1):125–129, 2012.

[50] Stefanie Lam, Nilufar Partovi, Lillian SL Ting, and Mary HH Ensom. Corticosteroid interactions with cyclosporine, tacrolimus, mycophenolate, and sirolimus: fact or fiction? Annals of Pharmacotherapy, 42(7-8):1037–1047, 2008.

[51] Ivan Stankovic, Alja Vlahovic-Stipac, Biljana Putnikovic, Zorica Cvetkovic, and Aleksandar N Neskovic. Concomitant administration of simvastatin and danazol associated with fatal rhabdomyolysis. Clinical Therapeutics, 32(5):909–914, 2010.

[52] Rebecca A Boyd, Ralph H Stern, Barhra H Stewart, Xiaochun Wu, Eric L Reyner, Elizabeth A Zegarac, Edward J Randinitis, and Lloyd Whitfield. Atorvastatin coadministration

may increase digoxin concentrations by inhibition of intestinal P-glycoprotein-mediated secretion. The Journal of Clinical Pharmacology, 40(1):91–98, 2000.

[53] L Fan, W Zhang, D Guo, Z-R Tan, P Xu, Q Li, Y-Z Liu, L Zhang, T-Y He, D-L Hu, D Wang, and H-H Zhou. The effect of herbal medicine baicalin on pharmacokinetics of rosuvastatin, substrate of organic anion-transporting polypeptide 1b1. Clinical Pharmacology & Therapeutics, 83(3):471–476, 2008.

[54] Jennifer J Kiser, John G Gerber, Julie A Predhomme, Pamela Wolfe, Devon M Flynn, and Dorie W Hoody. Drug/drug interaction between lopinavir/ritonavir and rosuvastatin in healthy volunteers. JAIDS Journal of Acquired Immune Deficiency Syndromes, 47(5):570–578, 2008.

[55] Eunjung Shin, Naree Shin, Ju-Hee Oh, and Young-Joo Lee. High-dose metformin may increase the concentration of atorvastatin in the liver by inhibition of Multidrug Resistance–Associated Protein 2. Journal of Pharmaceutical Sciences, 106(4):961–967, 2017.

[56] Brooke E Sipe, Ronald J Jones, and Gordon H Bokhart. Rhabdomyolysis causing AV blockade due to possible atorvastatin, esomeprazole, and clarithromycin interaction. Annals of Pharmacotherapy, 37(6):808–811, 2003.

[57] A Kalliokoski and M Niemi. Impact of OATP transporters on pharmacokinetics. British Journal of Pharmacology, 158(3):693–705, 2009.

[58] Mikko Niemi, Janne T Backman, Martin F Fromm, Pertti J Neuvonen, and Kari T Kivistö. Pharmacokinetic interactions with rifampicin. Clinical Pharmacokinetics, 42(9):819–850, 2003.

[59] Kazuya Maeda. Organic anion transporting polypeptide (OATP) 1B1 and OATP1B3 as important regulators of the pharmacokinetics of substrate drugs. Biological and Pharmaceutical Bulletin, 38(2):155–168, 2015.

[60] R Bhome and H Penn. Rhabdomyolysis precipitated by a sitagliptin–atorvastatin drug interaction. Diabetic Medicine, 29(5):693–694, 2012.

[61] Robert V DiGregorio and Yanina Pasikhova. Rhabdomyolysis caused by a potential sitagliptin-lovastatin interaction. Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy, 29(3):352–356, 2009.

[62] Huifen Wang, Jeffrey B Blumberg, C-Y Oliver Chen, Sang-Woon Choi, Michael P Corcoran, Susan S Harris, Paul F Jacques, Aleksandra S Kristo, Chao-Qiang Lai, Stefania Lamon-Fava, et al. Dietary modulators of statin efficacy in cardiovascular disease and cognition. Molecular Aspects of Medicine, 38:1–53, 2014.

[63] JB Schwartz. Effects of vitamin D supplementation in atorvastatin-treated patients: A new drug interaction with an unexpected consequence. Clinical Pharmacology & Therapeutics, 85(2):198–203, 2009.

Chapter 2

Rationale and Objectives

To best predict systemic exposure of atorvastatin and rosuvastatin, multiple linear regression models were used that control for a variety of patient factors including genetic polymorphisms in statin transporters and enzymes involved in their metabolism[1]. Although studies using the separate regression models for atorvastatin and rosuvastatin explained a large portion of the model variance (47% and 56% respectively)[1], there is still room for improving the fit of these models to better predict systemic exposure. We hypothesize that the systemic exposure models could be improved by both incorporating additional patient factors, and using non-linear modelling techniques to achieve a closer fit to the data.

The predictive model currently used to predict systemic exposure of statins does not include concomitant medications[2]. Given the wide range of comorbidities that are often present with hypercholesterolemia, most patients in this population will be prescribed concomitant medications that may interact with drug transporters or drug metabolizing enzymes and therefore affect drug exposure. Because of the concomitant medications that patients with hypercholesterolemia will be prescribed, there is an increased likelihood of drug interactions within this

population that could be used to predict systemic statin exposure. From a modelling perspective, the wide range of medicines that could be co-prescribed with statins poses an additional challenge of potentially having far more medications to account for than patient observations when training and validating models to predict statin plasma concentration. Importantly, atorvastatin has the potential to interact with many different commonly prescribed medications because of its pharmacokinetic properties, so creating a well-performing predictive model that incorporates information about concomitant medication use has the potential to improve the current systemic exposure model with potential clinical implications. In order to ameliorate the problem of having many potential relevant medications to include in a predictive model versus a relatively small sample size, feature selection algorithms can be adapted to take advantage of substantive subject knowledge in order to create robust selection criteria despite relatively few observations with which to train models.

In contrast, rosuvastatin undergoes very little metabolism before performing its intended action in the liver, and so metabolic drug interactions are of less concern for this modelling application[3]. Instead, thoroughly assessing genetic variation in patient drug transporters may be more relevant to the problem of predicting rosuvastatin plasma concentration in a clinical setting. It is possible that identifying additional genetic markers through the use of Next Generation Sequencing (NGS) of DNA from patients who have rosuvastatin plasma concentrations substantially higher than predicted by the original regression model (under-predicted patients) could account for additional variance in the rosuvastatin cohort, as it is less likely that including concomitant medications will cause substantial improvement in model fit for this group.

2.1 Research Objectives

The overarching goal of this thesis is to improve the predictive quality of previously developed regression models in the context of predicting statin systemic exposure. Improving predictive quality will be achieved by better accounting for model variance such that the predicted systemic exposure values are closer to the true systemic exposure values for each patient. This will be accomplished by two strategies: 1) identifying additional clinical factors that improve the predictive model fit quality; and 2) assessing modelling techniques beyond linear regression that can incorporate non-linear trends in the data. The two strategies can be further broken down into three main objectives:

2.1.1 Objective 1

- Adapt existing feature selection techniques for selecting concomitant medications relevant to the problem predicting atorvastatin and rosuvastatin plasma concentration; the new algorithm must be suitable for use with small datasets in which the number of features is much larger than the number of patient observations

- Assess the impact of the selected concomitant medications on the linear model fit for the atorvastatin and rosuvastatin patient cohorts, in comparison to the fit of the original systemic statin exposure model using the patient cohorts used to fit the original systemic exposure model

2.1.2 Objective 2

- Select and implement appropriate non-linear modelling techniques to explore the effect of incorporating non-linearity into the atorvastatin and rosuvastatin predictive models for statin systemic exposure
- Characterize the practical requirements and feasibility of achieving good model fit using these methods for guiding atorvastatin and rosuvastatin dosing in the context of predicting systemic exposure in order to minimize adverse drug events

2.1.3 Objective 3

- Develop selection criteria to identify rosuvastatin patients with severely under-predicted statin plasma concentrations based on the original systemic exposure model, as well as well-predicted controls; gather and process next generation sequencing data from the identified case and control groups
- Describe the process and analytic/software requirements for cleaning and formatting next generation sequencing information from a data-science perspective
- Identify and apply an appropriate variant prioritization method to select relevant genes for further biological analysis

This dissertation is presented in integrated article format, with separate background chapters for each substantive clinical research topic. There are a total of seven chapters. Chapter 1 gave an overview of the use of statins for hypercholesterolemia, potential adverse events associated with statin treatment, and the mechanisms by which statins are metabolized and

transported to the liver. A narrative review of the literature on concomitant medication use with statins is also presented. Chapter 2 outlines the main research objectives and organization of this thesis. Chapter 3 provides background on linear regression modelling and feature selection techniques; the concomitant medication selection algorithm is described, as well as results from the selection and linear modelling of concomitant medications for atorvastatin and rosuvastatin. Chapter 4 provides background information on two popular statistical techniques that are capable of modelling non-linear relationships in data. The implementation and results of these methods are described and compared to the fit of the original systemic exposure prediction model. Emphasis is placed on the practical aspects of implementing these techniques for clinical use, particularly with respect to tuning the models and choosing the best possible model parameters for each method. Chapter 5 is a background chapter giving an overview of the process of Next Generation Sequencing (NGS), methods for effective phenotype-based patient sampling, and available techniques for variant selection and prioritization. Chapter 6 characterizes the selection criteria for choosing patients to undergo genetic sequencing, as well as characteristics of the data obtained from this process. It also contains the workflow for processing the NGS data for this clinical modelling problem, as well as genes identified as relevant by the analysis technique chosen. Chapter 7 provides a summary of the key findings, strengths and limitations of this thesis. Implications of the findings described herein are also discussed, as well as future directions for this research.

References

[1] Marianne K DeGorter, Rommel G Tirona, Ute I Schwarz, Yun-Hee Choi, George K Dresser, Neville Suskin, Kathryn Myers, GuangYong Zou, Otito Iwuchukwu, Wei-Qi Wei, et al. Clinical and pharmacogenetic predictors of circulating atorvastatin and rosuvastatin concentration in routine clinical care. Circulation: Cardiovascular Genetics, 6(4):400–408, 2013.

[2] Marianne K DeGorter. Statin Transport by Hepatic Organic Anion-Transporting Polypeptides (OATPs). PhD thesis, The University of Western Ontario, 2012.

[3] Satoshi Kitamura, Kazuya Maeda, Yi Wang, and Yuichi Sugiyama. Involvement of multiple transporters in the hepatobiliary transport of rosuvastatin. Drug Metabolism and Disposition, 36(10):2014–2023, 2008.

Chapter 3

Modelling Atorvastatin and Rosuvastatin Plasma Concentration Using Selected Concomitant Medications

3.1 Introduction

Increased systemic exposure to statin drugs is thought to be a risk factor for the development of myalgia (muscle pain) in patients with hypercholesterolemia participating in lipid-lowering therapy[1].

Concomitant medications are an important consideration when modelling systemic exposure of statins as these interactions have the potential to increase systemic exposure. A common side-effect of statin treatment for hypercholesterolemia is myalgia; this is thought to be related to the concentration of statins in blood plasma, which can be affected by the concomitant ingestion of many different substances. Because hypercholesterolemia tends to cluster with other

chronic illnesses and medications are often prescribed for these comorbid conditions, Drug-Drug Interactions (DDIs) are likely in the context of statin treatment for hypercholesterolemia.

Advances have been made in previous work to establish an accurate model to predict statin plasma concentration[2]; however, concomitant medications were not taken into account in these models. Incorporating concomitant medications which have the potential to impact systemic statin exposure has a high potential to increase accuracy in this context. Establishing the most accurate model possible for predicting systemic exposure is an important problem, since it could in future be used to guide statin dosing with the goal of minimizing adverse events such as myalgia.

3.2 Methods: Linear Regression and Feature Selection

3.2.1 Linear Regression

In linear regression we seek to use the values of one or more independent (predictor) variables to predict the value of an outcome variable. For example, we could use a patient's age, sex and BMI to try to predict heart rate. We could then construct a model from an existing data set, and then use this model to predict heart rate values for future patients. This is often done via Ordinary Least Squares (OLS), which weights each variable according to its relation to the outcome variable while minimizing the resulting magnitude of the error of the predicted values compared to the observed outcome values. This is also useful because it allows us to make inferences about the relationship between the independent (predictor) variables and the dependent (outcome) variable. For an outcome variable $\boldsymbol{y}$ and a matrix of predictor variables

X, our predictions would take the form: $\hat{y} = X\hat{\boldsymbol{\beta}}$. For this simple example, the residuals (error) that we are trying to minimize would take the form $(y - X\beta)^2$. OLS computes weights β for each predictor variable as follows:

$$\hat{y} = \underset{\beta}{\operatorname{argmin}} \|\boldsymbol{y} - X\beta\|_2^2.$$

In our specific example predicting heart rate (HR) using multiple covariates, the predictions would take the following form:

$$\hat{\texttt{HR}} = \beta_0 + \beta_1 \times \texttt{age} + \beta_2 \times \texttt{sex} + \beta_3 \times \texttt{BMI}.$$

3.2.2 Variable Selection

Often in clinical modelling applications, we have the choice of including many different types of information in regression models and predictive models in general. Some examples of types of information that could be used in clinical modelling include demographic data such as age, sex, ethnicity, and marital status; clinical measurements like blood pressure, heart rate, respiration; laboratory results like plasma concentrations of drugs or endogenous biomarkers of organ function; waveform and longitudinal signal data from instruments monitoring the patient; and genetic information. It is sometimes tempting to include as much information as possible in a model to give the best predictions, but this can lead to several pitfalls: including unnecessary covariates in models can add noise which decreases the accuracy of the predictive model, it can be difficult to interpret models with a very large number of covariates[3], and unique modelling challenges arise when there are more covariates than data observations in the data set to

be modelled. Additionally, using all possible covariates becomes problematic if they are not routinely available, are hard to collect, or frequency contain missing data. For these reasons, it is often desirable to perform feature selection to prune the possible model covariates to only include those most relevant to the prediction problem at hand. A commonly seen example of this occurs in clinical pharmacogenetics where thousands of candidate single nucleotide polymorphisms (SNPs) may be tested for associations with differences in drug metabolism between patients. If we want to include information about a patient's ability to metabolize a particular drug based on their genetic information, we want to include at most a few different SNPs, and not the whole array of candidate genes. Many feature selection methods exist for the purpose of cutting out variables that don't add predictive value to models by selecting a subset of the covariates to include in the model. Some of these include best-subset selection, forward- and backward- stepwise selection, forward-stagewise selection[3], penalized regression, the lasso and group lasso, and more generally, composite absolutely penalties (CAP) used in regression[4].

Best-Subset Selection

Best-subset selection chooses a subset of size $k \in \{1, ..., p\}$ from the p available covariates by exploring all possible subsets of this size and picking the one with the smallest residual error. The choice of the size of the subset size k must be determined by the user; this is usually done to balance bias and variance while maintaining the desired sparsity of the model. Additionally, smaller models are not necessarily subsets of larger models in this framework: a subset of size two may not include the same variables found in a subset of size one[3]. This can make interpretation difficult within the context of the model that is chosen. A major downside of best-

subset selection is that it is limited to datasets that only have a moderate number of variables, because it must search through all possible combinations of the covariates to find the best model for each k. Even with an efficient algorithm to speed the selection process, for datasets with more than 30 or 40 variables the computational time required becomes prohibitive[3].

Forward- and Backward-Stepwise Selection

In contrast, forward- and backward-stepwise selection proceed sequentially from either the model with only the intercept and no covariates for forward-stepwise selection, or from the full model including all of the possible covariates for backward-stepwise selection. For forward-stepwise selection, the variable with the most impact on the model fit is added at each step k, where k corresponds to the size of the subset. Similarly, for backward-stepwise selection the variable that contributes the least to the model fit is removed from the pool of candidate variables at each step k[3]. The resulting output of these procedures is a set of nested models indexed by k, the number of covariates included in the model. Although this may seem less thorough than looking through all of the possible models of each size and choosing based on residual error, stepwise selection procedures have the advantage of being applicable to datasets with a very large number of variables. Additionally, because fewer candidate models are being compared in terms of error, model variance is reduced (although bias may be increased)[3]. A variant of this procedure is called Forward-Stagewise selection, and is a constrained version of forward-stepwise selection in the sense that the overall model fit does not change when additional variables are added. At each step in the selection procedure, the model adds the variable to the model that is most correlated with the current model residual; the value given to this variable in the model is found by computing the simple linear regression coefficient for

this model and the current residual and adding that to the variable's current coefficient value. Instead of only computing k model fits, forward-stagewise selection continues until none of the variables are correlated with the residuals. This means that it takes longer to reach the best-fitting model, but the procedure has computational advantages over stepwise selection methods when the number of variables to choose from becomes increasingly large[3].

Penalized Regression and the Lasso

Another method for selecting variables to include in a model is by modelling all of the candidate variables using regression and then shrinking the coefficients using a penalty term. The resulting regression estimates are smaller and some are shrunk to the point where they have a value of 0; those variables are then discarded. This is usually done by including a penalty term in the regression that penalizes the size of the coefficients. These models have the advantage of reducing the variance of the model estimates that can be problematic in ordinary least squares (OLS) models. Excessive variance is especially a problem when the number of variables far exceeds the number of available observations (ie. $p >> n$). The lasso[5] is an example of a penalized regression model capable of variable selection, and is defined as

$$\hat{\beta}_\lambda = \underset{\beta}{\operatorname{argmin}} \left\{ \tfrac{1}{2} \|\boldsymbol{y} - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\}.$$

The ℓ_1 norm is used to penalize the coefficients in the lasso, and minimizes the size of the absolute differences. Looking at the geometry of this norm can help to clarify the effect: instead of travelling along the diagonal of a grid, the ℓ_1 norm only moves in the horizontal and vertical directions. This can be likened to a taxicab navigating a grid of streets instead of

travelling to the destination as the crow flies[6].

The lasso is a popular example of a bridge regression model, in which a general ℓ_γ norm penalty with γ values ranging from 0 to ∞ may be applied to the model coefficients. The general form of the bridge penalty is[7]

$$T(\beta) = \left[\sum_{j=1}^{p} |\beta|^\gamma\right]^{\frac{1}{\gamma}} = ||\beta||_\gamma \quad .$$

The lasso is a special case of bridge regression ($\gamma = 1$), which has attractive computational properties for optimization and the regularization path is piecewise linear[7]. It is also special in that it can be formulated to calculate the full regularization path in a single step via Least Angle Regression (LARS)[8]. LARS was in part inspired by forward-stagewise regression, which is reflected in the algorithm that computes the full regularization path. Similar to forward-stagewise regression, LARS begins with the model including only the intercept. At each step, it chooses the variable that is most correlated with the current residual; however, instead of only moving a short distance in that direction, LARS takes the largest step possible in that direction until it encounters another variable that is just as correlated with the current residual. Once the other variable is encountered, the direction of the solution path changes, and continues in the direction equiangular to both predictors. This method is then repeated and the direction changes at each step to be equiangular with all of the variables seen up until that point. The procedure is complete when all variables are included in the model[8].

Composite Absolute Penalties (CAP) and the Group Lasso

Composite Absolute Penalty (CAP) models are a general type of penalized regression model (of which bridge regression is an example), but importantly allow for hierarchical variable selection. Models of this type allow for prespecified groups of variables (such as sets of dummy variables, or variables that naturally cluster together, like genomic protein expression data[9]) to enter the model simultaneously and can also encourage variables entering the model together to have coefficients with similar absolute values, depending on the penalties used. A common example of a hierarchical CAP model is the group lasso. Given an $n \times p$ design matrix X, an $n \times 1$ binary response variable $\boldsymbol{y}$, and a vector G of group indices of length k, the group lasso is defined as[4]:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ \tfrac{1}{2} ||\boldsymbol{y} - X\beta||_2^2 + \lambda \sum_{k=1}^{K} ||\beta_{G_k}||_2 \right\}.$$

The group lasso uses a hybrid ℓ_1/ℓ_2 penalty; instead of encouraging overall sparsity as in the lasso, groups are encouraged to enter the model sparsely while individual features in the groups included in the model are not penalized from inclusion in the model. In effect, this form of the group lasso performs feature selection at the level of factors, or clusters of features. CAP models generalize this type of feature selection by allowing different types of penalties to be applied across different groups, while maintaining an overall penalty. As described by Zhao et al.[7], the CAP model is constructed by designating G_k groups within the design matrix X with regression targets y and coefficients β, where $k = 1, \ldots, K$ denotes the group index. Generally these groupings are constructed to reflect the natural grouping structure of the covariates. For each group G_k we calculate coefficients β, and then take the norm N_k of the coefficients for this

group, where:

$$N_k = ||\beta_{G_k}||_{\gamma_k}$$

The norms from each group are then aggregated into a K-dimensional vector $N = N_1, \ldots, N_K$, and then the CAP penalty is calculated using the following formula[7]:

$$T(\beta) = ||N||_{\gamma_0}^{\gamma_0} = \sum_k |N_k|^{\gamma_0}$$

and the corresponding CAP estimate is given by:

$$\hat{\beta}(\lambda) = \underset{\beta}{\operatorname{argmin}} \sum_i L(Y_i, X_i, \beta) + \lambda \cdot T(\beta)$$

where L represents the loss function (often squared loss or logistic loss), and T is a CAP penalty.

An advantage of the Lasso and CAP penalties in general for this type of research is that they are primarily meant to be used in the context of predictive modelling, which makes it appropriate for this particular research problem. This is not necessarily true of forward selection, for which model fit is not directly motivated by prediction quality.

3.3 Linear Regression Models for Predicting Systemic Exposure of Statins

The current analyses are based on the original model predicting statin systemic exposure by DeGorter et al. described in Section 1.4. The dataset used for this analysis is limited to a subset

of patients and covariates that were used to train the final linear models in the original prospective cohort study (rosuvastatin $n = 130$; atorvastatin $n = 128$). The outcome of the original dose model is log plasma concentration, as in the original model[2]; looking at log plasma concentration is a relatively common practice in clinical pharmacology. Patient characteristics in the atorvastatin and rosuvastatin cohorts used for analysis in the current work are presented in Tables 3.1 and 3.2 respectively. Patients did not overlap between the two groups.

Table 3.1: Population characteristics of atorvastatin-prescribed prospective cohort ($n = 128$)

Patient Characteristic	Mean/Proportion	SD/Percentage
Age (years)	59.2	13.0
Body Mass Index (kg/m^2)	29.1	5.3
Time Pose Dose (hours)	13.0	5.0
4β-Hydroxycholesterol (ng/mL)	21.1	12.9
No. Concomitant Medications	6.7	3.3
Statin Dose		
10 (mg)	21	16.4%
20 (mg)	30	23.4%
40 (mg)	54	42.2%
60 (mg)	1	0.8%
80 (mg)	22	17.2%
Sex (Male=1)	78	60.9%
Ethnicity (Non-Caucasian=1)	20	15.6%
Minor Allelic Frequency		
OATP1B1 c.521C	29/256	11.3%
OATP1B1 c.388G	113/256	44.1%

3.3.1 Reassessment of the Original Statin Systemic Exposure Model

In order to more easily compare model fit quality between the original systemic exposure model and models that are developed in the current work, we replicated the original models to obtain the standardized coefficient estimates and confidence intervals (shown in Tables 3.3 and 3.4).

For the replicated original linear regressions, five-fold cross-validation (CV) was performed

Table 3.2: Population characteristics of rosuvastatin-prescribed prospective cohort (n = 130)

Patient Characteristic	Mean/Proportion	SD/Percentage
Age (years)	56.9	12.9
Body Mass Index (kg/m^2)	30.4	7.1
Time Pose Dose (hours)	13.6	3.4
No. Concomitant Medications	6.5	3.4
Statin Dose		
5 (mg)	20	15.4%
10 (mg)	38	29.2%
15 (mg)	1	0.8%
20 (mg)	36	27.7%
30 (mg)	2	1.5%
40 (mg)	33	25.4%
Sex (Male=1)	90	69.2%
Ethnicity (Non-Caucasian=1)	21	16.2%
Minor Allelic Frequency		
OATP1B1 c.521C	49/260	18.8%
ABCG2 c.421A	25/260	9.6%

to calculate Adjusted R^2, RMSE and AIC for both atorvastatin and rosuvastatin, in order to more easily compare them to the models subsequently developed in this thesis. The CV results for the original models are shown in Table 3.5 The adjusted R^2 for the atorvastatin model was 0.47 (n = 128); of this, the genetic component comprised only 38% of the explainable variability. The adjusted R^2 for the rosuvastatin model was higher than that of the atorvastatin model: R^2 = 0.56 (n = 130), with the transporter gene polymorphisms accounting for 88% of the explainable variability in this model, in contrast to the much lower value in the atorvastatin model.

Table 3.5: Original linear regression models CV performance results

Model	Adjusted R^2 (± SD)	RMSE (± SD)	AIC (± SD)
Atorvastatin	0.47 ± 0.03	21.06 ± 10.10	246.07 ± 7.88
Rosuvastatin	0.56 ± 0.03	16.13 ± 4.34	198.69 ± 7.87

Because so few unique values of dose were seen within the patient populations for atorvas-

Table 3.3: Atorvastatin regression with original covariates (n=128)

	Estimate	Confidence Interval	P-Value	Sig.
(Intercept)	-0.473	-1.752 to 0.806	0.466	
OATP1B1 c.521T>C	0.339	0.055 to 0.624	0.020	*
OATP1B1 c.388A>G	-0.278	-0.485 to -0.072	0.009	**
Age (yr)	0.018	0.007 to 0.029	0.002	**
4β-Hydroxycholesterol	-0.015	-0.026 to -0.005	0.006	**
Dose (mg)	0.021	0.014 to 0.027	<0.001	***
Time from last dose (hr)	-0.089	-0.117 to -0.062	<0.001	***
Body Mass Index (kg/m^2)	0.025	-0.002 to 0.052	0.065	.
Sex (Male = 1)	0.132	-0.149 to 0.412	0.355	
Ethnicity (Non-Caucasian = 1)	0.295	-0.094 to 0.684	0.136	

Table 3.4: Rosuvastatin regression with original covariates (n=130)

	Estimate	Confidence Interval	P-Value	Sig.
(Intercept)	0.544	-0.286 to 1.373	0.197	
OATP1B1 c.521T>C	0.425	0.233 to 0.618	<0.001	***
ABCG2 c.421C>A	0.301	0.043 to 0.559	0.023	*
Age	0.012	0.004 to 0.02	0.005	**
Dose (mg)	0.048	0.039 to 0.056	<0.001	***
Time Post Dose (hr)	-0.06	-0.092 to -0.029	<0.001	***
Body Mass Index (kg/m^2)	-0.01	-0.025 to 0.005	0.179	
Sex (Male = 1)	-0.159	-0.396 to 0.078	0.186	
Ethnicity (Non-Caucasian = 1)	0.051	-0.241 to 0.342	0.732	

tatin and rosuvastatin, the original regressions were also replicated with dose represented as a categorical variable. In order to perform this analysis a single patient on an outlying dose of atorvastatin (60mg) was removed to facilitate CV performance evaluation (final n =127). No other obvious outlying characteristics besides dose level were present for this patient. When dose was represented categorically instead of as a continuous variable, the Adjusted R^2 in the atorvastatin model increased from 0.474 to 0.488. To test whether this difference was due solely to the lowered variance from removing the patient with the outlying dose, a regression model with dose as continuous was conducted using the reduced atorvastatin dataset (shown in Table 3.6). Note that the models are nested in this case, as the continuous-dose linear model

is nested within the categorical dose model. As such, the linear dose-encoding model could be replicated in the continuous covariate paradigm with the correct coefficients. When the reduced continuous-dose model was compared to the categorical dose model with the same patient population, the change in model fit due to representing dose categorically was statistically significant (Adjusted R^2=0.458 to 0.488, $F(2, 115)$=4.375, p=0.015). Given that the amount of variance accounted for the model was reduced by removing the patient with the outlying dose, it appears likely that the improvement in model fit was caused by the change in representation of dose from continuous to categorical. For this reason, atorvastatin doses will be represented categorically in the models subsequently developed in the current work.

Table 3.6: Atorvastatin regression with dose-outlying patient removed and dose as continuous (n=127)

	Estimate	Confidence Interval	P-Value	Sig.
(Intercept)	-0.485	-1.749 to 0.78	0.449	
OATP1B1 c.521T>C	0.356	0.074 to 0.637	0.014	*
OATP1B1 c.388A>G	-0.287	-0.491 to -0.083	0.006	**
Age (yr)	0.017	0.006 to 0.028	0.003	**
4β-Hydroxycholesterol (ng/mL)	-0.014	-0.025 to -0.004	0.009	**
Dose (mg)	0.02	0.014 to 0.026	<0.001	***
Time from last dose (hr)	-0.084	-0.112 to -0.057	<0.001	***
Body Mass Index (kg/m^2)	0.024	-0.003 to 0.05	0.076	.
Sex (Male = 1)	0.164	-0.115 to 0.444	0.246	
Ethnicity (Non-Caucasian = 1)	0.291	-0.094 to 0.676	0.137	

Table 3.7: Atorvastatin regression with dose-outlying patient removed and dose as categorical (n=127)

	Estimate	Confidence Interval	P-Value	Sig.
(Intercept)	-0.435	-1.661 to 0.79	0.483	
OATP1B1 c.521T>C	0.402	0.126 to 0.678	0.005	**
OATP1B1 c.388A>G	-0.213	-0.417 to -0.008	0.042	*
Age (yr)	0.017	0.006 to 0.028	0.002	**
4β-Hydroxycholesterol (ng/mL)	-0.015	-0.026 to -0.005	0.005	**
Dose (20mg)	0.718	0.259 to 1.178	0.002	**
Dose (40mg)	1.154	0.742 to 1.566	<0.001	***
Dose (80mg)	1.615	1.141 to 2.089	<0.001	***
Time from last dose (hr)	-0.082	-0.109 to -0.055	<0.001	***
Body Mass Index (kg/m^2)	0.015	-0.011 to 0.041	0.263	
Sex (Male = 1)	0.076	-0.202 to 0.355	0.587	
Ethnicity (Non-Caucasian = 1)	0.189	-0.194 to 0.572	0.330	

Similarly, very few unique values were seen between patients with respect to rosuvastatin dose; the most commonly seen doses were 5 mg (n=20; 15.4%), 10 mg (n=38; 29.2%), 20 mg (n=36; 27.7%) and 40 mg (n=33; 25.4%). Only one patient in the dataset used to train the original systemic exposure regression model was on a dose of 15 mg, and only two patients were on a dose of 30 mg. In order to be able to represent dose as a categorical variable instead of a continuous variable (as was done in the original regression model), these three patients were excluded from all future analysis in the current work. No other obvious outlying characteristics besides dose level were present for these patients. When dose was represented as a categorical variable instead of a continuous variable, the amount of explained variance in the model (adjusted R^2) increased from 0.562 to 0.635. To tested whether this change in model fit was due to decreased variance as a result of taking out the three patients with outlying doses, the regression was also performed on the reduced rosuvastatin cohort with the dose represented continuously. The adjusted R^2 of the original model with the reduced cohort was 0.561; this was found to be a statistically significant improvement in model fit when compared to the

categorical-dose regression model with the same patient cohort ($F(2, 116)$=12.662, $p < 0.001$).

Table 3.8: Rosuvastatin linear regression with original covariates excluding dose outliers

	Estimate	Confidence Interval	P-Value	Sig.
(Intercept)	1.069	0.066 to 2.072	0.037	*
OATP1B1 c.521T>C	0.442	0.241 to 0.643	<0.001	***
ABCG2 c.421C>A	0.296	0.036 to 0.557	0.026	*
Age	0.013	0.004 to 0.021	0.004	**
Dose (mg)	0.047	0.039 to 0.056	<0.001	***
Time Post Dose (hr)	-0.059	-0.091 to -0.027	<0.001	***
BMI (kg/m^2)	-0.01	-0.025 to 0.005	0.186	
Sex (Male = 1)	-0.126	-0.369 to 0.116	0.305	
Ethnicity (Non-Caucasian = 1)	0.005	-0.296 to 0.305	0.976	

Table 3.9: Rosuvastatin regression with original covariates excluding dose-outliers and dose as categorical (n=127)

	Estimate	Confidence Interval	P-Value	Sig.
(Intercept)	0.32	-0.454 to 1.093	0.415	
OATP1B1 c.521T>C	0.398	0.213 to 0.583	<0.001	***
ABCG2 c.421C>A	0.341	0.102 to 0.58	0.006	**
Age	0.013	0.005 to 0.021	0.002	**
Dose (10mg)	0.852	0.544 to 1.161	<0.001	***
Dose (20mg)	1.429	1.117 to 1.741	<0.001	***
Dose (40mg)	1.999	1.677 to 2.321	<0.001	***
Time Post Dose (hr)	-0.058	-0.087 to -0.029	<0.001	***
BMI (kg/m^2)	-0.012	-0.026 to 0.002	0.086	.
Sex (Male = 1)	-0.179	-0.402 to 0.043	0.114	
Ethnicity (Non-Caucasian = 1)	-0.069	-0.345 to 0.207	0.62	

3.4 Selection Algorithm

A substantial challenge in choosing concomitant medications for inclusion in a predictive model for statin plasma concentration is the fact that there are many more potential medications to choose from than available patient observations. In the setting when the number of features p is much greater than the number of observations n ((ie. $p >> n$), leveraging informa-

tion shared between covariates can help to improve feature selection and the resulting model fit. In the setting of concomitant medication selection, shared information between individual drugs can be represented as a grouping of generic medication names based on drug function or primary method of action. In order to leverage this information, an ontology mapping generic medications to their corresponding classes of drug function was created by two clinical pharmacologists in the Division of Clinical Pharmacology at Western University; disagreements on generic medication classification were discussed until the experts came to an agreement of the most likely use case based on what the drugs are usually prescribed for in terms of primary and secondary uses. Because medications can have multiple uses or mechanisms of action, the ontology created for this analysis was specific to the context of medication uses most likely to be relevant to comorbidities associated with hypercholesterolemia.

The group lasso is a classic example of this type of modelling method that leverages shared information between covariates in the model (a more detailed overview is provided in Chapter 3.) In the current application, the generic medications to choose from would be labelled as belonging to a medication class as specified in the ontology. However, this method did not give very informative results compared to performing separate analyses for generic medication names, and the generic medications re-coded to correspond to their respective medication classes. At a high level, the group lasso penalizes individual groups from entering the model (having a non-zero coefficient), but once a group - or in this context, class of medications - has come into the model, the individual medications within the medication class are not penalized from entering. Because the group lasso does not encourage sparsity within groups, the output of the model does not necessarily indicate which medications within a particular class are driving the association with statin plasma concentration. Since the mapping between individual

generic medications and medication classes could change depending on the substantive clinical research topic, having a model that chooses generic medications only while still leveraging the information about medication classes specific to the modelling application is more useful for this application than relying on a classification scheme that is not universally applicable. To take this into account, the medication class information was used in the decision criteria for choosing individual generic medications without being included in the final selection output.

A number of standard methods of feature selection were described earlier in this chapter. Unfortunately, many of these methods lack stability when used on datasets with few patient observations, such as many of those seen in clinical pharmacology applications. The advantage of using such feature selection techniques is that many have efficient existing implementations that are capable of handling a large number of covariates. While the group lasso was not an optimal method for this modelling problem, the plain lasso is a powerful tool; its full solution can be calculated instead of having to work over a grid of possible shrinkage parameters. In order to combat the instability of using the lasso on such a small data set, a resampling method similar to the bootstap was used to determine which concomitant medications should be assessed for predictive ability in a linear regression model with the original model covariates. In each repetition of the selection algorithm, the dataset was randomly divided into five folds for the purposes of CV to choose the optimal value of shrinkage for the lasso, the penalized regression method used to select concomitant medications. Five-fold CV was used instead of ten-fold because of the prohibitively small size of the dataset; five-fold CV also allowed a greater number of potential permutations than leave-one-out CV (LOOCV). For each repetition, the variables in the active set were recorded; the active set contains all of the variables with non-zero coefficients in the lasso output for the shrinkage parameter λ chosen via CV with the specific

random permutation sampled for that repetition. The resampled penalized regression protocol was repeated 1000 times.

The decision criteria used the proportion of times a coefficient for a particular medication was non-zero in the resampling process to determine whether it would be included in the predictive model. Generic medications that were present in a greater proportion of model selection repetitions than the designated selection threshold were automatically included in the set of covariates to be put in the final model. The rationale for this is that if a generic drug enters the model more often than the threshold, it is relatively good evidence that it should be selected for inclusion in the final predictive model, because the model including all of the generic drugs is much noisier than the medication class model, given the number of variables to choose from. Generic medications with non-zero proportions of inclusion in the resampling process were also selected for inclusion in the model if their respective medication classes were present in a greater proportion of models than the designated threshold. Choosing based on the medication class offers more power because there are fewer categories to choose from, and thus a clearer signal for the detection of relevant model covariates. The full mapping of generic medications to their corresponding functional classes can be found in Appendix A.

3.5 Concomitant Medication Selection Results

3.5.1 Concomitant Medications in the Prospective Cohort

A full table of the concomitant medications taken by the patients in the atorvastatin and rosuvastatin prospective cohorts can be found in Tables A.2, A.4, A.5, A.6 and A.7. On average,

patients in the atorvastatin cohort had 6.734 (± 3.259) concomitant medications. Patients in the rosuvastatin cohort had 6.485 (± 3.406). 52 generic medications of 132 total were removed from further analysis in the atorvastatin concomitant medication selection process; 49 were removed because they were only taken by one person in the atorvastatin cohort (singleton medications), and 3 drugs were removed because the drug name originally used in the patient chart could not be matched with a generic equivalent. In the rosuvastatin cohort 71 of 157 total concomitant medications were removed because they were singleton medications, and 2 were removed because the drug originally named in the dataset could not be identified with a generic equivalent. The full list of concomitant medications present in both cohorts is presented in Appendix A.

3.5.2 Atorvastatin

To decide on inclusion of concomitant medications for the models predicting atorvastatin plasma concentration, we conducted five-fold CV on the resulting linear models with the selected medications included, as well as the covariates from the original regression model; this process was repeated 100 times. Adjusted R^2, RMSE and AIC were calculated by averaging over the 500 resulting folds from the linear regressions including the medications selected at thresholds of 90%, 95% and 99% (Table 3.10). Based on a cutoff proportion of concomitant medications being represented in at least 99% percent of the models in the selection procedure, the following generic drugs were selected for inclusion in the atorvastatin predictive model: acetylsalicylic acid, atenolol, candesartan, diclofenac, digoxin, esomeprazole, gliclazide, glucosamine, levothyroxine, losartan, metformin, misoprostol, nifedipine, tamsulosin, valsartan,

and venlafaxine. When the threshold was lowered to 95% inclusion, hydrochlorothiazide was also selected. When the threshold was further lowered to 90%, Vitamin B3 was added to the concomitant medications selected for inclusion. Restricting the number of covariates in a predictive model is generally desirable for decreasing noise and increasing power, however using a less conservative selection in this context gives us the opportunity to look at biological mechanisms for all potential relevant interactions between atorvastatin plasma concentration and concomitant medications. Because the RMSE and AIC were so similar for all thresholds tested, medications selected in 90% or more of the models in the concomitant medication selection procedure will be included in subsequent predictive models in the current work.

Table 3.10: Atorvastatin selection threshold cross-validation results

Threshold	Adjusted R^2	RMSE	AIC
90%	0.652 ± 0.030	20.376 ± 8.627	211.752 ± 8.028
95%	0.643 ± 0.030	19.524 ± 8.773	213.655 ± 7.811
99%	0.645 ± 0.029	19.122 ± 8.666	212.333 ± 7.419

3.5.3 Rosuvastatin

Ranitidine was the sole medication selected by the concomitant medication algorithm for inclusion in the rosuvastatin regression model. The dearth of concomitant medications relevant to rosuvastatin plasma concentration in this context is unsurprising given that rosuvastatin undergoes minimal metabolism before performing its intended action in the liver. Ranitidine was the only medication selected for all threshold values above 90%, and as such was included in the concomitant linear regression analysis without further selection protocols.

3.6 Linear Regression With Concomitant Medications

3.6.1 Atorvastatin

When the medications chosen by the selection algorithm were included in a standard linear regression model of atorvastatin plasma concentration along with the original model covariates, the adjusted amount of variance accounted for in the model (adjusted R^2) was significantly increased from 0.488 to 0.649 ($F(18, 97) = 3.940$, $p < 0.001$). When the concomitant medications were added to the model, the effect of OATP1B1 c.388A>G was attenuated compared to its effect with only the original covariates included, although the confidence interval was too wide to achieve statistical significance ($\hat{\beta}$ = -0.141, 95% CI = -0.343 to 0.061, $p > 0.05$). As in the model with the original covariates, all dose categories, age and time post dose were significant predictors of atorvastatin concentration; BMI, sex and ethnicity remained not statistically significantly associated with atorvastatin plasma concentration.

Losartan, metformin and tamsulosin were found to be statistically significant predictors of atorvastatin plasma concentration. Specifically, for patients on losartan compared to patients not taking losartan, the predicted atorvastatin plasma concentration increased by a factor of 2.659 (95% CI = 1.480 to 4.778, p=0.001); the predicted atorvastatin plasma concentration of patients taking metformin decreased by factor of 0.705 (95% CI = 0.510 to 0.977, p=0.036) compared to patients not taking metformin; and taking tamsulosin increased the predicted value of atorvastatin plasma concentration by a factor of 2.933 compared to patients not taking tamsulosin (95% CI = 1.284 to 6.706, p=0.011).

The coefficient estimates for candesartan, diclofenac, digoxin, levothyroxine, and niacin trended towards significance ($p < 0.1$). Specifically, for patients taking candesartan, predicted

atorvastatin plasma concentration increased by a factor of 2.130 (95% CI = 0.996 to 4.559, p=0.051) compared to patients without; predicted atorvastatin plasma concentration increased by a factor of 2.924 (95% CI = 0.960 to 8.908, p=0.059) for patients taking diclofenac compared to patients not taking this medication; patients taking digoxin had predicted atorvastatin concentrations increase by a factor of 1.795 (95% CI = 0.941 to 3.425, p=0.075) compared to those patients not on digoxin; patients taking levothyroxine saw predicted atorvastatin plasma concentration decrease by a factor of 0.674 compared to those without (95% CI = 0.425 to 1.067, p=0.091); and predicted values of atorvastatin plasma concentration decreased for patients taking niacin by a factor of 0.676 (95% CI = 0.432 to 1.055, p=0.084).

The remainder of the confidence intervals for the concomitant medication coefficient estimates were too wide to achieve statistical significance as predictors of atorvastatin plasma concentration: acetylsalicylic acid, atenolol, candesartan, esomeprazole, gliclazide, glucosamine, hydrocholorothiazide, misoprostol, nifedipine, valsartan and venlafaxine all had p values greater than 0.1 for their respective coefficient estimates. The regression coefficients resulting from the atorvastatin standard linear regression with inclusion of concomitant medications are shown in Table 3.11.

Table 3.11: Atorvastatin linear model including concomitant medications (n=127)

	Estimate	Confidence Interval	P-Value	Sig.
(Intercept)	-0.338	-1.454 to 0.779	0.55	
OATP1B1 c.521T>C	0.411	0.166 to 0.656	0.001	**
OATP1B1 c.388A>G	-0.141	-0.343 to 0.061	0.169	
Age	0.016	0.006 to 0.026	0.002	**
4β-hydroxholesterol	-0.02	-0.03 to -0.01	<0.001	***
Dose (20mg)	0.741	0.323 to 1.159	0.001	***
Dose (40mg)	1.1	0.729 to 1.472	<0.001	***
Dose (80mg)	1.601	1.175 to 2.027	<0.001	***
Time Post Dose (hr)	-0.085	-0.109 to -0.06	<0.001	***
BMI (kg/m^2)	0.015	-0.009 to 0.04	0.225	
Sex (Male = 1)	-0.051	-0.295 to 0.192	0.677	
Ethnicity (Non-Caucasian = 1)	0.063	-0.296 to 0.422	0.729	
Acetylsalicylic Acid	0.186	-0.089 to 0.461	0.183	
Atenolol	-0.304	-0.741 to 0.134	0.172	
Candesartan	0.756	-0.004 to 1.517	0.051	.
Diclofenac	1.073	-0.041 to 2.187	0.059	.
Digoxin	0.585	-0.061 to 1.231	0.075	.
Esomeprazole	0.602	-0.231 to 1.435	0.155	
Gliclazide	-0.347	-1.083 to 0.39	0.353	
Glucosamine	0.448	-0.401 to 1.297	0.298	
Hydrochlorothiazide	0.095	-0.277 to 0.467	0.613	
Levothyroxine	-0.395	-0.855 to 0.065	0.091	.
Losartan	0.978	0.392 to 1.564	0.001	**
Metformin	-0.349	-0.674 to -0.023	0.036	*
Misoprostol	0.162	-0.905 to 1.23	0.764	
Nifedipine	0.34	-0.348 to 1.028	0.329	
Tamsulosin	1.076	0.25 to 1.903	0.011	*
Valsartan	-0.114	-0.81 to 0.582	0.746	
Venlafaxine	-0.247	-0.967 to 0.472	0.497	
Vitamin B3 (Niacin)	-0.392	-0.839 to 0.054	0.084	.

Table 3.12: Atorvastatin linear regression model cross-validation performance results

Model	Adjusted R^2	RMSE	AIC
Atorvastatin	0.474 ± 0.038	21.064 ± 10.103	246.075 ± 7.877
Atorvastatin reduced cohort (RC)	0.458 ± 0.038	18.899 ± 9.219	241.899 ± 7.752
Atorvastatin RC + categorical dose (CD)	0.488 ± 0.040	18.570 ± 9.691	237.875 ± 9.452
Atorvastatin RC + CD + concomitant	0.652 ± 0.027	20.480 ± 8.642	211.646 ± 7.534

3.6.2 Rosuvastatin

When ranitidine, the sole medication selected by the algorithm, was included in a linear regression of rosuvastatin along with the covariates from the original linear regression model, the amount of explained variance (adjusted R^2) did not significantly change (0.635 vs. 0.633, $F(1,115) = 0.463$, $p = 0.497$)

Table 3.13: Rosuvastatin linear model including concomitant medications (n=127)

	Estimate	Confidence Interval	P-Value	Sig.
(Intercept)	1.02	0.099 to 1.941	0.03	*
OATP1B1 c.521T>C	0.403	0.217 to 0.590	<0.001	***
ABCG2 c.421C>A	-0.346	-0.586 to -0.106	0.005	**
Age	0.013	0.005 to 0.021	0.002	**
Dose (10mg)	0.844	0.534 to 1.154	<0.001	***
Dose (20mg)	1.428	1.115 to 1.741	<0.001	***
Dose (40mg)	1.990	1.666 to 2.314	<0.001	***
Time Post Dose (hr)	-0.057	-0.087 to -0.028	<0.001	***
BMI (kg/m^2)	-0.013	-0.026 to 0.001	0.077	.
Sex (Male = 1)	-0.175	-0.399 to 0.049	0.124	
Ethnicity (Non-Caucasian = 1)	-0.064	-0.341 to 0.213	0.648	
Ranitidine	0.273	-0.522 to 1.069	0.497	

Table 3.14: Rosuvastatin linear regression model cross-validation performance results

Model	Adjusted R^2	RMSE	AIC
Rosuvastatin	0.560 ± 0.030	16.129 ± 4.341	198.693 ± 7.866
Rosuvastatin reduced cohort (RC)	0.561 ± 0.031	16.040 ± 4.392	195.079 ± 8.051
Rosuvastatin RC + categorical dose (CD)	0.634 ± 0.028	16.314 ± 4.635	178.463 ± 7.608
Rosuvastatin RC + CD + concomitant	0.635 ± 0.028	17.047 ± 5.315	178.968 ± 7.254

3.7 Discussion

3.7.1 Model Fit with Original Covariates

The fit of both of the linear regression models for the atorvastatin and rosuvastatin prospective cohorts was significantly improved by modelling dose categorically instead of as a continuous variable. This improvement could also be explained in part by the probable reduction in variability caused by the exclusion of the patients on non-standard doses. The doses prescribed to the patients in each cohort were largely homogenous, with too few unique values to allow a linear relationship to be accurately modelled, and the individual dose categories capture the variation lost by the assumption of linearity. Of the two models, the rosuvastatin model was more impacted by the conversion of the dose encoding. A possible reason for this is that the relationship between plasma concentration and dose is not linear, and this is the assumption for continuous variables. Binning the values into a categorical variable allows for the relationship to be expressed without the assumption of linearity. Alternatively, the signal in the model could have been improved by the removal of the three patients with outlying doses if they had characteristics that were also inconsistent with the rest of their cohort.

3.7.2 Atorvastatin and Concomitant Medications

A number of medications were chosen by the concomitant selection algorithm for inclusion in the atorvastatin linear regression model, and several were consistent with previous literature suggesting a relationship between the medications and atorvastatin plasma concentration. Among these were candesartan[10], digoxin[11;12;13;14], esomeprazole[15], losartan[10], metformin[16],

nifedipine[12], valsartan[13], venlafaxine[11], and vitamin B (niacin)[17;18;19;20] . Three angiotensin II receptor blockers (ARBs) were selected for inclusion in the predictive model. Losartan had a significant effect in the linear model of increasing predicted plasma concentration, and candesartan increased predicted plasma concentration in the final model, although the 95% confidence interval for the estimated effect was too wide by a small margin to achieve statistical significance. Valsartan was included in the model, but the effect estimate was not statistically significant. Digoxin had the effect of increasing predicted atorvastatin plasma concentration in the model, but the 95% confidence interval for the effect estimate was too wide by a small margin to achieve statistical significance. The direction of the effect was positive, indicating that in this model the predicted atorvastatin plasma concentration increases for patients who take digoxin. In previous literature there is mixed evidence on the relationship between dixogin and atorvastatin pharmacokinetics[11;12;13;14]. Studies have suggested that digoxin concentration increases in the presence of atorvastatin via ABCB1 or CYP3A4 inhibition[14;11]; however, further study is required to determine whether digoxin in turn affects atorvastatin plasma concentration. While little literature exists on the potential for interactions between esomeprazole and atorvastatin, a case report was published documenting an incidence of rhabdomyolysis following treatment with atorvastatin, esomeprazole and clarithromycin[15]. While esomeprazole is mainly metabolized by CYP2C19, a small amount of metabolism occurs with CYP3A4; additionally, the authors speculate that ABCB1 may have played a role in generating the myotoxicity observed in this case[15]. Metformin has been shown to increase the risk of hepatotoxicity when taking atorvastatin because it is an inhibitor of the efflux protein MRP2[16]. This is reflected in the predictive model for atorvastatin concentration ($\hat{\beta}$ = -0.349, 95% CI = -0.674 to -0.023, p = 0.036). Nifedipine had the effect of increasing predicted plasma concentration

in the regression model, but the confidence interval was relatively wide (95% CI = -0.348 to 1.028, $p = 0.329$); the width of the confidence intervals in this instance could potentially be due to the small number of patients taking nifedipine in the atorvastatin prospective cohort (n = 4, 3.1%). A positive effect on plasma concentration would be consistent with previous literature suggesting that nifedipine inhibits CYP3A4 metabolism[12]. Venlafaxine increased predicted plasma concentration in the atorvastatin predictive model, but the effect had a wide confidence interval (95% CI = -0.967 to 0.472, $p = 0.497$). Previous studies have found that venlafaxine is metabolized in part by CYP3A4[11;21], suggesting a possible pharmacokinetic relationship between venlafaxine and atorvastatin. Further work is required to confirm whether these two drugs have a clinically meaningful interaction, and what the metabolic consequences of such an interaction would be. The final drug selected for inclusion in the model that is consistent with previous work on statin metabolism was niacin, or Vitamin B3. Niacin has been found to alter lipid metabolism, raising protective HDL cholesterol levels[22;23;17]. Indeed, before the discovery of statin drugs, it was used independently as a treatment for hypercholesterolemia[24]. It is unknown whether niacin is associated with other changes in the blood cholesterol profile that would impact the metabolism of atorvastatin; further study would be required to establish a causal relationship between concomitant niacin administration and atorvastatin plasma level. The current evidence on whether niacin affects atorvastatin plasma concentration and the risk of myalgia is inconsistent; results range from weak increases in specific patient populations, to niacin having no significant impact on statin plasma concentration[17;18;19;20].

A number of medications were also included as predictors in the linear regression model that have not previously been known to have an affect on atorvastatin plasma concentration. These medications include acetylsalicylic acid, atenolol, diclofenac, gliclazide, glucosamine,

hydrochlorothiazide, levothyroxine, misoprostol and tamsulosin. Of these, the only statistically significant effect in the model was of tamsulosin, a medication commonly prescribed to men suspected of having benign prostatic hyperplasia; it is used to treat lower urinary tract symptoms that can often result from this condition[25]. Tamsulosin increased the predicted plasma concentration in the atorvastatin model by a relatively large margin compared to some of the other statistically significant effect sizes of the original model covariates ($\hat{\beta}$ = 1.076, 95% CI = 0.25 to 1.903, p = 0.011). Notably, tamsulosin is metabolized in part by CYP3A4, and plasma concentration of this medication increases in the presence of statin drugs[25]. It is unknown whether tamsulosin in turn affects statin plasma concentrations, but given that tamsulosin is a significant predictor of atorvastatin plasma concentration in the linear model, further research investigating the pharmacokinetic relationship between these two drugs is warranted. It is important to note that these relationships may be confounded by the indication, such that it is really the underlying illness affecting the variation in statin plasma concentration, and not the medications used to treat it. This is not a problem for the purposes of predictive modelling since the goal is not inference primarily, but should be closely examined in future work looking at causal relationships between individual medications and statin plasma concentration.

3.7.3 Rosuvastatin and Concomitant Medications

The identical feature selection procedures for atorvastatin were used to identify potentially predictive concomitant generic medications for rosuvastatin plasma concentration. Using the developed concomitant medication selection algorithm, only one medication (ranitidine) was selected for inclusion in the predictive model. Ranitidine suppresses the production of stom-

ach acid, and used for the treatment of gastrointestinal issues such as acid reflux, dyspepsia and gastroduodenal mucosal lesions[26;27]. These can be caused by stress from critical illness, or adverse events related to using NSAIDs for the treatment of arthritis; however, it is not as effective as some other medications used for this purpose such as omeprazole[26;28]. Ranitidine is a CYP2C9 inhibitor[17]; as rosuvastatin is minimally metabolized by this enzyme, it is possible that ranitidine has a small impact on rosuvastatin pharmacokinetics by way of CYP2C9 inhibition. The ranitidine coefficient estimate in the linear model for rosuvastatin plasma concentation was positive with a wide confidence interval; it did not reach statistical significance ($\hat{\beta}$ = 0.273, 95% CI = -0.522 to 1.069, p = 0.497).

When included in a linear regression with the original rosuvastatin model covariates, model fit was not significantly impacted ($F(1, 115) = 0.463$, $p = 0.497$). Based on the model fit performance results from the linear regression models with and without the addition of the selected concomitant medication, it seems likely that the addition of concomitant medications does not improve model fit when predicting rosuvastatin plasma concentration. It is possible that ranitidine was present in 99% of the models generated by the concomitant medication selection algorithm because it truly has a small impact on rosuvastatin; however this may be an artefact of small sample size, given that only two patients in the dataset were on this medication (1.5% of the patients in the cohort). Based on this limitation of small sample size, it is more likely that the inclusion of ranitidine in the model is reflective of potential outlying characteristics of the two individuals in the rosuvastatin cohort taking this medication.

3.7.4 General Discussion

Model fit for the original systemic statin exposure predictive model was significantly improved for both the atorvastatin and rosuvastatin linear models predicting plasma concentration.

The patients in the atorvastatin and rosuvastatin prospective cohorts are perhaps not representative of statin users in the general population, as they were receiving specialized care in controlling their hypercholesterolemia. The patients recruited for the original study performed by DeGorter et al. received additional testing to tailor statin medication dosing based on their genetic profile. The patients used to generate the concomitant medications model may be more complex than general population of hypertensive patients in the community; however, patients with advanced hypercholesterolemia that is not responding adequately to statin treatment are often prescribed more than one drug to maximize cholesterol-lowering efficacy[29]. Patients on statins, especially with complicated illness, often experience polypharmacy which makes this population advantageous for studying the effects of concomitant medications on statin plasma concentration.

3.8 Conclusions

In the current work we aimed to characterize concomitant medication use in the atorvastatin and rosuvastatin prospective cohorts, and to develop a selection algorithm to identify concomitant medications that would improve the prediction of statin plasma levels. The medications selected by the algorithm were then added to the original covariates used to train the original statin systemic-exposure linear regression models to guide statin dosing. Before proceeding with the concomitant medication selection, we recoded dose as a categorical variable instead

of being represented continuously. This significantly improved the model fits for both atorvastatin and rosuvastatin. Rosuvastatin is minimally metabolized and has a smaller probability of concomitant medication interactions, so it is relatively unsurprising that we were unable to improve the prediction quality in the rosuvastatin model using this data. Instead, we suggest that the rosuvastatin predictive model could be further improved by identifying novel SNPs that are predictive of statin plasma concentration, and using these to augment the original linear regression model for rosuvastatin.

References

[1] Marianne K DeGorter. Statin Transport by Hepatic Organic Anion-Transporting Polypeptides (OATPs). PhD thesis, The University of Western Ontario, 2012.

[2] Marianne K DeGorter, Rommel G Tirona, Ute I Schwarz, Yun-Hee Choi, George K Dresser, Neville Suskin, Kathryn Myers, GuangYong Zou, Otito Iwuchukwu, Wei-Qi Wei, et al. Clinical and pharmacogenetic predictors of circulating atorvastatin and rosuvastatin concentration in routine clinical care. Circulation: Cardiovascular Genetics, 6(4):400–408, 2013.

[3] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. The elements of statistical learning. Springer-Verlag, 2009.

[4] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 68(1):49–67, 2006.

[5] Robert Tibshirani. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), 58(1):267–288, 1996.

[6] Eugene F. Krause. Taxicab Geometry: An Adventure in Non-Euclidean Geometry. Dover Publications, 1986.

[7] Peng Zhao, Guilherme Rocha, and Bin Yu. Grouped and hierarchical model selection

through composite absolute penalties. Department of Statistics, UC Berkeley, Tech. Rep, 703, 2006.

[8] Bradley Efron, Trevor Hastie, Iain Johnstone, Robert Tibshirani, et al. Least angle regression. The Annals of Statistics, 32(2):407–499, 2004.

[9] Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(1):91–108, 2005.

[10] Hideki Fujino, Tsuyoshi Saito, Yoshihiko Tsunenari, and Junji Kojima. Interaction between several medicines and statins. Arzneimittelforschung, 53(03):145–153, 2003.

[11] Debasish Maji, Shehla Shaikh, Dharmesh Solanki, and Kumar Gaurav. Safety of statins. Indian Journal of Endocrinology and Metabolism, 17(4):636–646, 2013.

[12] Anida Čaušević Ramosevac and Sabina Semiz. Drug interactions with statins. Acta Pharmaceutica, 63(3):277–293, 2013.

[13] Kenneth Kellick. Organic ion transporters and statin drug interactions. Current Atherosclerosis Reports, 19(65), 2017.

[14] Rebecca A Boyd, Ralph H Stern, Barhra H Stewart, Xiaochun Wu, Eric L Reyner, Elizabeth A Zegarac, Edward J Randinitis, and Lloyd Whitfield. Atorvastatin coadministration may increase digoxin concentrations by inhibition of intestinal P-glycoprotein-mediated secretion. The Journal of Clinical Pharmacology, 40(1):91–98, 2000.

[15] Brooke E Sipe, Ronald J Jones, and Gordon H Bokhart. Rhabdomyolysis causing AV blockade due to possible atorvastatin, esomeprazole, and clarithromycin interaction. Annals of Pharmacotherapy, 37(6):808–811, 2003.

[16] Eunjung Shin, Naree Shin, Ju-Hee Oh, and Young-Joo Lee. High-dose metformin may increase the concentration of atorvastatin in the liver by inhibition of Multidrug Resistance–Associated Protein 2. Journal of Pharmaceutical Sciences, 106(4):961–967, 2017.

[17] Yiannis S Chatzizisis, Konstantinos C Koskinas, Gesthimani Misirli, Chris Vaklavas, Apostolos Hatzitolios, and George D Giannoglou. Risk factors and drug interactions predisposing to statin-induced myopathy. Drug Safety, 33(3):171–187, 2010.

[18] Sivakumar Sathasivam. Statin induced myotoxicity. European Journal of Internal Medicine, 23(4):317–324, 2012.

[19] Huifen Wang, Jeffrey B Blumberg, C-Y Oliver Chen, Sang-Woon Choi, Michael P Corcoran, Susan S Harris, Paul F Jacques, Aleksandra S Kristo, Chao-Qiang Lai, Stefania Lamon-Fava, et al. Dietary modulators of statin efficacy in cardiovascular disease and cognition. Molecular Aspects of Medicine, 38:1–53, 2014.

[20] Russell A Wilke, Jason H Moore, and James K Burmester. Relative impact of CYP3A genotype and concomitant medication on the severity of atorvastatin-induced muscle damage. Pharmacogenetics and Genomics, 15(6):415–421, 2005.

[21] Steven M Fogelman, Jürgen Schmider, Karthik Venkatakrishnan, Lisa L von Moltke, Jerold S Harmatz, Richard I Shader, and David J Greenblatt. O-and N-demethylation of venlafaxine in vitro by human liver microsomes and by microsomes from

cDNA-transfected cells: effect of metabolic inhibitors and SSRI antidepressants. Neuropsychopharmacology, 20(5):480–490, 1999.

[22] Stefania Lamon-Fava, Margaret R Diffenderfer, P Hugh R Barrett, Aaron Buchsbaum, Mawuli Nyaku, Katalin V Horvath, Bela F Asztalos, Seiko Otokozawa, Masumi Ai, Nirupa R Matthan, et al. Extended-release niacin alters the metabolism of plasma apolipoprotein (Apo) AI and ApoB-containing lipoproteins. Arteriosclerosis, Thrombosis, and Vascular Biology, 28(9):1672–1678, 2008.

[23] Pattie S Green, Tomas Vaisar, Subramaniam Pennathur, J Jacob Kulstad, Andrew B Moore, Santica Marcovina, John Brunzell, Robert H Knopp, Xue-Qiao Zhao, and Jay W Heinecke. Combined statin and niacin therapy remodels the high-density lipoprotein proteome. Circulation, 118(12):1259–1267, 2008.

[24] Jonathan A Tobert. Lovastatin and beyond: the history of the HMG-CoA reductase inhibitors. Nature reviews Drug discovery, 2(7):517–526, 2003.

[25] Gabriela Franco-Salinas, Jean JMCH de la Rosette, and Martin C Michel. Pharmacokinetics and pharmacodynamics of tamsulosin in its modified-release and oral controlled absorption system formulations. Clinical Pharmacokinetics, 49(3):177–188, 2010.

[26] Michael J Levy, Charles B Seelig, Norman J Robinson, and Jane E Ranney. Comparison of omeprazole and ranitidine for stress ulcer prophylaxis. Digestive Diseases and Sciences, 42(6):1255–1259, 1997.

[27] PC Alexander, S Ramya, Rajkumar T Solomon, S Raja, M Priyadarshini, R Geetha,

Vijaya Srinivasan, V Jayanthi, et al. Effects of long-term acid suppressants with ranitidine and omeprazole on gastric mucosa. Journal of Digestive Endoscopy, 4(1–5):1, 2013.

[28] Neville D Yeomans, Zsolt Tulassay, László Juhász, István Rácz, John M Howard, Christoffel J Van Rensburg, Anthony J Swannell, and Christopher J Hawkey. A comparison of omeprazole with ranitidine for ulcers associated with nonsteroidal antiinflammatory drugs. New England Journal of Medicine, 338(11):719–726, 1998.

[29] Allan D Sniderman, Sotirios Tsimikas, and Sergio Fazio. The severe hypercholesterolemia phenotype: clinical diagnosis, management, and emerging therapies. Journal of the American College of Cardiology, 63(19):1935–1947, 2014.

Chapter 4

Non-linear Modelling of Statin Plasma Concentration

Augmenting the atorvastatin model by including concomitant medication information was greatly successful in improving model accuracy. However, even with taking the log of plasma concentration, both models had covariates that had a high possibility of a non-linear relationship with the outcome variable, such as the time post-dose that plasma concentration was measured. We hypothesize that using non-linear modelling techniques will increase the predictive performance of the atorvastatin and rosuvastatin systemic-exposure models originally developed by DeGorter et al.[1].

4.1 Background: Methods for Modelling Non-linearity

Ordinary linear regression models are very popular because they are relatively easy to implement and interpret. For each covariate used to predict the outcome variable, it is possible to

obtain a coefficient value that gives a clear depiction of the direction, magnitude and confidence intervals for the coefficient effect estimates. Unfortunately, much of the data generated in health care settings has non-linear relationships between the predictor and the outcome variable. One example of a non-linear predictor variable in a healthcare setting is the number of hours since a dose of medication if one is trying to accurately model drug plasma concentration[2]. Another example in a similar vein is trying to predict hot flashes in patients on tamoxifen over the course of therapy using patient age as a predictor; age changes with menopausal status, and is strongly associated with hot flash severity[3]. While linear regression is not capable of leveraging these relationships to improve predictive performance, many non-linear methods have been developed to more accurately model these variables.

4.1.1 Generalized Linear Models (GLMs)

Generalized Linear Models (GLMs) extend the functionality of linear regression modelling by allowing the use of any exponential distribution via a link function or kernel function that transforms the linear model into the exponential model. A basic GLM can be written in the following form:

$$g(\mu_i) = \boldsymbol{X}_i\boldsymbol{\beta}_i$$

where μ_i is the expected value of the response variable for the i-th observation (corresponding to row i in the design matrix $\boldsymbol{X}$; g transforms the expected outcome value using the link function specified in accordance with the model distribution; and $\boldsymbol{\beta}$ is the vector of coefficients for each predictor variable[4]. An example of a potential link function is the identity link function. With

the identity link function, the expected values of the outcome are left unchanged:

$$g(\mu_i) = \mu_i.$$

Linear regression is a special case of a GLM which uses the Normal distribution and the identity link function. Another popular type of GLM is the logistic regression model, which is used to model binary outcomes. The logistic regression model uses a Binomial distribution and the expected outcome values are transformed using the logit link function:

$$g(\mu_i) = ln\left(\frac{\mu_i}{1-\mu_i}\right).$$

GLMs differ from ordinary linear regression in that model fitting using an exponential distribution and link cannot be done in a single step, and the solution is not exact. Instead of solving for the exact coefficient values using least squares, maximum likelihood estimation (MLE) is used to come to the solution over a number of iterations[4]. MLE relies on the assumption that the data that is being used to generate the model (y_i, where $i = 1, ..., n$) is a random sample from some probability distribution. This distribution has a probability density of $\Pr_\theta(y)$, where θ is a vector of parameters characterizing the density. In MLE, we try to approximately solve for these parameters θ under the assumption that the best values for θ are those that give the highest probability of generating our random data sample outcome (y)[5]. Formally, MLE maximizes the log-probability $\mathcal{L}$ of our random sample y:

$$\mathcal{L}(\theta) = \sum_{i=1}^{N} \log \Pr_\theta(y_i)$$

GLMs use this method to find the each value of β in our regression model for the distribution and link function we have specified. At each step or iteration, the value of the parameters are updated, and the MLE is calculated. This process continues until no more adjustments to the parameters increase the value of the maximum likelihood above a specified threshold (generally very small), and the algorithm converges.

4.1.2 Generalized Additive Models (GAMs)

Generalized Additive Models (GAMs) are very similar to generalized linear models, with the difference being that the continuous variables in the dataset are modelled using a set of smooth functions instead of a single coefficient. In the regression setting, predictions generated by a GAM usually take the following form[5]:

$$E(Y|X_1, X_2, ...X_p) = \alpha + f_1(X_1) + f_2(X_2) + ... + f_p(X_p)$$

where Y represents the outcome variable, $X_1, X_2, ..., X_p$ represent the continuous independent predictor variables, and α is a constant similar to an intercept. Alternatively, this can be written as

$$\hat{Y} = \alpha + \sum_{j=1}^{p} f_j(X_j) + \epsilon$$

where ϵ is the prediction error[5].

The unspecified smooth, non-parametric curves represented by these functions can be fit by dividing the variable into segments and modelling each segment separately with a smooth function. This allows the function representing the relationship between the predictor and

dependent variable to have different trajectories for different ranges of values, and is why the effect of these covariates cannot be represented by a single coefficient value.

There are many different methods for generating the non-linear representation of a predictor variable for inclusion in a GAM, some of which require the modeller to input domain-specific knowledge about the modelling problem at hand, and some of which are automatic, requiring no additional input[6]. Some options for the former include using linear basis expansions to give a global non-linear representation of the covariate, and piecewise polynomial functions or splines to give local trajectories for different ranges of values of the predictor variable. A simple example of this is a linear spline. The first step to building this type of spline model is to choose the points (called knots) that separate the range of the predictor into different sections of lines with different slopes; the slope can change at the knot between sections[7]. This results in a piecewise linear function[8].

Knot locations can be chosen based on domain knowledge; however, this type of domain knowledge is not always available, and/or there is no good justification for exact locations for the placement of individual knots. Often in this case, knots are placed at the quartile values and the data points between them are smoothed with separate smoothing functions[7]. The choice of how many knots to include in a model can also be difficult; it is often best to compromise between having a model that is as simple as possible, but also does not model too much of the noise or error in the variable.

Polynomial and Natural Cubic Splines

Another example of a commonly used spline in a GAM is the natural cubic spline. Between each knot, the data is modelled using a cubic polynomial function; these are then connected

at the knots such that they are continuous at both the first and second derivatives[4]. Additionally, the final shape of the model is constrained such that the trajectory beyond the boundary knots (the lowest and highest specified knot values) must be linear[5]. The linearity constraint is included because without this constraint, such curved splines tend not to give accurate representations of the trend of the data at the tails (the outer ranges of values of the predictor variable)[8]. The inability of polynomial splines to give reasonable predictions beyond the boundary knots usually makes it very difficult to extrapolate from the data at hand to data with a wider range of values[7].

Instead, more complicated methods can be used to obtain the smoothing functions for each continuous covariate; many of these can be characterized as scatterplot smoothers, referring to two-dimensional scatterplot graphs with the dependent variable on the y axis, and the independent variable on the x axis[6]. An example of one such method that is capable of more accurately modelling the tails of the sample distribution is the use of thin plate regression splines.

Thin Plate Regression Splines

One of the largest challenges of using cubic splines is finding the optimal number and placement of knots to define the smoothing basis functions[8] One way to avoid this difficulty is to use a smoother that does not require the specification of knot placement, such as thin plate splines[9]. Instead, thin plate splines can be calculated using a closed-form solution; the main problem to be solved in this formulation is the tradeoff between how well the spline function should fit the data, and how smooth it should be[8]. A spline with a lot of curves that is very wiggly might be an excellent fit for the specific data set it is being trained on, but generalizability may suffer because the model does not perform well on other datasets generated from the same underlying

distribution. The wiggliness of the thin plate spline function is controlled by a penalization parameter λ, which is solved for by optimization[8]. Thin plate splines have the capacity to achieve an optimal smoothing solution in terms of minimizing error and maintaining good model fit[8]. Unfortunately, the optimization procedure required to solve the original formulation of thin plate spline functions was very computationally expensive, and using this smoothing method with a large number of variables was infeasible, since the resourced required were a cubic function of the number of model parameters[8].

Thin plate *regression* splines were developed as a way to take advantage of the attractive theoretical properties of thin plate splines while being computationally efficient enough to be used with large datasets[10]. They decrease the computational requirements of the optimization problem by restricting the space of potential solutions in terms of how much wiggle can be present in the spline function[8]. The reformulation of thin plate splines to thin plate regression splines provides a useful method for smoothness estimation based in statistical theory instead of heuristic methods[10]; as such, they are a popular method for use with GAMs.

4.1.3 Support Vector Regression (SVR)

Support Vector Regression (SVR) is a popular machine learning technique for predictive modelling problems with continuous outcome variables; this paradigm is called Support Vector Machine (SVM) or Support Vector Classification (SVC) for binary outcomes. The concept behind SVR is most easily demonstrated using the classic SVM binary classification paradigm: in this scenario the goal is to create a classification model that uses a line to separate the two classes of the binary outcome variable, leaving the widest margin possible between the two

classes. The term "support vectors" in SVM refers to the observations or training examples that determine the size and placement of the margins between the hyperplane separating the observations from each class, and the nearest observation to it for either class, as shown in Figure 4.1.

The concept between the SVM binary classifier can also be applied to regression problems; in this paradigm, the support vectors are the observations forming the outer envelope or margin in which the majority of observations are enclosed. Two model parameters are particularly important to consider when training a SVR model: epsilon, and cost. Cost in the context of SVR tuning refers to the desired ratio of training errors in positive versus negative training examples[11]. In the context of clinical research, a higher cost value selected during CV would indicate a larger penalty on errors incurred on training examples for cases versus training error for observations in the control group. The width of this envelope is given by epsilon (ϵ), which controls how much error is permissible in the model between the model fit and training observations[12]. A higher value of epsilon indicates a larger margin of error between the training data and model fit; a wider epsilon envelope is less sensitive to variations in the training data and thus less prone to overfitting. A depiction of the basic structure of a SVR model fit is shown in Figure 4.2.

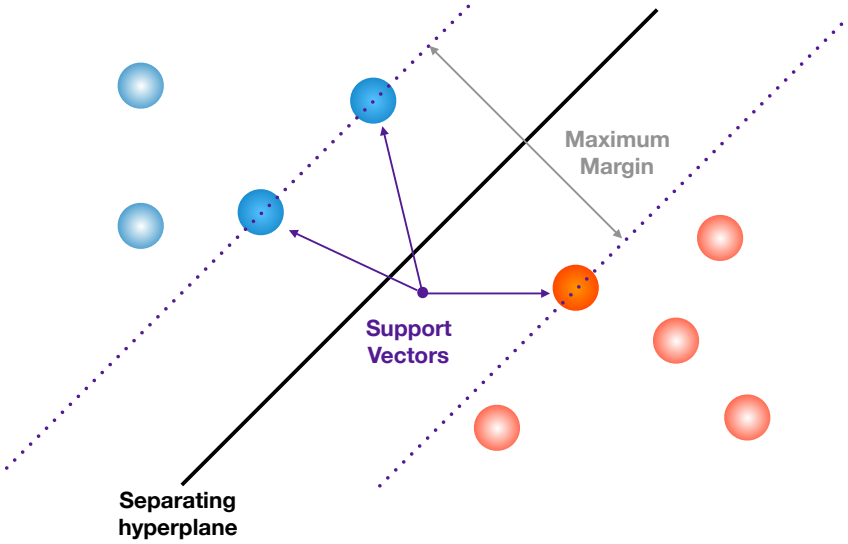


Figure 4.1: Simple support vector machine (SVM) binary classifier

SVMs and SVR are especially well-suited to non-linear models because they allow the user to transform the data into a higher dimension via a kernel, after which it is possible to fit a hyperplane to separate classes in the high dimensional space, or construct a more accurate SVR for models in which covariates are non-linear. Linear, radial and polynomials are especially popular for use in SVR.

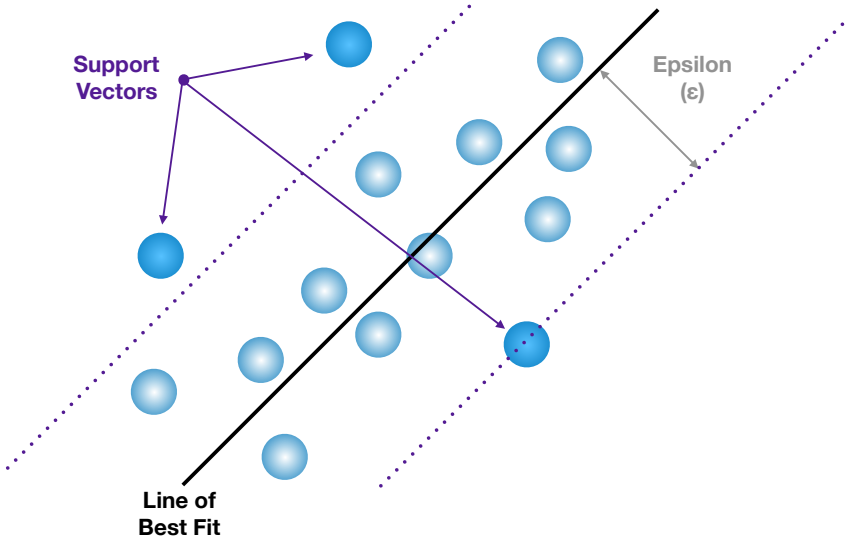


Figure 4.2: Simple support vector regression (SVR) model structure

4.2 Background: Model Performance Evaluation Metrics

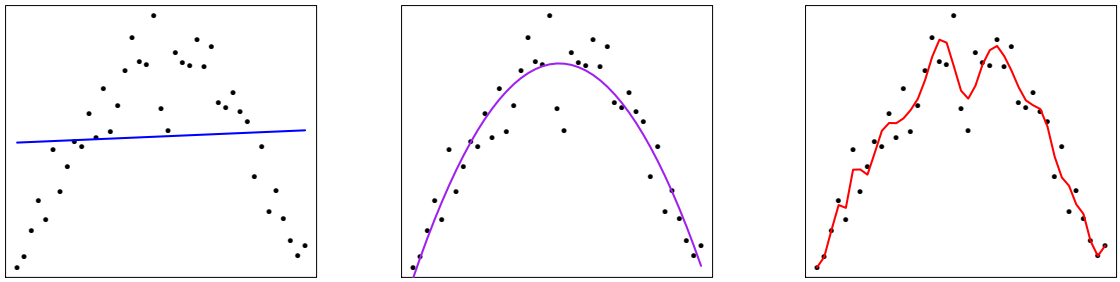
4.2.1 Model Fit

Choosing the right method to model a dataset is a crucial step in building a predictive model. However, making sure that the model fits the data well based on the purpose of prediction is equally important. Predictive modelling is an art as much as a science, and requires balancing competing goals. For example, a model that fits a dataset perfectly will probably perform poorly when trying to predict outcomes from new data observations. Similarly, often there are many candidate models that could be used for a particular data set. Several metrics of model fit quality exist that can be used to choose the best model for the prediction problem at hand.

4.2.2 Overfitting and Underfitting

Predictive modelling often involves a choice between using a simple model, a complicated model, or something in between. An example of a simple model is plain linear regression; all of the covariates in the model are linearly modelled with respect to the outcome variable, and no extra parameters besides the regression coefficients are needed to specify the model fit. The advantage of using a simple model is that they are generally easy to use and interpret. However, a risk that comes with simple modelling techniques is under-fitting. Under-fitting occurs when the model does not have enough parameters to capture the variability in the model and therefore cannot generate accurate predictions.

An example of a situation involving under-fitting would be trying to model a parabolic response using linear regression; the line cannot capture the curved shape of the relationship, and would not be as informative as a non-linear model. On the other hand, over-fitting occurs when the model captures the noise or error in the data instead of the true underlying trend. If the data is non-linear, a very simple model will not be as informative as a more complicated model that captures variation outside the boundaries of linear modelling so a balance between under- and over-fitting the data is required. Examples of what under-fitting and over-fitting look like compare to a good model fit are shown in Figure 4.3. One of the goals of achieving a good model fit is balancing model complexity with the quality of model fit with respect to the data.



(a) Under-fit (b) Well-fit (c) Over-fit

Figure 4.3: Visual examples of under- and over-fitting

4.2.3 Assessing Model Fit

Root Mean Squared Error (RMSE)

Perhaps the simplest way to assess model fit is to look at the magnitude of the differences between the predicted values and true values of the outcome variable generated by the predictive model; these are called the residuals. To calculate RMSE, we square the residuals, add them, and take the square root; squaring them alleviates the problem of errors in opposite directions cancelling each other out, and taking the square root after returns them to the original scale of the deviations. Formally,

$$\hat{\sigma}_e = \sqrt{\frac{1}{n-(p+1)\sum_{i=1}^{N} r_i^2}}$$

where r_i is the residual for observation i, n is the number of total observations, and p is the number of covariates in the model[7]. RMSE is also the squared deviation of the error term, which gives an idea of the spread of the errors from the predictions in the model. Unforunately, RMSE is only available for models with continuous outcome variables (ie. not logistic regression)[7].

Adjusted R^2

Adjusted R^2 is a measure specific to regression models with continuous outcomes. It describes the amount of variance that can be explained in the model by the covariates included. R^2 is also called the coefficient of determination; it can be thought of as the squared correlation between the true y values and the predicted outcome values generated by the model from the particular sample of data used. Formally, R^2 is calculated using the following formula[13]:

$$\mathrm{R}^2 = \frac{\sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

where $\bar{y}$ is the sample mean of the outcome variable; y_i is the value of the dependent variable for observation i; and $\hat{y}_i$ is the predicted outcome value for observation i. The quantity in the numerator is known as the regression sum of squares (the explained variance), and the denominator represents the total sum of squares (the total variance)[13].

While R^2 is a useful measure in theory, it is biased based on the size and specific sample used to calculate it; it tends to be an inflated estimate of the explained variance accounted for in the model[14]. Because of this, it is recommended to use an adjusted R^2 value when using it as a metric for model quality, or effect size. The adjustment is based on assumptions of the sampling error in the dataset used for the regression model, which depends on the number of observations, the number of covariates in the model, and the true size of the effect in the population[14]. Many different corrections are available to adjust for this bias, all of which make slightly different assumptions about the sample properties. One commonly used adjustment is

the Wherry method (*Wherry formula-1*)[15]:

$$\hat{R}^2_{\text{adj}} = 1 - \frac{N-1}{N-p-1}(1-\hat{R}^2)$$

where $N - p - 1$ is the residual degrees of freedom. This is the method used by the `lm`[16] command for linear modelling in R.

4.3 Implemented Methods

4.3.1 GAM

For both the atorvastatin and rosuvastatin models, Generalized Additive Models (GAMs) were fit using thin-plate to model non-linearity in the continuous covariates. Two models were fit for both atorvastatin and rosuvastatin, using smoothing parameters λ for each continuous variable chosen either by cross-validation or a fixed smoothing parameter. The fixed parameter models were performed to show the differences between tailored model fit and fit achieved by arbitrary parameter selection using a moderate smoothing parameter value. The R package `mgcv` was used for this process[10;4;17;8]. Conveniently, the `mgcv` package included an implemented CV strategy for the purpose of choosing the best smoothing parameters.

For each GAM fit, 5-fold CV was performed to calculate model error in order to compare the relative performance between models after the smoothing parameter selection was performed using the implemented package function. For each fold in the CV, the model was trained on a randomly selected (without replacement between folds) portion of data, comprising 80% of the available observations. Predictions were then generated using the newly trained

model for the remaining fifth of the data set, and root mean squared error (RMSE) was calculated for the difference between the predicted values generated by the trained model for that fold, and the the corresponding true plasma concentration values.

4.3.2 SVR

For both the atorvastatin/concomitant-medication model and the rosuvastatin plasma concentration models, support vector regression models were fit with a linear kernel, a low degree polynomial kernel (degree=3), a higher degree polynomial kernel (degree=5), and a radial kernel. This range of kernels was chosen to give the best picture of how each method might perform, and what the strengths and weaknesses of each model type are for these data sets. The R package `e1071`[18] was used to tune and fit the model for each SVR.

Similar to the procedure used to assess model fit for the GAMs, 5-fold CV was used to obtain RMSE for the fitted SVR models; however, the SVR models required an additional step to tune the model for each fold prior to assessing fit. A ready-implemented CV procedure for choosing epsilon and cost values was available with the software packaged used; this procedure used 10-fold CV over a grid search of epsilon specified by the user. The set of parameter values used in the current work for the grid search over epsilon was $\epsilon = (0.1, 0.2, ..., 1.0)$ and the set of cost values used to tune the models was $c = (2^2, 2^3, ..., 2^9)$. The final parameter values chosen for each fold were those that produced the lowest error over the grid search implemented in the `e1071` R package. RMSE was calculated using the predictions from the tuned model for each test fold. The tuned values of epsilon and cost for each model were also recorded. Earlier in this dissertation, it was observed that model parameter selection was unstable for such a small

sample size when choosing concomitant medications for atorvastatin and rosuvastatin. To increase robustness, the CV method was repeated 100 times; the mean and standard deviation of both cost and epsilon were calculated from the resultant 500 total folds. The average values of cost and epsilon chosen by the CV procedure were then used to train a final model fit for each type of kernel for both the atorvastatin and rosuvastatin cohorts.

4.4 Results

4.4.1 GAMs

Atorvastatin

The overall results of the GAM models for the atorvastatin cohort with the addition of all concomitant medications chosen by the selection algorithm described previously were similar to those found in the original linear regression model in the magnitude and direction of effect size for the parametric covariate estimates, but explained slightly more of the variance. The linear regression model is nested withing the generalized additive model: the results can be replicated by choosing linear functions for the continuous covariates smoothed in the non-linear model. This allows direct comparison of the model fits for the linear regression and GAM models using the F test[19]. The adjusted R^2 for the GAM with smoothing parameters chosen by CV ($GAM_{CV\lambda}$) was 0.663 ($\pm$ 0.028), compared to the adjusted R^2 value of 0.677 ($\pm$ 0.031) for the GAM with fixed smoothing parameter values ($GAM_{fixed\lambda}$). For reference, the original atorvastatin linear model without concomitant medications (with dose represented categorically) had an adjusted R^2 of 0.489 $\pm$ 0.040 (RMSE = 18.570 $\pm$ 9.691); the atorvastatin

linear regression model including all concomitant medications chosen using the previously described selection algorithm had an adjusted R^2 of 0.652 ± 0.027 (RMSE = 20.480 ± 8.642). While the $\text{GAM}_{\text{fixed}\lambda}$ explained slightly more of the model variation than the $\text{GAM}_{\text{CV}\lambda}$, the error was lower for the latter (RMSE_{CV} = 20.474 ± 8.626 versus $\text{RMSE}_{\text{fixed}\lambda}$ = 21.186 ± 9.066). Overall, the $\text{GAM}_{\text{CV}\lambda}$ fit was statistically significantly different from the fit of the linear model ($F(0.509, 96.491) = 8.163$, $p = 0.017$), but the $\text{GAM}_{\text{fixed}\lambda}$ fit was not ($F(5.701, 90.790) = 1.502$, $p = 0.189$). The difference in fit between the GAMs was not statistically significant after adjusting for multiple comparisons ($F(6.210, 90.790) = 2.048$, $p = 0.189$, method: Benjamini & Hochberg[20]), while the difference between the fit of the linear model and the $\text{GAM}_{\text{CV}\lambda}$ was only marginally short of statistical significance ($p = 0.052$) after applying the adjustment for multiple comparisons. The Benjamini & Hochberg method was chosen to gain additional statistical power, as it is a less conservative test than the Bonferroni correction[21].

Other differences in model fit were observed between the model fit with smoothing parameters chosen via CV, and the model fit using arbitrary fixed smoothing parameters. The parametric covariate estimates that were precise enough to achieve statistical significance in the linear model and both GAMs were *SLCO1B1* c.521T>C, atorvastatin dose (20mg, 40mg and 80mg), and concomitant use of losartan, metformin and tamsulosin. In the linear regression model, the coefficient estimates for candesartan, diclofenac, digoxin, levothyroxine and niacin trended towards statistical significance. In the GAMs, the estimates for candesartan and diclofenac confidence intervals sufficiently narrowed to achieve statistical significance. Specifically, the estimated atorvastatin plasma concentration for patients with concomitant use of candesartan compared to those without increased by a factor of 2.255 ($p = 0.034$) in the $\text{GAM}_{\text{CV}\lambda}$ fit compared to a non-statistically significant increase of a factor of 2.130 ($p = 0.051$)

in the linear regression model. A narrower confidence interval for this estimate was achieved in the $GAM_{fixed\lambda}$ the estimated atorvastatin plasma concentration increased by a factor of 2.377 in this model ($p = 0.027$). The estimated atorvastatin plasma concentration for patients taking diclofenac concomitantly with atorvastatin increased by a factor of 3.149 in the $GAM_{CV\lambda}$ ($p = 0.041$) compared to the non-statistically significant increase in the estimated plasma concentration by a factor of 2.924 ($p = 0.059$) in the linear regression model. Similar to the findings for candesartan, a narrower confidence interval was seen in the diclofenac coefficient estimate generated by the $GAM_{fixed\lambda}$ ($\hat{\beta} = 1.323$, $p = 0.019$). Finally, the relative confidence of the estimated effect of niacin on atorvastatin plasma concentration value differed between the linear regression model and the GAMs. The estimated effect size of niacin on predicted atorvastatin plasma concentration was -0.446 in the $GAM_{CV\lambda}$ ($p = 0.049$) compared to the estimated effect size of -0.392 ($p = 0.084$) seen in the linear regression model. Additional statistical significance was seen in the estimated effect size generated by the $GAM_{fixed\lambda}$ ($\hat{\beta} = -0.591$, $p = 0.012$). The model summary for the $GAM_{CV\lambda}$ parametric covariates is shown in Table 4.1 The $GAM_{fixed\lambda}$ parametric covariate summary is shown in Table 4.3.

The greatest differences observed between the $GAM_{CV\lambda}$ and $GAM_{fixed\lambda}$ fits were in the approximate significance and shape of the smoothed continuous covariates, which was expected. The model summary for the R `mgcv` package includes an estimate of how complex (wiggly) the smoothing function is for each non-parametric covariate, in the form of estimated degrees of freedom (eDF). In the $GAM_{CV\lambda}$, age, 4β-hydroxycholesterol, and BMI were not given complex smoothing functions, and were treated like parametric covariates (eDF = 1.000); despite their lack of a complex smoothing function, the age and 4β-hydroxycholesterol terms achieved statistical significance ($p < 0.05$). The only covariate given a non-linear smoothing function in

this model was time post dose, which also had a statistically significant effect (eDF = 1.509, F = 27.159 p <0.001). In contrast, all of the continuous covariates included in the $GAM_{fixed\lambda}$ for atorvastatin were given curved smoothing functions, because no penalty was applied to the use of these additional degrees of freedom when considering smoothing parameter value selection.

The model summary for the $GAM_{CV\lambda}$ non-parametric covariates is shown in Table 4.2, and the $GAM_{fixed\lambda}$ non-parametric covariate summary is shown in Table 4.4.

Table 4.1: Atorvastatin CV-smooth GAM parametric coefficients

	Estimate	Std. Error	P-Value	Sig.
(Intercept)	-0.443	0.200	0.029	*
SLCO1B1 c.521T>C	0.414	0.121	<0.001	***
SLCO1B1 c.388C>A	-0.150	0.100	0.136	
Dose (20mg)	0.757	0.207	<0.001	***
Dose (40mg)	1.117	0.184	<0.001	***
Dose (80mg)	1.602	0.211	<0.001	***
Gender (Male = 1)	-0.082	0.122	0.505	
Ethnicity (Non-Caucasian = 1)	0.057	0.178	0.750	
Acetylsalicylic Acid	0.194	0.136	0.158	
Atenolol	-0.268	0.218	0.222	
Candesartan	0.813	0.378	0.034	*
Diclofenac	1.147	0.553	0.041	*
Digoxin	0.580	0.320	0.073	.
Esomeprazole	0.491	0.418	0.243	
Gliclazide	-0.370	0.366	0.314	
Glucosamine	0.416	0.421	0.325	
Hydrochlorothiazide	0.115	0.185	0.535	
Levothyroxine	-0.380	0.228	0.098	.
Losartan	0.984	0.290	0.001	**
Metformin	-0.388	0.164	0.020	*
Misoprostol	0.153	0.529	0.773	
Nifedipine	0.328	0.340	0.337	
Tamsulosin	1.134	0.411	0.007	**
Valsartan	-0.076	0.345	0.826	
Venlafaxine	-0.258	0.356	0.470	
Vitamin B3	-0.446	0.224	0.049	*

Table 4.2: Atorvastatin GAM CV-smooth covariates (approximate significance)

Smoothed Covariate	Estimated DF	Reference DF	F	P-value	Sig.
s(Age)	1.000	1.000	10.516	0.002	**
s(4β-hydroxycholesterol)	1.000	1.000	18.265	<0.001	***
s(Time Post Dose)	1.509	1.851	27.159	<0.001	***
s(BMI)	1.000	1.000	1.911	0.170	

Table 4.3: Atorvastatin fixed-smooth GAM parametric coefficients

	Estimate	Std. Error	P-Value	Sig.
(Intercept)	-0.402	0.202	0.049307	*
SLCO1B1 c.521T>C	0.399	0.123	0.002	**
SLCO1B1 c.388C>A	-0.169	0.100	0.093	.
Dose (20mg)	0.735	0.209	<0.001	***
Dose (40mg)	1.118	0.187	<0.001	***
Dose (80mg)	1.544	0.215	<0.001	***
Gender (Male = 1)	-0.117	0.135	0.387	
Ethnicity (Non-Caucasian = 1)	0.0245	0.182	0.893	
Acetylsalicylic Acid	0.223	0.138	0.109	
Atenolol	-0.178	0.221	0.424	
Candesartan	0.866	0.385	0.027	*
Diclofenac	1.323	0.556	0.019	*
Digoxin	0.542	0.320	0.094	.
Esomeprazole	0.335	0.428	0.436	
Gliclazide	-0.411	0.367	0.266	
Glucosamine	0.414	0.425	0.332	
Hydrochlorothiazide	0.170	0.192	0.377	
Levothyroxine	-0.363	0.235	0.125	
Losartan	0.994	0.289	<0.001	***
Metformin	-0.419	0.170	0.015	*
Misoprostol	0.024	0.536	0.964	
Nifedipine	0.279	0.339	0.413	
Tamsulosin	1.191	0.421	0.006	**
Valsartan	-0.002	0.345	0.996	
Venlafaxine	-0.225	0.367	0.542	
Vitamin B3	-0.591	0.231	0.012	*

Table 4.4: Atorvastatin GAM fixed-smooth covariates (approximate significance)

Smoothed Covariate	Estimated DF	Reference DF	F	P-value	Sig.
s(Age)	2.686	3.356	3.381	0.020	*
s(4β-hydroxycholesterol)	2.521	3.165	5.684	0.001	**
s(Time Post Dose)	2.640	3.255	17.842	<0.001	***
s(BMI)	2.364	2.987	1.182	0.328	

Rosuvastatin

The linear model fit for the rosuvastatin cohort differed statistically significantly from the model fits of the rosuvastatin $\text{GAM}_{\text{CV}\lambda}$ ($F(0.331, 115.67) = 8.715$, $p = 0.038$) and the rosuvastatin $\text{GAM}_{\text{fixed}\lambda}$ ($F(4.856, 110.81) = 2.709$, $p = 0.038$), even after adjusting for multiple comparisons (method: Benjamini & Hochberg[20]). The rosuvastatin $\text{GAM}_{\text{CV}\lambda}$ and $\text{GAM}_{\text{fixed}\lambda}$ were not statistically significantly different, although this effect was marginal ($F(6.586, 108.82) = 1.997$, $p = 0.065$). The adjusted R^2 for the $\text{GAM}_{\text{CV}\lambda}$ was 0.640 ± 0.026 (RMSE = 16.259 ± 4.872) and 0.667 ± 0.028 for the $\text{GAM}_{\text{fixed}\lambda}$ (RMSE = 16.035 ± 4.614), while the adjusted R^2 for the linear rosuvastatin regression model was 0.634 ± 0.028 (RMSE = 16.314 ± 4.635).

The estimated parametric coefficient values for both GAM models were very similar to those seen in the linear regression for rosuvastatin; no additional covariate estimates achieved statistical significance that did not in the linear regression model, although the estimates differed slightly. The parametric covariate estimates that were statistically significant in all three models were for *SLCO1B1* c.521T>C, *ABCG2* c.421C>, and dose (10mg, 20mg and 40mg). The coefficient estimates for ethnicity and gender remained non-statistically significant as in the linear model, although in the rosuvastatin $\text{GAM}_{\text{fixed}\lambda}$ the gender coefficient estimate trended towards statistical significance ($\hat{\beta}$ = -0.204, $p = 0.065$), and was a great deal more precise than the estimate found in the linear regression model with dose represented categorically ($\hat{\beta}$ =

0.076, p = 0.587) and somewhat more precise than the estimate generated in the rosuvastatin $GAM_{CV\lambda}$ ($\hat{\beta}$ = -0.181, p = 0.107).

The treatment of the continuous variables in the rosuvastatin $GAM_{CV\lambda}$ was similar to that seen in the atorvastatin $GAM_{CV\lambda}$: the majority of the variables were treated linearly instead of being given smoothing curves that required more degrees of freedom. In contrast to the atorvastatin $GAM_{CV\lambda}$, time post dose was treated linearly for the rosuvastatin $GAM_{CV\lambda}$ (eDF = 1.000, p <0.001). BMI was also modelled using a linear function, and trended towards achieving statistical significance, unlike in the rosuvastatin linear regression model (eDF = 1.000, p = 0.094). The only continuous covariate given a non-linear smoothing function in the rosuvastatin $GAM_{CV\lambda}$ was age (eDF = 1.331, p = 0.005). As in the atorvastatin model, the rosuvastatin $GAM_{CV\lambda}$ did not penalize the use of curved smoothing functions to represent the continuous variables in the model. Qualitatively, the smoothing functions generated for the rosuvastatin cohort using fixed smoothing parameters were more curvy or "wiggly" than the smoothing functions in the atorvastatin $GAM_{fixed\lambda}$ (figures shown in Appendix C). Despite having a very slightly lower RMSE found via CV than the $GAM_{CV\lambda}$ (16.035 ± 4.614 versus 16.259 ± 4.872), the covariate estimates in the $GAM_{fixed\lambda}$ generally had wider 95% confidence intervals than those in the $GAM_{CV\lambda}$, based on the approximate p values provided in the GAM output. The optimization model error (ML) was also higher for the $GAM_{fixed\lambda}$ than the $GAM_{CV\lambda}$ (79.990 ± 3.563 versus 77.155 ± 3.707), suggesting that overfitting may be present in the model with arbitrary smoothing parameter values. The $GAM_{CV\lambda}$ results for the parametric and smoothed rosuvastatin covariates can be found in Tables 4.5 and 4.6 respectively. The $GAM_{fixed\lambda}$ results for the parametric and smoothed rosuvastatin covariates can be found in Tables 4.7 and 4.8 The results comparing the overall model fit for the atorvastatin and rosuvastatin cohorts is presented

in Table 4.9.

Table 4.5: Rosuvastatin CV-smooth GAM parametric coefficients

	Estimate	Std. Error	P-Value	Sig.
(Intercept)	1.446	0.2508	<0.001	***
SLCO1B1 c.521T>C	0.400	0.093	<0.001	***
ABCG2 c.421C>A	-0.348	0.119	0.004	**
Dose (10mg)	0.571	0.128	<0.001	***
Dose (20mg)	1.149	0.134	<0.001	***
Dose (40mg)	-0.863	0.154	<0.001	***
Gender (Male = 1)	-0.181	0.111	0.107	
Ethnicity (Non-Caucasian = 1)	-0.073	0.138	0.598	

Table 4.6: Rosuvastatin GAM CV-smooth covariates (approximate significance)

Smoothed Covariate	Estimated DF	Reference DF	F	P-value	Sig.
s(Age)	1.331	1.597	5.581	0.005	**
s(Time Post Dose)	1.000	1.000	15.646	<0.001	***
s(BMI)	1.000	1.000	2.852	0.094	.

Table 4.7: Rosuvastatin fixed-smooth GAM parametric coefficients

	Estimate	Std. Error	P-Value	Sig.
(Intercept)	1.506	0.245	<0.001	***
SLCO1B1 c.521T>C	0.397	0.092	<0.001	***
ABCG2 c.421C>A	-0.366	0.116	0.002	**
Dose (10mg)	0.557	0.126	<0.001	***
Dose (20mg)	1.145	0.131	<0.001	***
Dose (40mg)	-0.909	0.151	<0.001	***
Gender (Male = 1)	-0.204	0.109	0.065	.
Ethnicity (Non-Caucasian = 1)	-0.066	0.134	0.623	

Table 4.8: Rosuvastatin GAM fixed-smooth covariates (approximate significance)

Smoothed Covariate	Estimated DF	Reference DF	F	P-value	Sig.
s(Age)	2.930	3.683	4.294	0.003	**
s(Time Post Dose)	3.030	3.750	5.214	0.001	**
s(BMI)	2.228	2.750	1.285	0.270	

Table 4.9: CV results for atorvastatin and rosuvastatin GAMs

Model Cohort	Thin Plate CV Parameters			Thin Plate Fixed Parameters		
	RMSE	Adj. R^2	ML	RMSE	Adj. R^2	ML
Rosuvastatin	16.259	0.640	77.155	16.035	0.667	79.990
	± 4.872	± 0.026	± 3.707	± 4.614	± 0.028	± 3.563
Atorvastatin	20.474	0.663	74.698	21.186	0.677	80.087
	± 8.626	± 0.028	± 3.882	± 9.066	± 0.031	± 4.551

4.4.2 SVR

Atorvastatin

In order to find the optimal model parameters cost and epsilon for each kernel used to train the atorvastatin SVR models, a repeated 5-fold bootstrap CV method was employed. However, this procedure posed a problem with the atorvastatin data set including all selected concomitant medications. Because some of the atorvastatin model covariates were very sparse, it was not feasible to conduct cross validation with the dataset including the full complement of concomitant medications because of collinearity introduced by splitting the data into test and training sets. For a number of the medications that had particularly few patients with concomitant use, the randomly sampled test set comprising 20% of the available patient observations often failed to include a patient taking that particular medication. This in turn caused convergence issues because the column in the test dataset for that medication contained no variation. It was possible to train SVR models on the full atorvastatin dataset without CV, but this did not give as robust an estimate of model performance as assessing different permutations of the data.

In order to facilitate the SVR optimization to converge for models trained on the atorvastatin data set, concomitant medications with less than 10 patients taking them were excluded from the SVR to decrease model sparsity; the concomitant medications excluded were candesartan,

diclofenac, digoxin, esomeprazole, gliclazide, glucosamine, losartan, misoprostol, nifedipine, tamsulosin, valsartan and venlafaxine. The remaining concomitant medications included in the atorvastatin model were acetylsalicylic acid, atenolol, hydrochlorothiazide, levothyroxine, metformin, and niacin (vitamin B3). The performance of the full dataset was compared to the predictive of the reduced dataset using a two-step tuning procedure, since it was not feasible to obtain parameters estimated by CV for the full dataset. The first step of the manual tuning procedure was the same as that of the CV tuning procedure: a grid-search was conducted over epsilon values of $\epsilon = (0.1, 0.2, ..., 1.0)$ and cost values of $c = (2^2, 2^3, ..., 2^9)$. A graph was then generated of error values across the standard range of epsilon and cost values; darker areas of the graph represent parameter combinations resulting in lower error than light regions. Based on the graph colouring, the user then specified a more specific region for fine tuning of the parameter values. The second grid search was subsequently conducted between the minimum and maximum epsilon values specified by the user, increasing in intervals of 0.05, and between the minimum and maximum cost values specified by the user, increasing in intervals of 5. Another graph showing the model error for the range of epsilon and cost values in the grid search was generated from the second tuning procedure to show the size of the optimal regions for the more finely-tuned parameter combinations. The final values of epsilon and cost selected were those that produced the lowest error in the second grid-search of manual tuning procedure. Unsurprisingly, the reduced models tuned using the two-step procedure performed much more poorly than the models that included all of the concomitant medications. There are a number of potential reasons why this could be the case, the first of which is possible overfitting. In the linear regression model, the inclusion of the concomitant medications made a significant contribution to the amount of variability explained in the model, based on the adjusted R^2

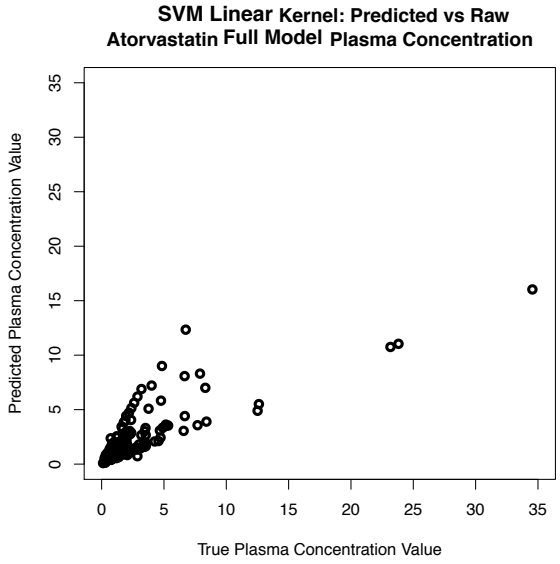
values from that analysis. However, additional parameters must be estimated in order to fit a SVR model. In combination with the relatively large number of sparse concomitant medication variables, it is possible that the model fits that were manually tuned fit our specific full dataset well, but would perform poorly on other datasets sampled from the same underlying distribution because the parameters are not generalizable. The polynomial kernel SVRs and the radial kernel SVR for the full concomitant medication model had performance comparable to that of the atorvastatin linear regression model: Degree 3 polynomial kernel RMSE = 23.533; Degree 5 polynomial kernel RMSE = 20.173; radial kernel RMSE = 16.805, versus RMSE = 20.480 ± 8.642 for the linear model. The final fit plots for the atorvastatin SVR models are shown in Figures 4.4, 4.64.7, 4.8 and 4.9.

Table 4.10: Atorvastatin SVR: manual tune model fit summary

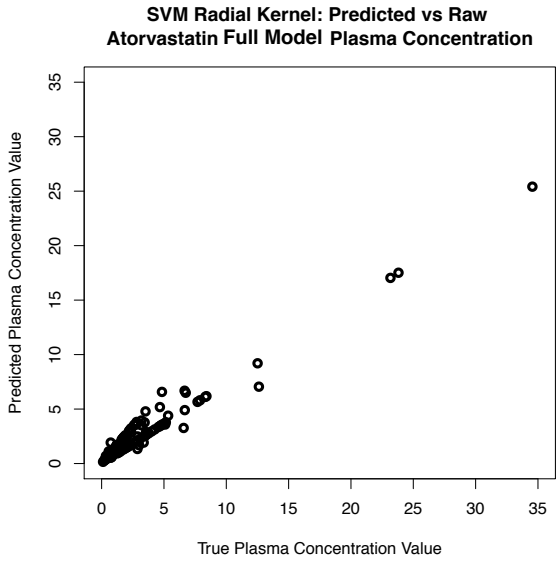
Kernel	Model	RMSE	Support Vectors	Epsilon	Cost
Linear	Reduced	45.532	57	0.55	30
	Full	32.342	37	0.75	320
Polynomial Degree 3	Reduced	38.381	60	0.55	5
	Full	23.533	86	0.4	15
Polynomial Degree 5	Reduced	39.048	72	0.50	10
	Full	20.173	86	0.35	345
Radial	Reduced	31.164	84	0.35	5
	Full	16.805	88	0.3	10

Table 4.11: Atorvastatin SVR CV summary (reduced model)

Kernel	RMSE	Support Vectors	Epsilon	Cost
Linear	18.964	65.464	0.348	157.848
	± 10.154	± 17.807	± 0.197	± 179.335
Polynomial Degree 3	20.553	56.372	0.546	22.304
	± 10.759	± 21.625	± 0.305	± 33.586
Polynomial Degree 5	20.561	59.898	0.514	144.466
	± 10.903	± 20.922	± 0.302	± 210.891
Radial	19.841	68.974	0.378	8.784
	± 10.376	± 19.492	± 0.265	± 9.132

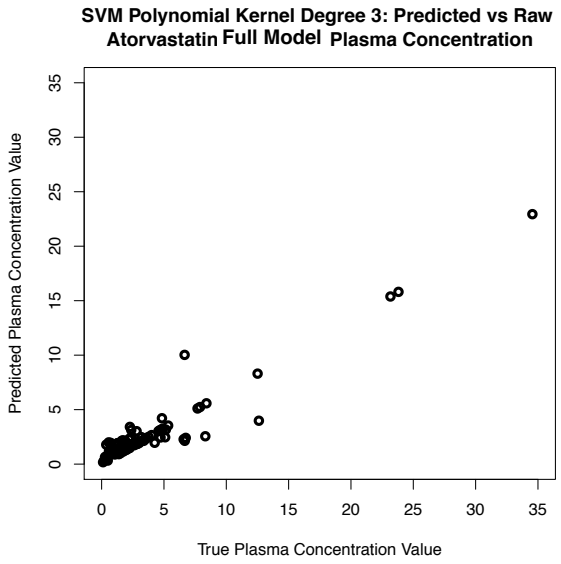


(a) Linear kernel

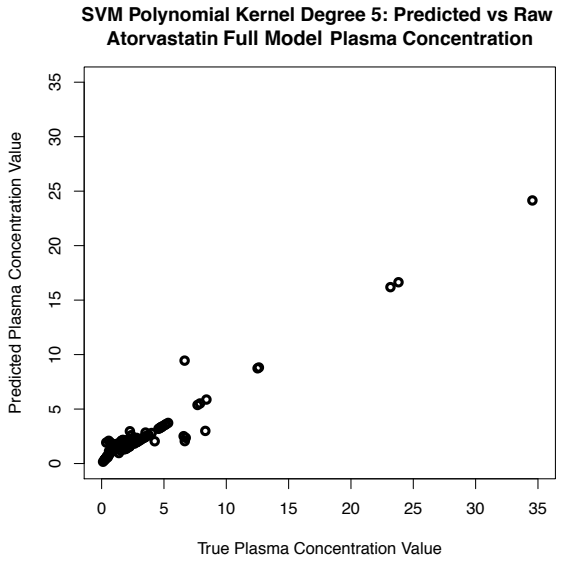


(b) Radial kernel

Figure 4.4: Atorvastatin SVR with all concomitant medications

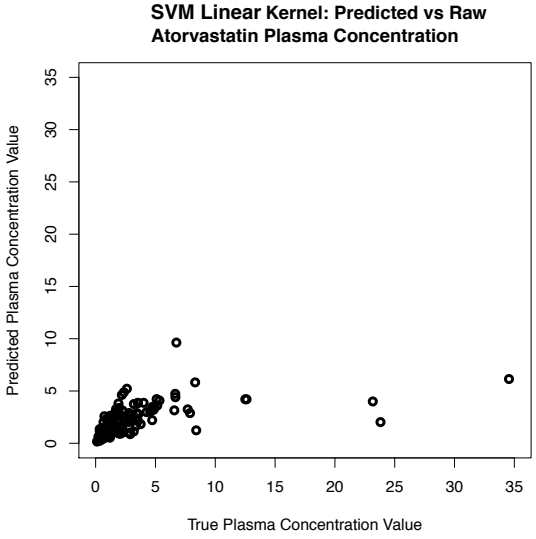


(a) Degree 3 polynomial kernel

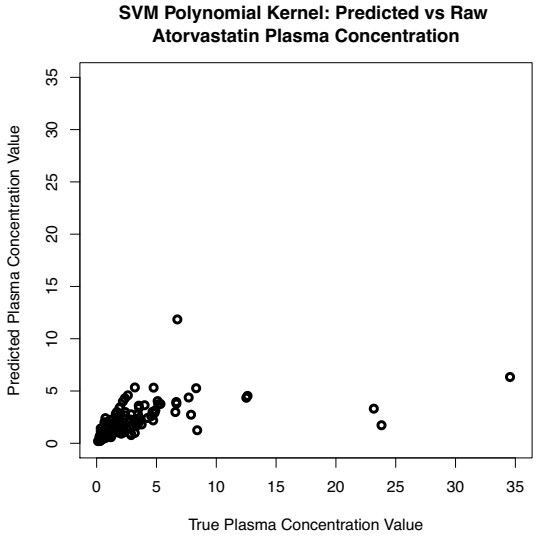


(b) Degree 5 polynomial kernel

Figure 4.5: Atorvastatin SVR with all concomitant medications

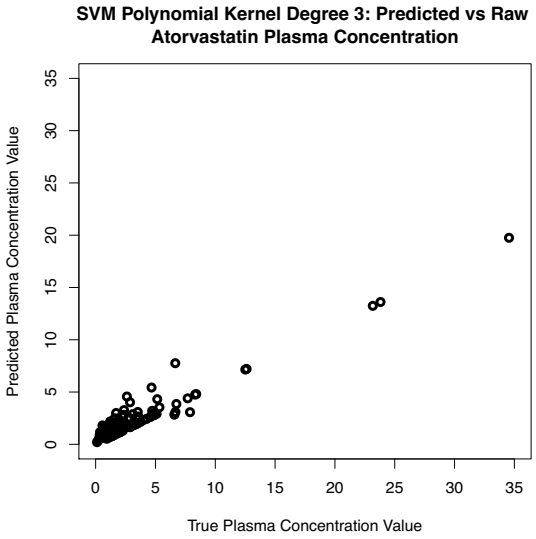


(a) CV parameter selection

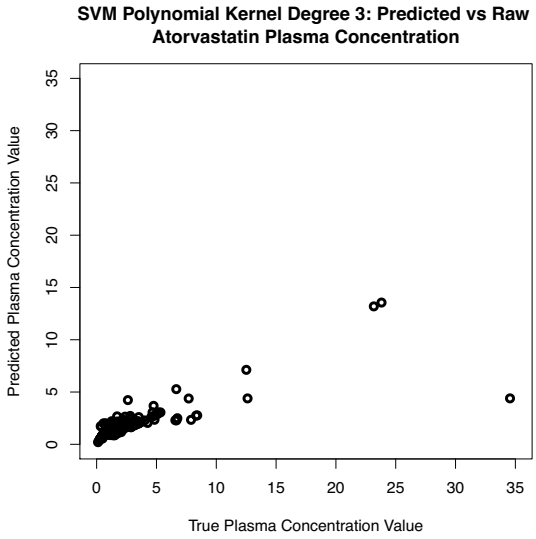


(b) Manual parameter selection

Figure 4.6: Atorvastatin reduced-model linear kernel SVR model fit

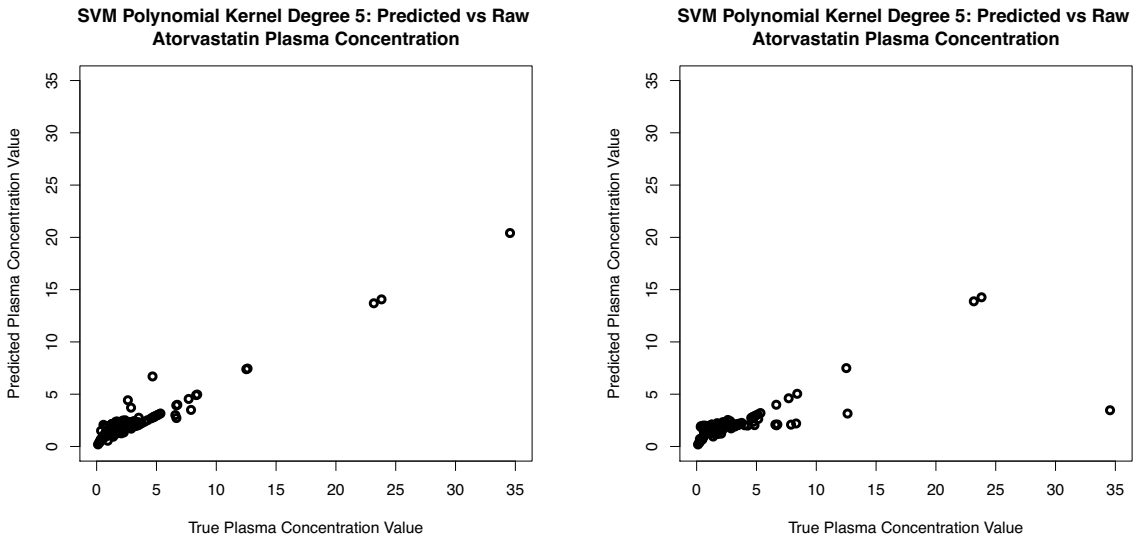


(a) CV parameter selection



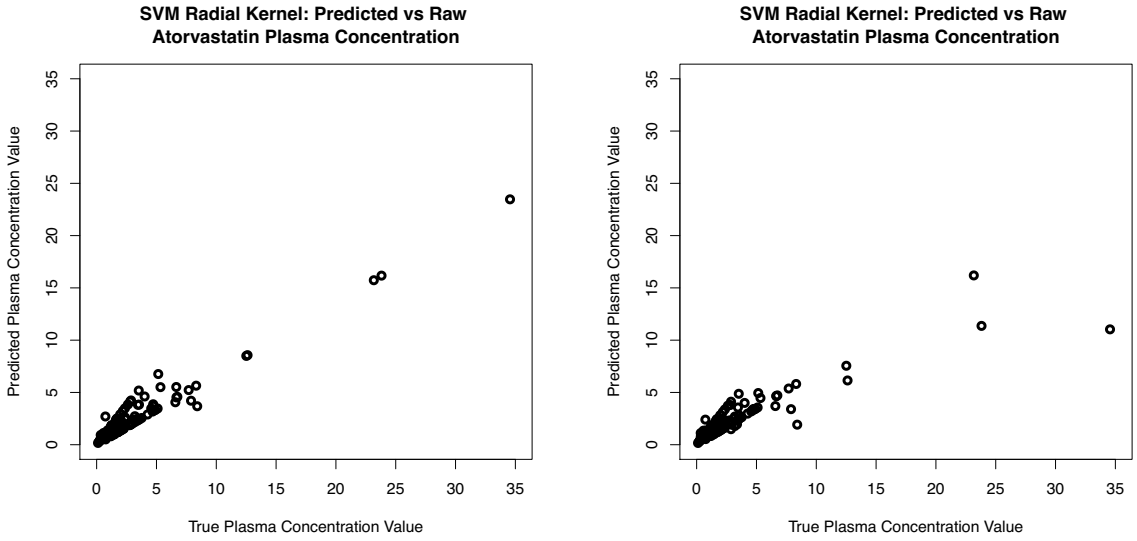
(b) Manual parameter selection

Figure 4.7: Atorvastatin reduced-model degree 3 polynomial kernel SVR model fit



(a) CV parameter selection (b) Manual parameter selection

Figure 4.8: Atorvastatin reduced-model degree 5 polynomial kernel SVR model fit



(a) CV parameter selection (b) Manual parameter selection

Figure 4.9: Atorvastatin reduced-model radial kernel SVR model fit

Rosuvastatin

The rosuvastatin SVR model fits assessed in the 5-fold CV procedure varied in quality in a manner similar to the model fits for the atorvastatin cohort. The model error for the rosuvastatin

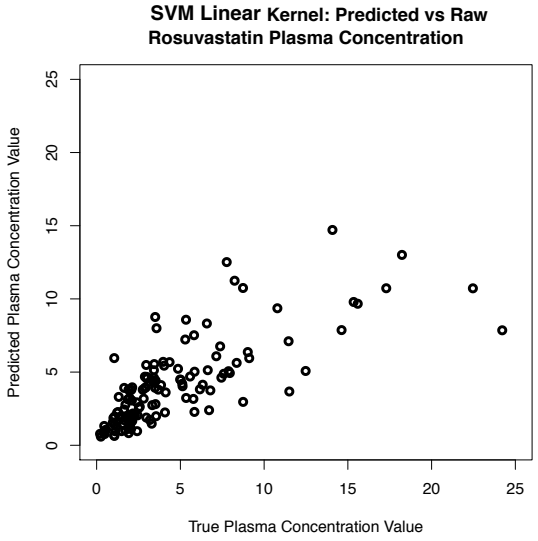
linear kernel SVR (RMSE = 16.864 ± 5.252) and the radial kernel SVR (RMSE = 17.715 ± 5.472) were comparable to the error seen in the rosuvastatin linear model using dose as a categorical covariate (RMSE = 16.314 ± 4.635). The linear kernel SVR had a very high amount of variability in the optimal cost parameters chosen during CV (c = 174.896 ± 179.610) in comparison to the average radial kernel model cost (c = 9.896 ± 13.574). In comparison to the linear and radial kernel SVRs for rosuvastatin, the polynomial kernel SVRs assessed in the CV analysis had substantially higher error, and comparably large variability in the quality of model fit (degree 3 polynomial kernel RMSE = 26.028 ± 21.760; degree 5 polynomial kernel RMSE = 25.907 ± 31.254). All of the rosuvastatin SVR model fits achieved by the two-step tuning procedure had substantially higher RMSE values than were found in the CV model fit assessment. The final model fits for the full rosuvastatin cohort are shown in Figures 4.10, 4.11, 4.12 and 4.13.

Table 4.12: Rosuvastatin SVR CV summary

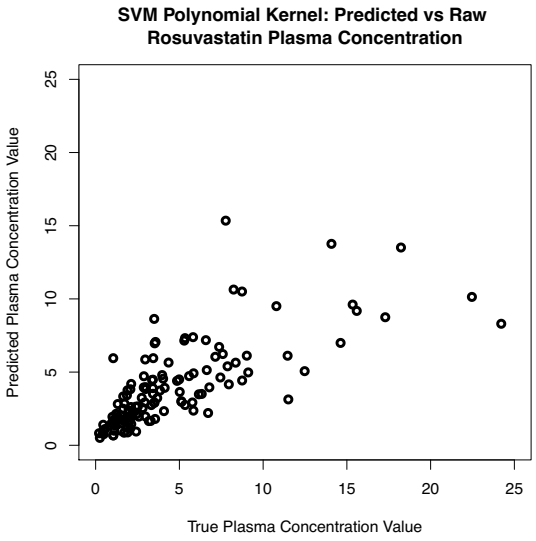
Kernel	RMSE	Support Vectors	Epsilon	Cost
Linear	16.864	43.636	0.526	174.896
	± 5.252	± 14.409	± 0.137	± 179.610
Polynomial Degree 3	26.028	68.250	0.379	17.912
	± 21.760	± 16.534	± 0.211	± 30.486
Polynomial Degree 5	25.907	64.298	0.489	23.976
	± 31.254	± 25.085	± 0.359	± 40.836
Radial	17.715	73.694	0.285	9.896
	± 5.472	± 13.268	± 0.127	± 13.574

Table 4.13: Rosuvastatin SVR: manual tune model fit summary

Kernel	RMSE	Support Vectors	Epsilon	Cost
Linear	34.384	29	0.75	70
Polynomial Degree 3	28.280	64	0.55	70
Polynomial Degree 5	34.861	65	0.60	80
Radial	29.151	83	0.35	5

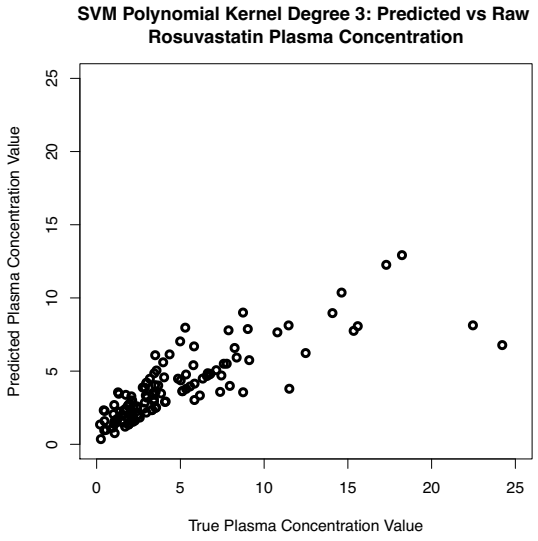


(a) CV parameter selection

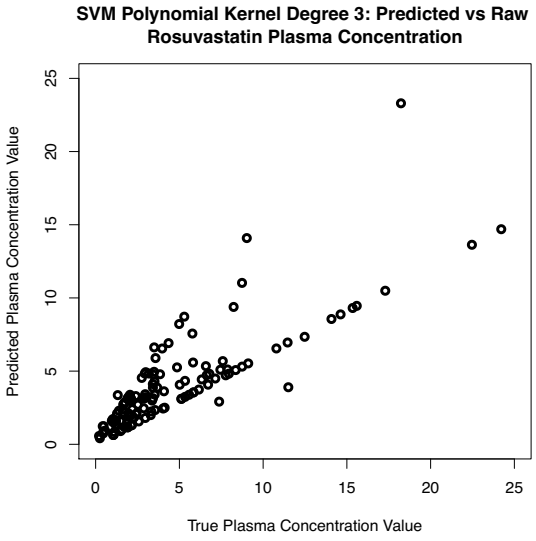


(b) Manual parameter selection

Figure 4.10: Rosuvastatin linear kernel SVR model fit

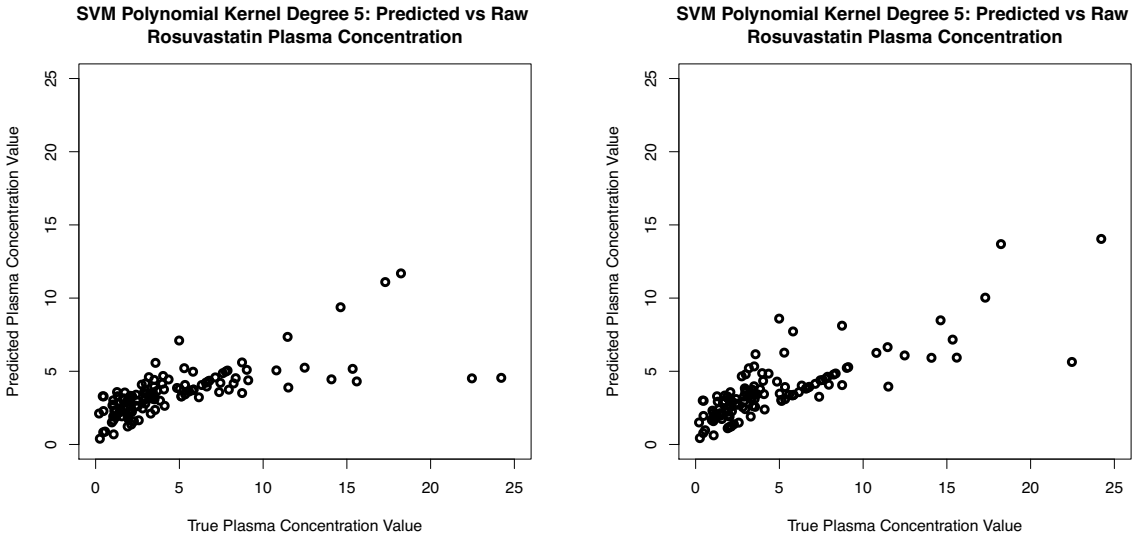


(a) CV parameter selection



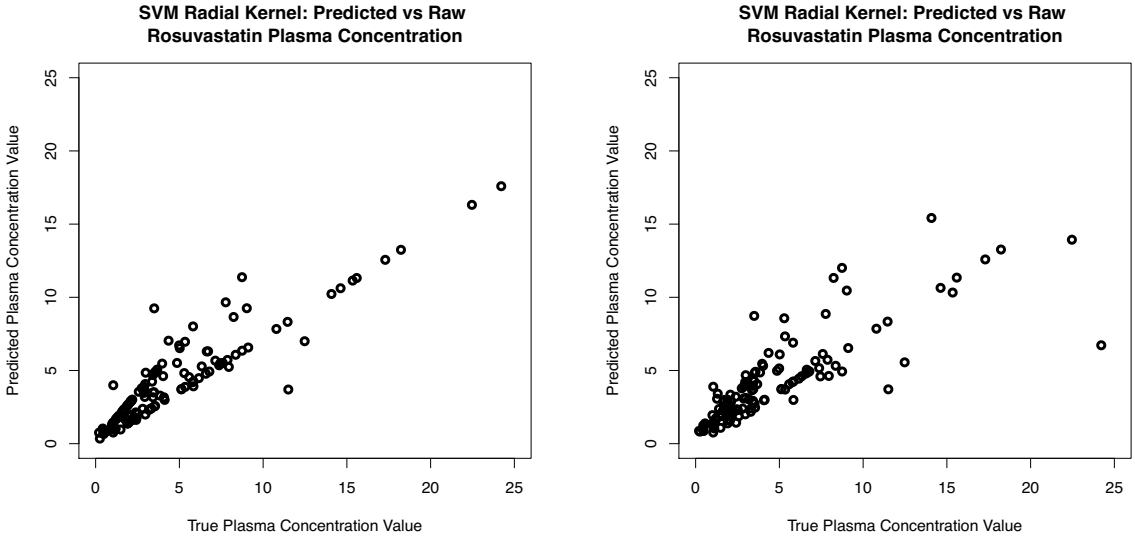
(b) Manual parameter selection

Figure 4.11: Rosuvastatin degree 3 polynomial kernel SVR model fit



(a) CV parameter selection (b) Manual parameter selection

Figure 4.12: Rosuvastatin degree 5 polynomial kernel SVR model fit



(a) CV parameter selection (b) Manual parameter selection

Figure 4.13: Rosuvastatin radial kernel SVR model fit

4.5 Discussion

4.5.1 GAMs

Most of the differences between the $GAM_{CV\lambda}$ fit compared to the $GAM_{fixed\lambda}$ fit appear to stem from the trade-off between the degrees of freedom used and the amount of smoothing given to the continuous covariates. In the CV parameter model, the use of non-linear smoothing parameters for the continuous covariates was penalized to a larger extent than in the $GAM_{fixed\lambda}$, in which no such penalization was present. The increase in quality of model fit for the atorvastatin $GAM_{CV\lambda}$ over the linear model fit can thus be largely explained by the non-linear adjustment to the modelling of time post dose. Based on the combined performance of the CV-smoothed and fixed-parameter smoothed GAMs, it appears that the true effect sizes for candesartan and diclofenac are larger than the coefficient estimates observed in the linear regression model. This can be inferred based on increased statistical significance seen in the GAMs that resulted in higher estimated changes in the predicted atorvastatin plasma concentration.

Both of the fitted GAMs for the rosuvastatin model statistically significantly improved model fit. While the two fitted GAMs did not differ significantly, the fixed smoothing parameter rosuvastatin GAM offered a slight improvement in model fit over the smoothing parameters chosen via CV, while this was not seen with the atorvastatin GAMs. A possible explanation for this is that the rosuvastatin model contains fewer parametric and non-parametric covariates than the atorvastatin GAM, thus using fewer degrees of freedom in the model to perform additional coefficient estimates. Because fewer coefficient estimates are needed in the rosuvastatin GAMs, using additional degrees of freedom on non-linear smoothing parameters for the continuous covariates has less of a negative impact on model error. This is not taken into

consideration when choosing smoothing parameters using CV, although this method is certainly the most consistent approach. A consequence of having arbitrarily smoothed functions for each covariate is additional noise in the model, which can be seen both in the increased RMSE found in the CV for the $\text{GAM}_{\text{fixed}\lambda}$, and the wider confidence intervals of the smoothing functions based on the resultant approximate p values. The sample size used to construct the GAMs for both cohorts was also small; it is possible that with a larger data set, using curved smoothing functions instead of linear functions for the continuous covariates would carry less risk of overfitting, and consequently be penalized less. The graphs comparing the smooth curves between the $\text{GAM}_{\text{CV}\lambda}$ and the $\text{GAM}_{\text{fixed}\lambda}$ for both atorvastatin and rosuvastatin may be found in Appendix B.

4.5.2 SVR

Both the atorvastatin and rosuvastatin SVRs had better overall fit when trained using parameter obtained via CV. However, performance varied widely between folds for some of the kernel types, like the polynomial kernels and the linear kernel. The linear kernel SVR for the atorvastatin cohort had a mean cost of 157.848 ± 179.335; similarly, the degree 5 polynomial kernel SVR had a mean cost of 144.466 ± 210.891. This phenomenon was also observed in the rosuvastatin linear kernel SVR, which had a mean cost of 174.896 ± 179.610. The RMSE for these models was comparable to the performance of the linear regression model, but the wide variation in cost across folds suggests that the optimal parameters for these kernel types are very sensitive to differences between individual data sets, which is not ideal for a predictive model. In contrast, the radial kernel SVR had relatively low variability in cost across folds for

both the atorvastatin and rosuvastatin datasets, and performance comparable to that of the linear regression models. Because of this, the radial kernel SVR would likely be the most stable choice for training a predictive model using SVR to predict statin plasma concentration in a clinical setting.

Overall, the performance of the SVRs was not significantly better than the performance of the linear regression models or GAMs. A potential reason for this is that it was possible to obtain a reasonably accurate estimate of the effect on plasma concentration for the full complement of selected concomitant medications for the atorvastatin cohort in the linear regression and GAM. The concomitant medications improved model fit, and seemed to explain a good deal of the variability of the upper tail, where the highest errors were for the reduced models. It was possible to obtain reasonable effect estimates for all of these covariates in the linear models because no extra parameters were necessary for model fit. However, this type of sparsity is problematic when training SVR models, because extra parameters specifying model fit must be selected in order to achieve a fit with reasonable accuracy. The large RMSE values observed for all kernels types of the atorvastatin SVR compared to the error in the linear regression model were likely due to one outlying patient with a particularly high atorvastatin plasma concentration. This patient's plasma concentration was predicted poorly by all of the SVR models trained on the reduced concomitant medication atorvastatin dataset.

In general, the performance of the model parameters chosen via CV resulted in less overall error than the models trained manually, perhaps because of the small size of the data set. Selecting overall tuning parameters on such a small dataset by CV did not give stable results for some of the kernels, where performance varied widely for different parameter combinations of cost and epsilon. At this time, using the linear regression model or GAM to guide clinical deci-

sions on statin dosing for new patients seems to be a better option than using SVR. If more data were available, it is possible that the SVR would offer increased predictive performance, since the ratio of parameter values to be estimated versus the number of available training observations would be more reasonable. Additionally, the SVR would have the advantage of being able to manually set the cost value higher to reduce model error for patients who were under-predicted using the original linear regression mode developed by DeGorter et al., in order to achieve conservative dosing recommendations.

4.6 Conclusions

Modelling atorvastatin plasma concentration with GAMs resulted in improved model fit over the linear regression models, as it allowed for modelling non-linearity in the continuous variables where advantageous. Most of the smoothing values chosen by CV had a low number of degrees of freedom, and were smoothed to the point of being linear. However, time post dose and age were given non-linear smooth fits for the atorvastatin and rosuvastatin groups, respectively.

At this time, SVR does not appear to be a feasible modelling strategy for this dataset because of the small number of observations available, compared to the required number of model parameters to be fit. At best, the predictive performance approached the accuracy of the linear regression model. However, the linear regression model and GAMs are more easily interpretable than the SVR results, which are a further reason to prefer using them over SVR when it does not offer a substantial advantage in predictive performance.

References

[1] Marianne K DeGorter, Rommel G Tirona, Ute I Schwarz, Yun-Hee Choi, George K Dresser, Neville Suskin, Kathryn Myers, GuangYong Zou, Otito Iwuchukwu, Wei-Qi Wei, et al. Clinical and pharmacogenetic predictors of circulating atorvastatin and rosuvastatin concentration in routine clinical care. Circulation: Cardiovascular Genetics, 6(4):400–408, 2013.

[2] Markus Gulilat, Anthony Tang, Steven E Gryn, Peter Leong-Sit, Allan C Skanes, Jeffrey E Alfonsi, George K Dresser, Sara L Henderson, Rhiannon V Rose, Daniel J Lizotte, et al. Interpatient variation in rivaroxaban and apixaban plasma concentrations in routine care. Canadian Journal of Cardiology, 33(8):1036–1043, 2017.

[3] Laura E Jansen, Wendy A Teft, Rhiannon V Rose, Daniel J Lizotte, and Richard B Kim. Cyp2d6 genotype and endoxifen plasma concentration do not predict hot flash severity during tamoxifen therapy. Breast cancer research and treatment, pages 1–8, 2018.

[4] Simon Wood. Generalized additive models. Chapman & Hall/CRC, 2006.

[5] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. The elements of statistical learning. Springer-Verlag, 2009.

[6] Trevor Hastie and Robert Tibshirani. Generalized additive models. Statistical Science, 1(3):297–310, 1986.

[7] Werner Vach. Regression models as a tool in medical research. CRC Press, 2012.

[8] Simon N Wood. Generalized additive models: an introduction with R. CRC press, 2017.

[9] Jean Duchon. Splines minimizing rotation-invariant semi-norms in sobolev spaces. In Lecture Notes in Mathematics, volume 571, pages 85–100. Springer, 1977.

[10] Simon N Wood. Thin plate regression splines. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 65(1):95–114, 2003.

[11] Katharina Morik, Peter Brockhausen, and Thorsten Joachims. Combining statistical learning with a knowledge-based approach: a case study in intensive care monitoring. Technical report, Technical Report, SFB 475: Komplexitätsreduktion in Multivariaten Datenstrukturen, Universität Dortmund, 1999.

[12] Raphael Pelossof, Andrew Miller, Peter Allen, and Tony Jebara. An SVM learning approach to robotic grasping. In Robotics and Automation, 2004. Proceedings. ICRA'04. 2004 IEEE International Conference on, volume 4, pages 3512–3518. IEEE, 2004.

[13] Inge S Helland. On the interpretation and use of r2 in regression analysis. Biometrics, 43(1):61–69, 1987.

[14] Lesley F Leach and Robin K Henson. The use and impact of adjusted r2 effects in published regression research. Multiple Linear Regression Viewpoints, 33(1):1–11, 2007.

[15] Ping Yin and Xitao Fan. Estimating r 2 shrinkage in multiple regression: a comparison of different analytical methods. The Journal of Experimental Education, 69(2):203–224, 2001.

[16] R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2014. Available online at `http://www.R-project.org/`.

[17] Simon N Wood. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73(1):3–36, 2011.

[18] David Meyer, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel, and Friedrich Leisch. e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien, 2015. R package version 1.6-7.

[19] Julian J Faraway. Extending the linear model with r, 2006.

[20] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological), 57(1):289–300, 1995.

[21] Shawn R Narum. Beyond Bonferroni: less conservative analyses for conservation genetics. Conservation Genetics, 7(5):783–787, 2006.

Chapter 5

Background: Next Generation Sequencing

Next Generation Sequencing (NGS) of patient genomes has gained popularity as a way to study genetic risk factors for disease, and to guide treatment regimes based on the prognosis suggested by genetic factors. It is only within the past decade that NGS technologies have become efficient, affordable, and accessible enough to allow researchers to study on a widespread scale. One of the largest and most well-known projects that has emerged in this area was the 1000 genomes project[1], in which researchers proposed to sequence the genomes of 1000 volunteers from around the globe. The objective of the project was to catalogue variability in the human genome, including rare variation. The now-completed final product includes the genetic sequencing information of 1092 individuals from 14 different populations, and is a resource for researchers studying genetic origins of disease[2].

The cost of genomic sequencing was originally prohibitive both in terms of financial resources and time required to produce results. Fortunately advances in technology have brought the cost of the first attempt to sequence the genome from 3 billion dollars over 12 years down to 1000 dollars and a single day of processing time using the current capabilities[3]. With genetic

sequencing becoming more feasible for large-scale research, some of the main challenges in the field are efficiently organizing, analysing and interpreting the huge amounts of data produced by this technology.

5.1 DNA Structure

Many NGS technologies function by synthesizing or amplifying genomic DNA and learning the sequence in the process. DNA synthesis is a complex process, and understanding the basic structure of DNA is critical for understanding the data generated using these technologies. The basic molecules and structure of DNA will be described in the following section, along with relevant information on protein coding and its applicability to detecting disease based on genetic variation between patients.

5.1.1 Basic Structure

Genomic DNA is two-stranded, and has a double helix structure that resembles a ladder. The sides of the ladder are composed of alternating deoxyribose sugar and phosphate groups, and and the rungs of the ladder are formed of the four nitrogenous bases adenosine (A), thymine (T), cytosine (C) and guanine (G). These are attached to the sides of the ladder in pairs: A pairs with T, and C pairs with G (Figure 5.1). Each pair of nitrogenous bases is held together by hydrogen bonds, but the two strands can be separated (like a zipper) for duplication and DNA synthesis[4].

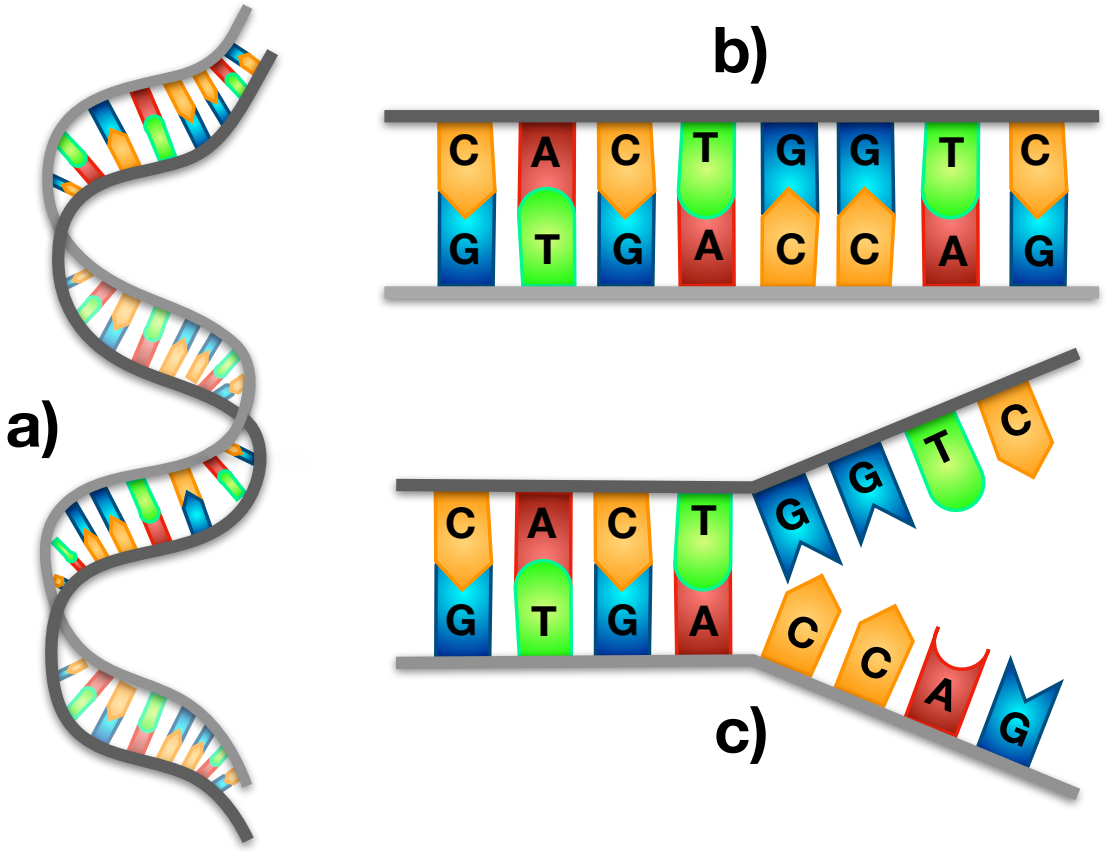


Figure 5.1: DNA Structure: a) double helix form, b) straightened, c) strands separated

5.1.2 Protein Coding and Polymorphisms

DNA contains the blueprint for proteins to be synthesized within the body; proteins consist of sequences of amino acid residues. Segments of DNA within the genome that encode proteins

to be expressed are termed exons; the intervening segments are intronic, or introns[5]. Both introns and genetic material between genes are called non-coding regions. The function of the eventual protein to be synthesized depends on its folded 3D shape, which in turn depends on the nucleotides in the exonic DNA sequence being in the correct order. When polymorphisms (alterations) occur in the sequence, it can cause the protein to fold improperly. Proteins are built by constructing a chain of amino acids, which are in turn coded for in DNA or RNA in chunks of three nucleotide bases. These sets of three nucleotide bases are termed "codons", and there are 64 in total. The majority of these code for amino acids; however, 3 of 64 codons are signals for termination[5]. Polymorphisms in DNA can change which amino acid is inserted into the peptide chain, which can in turn change the way the protein folds. If a polymorphism results in a terminal codon accidentally being read, the rest of the peptide chain to be folded into the final 3-D protein will not be produced.

Depending on which bases have been changed, the polymorphism may increase, decrease or negate the functionality of the protein relative to its "wild" (most common) type; however, polymorphisms are often synonymous and do not result in changes in the expressed protein. Polymorphisms in exonic DNA sequences that alter protein function are termed non-synonymous (in which case the protein is altered because of amino acids being changed) or non-sense (in which the protein is truncated prematurely because a stop codon has)[6]. When only one base in the sequence is changed, it is termed a single nucleotide polymorphism (SNP). Polymorphisms that occur in intronic DNA segments can alter the amount of protein produced, but do not alter the sequence of amino acids[6]. Longer sequences of genes can also be altered; these are termed structural variations[7].

Genetic polymorphisms can take different forms: some examples are substitutions, dele-

tions, insertions, copy number variations, translocations, frameshift alterations, and inverted sequences[5] (shown in Figure 5.2). A substitution occurs when a nucleotide in the sequence is exchanged for another in comparison to a reference genome. Deletions occur when the sequence is missing a base compared to a reference genome; similarly, insertions occur when an extra base is present in the sequence compared to a reference genome. In addition to single bases being inserted or deleted, larger sections of DNA can also be affected in this manner[5]. Together, these types of polymorphisms are termed indels. Importantly, insertions and deletions of bases in the reference genome can cause frame-shift polymorphisms, wherein the read frame is shifted forward or back, in turn causing subsequent codons to be misread as well. Insertions and deletions are less likely to cause as much change when they are in sets of three, as subsequent codons will not be misread. If the added codon(s) are not terminal, a single amino acid will be added or removed from the peptide chain for each inserted sequence of three[5].

Copy number variations occur when a single base or section of the reference genome has a different number of copies in the DNA sequence; deletions are a special case of copy number variations where fewer copies of the base are present in the DNA[6]. Translocation polymorphisms occur when a single base or larger section of DNA is in a different location in the sequence compared to the reference genome. Similarly, inverted sequence variations can occur where the bases in a larger region of DNA are reversed in the order found in the reference genome. By observing where in the genome these changes occur and their frequency in the population, individual sequence changes can be analyzed and associated with the risk of different diseases.

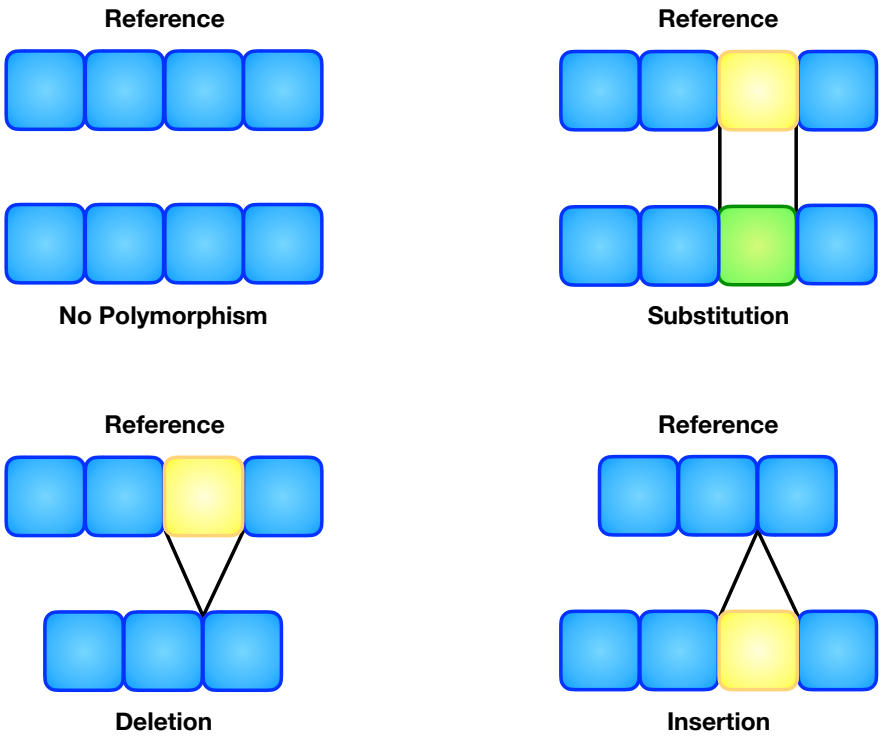


Figure 5.2: Different types of structural variation (polymorphisms)

5.2 NGS Data and Workflow

The process of acquiring NGS data requires three main processing and analysis components.

The first phase of NGS data acquisition involves the benchwork and preprocessing of the DNA

sample and obtaining DNA sequence reads using NGS[8]. The secondary step in NGS data processing uses various algorithms and computational processes to align the DNA sequence obtained in the primary analysis to a reference genome, and determine where variation is present in the sequenced DNA. The final component of NGS processing seeks to interpret the information generated in the previous two steps and determine whether variants present in the DNA have clinical relevance[8].

However, the low-level methodological details of how to use processed NGS data for the purpose of predictive modelling are not often discussed in the research literature concerning NGS, which can make using these techniques daunting for researchers that have not dealt extensively with NGS analysis. Additionally, fewer tools are available for prioritizing data from non-coding regions (introns) of the genome for predicting disease risk or drug response than for exomes, which have multiple established software options for variant annotation[5]. The following section provides a brief overview of the workflow required to obtain and process NGS data, and the resulting contents of the data available for analysis from a data-science perspective.

5.2.1 Primary Processing

During the genetic sequencing process, each DNA fragment may "read" a number of different times. After sequencing, all of the recorded fragment reads must be aligned and then compared to a reference genome to check for variants[5]. The number of times a particular nucleotide in a certain position is identified is called read depth or sequence coverage, and provides a measure of confidence in the sequencing accuracy for that particular location[9]. The read depth may vary

substantially between different nucleotides even within the same fragment of DNA. Overall indicators of the quality of coverage include average read depth, and minimum read depth; user of the latter was recommended by Muzzey et al. because 50 reads generally gives a very high confidence in the accuracy of the nucleotide identified, and further reads beyond this tend to increase the cost of sequencing more than they improve accuracy[3].

5.2.2 Secondary and Tertiary Processing

The second step in NGS analysis is to identify which nucleotides in the sequenced DNA have changed in comparison to the reference genome, and for each change, what type of variation is present[8]. This process is called variant calling, and developing optimal strategies for identifying these changes is an active area of research at present. Variant calling covers a wide range of methodologies, including the use of support vector machines (SVMs), Bayesian approaches[10], hidden Markov models (HMMs) and other machine learning techniques[11]. The exact statistical methods used for variant calling depends on the sample size available (often limited); whether the goal is to identify common variants (commonly defined as those found in >5% of the population[12]) or rare variation[6;13]; and the section of the genome to be sequenced[14]. This process is challenging because there are potentially many machine-produced artifacts from the sequencing process that could be falsely identified as relevant variants[7]. The Sequence/Alignment Mapping (SAM) format and its complementary binary format (BAM) are frequently-used alignment formats that were designed to perform efficient and accurate identification of true variants, and are produced in a typical NGS workflow for use in further processing[15]. The process of properly aligning the sequence to the reference genome and

calling relevant variants can also involve filtering out variants that do not result in a protein structure change, variants that do not cause a frameshift in alignment, duplicate segment data, and variants that have been previously studied in other databases[16].

After variants have been identified by comparing the newly sequenced DNA to the reference genome, the tertiary step of the workflow is to identify if any have clinical relevance to the substantive problem application. Various annotation techniques are available to help with this process, and they use existing genomic databases to flag the likelihood of each variant to cause changes to the structure of the protein that it is associated with. ANNOVAR is an example of a frequently-used annotation tool which is able to draw on previously disseminated genomic knowledge from a wide variety of sources that conform to a specific format (Generic Feature Format version 3 (GFF3); two of the major sources of annotation information are the 1000 Genomes Project and dbSNP[16]. Importantly, ANNOVAR also draws scores predicting the functional impact of polymorphisms from other software. One program used is SIFT (Sorting Intolerant from Tolerant[17]), which is an algorithm predicting the effect of amino acid substitutions in sequenced data. PolyPhen-2 is another algorithm used by ANNOVAR that detects missense variations that can have deleterious effects, and uses a Naive Bayes classifier to accomplish this task[18].

While the information provided by ANNOVAR allows clinical researchers to more easily prune the available SNPs for inclusion in statistical modelling and decision support, several drawbacks make using this information infeasible for some research goals. Firstly, ANNOVAR deals only with SNPs (substitution, insertion and deletion), and thus is not well suited to studying larger structural variation in DNA sequences. Additionally, annotation methods like SIFT and PolyPhen-2 do not actually predict disease risk, but protein nonfunction and while

there is significant overlap, protein nonfunction is not a wholly accurate proxy for the risk of disease acquisition[14]. Another difficulty in using previously annotated information about variation present in patient samples is that almost all of the annotated variants in these databases are for polymorphisms in coding regions of DNA. Because exonic DNA comprises only approximately a small fraction of the genome[17], much less information is available on the frequency and predicted effects of non-coding variants comparatively. Despite this, the CADD (Combined Annotation-Dependent Depletion) score evaluates the likelihood of variants to impact disease risk for both intronic and exonic regions of DNA[19;20]. Mutations in exonic regions of DNA are easier to study, and many tools have evolved for looking at only this portion of the genome; indeed, Whole Exome Sequencing (WES) is a commonly-used and less expensive alternative to sequencing an individual's entire genome[21]. A similar method, targeted exome sequencing (TES), further decreases the costs of sequencing by only sequencing a pre-selected panel of protein-coding genes relevant to the specific variant identification problem of interest[22]. Unfortunately, identifying clinically relevant variants is still a hard problem with exonic data, particularly when the variants observed in the data have not been previously annotated[14]. Additionally, it has been found that whole genome sequencing (WGS) actually has a higher level of uniformity in detecting SNPs and indels, and thus WGS may be more efficient despite the fact that it currently incurs greater financial cost to perform[21]. Ultimately, the choice to use WGS, WES or TES is dependent on the research question posed, the resources available for sequencing, the study design and individual researcher preferences; this is also true for the choice of statistical method used for variant calling and variant prioritization.

5.3 NGS Statistical Methods

Since the advances in cost- and time-effective genome sequencing using NGS techniques, more than 2000 polymorphisms in DNA that are associated with disease have been identified in genome wide association studies (GWAS)[23]. Polymorphisms that are present in a relatively large portion of the population are more likely to be identified in GWAS as being associated with disease risk or drug response because a higher minor allele frequency (MAF) offers more power when looking at the associating of each SNP individually with disease. In contrast, polymorphisms with MAF below 5% are much harder to detect in this type of analysis, because achieving adequate statistical power for these analyses would require a prohibitively large sample size. Additionaly, the technique of examining the association between each rare variant and disease risk separately requires thorough correction for multiple testing[23]. Fortunately many options for RV identification have been developed since the advent of lower resource-intensive sequencing techniques. Statistical techniques for identifying both rare and common variation in NGS data will be discussed in the following section, with an emphasis on challenges associated with small data sets.

5.3.1 Rare Variant Association Analysis

Two models for the effect of genetic variants on complex disease have been proposed: the Common Disease Common Variant (CDCV) hypothesis, and the Common Disease Rare Variant (CDRV) hypothesis. The CDCV theory postulates that many common variants with small to moderate effects are the main cause of common diseases; in contrast, the CDRV hypothesis states that complex disease is caused by rare variants with a large effect on phenotype. It has

been observed that a more likely causal scenario involves the contribution of both common and rare variants to the development of complex disease[24].

Although associating rare genetic variation with phenotype or disease state poses methodological challenges, researchers have hypothesized that RVs can add significant insight to the knowledge already gleaned from identifying common variants[13], particularly when the impact of the variant is strong[25]. Common variants are commonly defined as those with the MAF occurring in more than 5% of the population. However, there is no consensus as to the exact definition of what constitutes "rare" variation. In previous work, RVs have been variously defined as MAF below 0.5%[26], below 1%[13;27;28], below 3%[29], below 5%[12], or below an unspecified cutoff[24]. The practice of setting an exact threshold to define which variants should be considered rare has been noted to be arbitrary and is dependent on the characteristics of the disease and genes under study[25]. Regardless of the differing definitions of what constitutes rarity between studies, many methods have been developed to increase the power to identify RVs in sequenced data and relate these to disease status or phenotype. These techniques range from the use of different sampling practices to advanced statistical modelling and correction.

5.3.2 Extreme Phenotype Sampling

One of the challenges associated with studying rare genetic variation in association with phenotype or disease risk is the difficulty of detection; often in order to observe the rare variation, infeasibly large samples would be required[27]. A way of circumventing the need to sequence a prohibitive number of samples is to use Extreme Phenotype Sampling (EPS); in this method phenotypic extremes of the disease spectrum of interest are sequenced in order to increase MAF

of the RVs in the sampled DNA[30;27]. EPS is also alternatively known as trait dependent sampling[31]. EPS boosts the power of RV detection by increasing the likelihood that variants related to the condition of interest will be present in the sequenced DNA sampled and often results in much higher risk ratios observed for RVs than those observed in typical GWAS[32]. An example of a study that has successfully used EPS to identify RVs was performed by Lange and colleagues; the authors were able to identify RVs associated with low density lipoprotein (LDL-C) levels by sequencing patients with extremely high cholesterol and extremely low cholesterol and comparing their genetic profiles[33]. Extremes in this context are usually defined as the 5^{th} and 95^{th} percentiles of the trait distribution[28]. EPS has been promoted as much more powerful and cost effective alternative to random patient sampling to identify RVs[29;34]. However, it has been noted that Type I error rates and bias may be inflated when using EPS with classic linear regression, particularly when stratifying on more than one phenotypic trait. To deal with this, it has been suggested that the use of iterative maximum likelihood estimation (MLE) optimization can decrease the probability of false positives when detecting RVs in association with extreme phenotypes[31]. Other strategies have also been suggested to improve performance for RV identification with EPS: Li et al. proposed a two stage approach in which the extreme tails of the sample are tested, followed by the patients with non-extreme phenotypic values[34]. The authors of this work also suggested an "almost-extreme" sampling method wherein the most extreme values at the tails are discarded because of increased risk of measurement error and variation, and the analysis is performed on the tails of the truncated distribution[34].

5.3.3 Burden Tests

While testing the association between single SNPs and phenotype or disease risk works quite well for common variants, this method is severely underpowered for RVs because of the reduced likelihood of seeing the variant in any given patient population sampled[35]. Alternative methodologies for identifying rare variants can be roughly divided into either burden tests, or nonburden tests[29]. Both types of analysis combine the rare variation in a prespecified region of sequenced data (such as a gene, or moving window of a fixed size), but differ in how the variation is represented. Burden tests capture the number of variants in a given region, and generally assume that each variant in a region affects the risk of disease in the same direction (deleterious or protective) and with the same magnitude[29]. Often the variants in a region are thresholded based on their rarity, so that all of the variants collapsed into a single indicator for a given region have equivalent frequency[33]. The Combined Multivariate and Collapsing Method (CMC) is a commonly used burden test formulation developed by Li and colleagues[24] that combines a method of collapsing rare variation with a multivariate test. In the Cohort Allelic Sums Test (CAST) collapsing method[36], each case or control is given a variable that indicates whether that individual has any rare variation (one or more copies of the variant allele) at the specified location:

$$X_i = \begin{cases} 1 & \text{if rare variation present} \\ 0 & \text{otherwise} \end{cases}$$

The amount of variation in a region can be tested between cases and controls in this paradigm using the χ^2 test for the collapsed measure of variation[24]. In the multivariate test paradigm, all genetic variants in a given region are tested for association with disease risk or phenotype si-

multaneously. Genetic variation is defined slightly differently than in the collapsing paradigm, with specific indicators for one or two copies of the variant allele being present in a given location j for patient case i[24]:

$$X_{ij} = \begin{cases} 1 & \text{if 0 variant alleles} \\ 0 & \text{if 1 variant allele} \\ -1 & \text{if 2 variant alleles} \end{cases}$$

Y_{ij} is defined similarly for non-cases (controls) in the sample. Using this notation, the vector of genotypes for patient i is given by $X_i = (X_{i1}, ..., X_{iM})^T$, where M is the number of variant sites. Similarly, $Y_i = (Y_{i1}, ..., Y_{iM})^T$ denotes the vector of M genotypes for patient i of the control population. We can also define a vector of genotypes $\bar{X}_j$ and $\bar{Y}_j$ for the case population and control population respectively at variant site j, adjusted for the number of cases and controls present in the sample:

$$\bar{X}_j = \frac{1}{N_A} \sum_{i=1}^{N_A} X_{ij} \qquad \bar{Y}_j = \frac{1}{N_{\bar{A}}} \sum_{i=1}^{N_{\bar{A}}} Y_{ij}.$$

Hotelling's T^2 test can be used to determine whether the presence of variation differs between cases and controls in the multivariate test paradigm:

$$T^2 = \frac{N_A N_{\bar{A}}}{N_A + N_{\bar{A}}} (\bar{X} - \bar{Y})^T S^{-1} (\bar{X} - \bar{Y})$$

where S denotes the covariance matrix for the indicator variables across variants, $\bar{X} = (\bar{X}_1, ..., \bar{X}_M)$; $\bar{Y} = (\bar{Y}_1, ..., \bar{Y}_M)$, and N_A and $N_{\bar{A}}$ denote the number of cases and controls, respectively[24]. Conceptually, in the above formula $\bar{X}$ and $\bar{y}$ represent the "mean" genotype over the cases and

controls, respectively. The CMC method combines these strategies of capturing genetic variation by first collapsing over a group of markers of genetic variation according to predefined criteria such as MAF, and then applying the multivariate test to the groups of collapsed markers. The burden test score can alternatively be written more simply in the following form[37]:

$$Q = \left[\sum_{j=1}^{m} w_j \sum_{i=1}^{n} (Y_i - \widehat{\mu_{i,0}}) X_{ij} \right]^2$$

where n is the number of sequenced subjects, m is the number of variants in the region of study, Y_i is the phenotypic outcome, $\hat{\mu_0}$ is a vector containing the estimated probabilities of the outcome phenotype under the null hypothesis, and w_j is the pre-specified weight for variant j. Note that this form of the burden test score differs considerably from the above T^2 equation; in this formulation, Y indicates cases versus controls, and the covariate matrix X is defined for all subjects.

Burden tests offer a substantial increase in power to detect the effect of RVs in regions where the variants have the same effect on the outcome of disease. However, in cases where the RVs have effects in opposite directions using burden tests that collapse variation in this way can have a severely detrimental effect on detection power. This is because the effects of variants in a region (positive and negative) towards the risk of disease can cancel each other out and thus fail to be identified[27]. Additionally, the CMC method does not easily allow the inclusion of other covariates of interest (such as family history or other clinical factors that might affect patient phenotype), which is a substantial modelling drawback[24].

5.3.4 Nonburden Tests

Burden tests are a powerful tool in circumstances where RVs are predictive of a phenotype to the same extent and with the same effect (deleterious or protective)[24]. However, when RVs in a region of DNA are being aggregated, it is more probable that both deleterious and protective RVs will be present; at minimum, it would be very difficult to ensure that only RVs affecting phenotype in the same direction would be present in any given subsection of RVs to be aggregated. Nonburden tests aggregate variation over prespecified regions of DNA, but unlike burden tests they allow for expression of different qualitative and quantitative effects of RVs on disease risk. The Sequence Kernel Association Test (SKAT) is an example of a nonburden test that provides a measure of association between phenotype and rare genetic variation.

Sequence Kernel Association Test

The Sequence Kernel Association Test (SKAT) captures rare genetic variation by calculating a *P* value for each region of DNA over which variants are aggregated. For the linear modelling context, SKAT follows the classic regression formulation:

$$y_i = \alpha_0 + \boldsymbol{\alpha}' \boldsymbol{X_i} + \boldsymbol{\beta}' \boldsymbol{G_i} + \epsilon_i$$

where the phenotype of individual i is denoted by y_i; α_0 is a constant intercept; $\boldsymbol{X_i}$ is a vector of m covariates to be adjusted for; $\boldsymbol{G_i} = (G_{i1}, G_{i2}, ..., G_{ip})$ denotes the genotype for each individual at each of p variant locations; ϵ_i is an error term; $\boldsymbol{\alpha}'$ is the vector of regression coefficients for the m adjustment covariates; and $\boldsymbol{\beta}$ is our quantity of interest: the vector of regression coefficients for each of the p variants[38]. The null hypothesis $\boldsymbol{\beta} = 0$ is tested using a variance-

component score statistic Q, defined as:

$$Q = (\boldsymbol{y} - \hat{\mu})' \boldsymbol{K} (\boldsymbol{y} - \hat{\mu})$$

where $\hat{\mu}$ denotes the predicted mean of the outcome variable y under the null hypothesis and $\boldsymbol{K} = \boldsymbol{G}\boldsymbol{W}\boldsymbol{G}'$ is the kernel matrix where genotype is denoted by $\boldsymbol{G}$. Within the kernel matrix, $\boldsymbol{W}$ represents the weight given to the genotype for a particular variant. Alternatively, the SKAT score statistic can be written as[37]:

$$Q_{\rho=0} = \sum_{j=1}^{m} w_j^2 \left[\sum_{i=1}^{n} (Y_i - \widehat{\mu_{i,0}}) X_{ij} \right]^2$$

where m is the number of the number of variants in a region, n is the number of sequenced individuals, and w_j is the weight for variant j. In this formulation, ρ denotes the correlation structure of the variants and the outcome phenotype; in the case of SKAT, the variants need not all have the same effect on the outcome and so $\rho = 0$. The score statistics for burden and the SKAT described above look very similar, because the equations differ only by the treatment of the weighting on the variants; the weights are squared for the SKAT only. Conceptually, the score statistic describes the amount of variation in y that is associated with genetic polymorphisms.

Choosing a good pre-specified weight for each variant can increase the power to detect RV associations with phenotype; however, in practice it is unknown which variants should be weighted more heavily because they are more predictive of patient phenotype. Wu et al. propose to use a beta distribution density function with parameter values related to the MAF in

the observed sample for each variant[38]. Based on the CDRV theory, variants that are rare are assumed to have a larger effect and thus should be weighted more heavily than variants with a higher MAF; however, weights can be chosen according to whatever theory the user ascribes to. The effect of weighting is similar to a prior distribution; it is easier to detect effects of variants that are given high weight, but that makes variants with lower weight more difficult to detect.

SKAT Reformulations

SKAT is a generalization of the C-alpha test, which uses permutation to calculate the p values for aggregate variants in a region of DNA[38]. Either continuous or binary outcomes may be used with the SKAT. Permutation methods involve resampling the available data (similar to the bootstrap); while they can greatly enhance accuracy, they also incur significant computational burden. SKAT has the advantage of not relying on computationally expensive permutation methods to determine the association between variants and phenotype, and additionally allows the user to control for relevant covariates[38]. The SKAT offers substantial improvements in power over burden tests when variants in a region have varying qualitative effects on disease risk (deleterious, null and protective). However, when variants in a region all influence the risk of disease in a similar direction and magnitude, burden tests have the advantage of power and thus the original SKAT formulation is not optimal[39]. To maximize power in both situations, an optimal SKAT formulation (SKAT-O) was developed by Lee et al to modulate the test methodology depending on the correlation structure of the data, denoted by ρ[39]:

$$Q_\rho = (1-\rho)Q_{\text{SKAT}} + \rho Q_{\text{Burden}}$$

The SKAT-O is an additive linear combination of burden testing and the SKAT; in the SKAT-O, variant regions with more homogenous effects are calculated using the burden score statistic, and variant regions with heterogeneous effects are treated using the original SKAT score statistic[39]. Lee et al. give a strategy for searching over multiple values of ρ while controlling the Type I error rate[39].

Another extension of the SKAT accommodates testing both common and rare variants in a region. In this formulation (the RC-SKAT), the regression equation is further broken down with separate weighting schemes for rare and common variants[37]:

$$g[E(Y_i)] = \alpha_0 + \boldsymbol{\alpha}'\boldsymbol{X_i} + \boldsymbol{\beta_1}\boldsymbol{R_i} + \boldsymbol{\beta_2}\boldsymbol{C_i}$$

where $g(\cdot)$ is a link function; α_0 is the intercept for the linear model; $\boldsymbol{X}_i$ is the covariate design matrix for subject i; $\boldsymbol{R}_i$ is the matrix of rare variants for subject i; $\boldsymbol{C_i}$ is the matrix of common genetic variants; and $\boldsymbol{\alpha}'$,$\boldsymbol{R_i}$ and $\boldsymbol{C_i}$ are the coefficient vectors for the covariates, rare variants, and common variant effects respectively. The effects of the common and rare variants on disease risk are assessed jointly using an weighted sum of the score test statistics for each. Under the CDRV theory, genetic variations that are rare are hypothesized to have stronger causal effects on disease risk than common variations, and thus the common and rare variants in this model are given different weighting schemes that reflect this[37]. Other alternatives for modelling the joint effect of common and rare variants on disease risk proposed by the authors who developed the RC-SKAT are an adaptive sum test, or Fisher's Combination Method, instead of a simple weighted sum. Finally, SKAT has also been reformulated to accommodate small sample sizes, which are often used by necessity in WGS and WES because of the prohibitive costs

involved in sequencing a large number of subjects. The original formulation of the SKAT has been found to be conservative for small sample sizes, which already suffer from low power[40]. The SKAT adjustment for small sample sizes increases the power of the test by adjusting for the skew and kurtosis of the sample[40].

Kernel Choice

One of the fundamental strengths of the SKAT is the use of a kernel function to test variant association; the kernel test evaluates pairwise genetic and trait similarities for all of the patients and variants in the sample, and the level of similarity is then used as a proxy for genotype/phenotype association[35]. The linear kernel is a popular choice for constructing a similarity matrix. The linear kernel formula is given by

$$K(Z_i, Z_{i'}) = Z_i' Z_{i'}.$$

Although only linear kernel machines are usually discussed in the context of the SKAT, they are not always a good fit for the data. Many kernels can be used for this purpose, and examples of other candidate kernels for identifying associations in NGS data include the quadratic kernel:

$$K(Z_i, Z_{i'}) = (Z_i' Z_{i'} + 1)^2$$

and the IBS kernel[35]:

$$K(Z_i, Z_{i'}) = (2p)^{-1} \sum_{j=1}^{2} (2 - |Z_{ij} - Z_{i'j}|).$$

The associations between the SNPs in Z_i and y_i are tested using the test score statistic

$$Q = \frac{(\boldsymbol{y} - \widehat{\boldsymbol{y}_0})' \boldsymbol{K} (\boldsymbol{y} - \widehat{\boldsymbol{y}_0})}{\widehat{\sigma}_0^2}.$$

Using an inappropriate type of kernel can severely reduce the power of the test to identify associated SNPs. One potential method for choosing a kernel would be to simply test all of the kernels that may be appropriate given the distribution and assumptions underlying the data and use the kernel with the greatest statistical significance. However, as in most statistical paradigms, this type of multiple testing can lead to an increased risk of falsely identifying associations in the data. Additionally, it is possible to combine multiple kernels as well as choosing between them. Wu and colleagues developed a testing framework for choosing the best possible kernel or combination of kernels for use in genetic association studies[35]. In this paradigm, a P value can be calculated for a weighted average of kernels directly, or using perturbation-based inference to compare the performance of different candidate kernels for the set of SNPs being tested. In this case, P values are compared instead of a test statistic because the test scores can be scaled in very different ways depending on the kernel used, which makes direct comparison of those statistics infeasible[35].

5.3.5 DoEstRare Rare Variant Identification

The previously described methods seek to identify RVs associated with disease risk by evaluating their frequency and associated estimated burden irrespective of the exact position of the variant's location. However, it is also possible to take advantage of specific positional information for variants to identify associations with disease risk or phenotype. Several methods

have been developed for this purpose, but the method using this strategy found to be the most effective for identifying RVs most recently is the Density-oriented Estimation for Rare-Variant positions (DoEstRare) test[25]. DoEstRare accomplishes this by using a weighting function to represent the overall variant frequency and kernel methods to compare the position density of the variants in cases versus controls, in line with the following hypotheses:

$$H_0 : f^A = f^U \quad \text{AND} \quad p^A = p^U$$

$$H_1 : f^A \neq f^U \quad \text{OR} \quad p^A \neq p^U$$

where f denotes the mutation position density function for affected individuals A (cases) and unaffected individuals U (controls), and p denotes the average allele frequency for affected individuals A (cases) and unaffected individuals U (controls)[25]. The test score for DoEstRare is given by

$$\int_1^{Lg} |\widehat{p^A} \times \widehat{f^A}(pos) - \widehat{p^U} \times \widehat{f^U}(pos)| \quad dpos$$

where Lg is the gene's length, the mean allele frequencies are estimated by $\widehat{p^A}$ and $\widehat{p^U}$, and the position density functions are estimated by $\widehat{f^A}$ and $\widehat{f^U}$. While this method is extremely powerful for identifying disease associated RVs in large sample sizes that have a relatively good representation of rare MAFs in the data, the performance of this method deteriorates when faced with extremely small sample sizes. In the case where few RVs are present in the sample, the location information is irrelevant because there generally will not be enough positional overlap between variants in different individuals[25].

References

[1] Nayanah Siva. 1000 genomes project, 2008.

[2] 1000 Genomes Project Consortium et al. An integrated map of genetic variation from 1,092 human genomes. Nature, 491(7422):56–65, 2012.

[3] Dale Muzzey, Eric A Evans, and Caroline Lieber. Understanding the basics of NGS: from mechanism to variant calling. Current Genetic Medicine Reports, 3(4):158–165, 2015.

[4] Dee Unglaub Silverthorn, William C Ober, Claire W Garrison, Andrew C Silverthorn, and Bruce R Johnson. Human physiology: an integrated approach (6th edition). Pearson/Benjamin Cummings San Francisco, CA, USA:, 2013.

[5] Xinkun Wang. Next-generation sequencing data analysis. CRC Press, 2016.

[6] Stephen Chanock. Technologic issues in GWAS and follow-up studies, 2007.

[7] Mark A DePristo, Eric Banks, Ryan Poplin, Kiran V Garimella, Jared R Maguire, Christopher Hartl, Anthony A Philippakis, Guillermo Del Angel, Manuel A Rivas, Matt Hanna, et al. A framework for variation discovery and genotyping using next-generation dna sequencing data. Nature Genetics, 43(5):491–498, 2011.

[8] Amy S Gargis, Lisa Kalman, David P Bick, Cristina Da Silva, David P Dimmock, Birgit H Funke, Sivakumar Gowrisankar, Madhuri R Hegde, Shashikant Kulkarni, Christopher E Mason, et al. Good laboratory practice for clinical next-generation sequencing informatics pipelines. Nature Biotechnology, 33(7):689–693, 2015.

[9] Can Alkan, Bradley P Coe, and Evan E Eichler. Genome structural variation discovery and genotyping. Nature Reviews Genetics, 12(5):363–375, 2011.

[10] Christopher T Saunders, Wendy SW Wong, Sajani Swamy, Jennifer Becq, Lisa J Murray, and R Keira Cheetham. Strelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs. Bioinformatics, 28(14):1811–1817, 2012.

[11] Diksha Garg, Ankita Jiwan, and Shailendra Singh. Computational approaches for variant identification. International Journal of Computer Applications, 165(8):18–24, 2017.

[12] Chengqing Wu, Kyle M Walsh, Andrew T DeWan, Josephine Hoh, and Zuoheng Wang. Disease risk prediction with rare and common variants. In BMC Proceedings, volume 5, pages S61–S65. BioMed Central, 2011.

[13] Yao-Hwei Fang and Yen-Feng Chiu. A novel support vector machine-based approach for rare variant detection. PloS One, 8(8):e71114 (1–9), 2013.

[14] Manuel Giollo, David T Jones, Marco Carraro, Emanuela Leonardi, Carlo Ferrari, and Silvio CE Tosatto. Crohn disease risk prediction best practices and pitfalls with exome data. Human Mutation, 38:11931200, 2017.

[15] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The sequence alignment/map format and samtools. Bioinformatics, 25(16):2078–2079, 2009.

[16] Kai Wang, Mingyao Li, and Hakon Hakonarson. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Research, 38(16):e164–e164, 2010.

[17] Ngak-Leng Sim, Prateek Kumar, Jing Hu, Steven Henikoff, Georg Schneider, and Pauline C Ng. SIFT web server: predicting effects of amino acid substitutions on proteins. Nucleic Acids Research, 40(W1):W452–W457, 2012.

[18] Ivan A Adzhubei, Steffen Schmidt, Leonid Peshkin, Vasily E Ramensky, Anna Gerasimova, Peer Bork, Alexey S Kondrashov, and Shamil R Sunyaev. A method and server for predicting damaging missense mutations. Nature Methods, 7(4):248–249, 2010.

[19] Hui Yang and Kai Wang. Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. Nature Protocols, 10(10):1556–1566, 2015.

[20] Yuval Itan, Lei Shang, Bertrand Boisson, Michael J Ciancanelli, Janet G Markle, Ruben Martinez-Barricarte, Eric Scott, Ishaan Shah, Peter D Stenson, Joseph Gleeson, et al. The mutation significance cutoff: gene-level thresholds for variant predictions. Nature Methods, 13(2):109–110, 2016.

[21] Aziz Belkadi, Alexandre Bolze, Yuval Itan, Aurélie Cobat, Quentin B Vincent, Alexander Antipenko, Lei Shang, Bertrand Boisson, Jean-Laurent Casanova, and Laurent Abel. Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. Proceedings of the National Academy of Sciences, 112(17):5473–5478, 2015.

[22] Illumina Inc. Introduction to targeted gene sequencing, 2018.

[23] Seunggeung Lee, Gonçalo R Abecasis, Michael Boehnke, and Xihong Lin. Rare-variant association analysis: study designs and statistical tests. The American Journal of Human Genetics, 95(1):5–23, 2014.

[24] Bingshan Li and Suzanne M Leal. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. The American Journal of Human Genetics, 83(3):311–321, 2008.

[25] Elodie Persyn, Matilde Karakachoff, Solena Le Scouarnec, Camille Le Clézio, Dominique Campion, Jean-Jacques Schott, Richard Redon, Lise Bellanger, Christian Dina, French Exome Consortium, et al. DoEstRare: A statistical test to identify local enrichments in rare genomic variants associated with disease. PloS One, 12(7):e0179364 (1–21), 2017.

[26] Hayato Tada, Masa-aki Kawashiri, Tetsuo Konno, Masakazu Yamagishi, and Kenshi Hayashi. Common and rare variant association study for plasma lipids and coronary artery disease. Journal of Atherosclerosis and Thrombosis, 23(3):241–256, 2012.

[27] Ian J Barnett, Seunggeun Lee, and Xihong Lin. Detecting rare variant effects using extreme phenotype sampling in sequencing association studies. Genetic Epidemiology, 37(2):142–151, 2013.

[28] Paul L Auer and Guillaume Lettre. Rare variant association studies: considerations, challenges and opportunities. Genome Medicine, 7(1):16–26, 2015.

[29] Ya-Jing Zhou, Yong Wang, and Li-Li Chen. Detecting the common and individual effects of rare variants on quantitative traits by using extreme phenotype sampling. Genes, 7(1):2–13, 2016.

[30] Ron Do, Sekar Kathiresan, and Gonçalo R Abecasis. Exome sequencing and complex

disease: practical aspects of rare variant association studies. Human Molecular Genetics, 21(R1):R1–R9, 2012.

[31] Dan-Yu Lin, Donglin Zeng, and Zheng-Zheng Tang. Quantitative trait analysis in sequencing studies under trait-dependent sampling. Proceedings of the National Academy of Sciences, 110(30):12247–12252, 2013.

[32] David Gurwitz and Howard L McLeod. Genome-wide studies in pharmacogenomics: harnessing the power of extreme phenotypes. Pharmacogenomics, 14(4):337–339, 2013.

[33] Leslie A Lange, Youna Hu, He Zhang, Chenyi Xue, Ellen M Schmidt, Zheng-Zheng Tang, Chris Bizon, Ethan M Lange, Joshua D Smith, Emily H Turner, et al. Whole-exome sequencing identifies rare and low-frequency coding variants associated with ldl cholesterol. The American Journal of Human Genetics, 94(2):233–245, 2014.

[34] Dalin Li, Juan Pablo Lewinger, William J Gauderman, Cassandra Elizabeth Murcray, and David Conti. Using extreme phenotype sampling to identify the rare causal variants of quantitative traits in association studies. Genetic Epidemiology, 35(8):790–799, 2011.

[35] Michael C Wu, Arnab Maity, Seunggeun Lee, Elizabeth M Simmons, Quaker E Harmon, Xinyi Lin, Stephanie M Engel, Jeffrey J Molldrem, and Paul M Armistead. Kernel machine SNP-set testing under multiple candidate kernels. Genetic Epidemiology, 37(3):267–275, 2013.

[36] Stephan Morgenthaler and William G Thilly. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums

test (cast). Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis, 615(1):28–56, 2007.

[37] Iuliana Ionita-Laza, Seunggeun Lee, Vlad Makarov, Joseph D Buxbaum, and Xihong Lin. Sequence kernel association tests for the combined effect of rare and common variants. The American Journal of Human Genetics, 92(6):841–853, 2013.

[38] Michael C Wu, Seunggeun Lee, Tianxi Cai, Yun Li, Michael Boehnke, and Xihong Lin. Rare-variant association testing for sequencing data with the sequence kernel association test. The American Journal of Human Genetics, 89(1):82–93, 2011.

[39] Seunggeun Lee, Michael C Wu, and Xihong Lin. Optimal tests for rare variant effects in sequencing association studies. Biostatistics, 13(4):762–775, 2012.

[40] Seunggeun Lee, Mary J Emond, Michael J Bamshad, Kathleen C Barnes, Mark J Rieder, Deborah A Nickerson, David C Christiani, Mark M Wurfel, Xihong Lin, NHLBI GO Exome Sequencing Project, et al. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. The American Journal of Human Genetics, 91(2):224–237, 2012.

Chapter 6

Identifying Novel Genetic Polymorphisms to Model Rosuvastatin Plasma Concentration

6.1 NGS Patient Selection

In order to identify SNPs associated with prediction quality for patients taking rosuvastatin in our model, we ideally would perform next generation sequencing (NGS) on all of the patients in the sample so as to have the greatest power to detect rare variation. Unfortunately, while NGS has become much more affordable and thus accessible, the cost of sequencing a large number of patients can still be prohibitive depending on the resources available to the researcher. In order to leverage the most relevant information for our prediction problem, we employed a method of Extreme Phenotype Sampling (EPS) to identify differences in the extremes of the sample distribution for our outcome of interest[1]. Generally EPS is used when

large phenotypic variation is present for a particular substantive topic. In this case, we can use the extremes of predictive performance of the regression model to select patients for further study. This involves identifying patients whose plasma concentrations are predicted well using the model, and patients whose predicted plasma concentrations based on the model are very inaccurate, using a common metric of prediction quality. If a patient is poorly predicted, there are two possible outcomes for their predicted plasma concentration: it is either much lower than expected, or it could be much higher than expected. Over-prediction would be hypothesized to confer a lower risk of myalgia, since the high estimated statin plasma concentration would be associated with a more conservative dose of medication given to the patient, if used for clinical decision support. Under-prediction in this context may be more problematic however, because if a patient's plasma concentration were to be higher than their expected value, they could unknowingly be prescribed a dose that would have a higher possibility of producing undesirable side effects. We hypothesize that studying the differences in genetic variation between patients who are well-predicted versus patients whose actual plasma concentration values are much higher than expected given their model predictions would have the greatest impact for future studies modelling rosuvastatin plasma concentration as a potential indicator for the risk of adverse drug events.

6.1.1 Original Rosuvastatin Systemic Exposure Linear Regression Model Fit Assessment

A predictive model including common genetic markers for rosuvastatin plasma concentration has been previously developed by DeGorter et al.[2;3]. The specific details about the rosuvas-

tatin patient cohort characteristics can be found in Chapter 4. The outcome of the linear model was rosuvastatin plasma concentration; this value was log-transformed to make the distribution more normal. The clinical variables included in this analysis were age, dose (mg; represented continuously in the originally developed systemic exposure model and converted to a categorical variable in the augmented model we present), time post dose (hours), BMI (km/m^2). The genetic polymorphisms included in this model were *SLCO1B1* c. 521T>C and *ABCG2* c. 421C>A. The amount of variability in the outcome measure captured by the original linear regression model was relatively high: with the original covariates and dose represented categorically, the model had an R^2 value of 0.56 (SD=0.03), obtained by 5-fold cross-validation (CV), the full method details for which can be found in Chapter 4.

In order to more finely assess the predictive performance of the original systemic exposure model for the rosuvastatin cohort, we performed an analysis using leave-one-out CV (LOOCV) for the patients in the rosuvastatin cohort. We chose not do 5-fold CV because wanted to use the maximum amount of data for every individual prediction, since the dataset is relatively small. Using the covariates included in the original rosuvastatin systemic exposure model, predictions were generated for each patient in the rosuvastatin cohort, with the coefficients for each model obtained by training a linear regression model on the remainder of patients in the data set. The quality of the model was assessed for each patient by using their profile as a test case for the model trained on the remaining patients. The following predictive model fit quality metrics were obtained for each patient in the rosuvastatin cohort: the magnitude of deviation of the predicted concentration from the true plasma concentration, whether or not the concentration was under-predicted, squared error, and proportional difference between the predicted and actual values.

In addition to obtaining summary statistics from the LOOCV, we also conducted a visual analysis of the range of predicted versus actual plasma concentration values for the rosuvastatin predictive model. Using the predicted values obtained from the CV, we made comparative graphs of the predicted and actual plasma concentrations, for both the raw and log-transformed values. The log-transformed values for the rosuvastatin cohort were reasonably close to linear; however, even with the log transform, the tails of the distribution did not fit the assumption of normality. The skewness of the distribution is likely due to the log transformation and the fact that one cannot have a negative value for statin plasma concentration; this necessarily increases the number of patients underpredicted than overpredicted, as the statin plasma concentration cannot go below zero.

As postulated at the beginning of this section, poor prediction at the upper tail of the plasma concentration distribution is problematic for predicting statin dose in such a way so as to minimize the potential for augmented plasma concentration and thus myopathy, if used in the context of clinical decision support. The trade-off for this is that focusing on only the upper tail may carry the risk of having overly conservative dosing schemes, resulting in lowered statin efficacy. This risk could be balanced by periodic monitoring of the patient's statin plasma and LDL cholesterol levels. The rosuvastatin plasma concentration distributions are shown in Figures 6.1 and 6.2.

6.1.2 Selection Algorithm for Patient Sequencing

The quality metric of accuracy for selecting patients for NGS was the raw proportional difference between the actual value and the value predicted by the modified rosuvastatin systemic

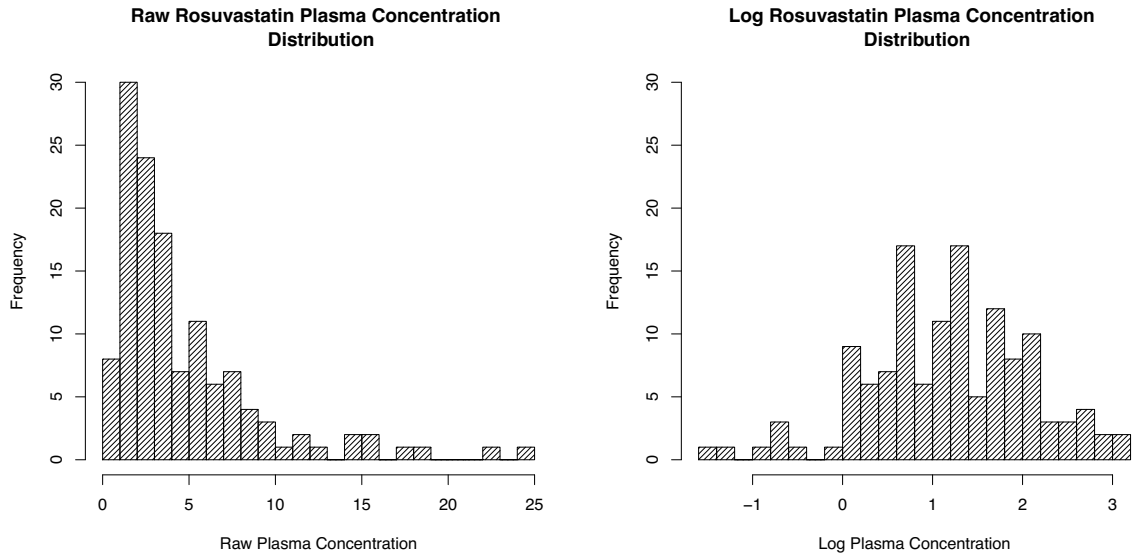


Figure 6.1: Raw vs log plasma concentration distribution histograms

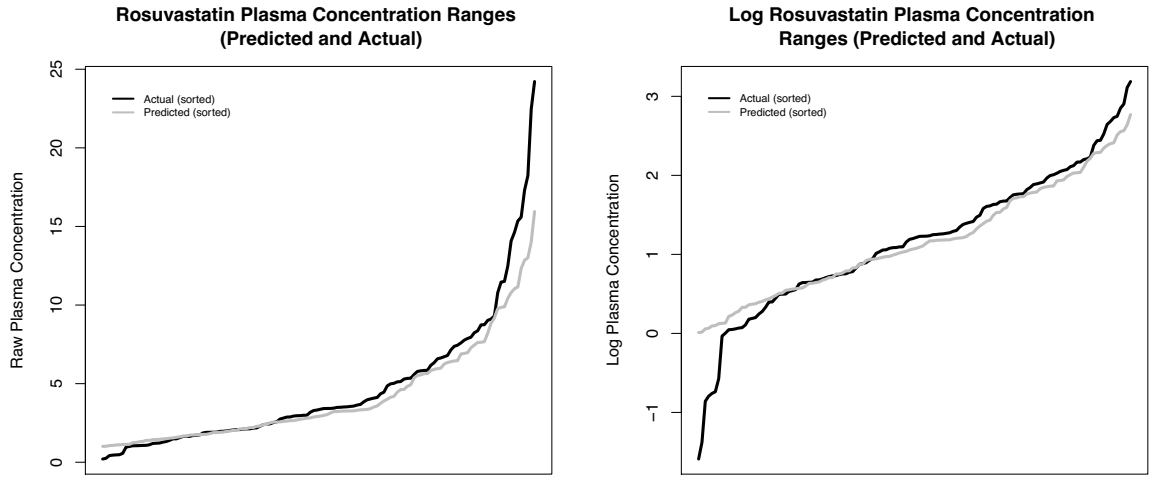


Figure 6.2: Raw vs log plasma concentration distribution sorted values

exposure model. This metric was chosen because for the purposes of this research question, we are much more interested in patients who are under-predicted using the original rosuvastatin systemic exposure prediction model, rather than those who are over-predicted. The proportional difference range of the "well predicted" values was defined as 0 plus or minus the absolute value of the maximum (positive) proportional difference seen in the sample (0 ± 0.778), in order to ensure that range of well-predicted values was symmetrical. The number of available patient slots for NGS processing was 48; 54 patients were chosen from the sample in case some of the samples could not be properly sequenced. All of the patients whose proportional difference values fell outside the specified range of well-predicted variation were to be sequenced; for the remaining patient slots, a random sample of patients within the specified range of well-predicted proportional differences were selected as a comparison group. A number of patients within the lab had already had their genomes analysed with NGS; these patients were excluded from the candidate patients in the sequencing selection process, but their results were included in the final analysis for identifying novel variants. The full distribution of proportional difference scores is shown in Figure 6.3, with the under-predicted patients selected for sequencing in red, the well-predicted patients randomly selected for sequencing in blue, and the unsequenced patients in white.

6.1.3 Patient Selection Results

48 patients in total were to be selected for NGS. In total, 21 patients in the rosuvastatin cohort were identified as being cases: these patients were under-predicted using the original model developed by DeGorter et al.[3], and the proportional difference values between their predicted

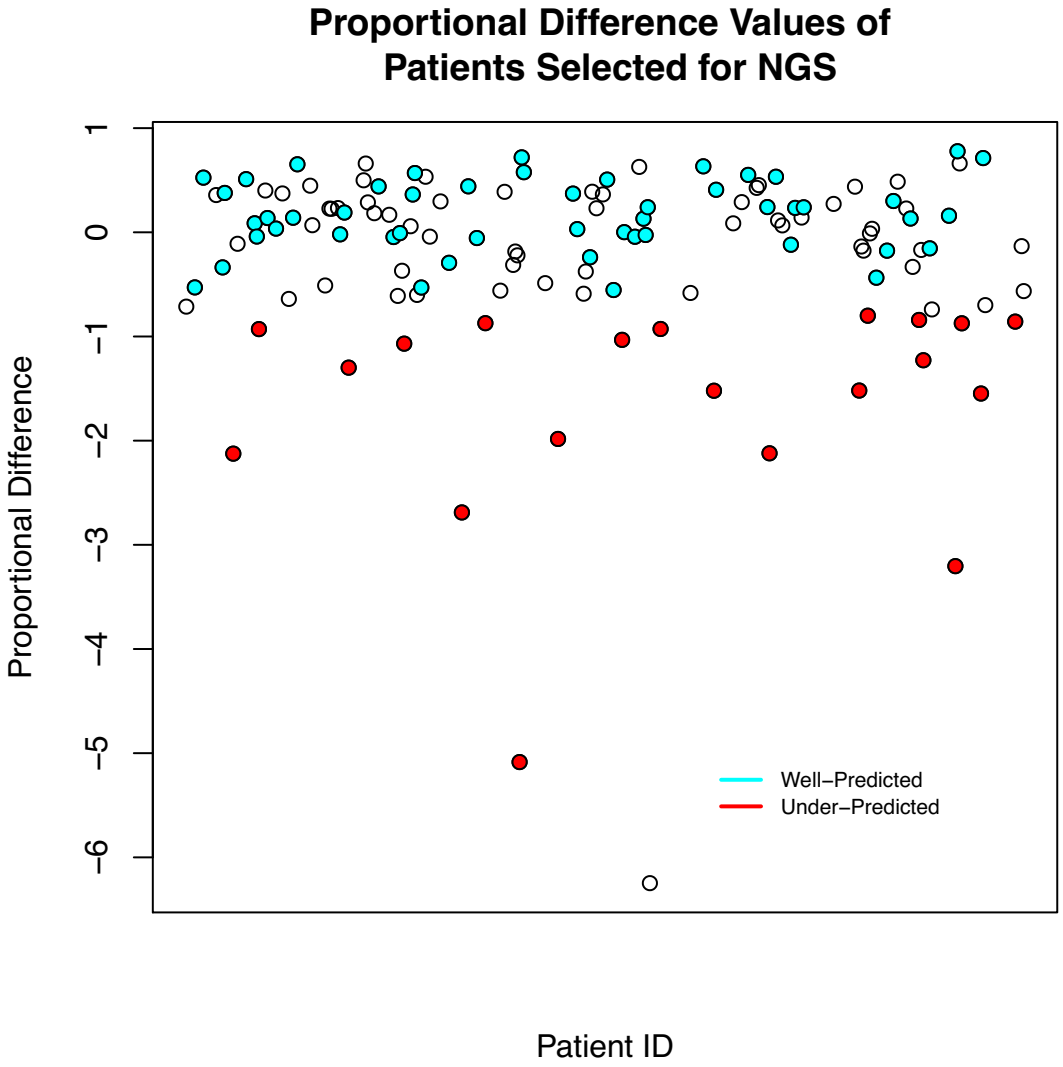


Figure 6.3: Rosuvastatin cohort proportional differences between predicted and raw values of plasma concentration based on the modified systemic exposure linear regression model

and actual plasma concentrations using that model were relatively large. Of these 21 patients, 5 had been previously sequenced using NGS technology, while one patient's DNA could not be sequenced, leaving a total of 20 under-predicted cases for this analysis. A table summarizing the population characteristics of the under-predicted cases can be found in Table 6.1. Furthermore, 33 individuals were selected for sequencing within the control group of patients who were well-predicted using the previously developed rosuvastatin systemic exposure regression model. An additional 17 patients whose proportional difference values fell within the range of well-predicted plasma concentrations had previously been sequenced by the lab; combining these, a total of 50 well-predicted control patients were included in the analysis (population characteristics also summarized in Table 6.1). The population characteristics for the cases and controls combined are shown in Table 6.2.

Table 6.1: Population characteristics of NGS processed rosuvastatin cases and controls (n=20)

	Cases (N=20)		Controls (N=50)	
Patient Characteristic	Mean/Prop.	SD%	Mean/Prop.	SD/%
Age (years)	58.60	11.85	56.10	13.69
Body Mass Index (kg/m^2)	30.64	5.88	29.60	5.60
Time Pose Dose (hours)	13.43	4.01	14.01	3.66
Rosuvastatin Dose (mg)				
5	9	45.0%	5	10.0 %
10	4	20.0%	14	28.0%
20	1	5.0 %	22	44.0%
30	0	0.0 %	2	4.0%
40	6	30 %	7	14.0%
Gender (Male=1)	10	50.0%	34	68.0%
Ethnicity (Non-Caucasian=1)	3	15.0%	4	8.0%
Minor Allelic Frequency				
SLCO1B1 c.521C	4/40	10.0%	22/100	22.0%
ABCG2 c.421A	4/40	10.0%	8/100	8.0%

Table 6.2: Population characteristics of all NGS processed rosuvastatin patients (n=70)

Patient Characteristic	Mean/Proportion	SD/Percentage
Age (years)	56.81	13.16
Body Mass Index (kg/m^2)	29.89	5.66
Time Pose Dose (hours)	13.85	3.75
Rosuvastatin Dose (mg)		
5	14	20.0%
10	18	25.7%
15	0	0.0%
20	23	32.9%
30	2	2.9%
40	13	18.6%
Gender (Male=1)	44	62.9%
Ethnicity (Non-Caucasian=1)	7	10.0%
Minor Allelic Frequency		
SLCO1B1 c.521C	26/140	18.6%
ABCG2 c.421A	12/140	8.6%

6.2 Novel SNP Identification via NGS

6.2.1 DNA Processing

DNA had been previously extracted from a blood sample. Targeted exome NGS was applied using the PGxSeq panel including 100 genes relevant to drug metabolism, absorption, distribution, excretion and response were targeted for sequencing; full details of these methods are described by Gulilat et al in as-yet unpublished data[4]. The Nextera Rapid Capture Custom Enrichment Kit (Illumina, San Diego, CA) was used to enrich these regions prior to the NGS process. The NGS procedure was conducted on the Illumina MiSeq Sequencer (Illumina, San Diego, CA), and took place at the London Regional Genomics Centre in London, Ontario.

Following sequencing, the sequenced genomic information was obtained in the form of FASTQ files, and then underwent a quality control assessment using FastQC[5]. Variant calling and sequence alignment were performed using the CLC Bio Genomics Workbench 7.0 (CLC

Bio, Aarhus, Denmark) using a custom-automatic workflow[4].

6.2.2 Data Processing

Ideally, we would attempt to identify specific novel genetic polymorphisms within each gene; however, this is infeasible given the small sample size and large number of variants present in the current work. Instead we aim to identify full genes that are relevant to the problem of predicting rosuvastatin plasma concentration, and subsequently perform more specific analyses within these genes to locate specific variations that are pertinent to changes in rosuvastatin plasma concentration. The sequenced and filtered variant information was exported for statistical analysis in the form of Variant Call Format (VCF) files for each individual patient. Following this, index files were generated for each patient VCF file in preparation for merging the files together. The index files were obtained using the `tabix` command from the SAMtools package[6]; the command performs indexing on compressed TAB-delimited position files and adds supplemental sequencing information[7]. Subsequently a shell script generated using R was used to automate the merging process on the command-line with the Genome Analysis ToolKit (GATK)[8]. Following this, the VCFtools software package[9] was used to remove longer indels from the merged data that would interfere with future processing, as well as genes on the X and Y chromosomes. Auxilliary files (in BIM, FAM, and SSD formats) were required for later processing steps; these were generated using PLINK on the command-line[10]; PLINK is an open-source C/C++ toolkit that was developed for processing large data sets for whole-genome association studies.

6.2.3 Methods

The Sequence Kernel Association Test (SKAT) was used to determine differences between the genetic profiles of patients who were under-predicted versus those who were well-predicted using the original rosuvastatin systemic exposure model[3]. The R `SKAT`[11] package was used for this analysis, using the optimal SKAT-O formulation which combines the burden and SKAT tests[12]. Other SKAT formulations could also have been used, and differences in the results of using these should be compared in future work. Because the dataset contained only a small number of individuals, the SKAT correction for small sample size was applied[13]. A linear kernel was used for this analysis.

Before the data could be processed using the `SKAT` function, it had to be imported into R and formatted into the gene matrix Z required as input for the analysis. The first stage of this process involved loading the zipped VCF file containing the merged data of all of the sequenced patients; the `vcfR` package was used for this purpose[14]. This package was extremely helpful for getting the data into R; however, the vcfR object was not in a compatible format to be used with the `SKAT` function. The second stage of the process of preparing the data for analysis in R was to reformat the vcfR object into a matrix of genotypes using the `vcfR2loci` function implemented in the `adegenet` R package[15;16]. This extracted the SNP data into the right structure, but had to be manually recoded from a matrix of factors to the numeric coding required by the `SKAT` function (2 = '1/1', 1 = '0/1', 0 = '0/0', 9 = missing). Once Z was in the correct format for the analysis, additional processing was required to generate an indexing scheme to map the SNPs to a gene or region; this was the step that required the BIM, BED, FAM and SSD auxilliary files. The ANNOVAR software package[17] was also required for the

generation of these files. The `Generate_SSD_SetID` function included with the `SKAT` package combined these files to create a SetID object. This SetID object then had to be cast to a variable using the `Open_SSD` function, also included in the `SKAT` package. In an ideal situation this elaborate data formatting would be unnecessary, as ANNOVAR is able to identify gene names based on chromosome location information; we had to perform this step manually due to software limitations.

Before running the actual SKAT analysis, we fit the null model for dichotomous outcomes (`SKAT_Null_Model`), controlling for the original covariates included in the linear regression model. At this time, all of the data was formatted properly for input into the `SKAT` function. The first time the `SKAT.SSD.All` function was run, it returned an output consisting of a series of *P* values for each gene or region that contained the individual SNPs as specified in the SetID file. However, it was discovered that one of the auxilliary files contained errors, which were manually changed in a text editor so as not to require regenerating all of the auxilliary files. Further attempts to use the `SKAT.SSD.All` function were fruitless, as were attempts to manually specify the set ID and use the `SKATBinary`. The entire data-formatting process was repeated from after the VCF merge, but no solution in R or the command-line could be found. The basic `SKAT` function for a SNPs in a single set was still operational, so in the end a `for` loop was used to manually iterate over all of the sets (>1000) and calculate the approximate *P* value for each gene. The *P* values were then compiled and the Holm method was applied to correct for multiple testing[18]. The choice of correcting method wound up being somewhat irrelevant, given that the only corrected *P* value below 1.0 was for NR1I2 ($P = 0.99$), and this did not differ meaningfully for the other available methods with this package.

6.2.4 Results

Most of the *P* values returned from the SKAT procedure were 1.0; however, 41 total genetic regions had *P* values below this level, of the 1212 genetic locations in the dataset located in the 100 gene targeted sequencing panel. Given the potential for an inflated family-wise error rate (Type I/ α error) due to multiple testing, it would not be meaningful to discuss the genes highlighted in this analysis in terms of traditional "statistical significance" (ie. $P < 0.05$). Instead, the top three genes with the lowest *P* values found in the SKAT analysis will be discussed. Other genes may also be relevant, but without a larger sample size for sequencing, it is difficult to determine with more precision which specific regions they include. The top 10 genes with the strongest signal for the current prediction problem are shown in Table 6.3, and subsequently described using information from the GeneCards database[19].

Table 6.3: Rank and unadjusted *P* values from SKAT procedure

Rank	Gene	Unadjusted *P* Value
1	*ABCC1*	0.022
2	*NR1I2*	0.031
3	*SLCO1B3*	0.035
4	*MTHFR*	0.058
5	*C1orf167*	0.058
6	*CBR3*	0.085
7	*SLCO2B1*	0.094
8	*ATIC*	0.108
9	*POR*	0.169
10	*SLC22A1*	0.171

The gene with the strongest signal resultant from the SKAT analysis was the ATP Binding Cassette Subfamily C Member 1 gene *ABCC1*, which is also known as Multidrug Resistance-Associated Protein 1[20]. Importantly, ABCC1 is an efflux transporter acknowledged to be strongly associated with intracellular statin accumulation, and has been shown to be expressed

in skeletal muscle tissue[21]. In the same study, it was shown that intracellular statin toxicity was ameliorated by over-expressing this efflux protein[21]. Individual variants of *ABCC1* and their corresponding frequencies in the case and control cohorts are shown in Table C.1.

The gene with the second strongest signal resultant from the SKAT analysis was the pregnane X receptor (*PXR*); this is also known as the human orphan nuclear receptor[22] or the Nuclear Receptor Subfamily 1 Group I Member 2 (*NR1I2*)[23;24]. *PXR* is a transcription factor of xenobiotic- and drug-inducible expression of key genes that encode members of the phase I and phase II metabolic enzymes and drug transporters, including ABCG2[24], of which rosuvastatin is a substrate. ABCG2 regulates systemic exposure to rosuvastatin by controlling its efflux and limiting its absorption from the gut[25]. No direct association between rosuvastatin pharmacokinetics and PXR was found in a recently performed study[26] examining the impact of polymorphisms in *PXR*, *ABCG2*, and other genes. However, given its control of ABCG2 expression in the gut and liver, it is possible that further study into PXR as a mediator of rosuvastatin pharmacokinetics is warranted. The variants present in the case and control cohorts as well as their MAFs are shown in Table C.2.

The gene with the third strongest signal produced by the SKAT analysis was Solute Carrier Organic Anion Transporter Family Member 1B3 (*SLCO1B3*), which codes for the drug transporter OATP1B3, which is known to transport rosuvastatin[27]. OATP1B3 has been discussed as an important factor in the hepatic clearance of its associated substrates, including rosuvastatin[28;29], which can contribute to systemic exposure.

6.2.5 Discussion

All of the top three genes identified in the SKAT analysis are associated with rosuvastatin drug transport, which provides a potential explanation of why these particular genes have sufficient signal to distinguish under-predicted rosuvastatin patients from well-predicted controls. Because of the relatively small sample size of the current study, the results are exploratory and should be used to guide further biological inquiry into specific polymorphisms in these genes that have the potential to affect rosuvastatin plasma concentration, and whether inclusion of these SNPs as predictors could be used to increase the accuracy of the original rosuvastatin systemic exposure model.

A major obstacle in the current work was obtaining results from the analysis because of difficulties in using software that is currently the state-of-the-art for this type of modelling. An amazing amount of work has gone into developing each piece of open-source software used to format and process the NGS data; however, coordinating these separate processes incurs a prohibitive technical burden. This is especially true for researchers who are not familiar with shell-scripting and using the command-line. Even installing the packages for use in processing the data was challenging, given that some operating systems (such as macOS) do not possess the necessary command-line tools by default.

In general, we were unable to find any thorough tutorials that guided the user from the data in FASTQ format to the eventual analysis we performed. The creation of a technically accessible open-source application to facilitate genetic analysis should be considered an open software engineering problem; until then, statistical analysis of NGS data may be performed suboptimally just because of the practical difficulties involved in its execution. For example,

in the current work we were unable to use the SKAT analysis specific to common and rare variants, and testing multiple kernels was infeasible due to the computational requirements of manually iterating over each collection of SNPs. Ease of use should be a primary consideration when designing software for a non-technical user base, as not all genetic researchers possess extensive skills in bioinformatics.

6.2.6 Conclusions

In the current work, we presented results from an exploratory analysis that highlights genes that may be of relevance for the problem of more accurately predicting rosuvastatin plasma concentration, and in particular identifying individuals who are at risk of having much higher than expected plasma concentrations that would put them at a higher risk for adverse drug events. The three genes with the strongest signal resulting from this analysis all have extensive ties to pathways regulating lipid metabolism. However, further biological research must be done to confirm this relationship and identify specific polymorphisms in these genetic regions that could be particularly predictive of rosuvastatin plasma concentration. A limitation of this work is the subjects for the NGS analysis were selected using the phenotypes generated original linear regression model, and not the slightly improved GAM model fit; this was simply because of research timing (the two analyses were run concurrently).

A substantial barrier to performing the analysis described in this chapter is the technical burden required to install and use the necessary individual software components. We propose that developing a user-oriented software interface that improves ease of use should be considered an open software engineering problem.

References

[1] David Gurwitz and Howard L McLeod. Genome-wide studies in pharmacogenomics: harnessing the power of extreme phenotypes. Pharmacogenomics, 14(4):337–339, 2013.

[2] Marianne K DeGorter. Statin Transport by Hepatic Organic Anion-Transporting Polypeptides (OATPs). PhD thesis, The University of Western Ontario, 2012.

[3] Marianne K DeGorter, Rommel G Tirona, Ute I Schwarz, Yun-Hee Choi, George K Dresser, Neville Suskin, Kathryn Myers, GuangYong Zou, Otito Iwuchukwu, Wei-Qi Wei, et al. Clinical and pharmacogenetic predictors of circulating atorvastatin and rosuvastatin concentration in routine clinical care. Circulation: Cardiovascular Genetics, 6(4):400–408, 2013.

[4] Markus Gulilat, Tyler Lamb, Wendy A. Teft, John F. Robinson, Rommel G. Tirona, Robert A. Hegele, Richard B. Kim, and Ute I. Schwarz. Targeted next generation sequencing as a tool for precision medicine. In submission.

[5] Simon Andrews et al. FastQC: a quality control tool for high throughput sequence data. 2010.

[6] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The sequence alignment/map format and samtools. Bioinformatics, 25(16):2078–2079, 2009.

[7] Heng Li. Tabix: fast retrieval of sequence features from generic tab-delimited files. Bioinformatics, 27(5):718–719, 2011.

[8] Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research, 20(9):1297–1303, 2010.

[9] Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A Albers, Eric Banks, Mark A DePristo, Robert E Handsaker, Gerton Lunter, Gabor T Marth, Stephen T Sherry, et al. The variant call format and vcftools. Bioinformatics, 27(15):2156–2158, 2011.

[10] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel AR Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul IW De Bakker, Mark J Daly, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics, 81(3):559–575, 2007.

[11] Seunggeun Lee, with contributions from Larisa Miropolsky, and Michael Wu. SKAT: SNP-Set (Sequence) Kernel Association Test, 2017. R package version 1.3.2.1.

[12] Seunggeun Lee, Michael C Wu, and Xihong Lin. Optimal tests for rare variant effects in sequencing association studies. Biostatistics, 13(4):762–775, 2012.

[13] Seunggeun Lee, Mary J Emond, Michael J Bamshad, Kathleen C Barnes, Mark J Rieder, Deborah A Nickerson, David C Christiani, Mark M Wurfel, Xihong Lin, NHLBI GO Exome Sequencing Project, et al. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. The American Journal of Human Genetics, 91(2):224–237, 2012.

[14] Brian J. Knaus and Niklaus J. Grünwald. VCFR: a package to manipulate and visualize variant call format data in R. Molecular Ecology Resources, 17(1):44–53, 2017.

[15] Thibaut Jombart. adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics, 24(11):1403–1405, 2008.

[16] Thibaut Jombart and Ismail Ahmed. adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. Bioinformatics, 27(21):3070–3071, 2011.

[17] Kai Wang, Mingyao Li, and Hakon Hakonarson. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Research, 38(16):e164–e164, 2010.

[18] Sture Holm. A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics, 6(2):65–70, 1979.

[19] Marilyn Safran, Irina Dalah, Justin Alexander, Naomi Rosen, Tsippi Iny Stein, Michael Shmoish, Noam Nativ, Iris Bahir, Tirza Doniger, Hagit Krug, et al. Genecards version 3: the human gene integrator. Database, 2010(1), 2010.

[20] GeneCards Human Gene Database. ABCC1 gene (protein coding), 2018.

[21] Michael J Knauer, Bradley L Urquhart, Henriette E Meyer zu Schwabedissen, Ute I Schwarz, Christopher J Lemke, Brenda F Leake, Richard B Kim, and Rommel G Tirona. Human skeletal muscle drug transporters determine local exposure and toxicity of statins. Circulation Research, 106(2):297–306, 2010.

[22] Jürgen M Lehmann, David D McKee, Michael A Watson, Timothy M Willson, John T

Moore, and Steven A Kliewer. The human orphan nuclear receptor pxr is activated by compounds that regulate CYP3A4 gene expression and cause drug interactions. The Journal of Clinical Investigation, 102(5):1016–1023, 1998.

[23] GeneCards Human Gene Database. NR1I2 gene (protein coding), 2018.

[24] Mei Liu, Xiu-Jun Wu, Gui-Lian Zhao, Ti Zhang, Shan-Sen Xu, Ya-Xin Sun, Feng Qiu, and Li-Mei Zhao. Effects of polymorphisms in nr1h4, nr1i2, slco1b1, and abcg2 on the pharmacokinetics of rosuvastatin in healthy chinese volunteers. Journal of Cardiovascular Pharmacology, 68(5):383–390, 2016.

[25] JE Keskitalo, O Zolk, MF Fromm, KJ Kurkinen, PJ Neuvonen, and M Niemi. ABCG2 polymorphism markedly affects the pharmacokinetics of atorvastatin and rosuvastatin. Clinical Pharmacology & Therapeutics, 86(2):197–203, 2009.

[26] Mei Liu, Xiu-Jun Wu, Gui-Lian Zhao, Ti Zhang, Shan-Sen Xu, Ya-Xin Sun, Feng Qiu, and Li-Mei Zhao. Effects of polymorphisms in NR1H4, NR1I2, SLCO1B1, and ABCG2 on the pharmacokinetics of rosuvastatin in healthy chinese volunteers. Journal of cardiovascular pharmacology, 68(5):383–390, 2016.

[27] Richard H Ho, Rommel G Tirona, Brenda F Leake, Hartmut Glaeser, Wooin Lee, Christopher J Lemke, Yi Wang, and Richard B Kim. Drug and bile acid transporters in rosuvastatin hepatic uptake: function, expression, and pharmacogenetics. Gastroenterology, 130(6):1793–1806, 2006.

[28] Satoshi Kitamura, Kazuya Maeda, Yi Wang, and Yuichi Sugiyama. In-

volvement of multiple transporters in the hepatobiliary transport of rosuvastatin. Drug Metabolism and Disposition, 36(10):2014–2023, 2008.

[29] Kazuya Maeda. Organic anion transporting polypeptide (OATP) 1B1 and OATP1B3 as important regulators of the pharmacokinetics of substrate drugs. Biological and Pharmaceutical Bulletin, 38(2):155–168, 2015.

Chapter 7

Discussion and Conclusions

The primary objective of this thesis was to explore avenues for improving a previously developed algorithm to predict atorvastatin and rosuvastatin plasma concentration using clinical and genetic determinants. This was achieved by 1) developing a selection algorithm to identify relevant concomitant medications to improve model fit; 2) exploring the use of non-linearar modelling techniques in the atorvastatin and rosuvastatin cohorts in comparison to the original linear model used to predict systemic exposure[1]; and 3) selecting patients for DNA sequencing and conducting exploratory analysis to identify potentially relevant genes and their associated variants for further biological study. This final chapter will identify the key contributions to the relevant literature made in this thesis, as well as the main strengths and limitations of the current work. Future directions for this research will also be discussed.

7.1 Summary of Key Contributions

7.1.1 Objective 1

The first objective of the current work was to review the literature on appropriate variable selection techniques that could be used to identify concomitant medications associated with changes in atorvastatin and rosuvastatin plasma concentration, and assess the effect of their inclusion in the quality of fit for the linear regression models. The initial product of this objective was a narrative review detailing which medications found in the literature could theoretically impact statin plasma concentration. Following this, we characterized concomitant medication use in the atorvastatin and rosuvastatin patient cohorts. The final methodological product of this research was the development of a robust concomitant medication selection paradigm, which used penalized regression to determine which medications had the strongest predictive signal in relation to changes in plasma concentration. The selection algorithm leverages information on the medication class to overcome the modelling challenges associated with having many more covariates than available observations.

A number of concomitant medications were identified as being associated with atorvastatin plasma concentration. These included: acetylsalicylic acid, atenolol, candesartan, diclofenac, digoxin, esomeprazole, gliclazide, glucosamine, hydrocholorothiazide, levothyroxine, losartan, metformin, misoprostol, nifedepine, tamsulosin, valsartan, venlafaxine and vitamin B3. Many of these medications had been identified in the narrative review as having the potential to affect atorvastatin plasma concentration; however, several medications had not been previously mentioned in the body of literature examined in the current work. These concomitant medications included acetylsalicylic acid, atenolol, diclofenac, esomeprazole, gliclazide, glu-

cosamine, hydrochlorthiazide, levothyroxine, misoprostol and tamsulosin. We recommend that further biological research be conducted to determine if there are causal effects between these medications and changes in atorvastatin plasma concentration. The inclusion of all of the selected medications in the atorvastatin linear model for dose-prediction had a significant impact on model fit, and greatly increased the variance explained in the linear regression.

In contrast, only one medication was identified as being associated with rosuvastatin plasma concentration. Ranitidine was identified by the concomitant medication algorithm as being associated with rosuvastatin plasma concentration. However, it is likely that this association is spurious. Reasons for this include a lack of evidence in previous literature, a lack of any plausible causal mechanisms between the pathways affected by both drugs, and that only a very small number of patients in the rosuvastatin cohort had concomitant use of ranitidine. The inclusion of this medication in the linear model had no significant impact on model fit. The lack of concomitant medications associated with rosuvastatin plasma concentration is consistent with the literature suggesting that genetic variation and polymorphisms in drug transporters may have a greater effect on changes in rosuvastatin plasma concentration than concomitant medications[2], because rosuvastatin is minimally metabolized, unlike atorvastatin[3]. Despite not identifying any concomitant medications that increased the predictive ability of the model, the model fit of the linear regression model was improved by modelling rosuvastatin dose as a categorical variable instead of a continuous variable. This is likely because very few unique doses were present in the rosuvastatin cohort, which would make establishing a linear relationship between dose and plasma concentration very difficult.

7.1.2 Objective 2

The second objective of the current work was to explore whether non-linearar modelling techniques could be of use in improving the predictive capability of the dose-prediction linear regression model with the atorvastatin and rosuvastatin cohort data. The two techniques tested were generalized additive models (GAMs) with linear, degree 3 polynomial, degree 5 polynomial, and radial kernels; and support vector regression modelling (SVR) We also examined the effect of choosing smoothing parameters for the GAMs via cross-validation (CV) or using fixed parameters with a moderate smoothing value. We employed CV to assess model error and fit for each GAM both with fixed parameters and parameters chosen by CV within the GAM.

The atorvastatin GAM with smoothing parameters chosen by CV offered a significant improvement in model fit over the atorvastatin linear model including the concomitant medications, while the atorvastatin GAM with arbitrary smoothing parameters did not. This effect was attenuated when corrected for multiple comparisons, but still trended towards statistical significance. The parameters chosen by CV penalized the smoothing of age, 4β-hydroxycholesterol and BMI to the point where they were treated much like the parametric covariates and modelling using linear functions. However, a significant non-linearar relationship between time post dose and atorvastatin plasma concentration was identified in this analysis, based on the additional effective degrees of freedom used to generate the curve using the smoothing function. Both of the GAMs fitted for the rosuvastatin cohort differed significantly from the original linear model, even after adjusting for multiple comparisons. The two GAM fits were not statistically different from each other, although the rosuvastatin GAM with fixed smoothing parameters offered a marginally better fit than the GAM with parameters chosen by CV. Like

the atorvastatin GAM with parameters chosen by CV, the smoothing of the majority of continuous covariates was penalized to the point of linearity; however, time post dose and BMI were modelled linearly, while the relationship between age and rosuvastatin plasma concentration had a significant non-linearar component.

In contrast to the GAMs for atorvastatin and rosuvastatin, SVR does not appear to offer any substantial benefit in terms of model fit for any of the kernels tested at this time. This is likely due to the small sample size of our cohort; additional parameters of cost and epsilon must be tuned for the SVR to be fit, which was not possible with the amount of data available. In order to perform CV to assess the fit of the models for atorvastatin, all concomitant medications used by less than 10 patients were removed prior to fitting the SVR. It is possible that SVR could offer improvements in model fit over the linear model and GAM if more data were available to choose appropriate tuning parameters for the rosuvastatin and atorvastatin cohorts. However, we do not recommend using SVR for this dataset at present, particularly since both the GAM and linear models are easier to interpret from a clinical perspective.

7.1.3 Objective 3

The final objective of this thesis was to select patients in the rosuvastatin cohort to undergo more thorough genetic sequencing, in order to identify novel genetic variation that may play a role in increasing or decreasing rosuvastatin plasma concentration. A variation of extreme phenotype sampling[4;5] was employed in order to increase the likelihood of identifying relevant rare genetic variation despite our relatively small sample size. Proportional difference values between the predicted and actual rosuvastatin concentrations were used to identify patients

as being under-predicted or well-predicted using the original rosuvastatin systemic exposure model. Because we were limited to selecting 48 patients for next generation sequencing (NGS), all of the patients who were underpredicted using the current systemic exposure model were chosen for sequencing, and the remaining sequencing slots were filled by randomly sampling the patients in the well-predicted category. A number of patients had already undergone NGS; these were also included in the analysis.

The Sequence Kernel Association Test (SKAT) was used to analyse the data and identify genetic regions that were associated with distinctions between the under-predicted and well-predicted individuals who had undergone additional sequencing. This test suffered from a severe lack of power, given the large number of genes to be tested for association with case or control group membership, and the small number of patients who had undergone NGS. Despite this, 41 out of 1212 total candidate regions had P values that were lower than 1.0, indicating that they may have the potential to be predictive of whether or not a patient's plasma concentration on rosuvastatin was likely to be higher than anticipated based on the current dose-prediction model. The top three candidate genetic regions for prediction were within *ABCC1*, PXR (*NR1I2*), and *SLCO1B3*. All three genes have been shown to be associated with cellular rosuvastatin uptake (ABCC1 and SLCO1B3) or statin transporter expression (PXR), suggesting a plausible causal mechanism for directly impacting rosuvastatin plasma concentration. We recommend that further biological research be conducted to confirm a causal relationship between the genetic variation identified and rosuvastatin plasma concentration.. In particular, further research into the polymorphisms identified in these regions observed to have large differences in frequency between cases (under-predicted plasma concentration) and controls (well-predicted plasma concentration). Ideally, this research will be able to elucidate which

polymorphisms in these genes are most relevant for predicting rosuvastatin.

A significant challenge faced in performing the NGS analysis was installing and integrating all of the currently available tools for processing this type of data. We found that this type of analysis has an extremely high technical burden, which imposes barriers and decreases the accessibility of these statistical techniques to genetic researchers who want to use them to advance our understanding of the human genome. We suggest that the development of a user-friendly integrated open-source tool for performing analysis on genetic data be considered as an open software-engineering problem.

7.2 Strengths and Limitations

7.2.1 Strengths

A notable strength of the work performed in this thesis is the emphasis on addressing challenges of modelling patient data from a practical user-oriented perspective. Often the exact details for performing modelling in a clinical context are left out of publications, which creates barriers for other researchers to use the methods described for their own data. This is particularly a problem in the NGS variant prioritization literature, given how quickly research is progressing in this field, and the newness of the statistical tools appropriate for this type of analysis. Another user-oriented contribution in this work was that the concomitant medication algorithm was developed for use with both small and large sample sizes. The availability of statistical techniques appropriate for small sample sizes is an important consideration in this field of research, since patient populations are usually small due to the resources required to

collect and process the data.

7.2.2 Limitations

A key limitation of the current work was the sample size of the atorvastatin and rosuvastatin cohorts for the modelling activities performed in this thesis. While the sample size for both cohorts is relatively large from the perspective of clinical pharmacology research, the number of potential additional predictors we examined vastly outnumbered the patient observations available to us. This was particularly a problem for accurately gauging the performance of the non-linear modelling techniques using CV, since it was infeasible to include all of the covariates identified in the atorvastatin concomitant medication analysis with the extra model parameters that required fitting, such as cost and epsilon for SVR.

Another limitation of the current work was the difficulty encountered in using the software tools currently available for NGS analysis. Because the software did not work as intended with our data (and no possible causes for this were found), the NGS analysis used a simpler technique than would perhaps be optimal, given that it did not include the modification to allow examination of both common and rare variants.

7.3 Implication of Key Contributions

The findings of the work performed in the completion of this thesis have many potential implications for improving personalized medicine in the context of dose guidance for atorvastatin and rosuvastatin. These implications largely target the research activities of modelling and predictive dosing-guidance, given the emphasis on methodological improvements and identifying

potentially predictive factors for further biological analysis. The development of a concomitant medication selection algorithm will allow researchers to create more feature-rich datasets that include relevant clinical information, in a format that is accessible to clinical researchers. Additionally, the concomitant medications found to be associated with changes in atorvastatin plasma concentration may impact the future clinical use of these substances in order to better prevent adverse events from taking place. The genes identified in the NGS analysis have the potential to increase our understanding of the factors affecting lipid metabolism in the human body, which could in turn improve clinical care for patients with dyslipidemia.

7.4 Future Directions

A substantial amount of work presented in this thesis was exploratory, with the goal of hypothesis generation for confirmatory biological analysis. Further analyses that must be performed to validate the findings in this work include in vitro testing to confirm a causal relationship between the identified concomitant medications included in the atorvastatin systemic exposure model and changes in statin exposure. Similarly, further in vitro biological research is necessary to support the relationship between specific genetic variation in *ABCC1*, *NR1I2* and *SLCO1B3*, and changes in rosuvastatin plasma concentration.

The current work failed to find benefit in using more complex non-linear and machine learning techniques with the datasets from the atorvastatin and rosuvastatin cohorts. These analyses need to be repeated with larger datasets, as it is possible that the extra parameter estimates necessary to model non-linear relationships in the data were too numerous compared to the number of observations available for parameter selection and training.

A substantial amount of work remains to be done with respect to improving access to complex statistical techniques for non-experts in statistical analysis. An important next step in paving the way for the use of appropriate (though complex) statistical techniques in clinical pharmacology and other areas is the implementation of a user-friendly interface for non-linear analysis, as well as increased education on the availability of different statistical techniques for experts in other fields. Appendix D describes initial steps towards this work, in the form of an interactive statistical platform that was partially developed during the completion of this thesis.

Finally, one of the most surprising and concerning discoveries of this thesis was the difficulty of performing variant calling and prioritization with NGS data even for experienced bioinformaticians and experts in biostatistical analysis. As genetic data becomes increasingly available for decision making in healthcare, the more urgent the need for user-friendly, accurate and transparent tools for NGS data analysis becomes. As mentioned previously in this thesis, developing improved, stable, transparent and user-friendly techniques for this purpose should be considered an open software engineering problem. Without such tools, we may never be able to explore the full potential of genetic variation in the context of clinical decision support.

References

[1] Marianne K DeGorter, Rommel G Tirona, Ute I Schwarz, Yun-Hee Choi, George K Dresser, Neville Suskin, Kathryn Myers, GuangYong Zou, Otito Iwuchukwu, Wei-Qi Wei, et al. Clinical and pharmacogenetic predictors of circulating atorvastatin and rosuvastatin concentration in routine clinical care. Circulation: Cardiovascular Genetics, 6(4):400–408, 2013.

[2] Marianne K DeGorter. Statin Transport by Hepatic Organic Anion-Transporting Polypeptides (OATPs). PhD thesis, The University of Western Ontario, 2012.

[3] Hideki Fujino, Tsuyoshi Saito, Yoshihiko Tsunenari, and Junji Kojima. Interaction between several medicines and statins. Arzneimittelforschung, 53(03):145–153, 2003.

[4] Dan-Yu Lin, Donglin Zeng, and Zheng-Zheng Tang. Quantitative trait analysis in sequencing studies under trait-dependent sampling. Proceedings of the National Academy of Sciences, 110(30):12247–12252, 2013.

[5] Ian J Barnett, Seunggeun Lee, and Xihong Lin. Detecting rare variant effects using extreme phenotype sampling in sequencing association studies. Genetic Epidemiology, 37(2):142–151, 2013.

Appendix A

Concomitant Medication Selection

A.1 Mapping of Generic Drugs to Functional Classes

Table A.1: Drug class/ generic drug mapping

Medication Class	Generic Drug Name
5α Reductase Inhibitor	Dutasteride
α1 Blocker	Alfuzosin
α1 Blocker	Doxazosin
α1 Blocker	Tamsulosin
α1 Blocker	Terazosin
α2 Adrenergic Agonist	Brimonidine
α2 Adrenergic Agonist	Clonidine
Ace Inhibitor	Cilazapril
Ace Inhibitor	Enalapril
Ace Inhibitor	Fosinopril
Ace Inhibitor	Lisinopril
Ace Inhibitor	Perindopril
Ace Inhibitor	Quinapril
Ace Inhibitor	Ramipril
Ace Inhibitor	Trandolapril
Alkaloid	Quinine
Analgesic	Acetaminophen

Drug class/ generic drug mapping (A-A)

Medication Class	Generic Drug Name
Angiotensin II Receptor Agonist	Candesartan
Angiotensin II Receptor Agonist	Irbesartan
Angiotensin II Receptor Agonist	Losartan
Angiotensin II Receptor Agonist	Olmesartan
Angiotensin II Receptor Agonist	Telmisartan
Angiotensin II Receptor Agonist	Valsartan
Antiarrhythmic	Propafenone
Antibiotic	Amoxicillin
Antibiotic	Gentamicin
Antibiotic	Sulfacetamide
Antibiotic	Tetracycline
Anticholinergic	Oxybutynin
Anticholinergic Bronchodilator	Ipratropium
Anticholinergic Bronchodilator	Tiotropium Bromide
Anticholinesterase	Donepezil
Anticoagulant	Warfarin
Anticonvulsant	Phenobarbital
Anticonvulsant	Phenytoin
Anticonvulsant	Topiramate
Anticonvulsant	Trazodone
Anticonvulsant/Analgesic	Gabapentin
Antidiabetic	Metformin
Antiemetic	Dimenhydrinate
Antihistamine	Cetirizine
Antihistamine	Desloratadine
Antihistamine	Diphenhydramine
Anti-Inflammatory	Colchicine
Anti-Inflammatory	Sulfasalazine
Antimalarial	Hydroxychloroquine
Antimuscarinic	Tolterodine
Antiplatelet	Plavix
Antiviral	Famciclovir
Antiviral	Valacyclovir
Artificial Tears	Eye Lubricant
Atypical Antipsychotic	Risperidone

Drug class/ generic drug mapping (B-C)

Medication Class	Generic Drug Name
Benzodiazepine	Alprazolam
Benzodiazepine	Clonazepam
Benzodiazepine	Lorazepam
Benzodiazepine	Oxazepam
Benzodiazepine	Temazepam
Beta Agonist	Betaxolol
Beta Agonist	Formoterol
Beta Agonist	Salbutamol
Beta Agonist	Salmeterol Xinafoate
Beta Blocker	Acebutolol
Beta Blocker	Atenolol
Beta Blocker	Bisoprolol
Beta Blocker	Carvedilol
Beta Blocker	Cavedilol
Beta Blocker	Metoprolol
Beta Blocker	Nadolol
Beta Blocker	Sotalol
Beta Blocker	Timolol
Bile Acid Sequestrant	Cholestyramine
Bisphosphonate	Alendronate
Bisphosphonate	Etidronate
Bisphosphonate	Risedronate
Bisphosphonate	Zoledronate
Calcium Channel Blocker	Amlodipine
Calcium Channel Blocker	Diltiazem
Calcium Channel Blocker	Felodipine
Calcium Channel Blocker	Nifedipine
Calcium Channel Blocker	Verapamil
Carbonic Anhydrase Inhibitor	Dorzolamide
Cardiac Glycoside	Digoxin
Chemotherapy	Melphalan
Chemotherapy	Methotrexate

Drug class/ generic drug mapping (C-G)

Medication Class	Generic Drug Name
Cholesterol Absorption Blocker	Ezetimibe
Cholinergic	Bethanechol
Cholinesterase Inhibitor	Donepezil
Co Q10	Co Q10
Contraceptive	Contraceptive
Corticosteroid	Budesonide
Corticosteroid	Dexamethasone
Corticosteroid	Fluticasone
Corticosteroid	Fluticasone Propionate
Corticosteroid	Mometasone
Corticosteroid	Prednisone
COX Inhibitor	Celecoxib
Cyclopyrrolone	Zopiclone
Direct Renin Inhibitor	Aliskiren
Diuretic	Chlorthalidone
Diuretic	Furosemide
Diuretic	Hydrochlorothiazide
Diuretic	Indapamide
Diuretic	Spironolactone
Dopamine Serotonin Adrenergic Antagonist	Quetiapine
DPP4 Inhibitor	Sitagliptin
Estrogen Receptor Antagonist	Tamoxifen
Eye Drops	Multivitamin
Fatty Acid	Cetyl Myristoleate
Fibrate	Bezafibrate
Fibrate	Fenofibrate
Fibrate	Gemfibrozil
Gastroprokinetic	Domperidone
Glucosidase Inhibitor	Acarbose

Drug class/ generic drug mapping (H-O)

Medication Class	Generic Drug Name
Herbal	Curcumin
Herbal	Herbal
Histamine II Blocker	Ranitidine
Hormone	Estrogen
Hormone	Progesterone
Hormone	Testosterone Undecanoate
Immunosupressant	Mycophenolate Mofetil
Immunosupressant	Sirolimus
Insulin	Insulin
Ion Exchange Resin	Sodium Polystyrene Sulfonate
Laxative	Sodium Citrate
Leukotriene Receptor Antagonist	Montelukast
LHRH Agonist	Goserelin
Meglitinide	Repaglinide
Multivitamin	Multivitamin
Muscle Relaxant	Baclofen
Muscle Relaxant	Cyclobenzaprine
Muscle Relaxant	Methocarbamol
Nassa	Mirtazapine
NDRI	Bupropion
Nitrate	Glyceryl Trinitrate
NSAID	Acetylsalicylic Acid
NSAID	Diclofenac
NSAID	Ibuprofen
NSAID	Ketorolac
NSAID	Meloxicam
NSAID	Naproxen
Opioid	Fentanyl
Opioid	Hydromorphone
Opioid	Morphine
Opioid	Oxycodone
Opioid	Tramadol

Drug class/ generic drug mapping (P-S)

Medication Class	Generic Drug Name
PDE5 Inhibitor	Sildenafil
PDE5 Inhibitor	Tadalafil
PPARg Agonist	Rosiglitazone
Probiotic	Lactobacillus Acidophilus
Prostaglandin Analogue	Bimatoprost
Prostaglandin Analogue	Lanatoprost
Prostaglandin Analogue	Misoprostol
Proton Pump Inhibitor	Esomeprazole
Proton Pump Inhibitor	Lansoprazole
Proton Pump Inhibitor	Omeprazole
Proton Pump Inhibitor	Pantoprazole
Proton Pump Inhibitor	Rabeprazole
Proton Pump Inhibitor	Unknown
Selective Estrogen Receptor Modulator	Raloxifene
SNRI	Duloxetine
SSRI	Citalopram
SSRI	Escitalopram
SSRI	Fluoxetine
SSRI	Paroxetine
SSRI	Sertraline
SSRI/SNRI	Venlafaxine
Statin	Atorvastatin
Statin	Rosuvastatin
Stimulant	Dextroamphetamine
Sulfonylurea	Gliclazide
Sulfonylurea	Glyburide
Supplement	Calcium Supplement
Supplement	Chondritin
Supplement	Fish Oil
Supplement	Glucosamine
Supplement	Iron Supplement
Supplement	Magnesium Supplement
Supplement	Omega 3

Drug class/ generic drug mapping (T-Z)

Medication Class	Generic Drug Name
Thiazolidinedione	Pioglitazone
Thyroid Hormone	Levothyroxine
Tricyclic	Amitriptyline
Tricyclic	Nortriptyline
Triptan	Rizatriptan
Triptan	Sumatriptan
Unknown	Unknown
Urinary Alkalinizers	Potassium Citrate
Vasodilator	Isosorbide Dinitrate
Vasodilator	Nitroglycerin
Vitamin B	Vitamin B
Vitamin B	Vitamin B12
Vitamin B	Vitamin B2
Vitamin B	Vitamin B3
Vitamin B	Vitamin B6
Vitamin B	Vitamin B9
Vitamin C	Vitamin C
Vitamin D	Vitamin D
Vitamin D	Vitamin D3
Vitamin E	Vitamin E
Vitamin K	Vitamin K
Xanthine Oxidase Inhibitor	Allopurinol

A.2 Atorvastatin Concomitant Medication Analysis

Supplemental Tables

A.2.1 Overview of Concomitant Medications Present in the Prospective Cohorts

Table A.2: Generic drugs in the atorvastatin and rosuvastatin prospective cohorts (A-B)

	Atorvastatin (n=128)		Rosuvastatin (n=130)	
Acarbose			1	-
Acebutolol	3	2.3%	2	1.5%
Acetaminophen	8	6.2%	6	4.6%
Acetylsalicylic.acid	86	67.2%	73	56.2%
Alendronate	1	-	1	-
Alfuzosin	1	-		
Aliskiren	2	1.6%	1	-
Allopurinol	3	2.3%	6	4.6%
Alprazolam	1	-	2	1.5
Amitriptyline	3	2.3%	1	-
Amlodipine	23	18%	19	14.6%
Amoxicillin	1	-		
Atenolol	14	10.9%	10	7.7%
Atorvastatin	128	100%		
Baclofen			1	-
Betaxolol	1	-		
Bethanechol	1	-		
Bezafibrate	2	1.6%		
Bimatoprost	1	-	2	1.5%
Bisoprolol	12	9.4%	9	6.9%
Brimonidine			1	-
Budesonide			2	1.5%
Bupropion	2	1.6%	4	3.1%

Table A.3: Generic drugs in the atorvastatin and rosuvastatin prospective cohorts (C)

	Atorvastatin (n=128)		Rosuvastatin (n=130)	
Calcium Supplement	11	8.6%	11	8.5%
Candesartan	3	2.3%	3	2.3%
Carvedilol			1	-
Cavedilol	1	-		
Celecoxib	4	3.1%	1	-
Cetirizine			2	1.5%
Cetyl Myristoleate			1	-
Chlorthalidone	3	2.3%	1	-
Cholestyramine	1	-	1	-
Chondritin	1	-		
Cilazapril			1	-
Citalopram	2	1.6%	2	1.5%
Clonazepam	2	1.6%		
Clonidine	1	-		
Co-Q10	7	5.5%	9	6.9%
Colchicine			2	1.5%
Contraceptive	2	1.6%		
Curcumin			1	-
Cyclobenzaprine	2	1.6%		

Table A.4: Generic drugs in the atorvastatin and rosuvastatin prospective cohorts (D-H)

	Atorvastatin (n=128)		Rosuvastatin (n=130)	
Desloratadine			1	-
Dexamethasone			1	-
Dextroamphetamine	1	-		
Diclofenac	2	1.6%	3	2.3%
Digoxin	5	3.9%		
Diltiazem	6	4.7%	3	2.3%
Dimenhydrinate	1	-		
Diphenhydramine			1	-
Domperidone	1	-	7	5.4%
Donepezil	1	-	1	-
Doxazosin			3	2.3%
Duloxetine	1	-		
Dutasteride	1	-	1	-
Enalapril	1	-	3	2.3%
Escitalopram			2	1.5%
Esomeprazole	3	2.3%	1	-
Estrogen	1	-		
Etidronate	1	-		
Eye Lubricant	1	-	1	-
Ezetimibe	39	30.5%	52	40%
Famciclovir	1	-		
Felodipine			2	1.5%
Fenofibrate	15	11.7%	23	17.7%
Fentanyl			1	-
Fish Oil	6	4.7%	4	3.1%
Fluoxetine			1	-
Fluticasone	1	-	1	-
Formoterol	1	-	2	1.5%
Fosinopril	1	-		
Furosemide	7	5.5%	4	3.1%
Gabapentin	1	-	2	1.5%
Gemfibrozil	3	2.3%	1	-
Gentamicin			1	-
Gliclazide	5	3.9%	3	2.3%
Glucosamine	4	3.1%	3	2.3%
Glyburide	5	3.9%	5	3.8%
Glyceryl Trinitrate			1	-
Goserelin			1	-
Herbal	4	3.1%		
Hydrochlorothiazide	17	13.3%	16	12.3%
Hydromorphone	1	-	1	-
Hydroxychloroquine			3	2.3%

Table A.5: Generic drugs in the atorvastatin and rosuvastatin prospective cohorts (I-O)

	Atorvastatin (n=128)		Rosuvastatin (n=130)	
Ibuprofen	2	1.6%	2	1.5%
Indapamide	2	1.6%	3	2.3%
Insulin	9	7.0%	9	6.9%
Ipratropium	1	-		
Irbesartan	7	5.5%	7	5.4%
Iron Supplement	3	2.3%	4	3.1%
Isosorbide.dinitrate	1	-		
Ketorolac			1	-
Lactobacillus.acidophilus			1	-
Lanatoprost			1	-
Lansoprazole	5	3.9%	5	3.8%
Levothyroxine	10	7.8%	13	10.0%
Lisinopril	2	1.6%	2	1.5%
Lorazepam	3	2.3%	1	-
Losartan	5	3.9%	2	1.5%
Magnesium.supplement			1	-
Meloxicam	2	1.6%	1	-
Melphalan	1	-		
Metformin	24	18.8%	16	12.3%
Methotrexate	1	-	2	1.5%
Metoprolol	28	21.9%	21	16.2%
Mirtazapine	1	-	1	-
Misoprostol	2	1.6%		
Mometasone			1	-
Montelukast			1	-
Morphine			1	-
Multivitamin	14	10.9%	13	10.0%
Mycophenolate Mofetil			1	-
Nadolol			1	-
Naproxen	1	-	1	-
Nifedipine	4	3.1%	3	2.3%
Nitroglycerin	15	11.7%	16	12.3%
Nortriptyline	1	-	1	-
Olmesartan			1	-
Omega 3	7	5.5%	8	6.2%
Omeprazole	4	3.1%	6	4.6%
Oxazepam			1	-
Oxybutynin			1	-
Oxycodone	1	-	1	-

Table A.6: Generic drugs in the atorvastatin and rosuvastatin prospective cohorts (P-S)

	Atorvastatin (n=128)		Rosuvastatin (n=130)	
Pantoprazole	3	2.3%	11	8.5%
Paroxetine			1	-
Perindopril	9	7.0%	12	9.2%
Phenobarbital	1	-		
Phenytoin	2	1.6%	1	-
Pioglitazone	1	-	1	-
Plavix	23	18.0%	19	14.6%
Potassium Citrate			1	-
Prednisone	2	1.6%		
Progesterone	1	-		
Propafenone	1	-		
Quetiapine			3	2.3%
Quinapril	1	-	4	3.1%
Quinine	2	1.6%	2	1.5%
Rabeprazole	10	7.8%	13	10.0%
Raloxifene	1	-		
Ramipril	49	38.3%	30	23.1%
Ranitidine	2	1.6%	2	1.5%
Repaglinide			1	-
Risedronate			2	1.5%
Risperidone			1	-
Rizatriptan			1	-
Rosiglitazone	3	2.3%	1	-
Rosuvastatin	1	-	130	100%
Salbutamol	2	1.6%	2	1.5%
Salmeterol Xinafoate			2	1.5%
Sertraline	2	1.6%		
Sildenafil			1	-
Sirolimus			1	-
Sitagliptin	2	1.6%	1	-
Sodium Citrate	1	-		
Sodium Polystyrene Sulfonate			1	-
Sotalol			1	-
Spironolactone	8	6.2%	3	2.3%
Sulfacetamide			1	-
Sulfasalazine			1	-
Sumatriptan	1	-		

Table A.7: Generic drugs in the atorvastatin and rosuvastatin prospective cohorts (T-Z)

	Atorvastatin (n=128)		Rosuvastatin (n=130)	
Tadalafil	1	-	2	1.5%
Tamoxifen			1	-
Tamsulosin	3	2.3%	4	3.1%
Telmisartan	4	3.1%	2	1.5%
Temazepam	1	-	2	1.5%
Terazosin			2	1.5%
Testosterone Undecanoate			1	-
Tetracycline			1	-
Timolol			2	1.5%
Tiotropium Bromide	1	-	1	-
Tolterodine			2	1.5%
Topiramate	2	1.6%	1	-
Trandolapril			2	1.5%
Trazodone			3	2.3%
Unknown	3	2.3%	2	1.5%
Valacyclovir			1	-
Valsartan	5	3.9%	7	5.4%
Venlafaxine	4	3.1%	2	1.5%
Verapamil			1	-
Vitamin B	3	2.3%		
Vitamin B12	11	8.6%	13	10.0%
Vitamin B2			1	-
Vitamin B3	10	7.8%	11	8.5%
Vitamin B6	4	3.1%	4	3.1%
Vitamin B9	9	7.0%	10	7.7%
Vitamin C	8	6.2%	4	3.1%
Vitamin D	22	17.2%	22	16.9%
Vitamin D			2	1.5%
Vitamin E	4	3.1%	3	2.3%
Vitamin K			1	-
Warfarin	11	8.6%	1	-
Zoledronate	1	-		
Zopiclone	1	-		

A.2.2 Initial Approach: Group Lasso

Table A.8: Initial concomitant medication coefficient values for atorvastatin with group lasso

Medication	Coefficient
Ace Inhibitor	
Lisinopril	0
Perindopril	0
Ramipril	0
Alkaloid	
Quinine	0
Alpha I Blocker	
Tamsulosin	0.683
Analgesic	
Acetaminophen	0.075
Angiotensin II Receptor Antagonist	
Candesartan	0.305
Irbesartan	-0.400
Losartan	0.564
Telmisartan	0.191
Valsartan	0.002
Anticoagulant	
Warfarin	-0.024
Anticonvulsant	
Phenytoin	0
Topiramate	0
Antidiabetic	
Metformin	-0.119
Antiplatelet	
Plavix	0
Benzodiazepine	
Clonazepam	0
Lorazepam	0
Beta Agonist	
Salbutamol	0
Beta Blocker	
Acebutolol	0
Atenolol	0
Bisoprolol	0
Metoprolol	0
Calcium Channel Blocker	
Amlodipine	0
Diltiazem	0
Nifedipine	0
Cardiac Glycoside	
Digoxin	0.201
Cholesterol Absorption Blocker	
Ezetimibe	0

Initial concomitant medication coefficient values for atorvastatin with group lasso

Medication	Coefficient
Co-Q10	
Co-Q10	0
Contraceptive	
Contraceptive	0
Corticosteroid	
Prednisone	0
Cox Inhibitor	
Celecoxib	0.369
Direct Renin Inhibitor	
Aliskiren	0
Diuretic	
Chlorthalidone	0
Furosemide	0
Hydrochlorothiazide	0
Indapamide	0
Spironolactone	0
DPP4 Inhibitor	
Sitagliptin	0
Eye Drops	
Multivitamin	0
Fibrate	
Bezafibrate	0
Fenofibrate	0
Gemfibrozil	0
Histamine II Blocker	
Ranitidine	0
Insulin	
Insulin	0
Muscle Relaxant	
Cyclobenzaprine	0
NDRI	
Bupropion	0
NSAID	
Acetylsalicylic Acid	0.070
Diclofenac	0.330
Ibuprofen	0.040
Meloxicam	-0.060
PPARg Agonist	
Rosiglitazone	0
Prostaglandin Analogue	
Misoprostol	0.395

Initial concomitant medication coefficient values for atorvastatin with group lasso

Medication	Coefficient
Proton Pump Inhibitor	
Esomeprazole	0
Lansoprazole	0
Omeprazole	0
Pantoprazole	0
Rabeprazole	0
SSRI	
Citalopram	0
Sertraline	0
SSRI-SNRI	
Venlafaxine	-0.262
Sulfonylurea	
Gliclazide	0
Glyburide	0
Supplement	
Calcium Supplement	0
Fish Oil	0
glucosamine	0
Herbal	0
Iron Supplement	0
Omega 3	0
Thyroid Hormone	
Levothyroxine	-0.232
Tricyclic	
Amitriptyline	0
Vasodilator	
Nitroglycerin	0
Vitamin B	
Vitamin B	0
Vitamin B12	0
Vitamin B3	0
Vitamin B6	0
Vitamin B9	0
Vitamin C	
Vitamin C	0
Vitamin D	
Vitamin D	0
Vitamin E	
Vitamin E	0
Xanthine Oxidase Inhibitor	
Allopurinol	0

A.2.3 Concomitant Selection Algorithm Proportions

Table A.9: Atorvastatin selection algorithm proportions - 1000 repetitions

	Proportion Nonzero	Coefficient Mean	Coefficient SD
Beta Blocker	**0.001**	**0**	**0**
Acebutolol	0	0	0
Atenolol	0.995	-0.14	0.031
Bisoprolol	0	0	0
Metoprolol	0	0	0
Analgesic	**0**	**0**	**0**
Acetaminophen	0	0	0
NSAID	**0.999**	**0.182**	**0.021**
Acetylsalicylic Acid	0.998	0.086	0.02
Diclofenac	0.998	0.406	0.085
Ibuprofen	0	0	0
Meloxicam	0	0	0
Direct Renin Inhibitor	**0**	**0**	**0**
Aliskiren	0	0	0
Xanthine Oxidase Inhibitor	**0**	**0**	**0**
Allopurinol	0	0	0
Tricyclic	**0.001**	**0**	**0.001**
Amitriptyline	0	0	0
Calcium Channel Blocker	**0.999**	**0.068**	**0.011**
Amlodipine	0	0	0
Diltiazem	0	0	0
Nifedipine	0.035	0.001	0.005
Fibrate	**0**	**0**	**0**
Bezafibrate	0	0	0
Fenofibrate	0	0	0
Gemfibrozil	0	0	0
NDRI	**0**	**0**	**0**
Bupropion	0	0	0

Atorvastatin selection algorithm proportions - 1000 repetitions (2/4)

	Proportion Nonzero	Coefficient Mean	Coefficient SD
Supplement	**0**	**0**	**0**
Calcium Supplement	0	0	0
Fish Oil	0	0	0
Glucosamine	0.999	0.273	0.042
Iron Supplement	0	0	0
Omega 3	0	0	0
Angiotensin II Receptor Agonist	**0.997**	**0.103**	**0.025**
Candesartan	0.249	0.014	0.026
Irbesartan	0	0	0
Losartan	0.999	0.853	0.026
Telmisartan	0	0	0
Valsartan	0.816	-0.017	0.015
COX Inhibitor	**0**	**0**	**0**
Celecoxib	0	0	0
Diuretic	**0.989**	**0.043**	**0.015**
Chlorthalidone	0	0	0
Furosemide	0	0	0
Hydrochlorothiazide	0.001	0	0
Indapamide	0	0	0
Spironolactone	0	0	0
SSRI	**0.085**	**0.004**	**0.017**
Citalopram	0	0	0
Sertraline	0	0	0
Benzodiazepine	**0**	**0**	**0**
Clonazepam	0	0	0
Lorazepam	0	0	0
Co Q10	**0**	**0**	**0**
	0	0	0
Contraceptive	**0.382**	**0.032**	**0.056**
	0.035	0	0.002
Muscle Relaxant	**0**	**0**	**0**
Cyclobenzaprine	0	0	0
Cardiac Glycoside	**0.997**	**0.252**	**0.068**
Digoxin	0.816	0.066	0.047

Atorvastatin selection algorithm proportions - 1000 repetitions (3/4)

	Proportion Nonzero	Coefficient Mean	Coefficient SD
Proton Pump Inhibitor	**0**	**0**	**0**
Esomeprazole	0.999	0.34	0.052
Lansoprazole	0.249	-0.007	0.014
Omeprazole	0.249	-0.004	0.009
Pantoprazole	0	0	0
Rabeprazole	0	0	0
Cholesterol Abs. Blocker	**0.085**	**-0.001**	**0.002**
Ezetimibe	0	0	0
Sulfonylurea	**0**	**0**	**0**
Gliclazide	0.999	-0.095	0.004
Glyburide	0	0	0
Insulin	**0**	**0**	**0**
	0	0	0
Thyroid Hormone	**0.999**	**-0.152**	**0.009**
Levothyroxine	0.999	-0.134	0.02
Ace Inhibitor	**0**	**0**	**0**
Lisinopril	0	0	0
Perindopril	0	0	0
Ramipril	0	0	0
Antidiabetic	**0.999**	**-0.214**	**0.041**
Metformin	0.995	-0.125	0.029
Prostaglandin Analogue	**0.989**	**0.21**	**0.068**
Misoprostol	0.001	0	0.001
Eye Drops	**0**	**0**	**0**
Multivitamin	0	0	0
Vasodilator	**0**	**0**	**0**
Nitroglycerin	0	0	0
Anticonvulsant	**0**	**0**	**0**
Phenytoin	0	0	0
Topiramate	0	0	0
Antiplatelet	**0.001**	**0**	**0**
Plavix	0	0	0

Atorvastatin selection algorithm proportions - 1000 repetitions (4/4)

	Proportion Nonzero	Coefficient Mean	Coefficient SD
Corticosteroid	**0.008**	**0**	**0.003**
Prednisone	0	0	0
Alkaloid	**0**	**0**	**0**
Quinine	0	0	0
Histamine II Blocker	**0.008**	**0**	**0.005**
Ranitidine	0.001	0	0
PPARg Agonist	**0**	**0**	**0**
Rosiglitazone	0	0	0
Beta Agonist	**0.188**	**-0.005**	**0.016**
Salbutamol	0.001	0	0
Alpha1 Blocker	**0.999**	**0.588**	**0.087**
Tamsulosin	0.999	0.507	0.059
SSRI/SNRI	**0.998**	**-0.127**	**0.037**
Venlafaxine	0.995	-0.116	0.024
Vitamin B	**0.906**	**-0.046**	**0.034**
Vitamin B	0	0	0
Vitamin B12	0	0	0
Vitamin B3	0.551	-0.013	0.019
Vitamin B6	0	0	0
Vitamin B9	0	0	0
vitamin C	**0**	**0**	**0**
Vitamin C	0	0	0
Vitamin D	**0**	**0**	**0**
	0	0	0
Vitamin E	**0.001**	**0**	**0.001**
	0	0	0
Anticoagulant	**0.03**	**0**	**0.003**
Warfarin	0	0	0

A.2.4 Regression Results for Different Selection Thresholds

Table A.10: Atorvastatin linear regression: 90% cutoff inclusion threshold

	Estimate	Confidence Interval	P-Value	Sig.
(Intercept)	-0.338	-1.454 to 0.779	0.55	
SLCO1B1 c.521T>C	0.411	0.166 to 0.656	0.001	**
SLCO1B1 c.388A>G	-0.141	-0.343 to 0.061	0.169	
Age	0.016	0.006 to 0.026	0.002	**
4β-hydroxholesterol	-0.02	-0.03 to -0.01	<0.001	***
Dose (20mg)	0.741	0.323 to 1.159	0.001	***
Dose (40mg)	1.1	0.729 to 1.472	<0.001	***
Dose (80mg)	1.601	1.175 to 2.027	<0.001	***
Time Post Dose (hr)	-0.085	-0.109 to -0.06	<0.001	***
BMI (kg/m^2)	0.015	-0.009 to 0.04	0.225	
Gender (Male = 1)	-0.051	-0.295 to 0.192	0.677	
Ethnicity (Non-Caucasian = 1)	0.063	-0.296 to 0.422	0.729	
Acetylsalicylic Acid	0.186	-0.089 to 0.461	0.183	
Atenolol	-0.304	-0.741 to 0.134	0.172	
Candesartan	0.756	-0.004 to 1.517	0.051	.
Diclofenac	1.073	-0.041 to 2.187	0.059	.
Digoxin	0.585	-0.061 to 1.231	0.075	.
Esomeprazole	0.602	-0.231 to 1.435	0.155	
Gliclazide	-0.347	-1.083 to 0.39	0.353	
Glucosamine	0.448	-0.401 to 1.297	0.298	
Hydrochlorothiazide	0.095	-0.277 to 0.467	0.613	
Levothyroxine	-0.395	-0.855 to 0.065	0.091	.
Losartan	0.978	0.392 to 1.564	0.001	**
Metformin	-0.349	-0.674 to -0.023	0.036	*
Misoprostol	0.162	-0.905 to 1.23	0.764	
Nifedipine	0.34	-0.348 to 1.028	0.329	
Tamsulosin	1.076	0.25 to 1.903	0.011	*
Valsartan	-0.114	-0.81 to 0.582	0.746	
Venlafaxine	-0.247	-0.967 to 0.472	0.497	
Vitamin B3	-0.392	-0.839 to 0.054	0.084	.

Table A.11: Atorvastatin linear regression: 95% cutoff inclusion threshold

	Estimate	Confidence Interval	P-Value	Sig.
(Intercept)	-0.53	-1.636 to 0.577	0.344	
SLCO1B1 c.521T>C	0.385	0.14 to 0.631	0.002	**
SLCO1B1 c.388A>G	-0.136	-0.34 to 0.068	0.188	
Age	0.017	0.007 to 0.027	0.001	***
4β-hydroxholesterol	-0.019	-0.028 to -0.009	<0.001	***
Dose (20mg)	0.773	0.352 to 1.194	<0.001	***
Dose (40mg)	1.137	0.764 to 1.51	<0.001	***
Dose (80mg)	1.637	1.208 to 2.066	<0.001	***
Time Post Dose (hr)	-0.081	-0.105 to -0.056	<0.001	***
BMI (kg/m^2)	0.015	-0.01 to 0.04	0.227	
Gender (Male = 1)	-0.031	-0.276 to 0.214	0.802	
Ethnicity (Non-Caucasian = 1)	0.068	-0.295 to 0.431	0.712	
Acetylsalicylic Acid	0.17	-0.107 to 0.447	0.227	
Atenolol	-0.39	-0.821 to 0.04	0.075	.
Candesartan	0.639	-0.117 to 1.395	0.097	.
Diclofenac	1.012	-0.111 to 2.136	0.077	.
Digoxin	0.582	-0.07 to 1.235	0.08	.
Esomeprazole	0.703	-0.13 to 1.536	0.097	.
Gliclazide	-0.322	-1.065 to 0.422	0.393	
Glucosamine	0.563	-0.285 to 1.41	0.191	
Hydrochlorothiazide	0.086	-0.29 to 0.462	0.649	
Levothyroxine	-0.352	-0.814 to 0.11	0.134	
Losartan	1.019	0.429 to 1.61	0.001	***
Metformin	-0.323	-0.651 to 0.005	0.054	.
Misoprostol	0.14	-0.938 to 1.219	0.797	
Nifedipine	0.353	-0.342 to 1.047	0.316	
Tamsulosin	0.953	0.13 to 1.776	0.024	*
Valsartan	-0.155	-0.856 to 0.547	0.663	
Venlafaxine	-0.252	-0.979 to 0.475	0.493	

Table A.12: Atorvastatin linear regression: 99% cutoff inclusion threshold

	Estimate	Confidence Interval	P-Value	Sig.
(Intercept)	-0.534	-1.635 to 0.568	0.339	
SLCO1B1 c.521T>C	0.388	0.144 to 0.633	0.002	**
SLCO1B1 c.388A>G	-0.141	-0.343 to 0.061	0.169	
Age	0.017	0.007 to 0.027	0.001	***
4β-hydroxholesterol	-0.019	-0.028 to -0.009	<0.001	***
Dose (20mg)	0.764	0.347 to 1.182	<0.001	***
Dose (40mg)	1.128	0.759 to 1.497	<0.001	***
Dose (80mg)	1.627	1.202 to 2.051	<0.001	***
Time Post Dose (hr)	-0.08	-0.104 to -0.056	<0.001	***
BMI (kg/m^2)	0.016	-0.009 to 0.04	0.217	
Gender (Male = 1)	-0.023	-0.265 to 0.218	0.85	
Ethnicity (Non-Caucasian = 1)	0.066	-0.295 to 0.428	0.717	
Acetylsalicylic Acid	0.166	-0.11 to 0.441	0.235	
Atenolol	-0.381	-0.808 to 0.046	0.08	.
Candesartan	0.637	-0.116 to 1.39	0.096	.
Diclofenac	1.011	-0.107 to 2.13	0.076	.
Digoxin	0.584	-0.066 to 1.234	0.078	.
Esomeprazole	0.731	-0.09 to 1.552	0.08	.
Gliclazide	-0.307	-1.045 to 0.431	0.411	
Glucosamine	0.615	-0.199 to 1.428	0.137	
Levothyroxine	-0.357	-0.817 to 0.102	0.126	
Losartan	1.026	0.438 to 1.613	0.001	***
Metformin	-0.325	-0.652 to 0.001	0.051	.
Misoprostol	0.154	-0.918 to 1.226	0.776	
Nifedipine	0.382	-0.298 to 1.062	0.268	
Tamsulosin	0.935	0.119 to 1.75	0.025	*
Valsartan	-0.163	-0.86 to 0.534	0.643	
Venlafaxine	-0.256	-0.98 to 0.468	0.484	

A.3 Rosuvastatin Concomitant Medication Analysis Supplemental Tables

A.3.1 Concomitant Selection Algorithm Proportions

Table A.13: Rosuvastatin selection algorithm proportions

	Proportion Nonzero	Coefficient Mean	Coefficient SD
Beta Blocker	**0.022**	**0**	**0.005**
Acebutolol	0	0	0
Atenolol	0	0	0
Bisoprolol	0	0	0
Metoprolol	0	0	0
Timolol	0	0	0
Analgesic	**0.003**	**0.002**	**0.061**
Acetaminophen	0	0	0
Xanthine Oxidase Inhibitor	**0.001**	**0**	**0.005**
Allopurinol	0	0	0
Benzodiazepine	**0.091**	**-0.011**	**0.057**
Alprazolam	0	0	0
Temazepam	0	0	0
Calcium Channel Blocker	**0.002**	**0**	**0.004**
Amlodipine	0	0	0
Diltiazem	0	0	0
Felodipine	0	0	0
Nifedipine	0	0	0
Prostaglandin Analogue	**0.001**	**0**	**0.012**
Bimatoprost	0	0	0
Corticosteroid	**0.002**	**0.001**	**0.02**
Budesonide	0	0	0
NDRI	**0.003**	**0.001**	**0.03**
Bupropion	0	0	0
Angiotensin II Receptor Agonist	**0.003**	**0**	**0.003**
Candesartan	0	0	0
Irbesartan	0.019	-0.004	0.027
Losartan	0	0	0
Telmisartan	0	0	0
Valsartan	0	0	0
Antihistamine	**0.001**	**0.001**	**0.017**
Cetirizine	0	0	0

Rosuvastatin selection algorithm proportions (2/3)

	Proportion Nonzero	Coefficient Mean	Coefficient SD
SSRI	**0.035**	**0.003**	**0.016**
Citalopram	0	0	0
Escitalopram	0	0	0
Antiinflammatory	**0.002**	**0**	**0.006**
Colchicine	0	0	0
NSAID	**0.144**	**-0.023**	**0.075**
Diclofenac	0	0	0
Ibuprofen	0.149	-0.079	0.226
Gastroprokinetic	**0.002**	**-0.001**	**0.017**
Domperidone	0	0	0
Alpha1 Blocker	**0.014**	**0.001**	**0.014**
Doxazosin	0	0	0
Tamsulosin	0	0	0
Terazosin	0	0	0
Ace Inhibitor	**0.001**	**0**	**0.005**
Enalapril	0	0	0
Lisinopril	0	0	0
Perindopril	0	0	0
Quinapril	0	0	0
Ramipril	0	0	0
Trandolapril	0	0	0
Cholesterol Absorption Blocker	**0.001**	**0**	**0.001**
Ezetimibe	0	0	0
Fibrate	**0.001**	**0**	**0.005**
Fenofibrate	0	0	0
Beta Agonist	**0.001**	**0**	**0.002**
Formoterol	0	0	0
Salbutamol	0	0	0
Diuretic	**0.153**	**0.013**	**0.035**
Furosemide	0	0	0
Hydrochlorothiazide	0.013	0.001	0.005
Indapamide	0	0	0
Spironolactone	0	0	0
Anticonvulsant Analgesic	**0.002**	**0.003**	**0.069**
Gabapentin	0	0	0
Sulfonylurea	**0.121**	**0.006**	**0.021**
Gliclazide	0	0	0
Glyburide	0.006	0	0.002
Supplement	**0.003**	**0.002**	**0.035**
Glucosamine	0	0	0

Rosuvastatin selection algorithm proportions (3/3)

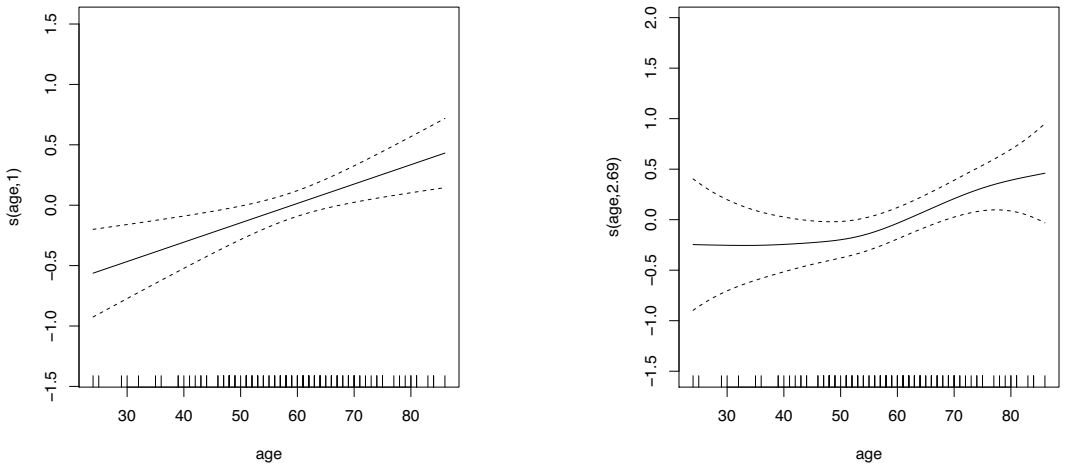
	Proportion Nonzero	Coefficient Mean	Coefficient SD
Antimalarial	**0.001**	**0**	**0.008**
Hydroxychloroquine	0	0	0
Insulin	**0.035**	**0.001**	**0.011**
	0	0	0
Proton Pump Inhibitor	**0.001**	**0**	**0.003**
Lansoprazole	0	0	0
Omeprazole	0	0	0
Pantoprazole	0	0	0
Rabeprazole	0	0	0
Thyroid Hormone	**0.002**	**0**	**0.005**
Levothyroxine	0	0	0
Antidiabetic	**0.001**	**0**	**0.003**
Metformin	0	0	0
Chemotherapy	**0.001**	**0.002**	**0.05**
Methotrexate	0	0	0
Eye Drops	**0.167**	**-0.054**	**0.123**
Multivitamin	0.112	-0.018	0.06
Vasodilator	**0.001**	**0**	**0.001**
Nitroglycerin	0	0	0
Antiplatelet	**0.002**	**0**	**0.006**
Plavix	0	0	0
Dopamine Serotonin Adrenergic Antagonist	**0.002**	**-0.006**	**0.154**
Quetiapine	0	0	0
Alkaloid	**0.001**	**0.001**	**0.016**
Quinine	0	0	0
Histamine II Blocker	**0.999**	**-0.43**	**0.576**
Ranitidine	0.896	-0.221	0.335
Bisphosphonate	**0.035**	**-0.006**	**0.059**
Risedronate	0	0	0
PDE-5 Inhibitor	**0.001**	**0**	**0.012**
Tadalafil	0	0	0
Antimuscarinic	**0.001**	**0**	**0.013**
Tolterodine	0	0	0
Anticonvulsant	**0.035**	**0.01**	**0.152**
Trazodone	0	0	0
SSRI/SNRI	**0.001**	**0**	**0.004**
Venlafaxine	0	0	0

Appendix B

Non-Linear Modelling Supplemental Results

B.1 GAM Smoothing Parameter Graphs

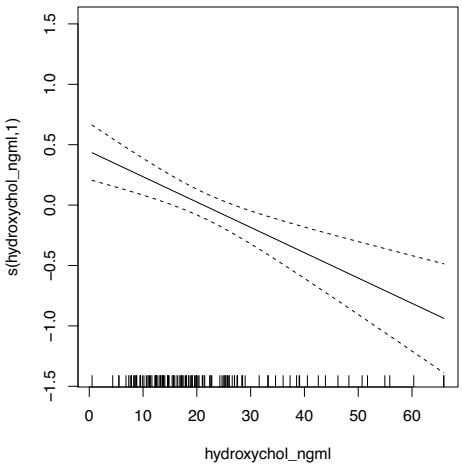
B.1.1 Atorvastatin



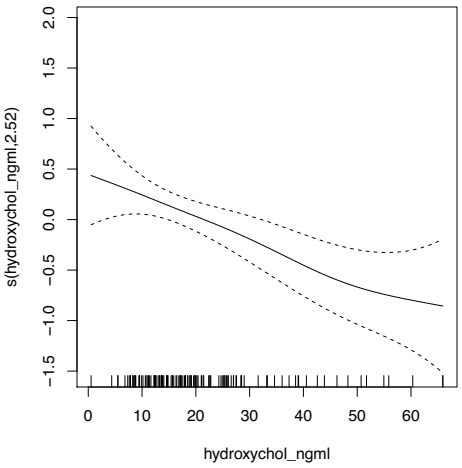
(a) CV smoothing parameters

(b) Fixed smoothing parameters

Figure B.1: Atorvastatin GAM smoothing for Age covariate

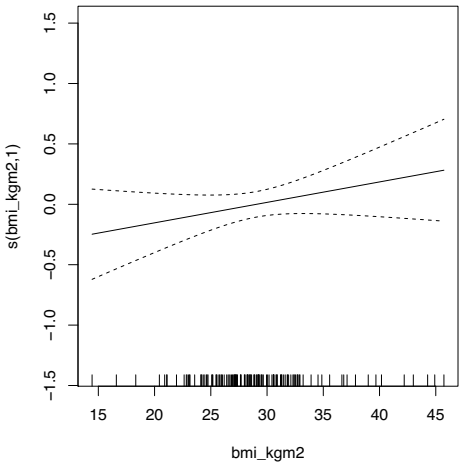


(a) CV smoothing parameters

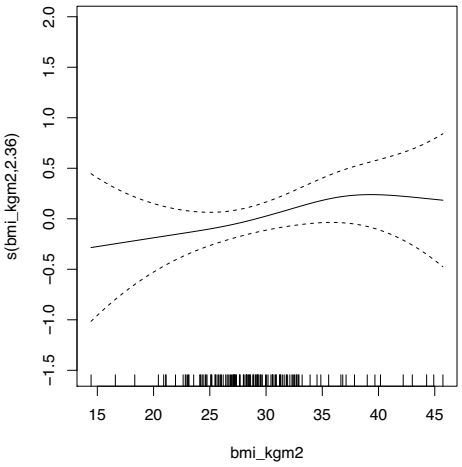


(b) Fixed smoothing parameters

Figure B.2: Atorvastatin GAM smoothing for 4β-hydroxycholesterol covariate

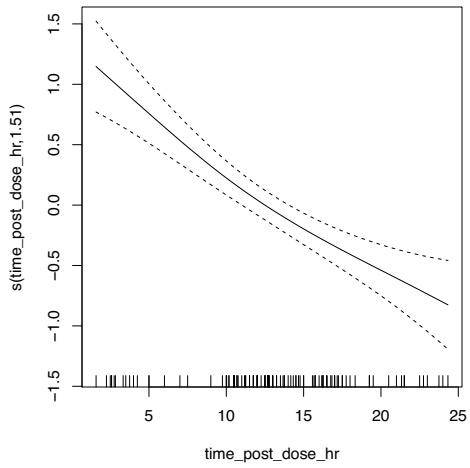


(a) CV smoothing parameters

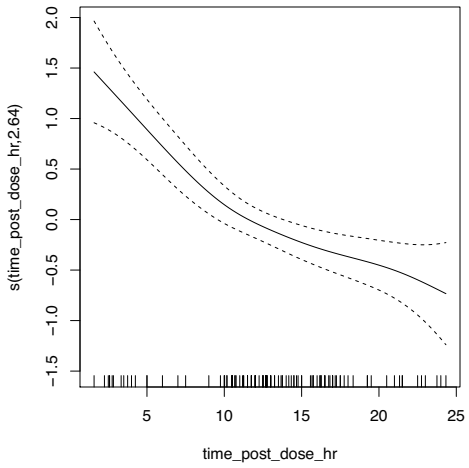


(b) Fixed smoothing parameters

Figure B.3: Atorvastatin GAM smoothing for BMI covariate



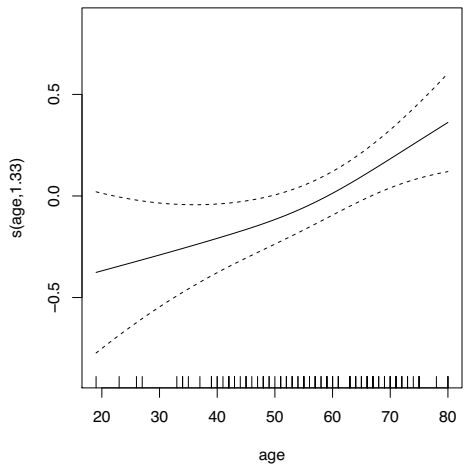
(a) CV smoothing parameters



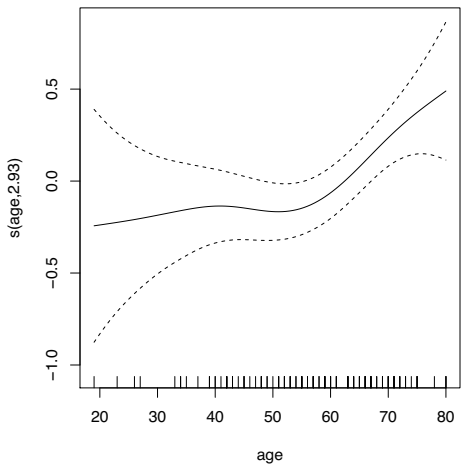
(b) Fixed smoothing parameters

Figure B.4: Atorvastatin GAM smoothing for Time Post Dose (h) covariate

B.1.2 Rosuvastatin

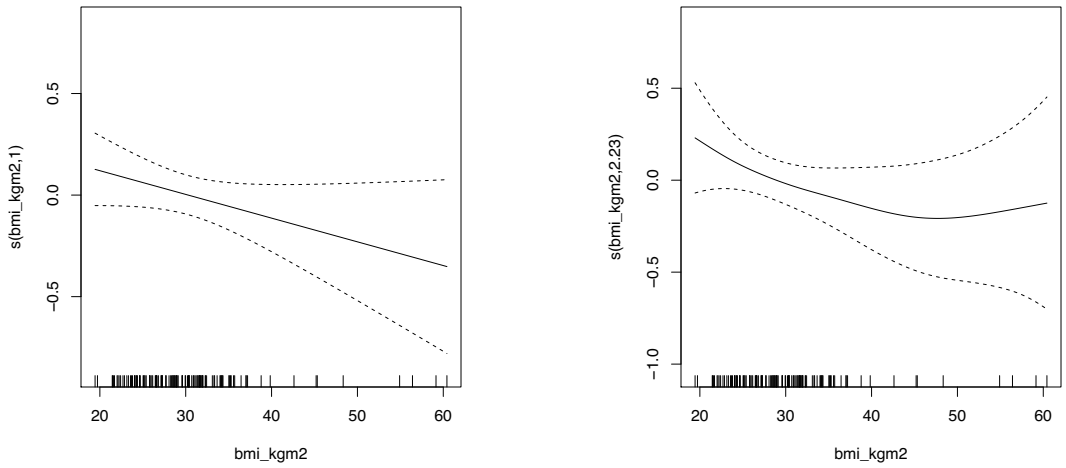


(a) CV smoothing parameters



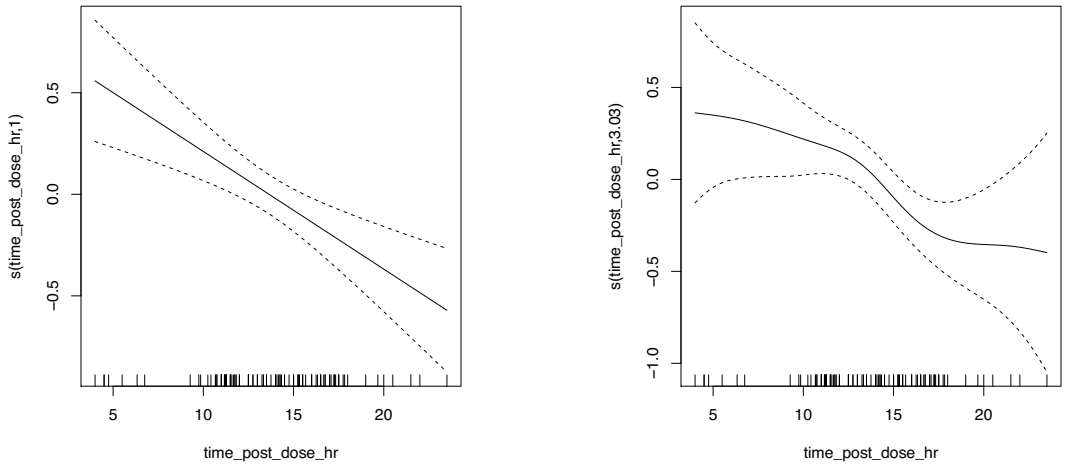
(b) Fixed smoothing parameters

Figure B.5: Rosuvastatin reduced-model linear kernel SVR tuning



(a) CV smoothing parameters (b) Fixed smoothing parameters

Figure B.6: Rosuvastatin GAM smoothing for BMI covariate

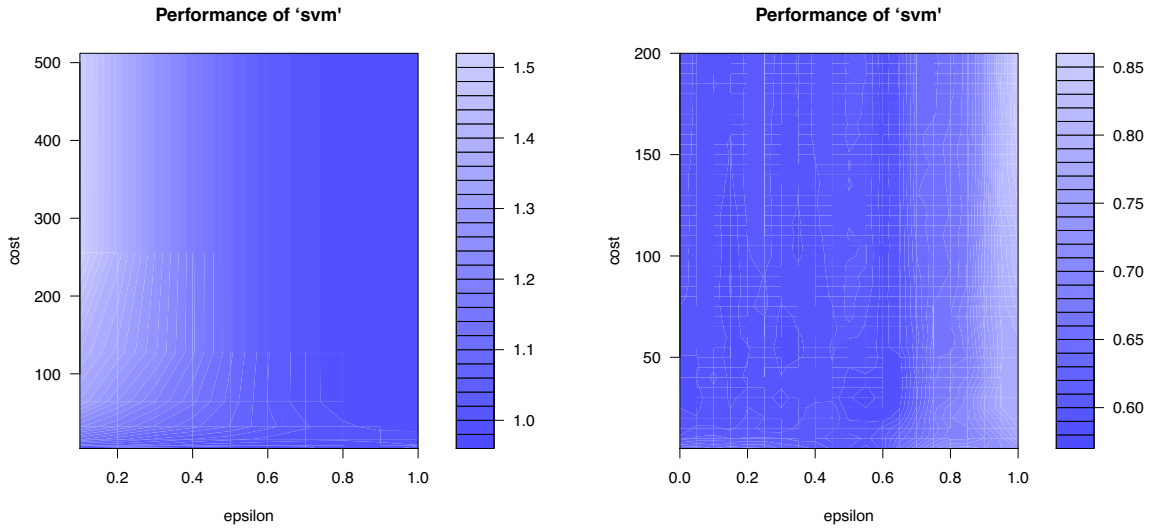


(a) CV smoothing parameters (b) Fixed smoothing parameters

Figure B.7: Rosuvastatin GAM smoothing for Time Post Dose (h) covariate

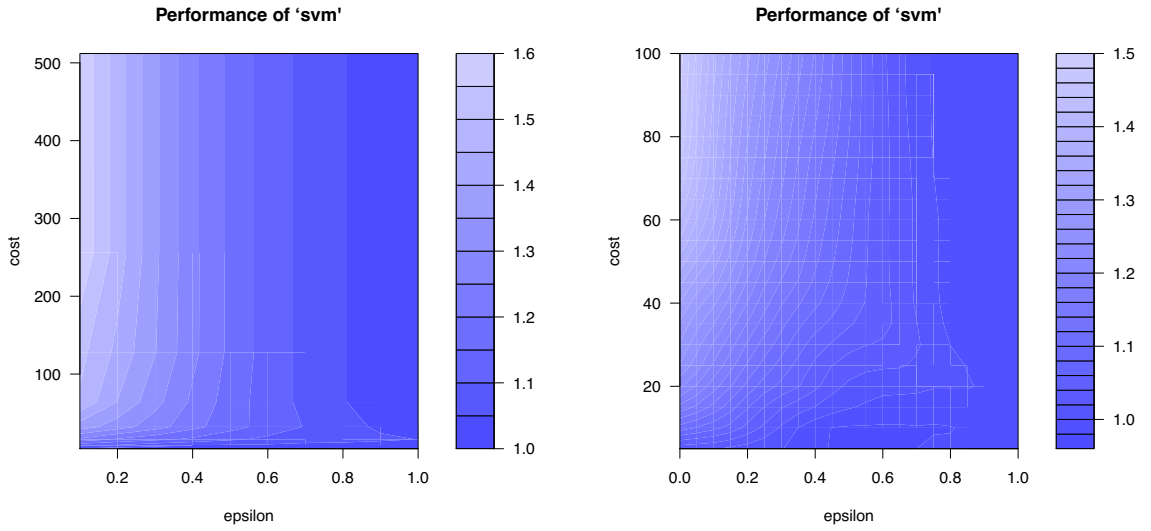
B.2 SVR Model Tuning Graphs

B.2.1 Atorvastatin



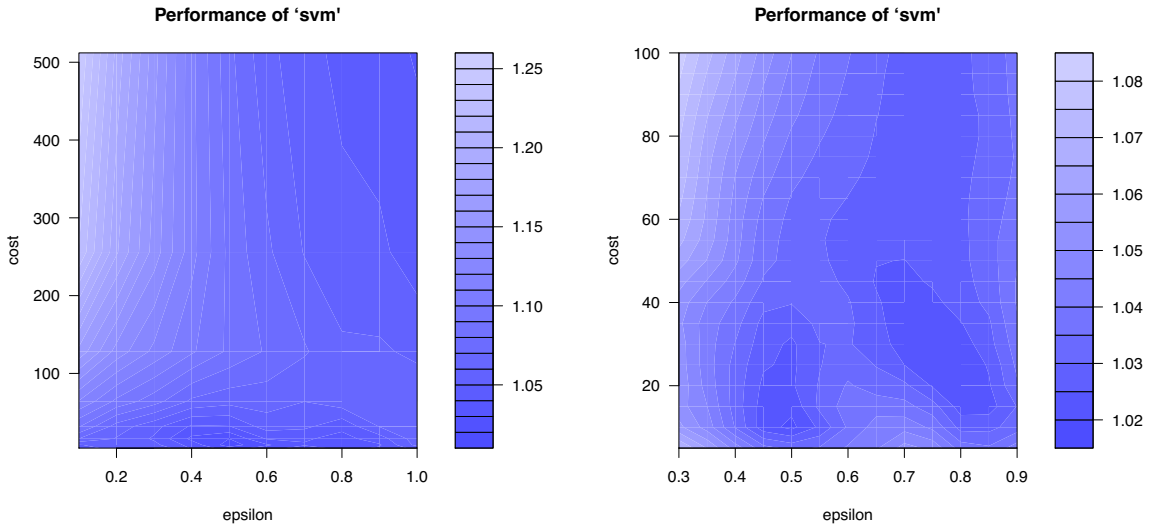
(a) SVR coarse tune (b) SVR fine tune

Figure B.8: Atorvastatin reduced-model linear kernel SVR tuning



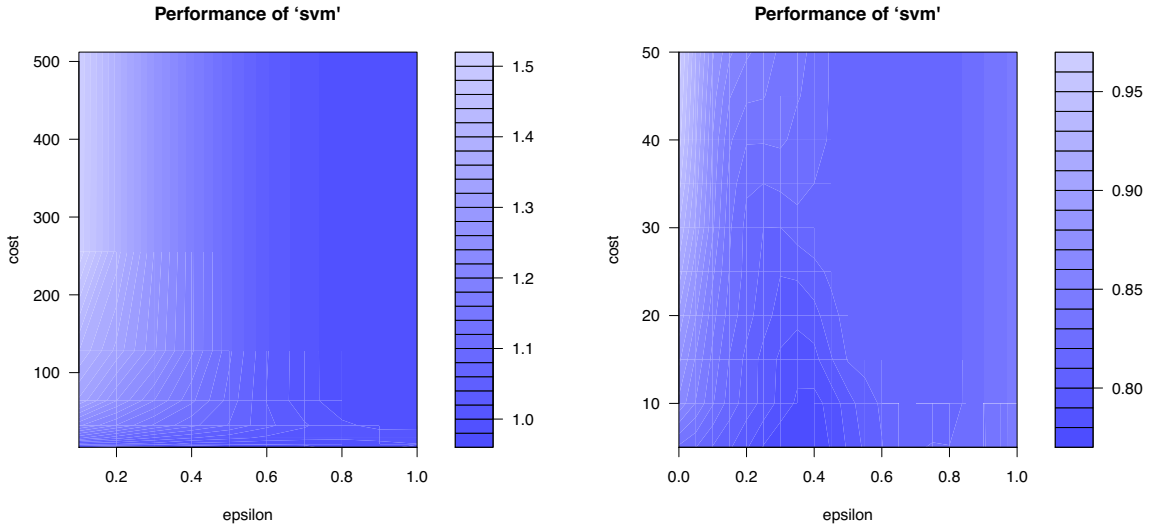
(a) SVR coarse tune (b) SVR fine tune

Figure B.9: Atorvastatin reduced-model degree 3 polynomial SVR tuning



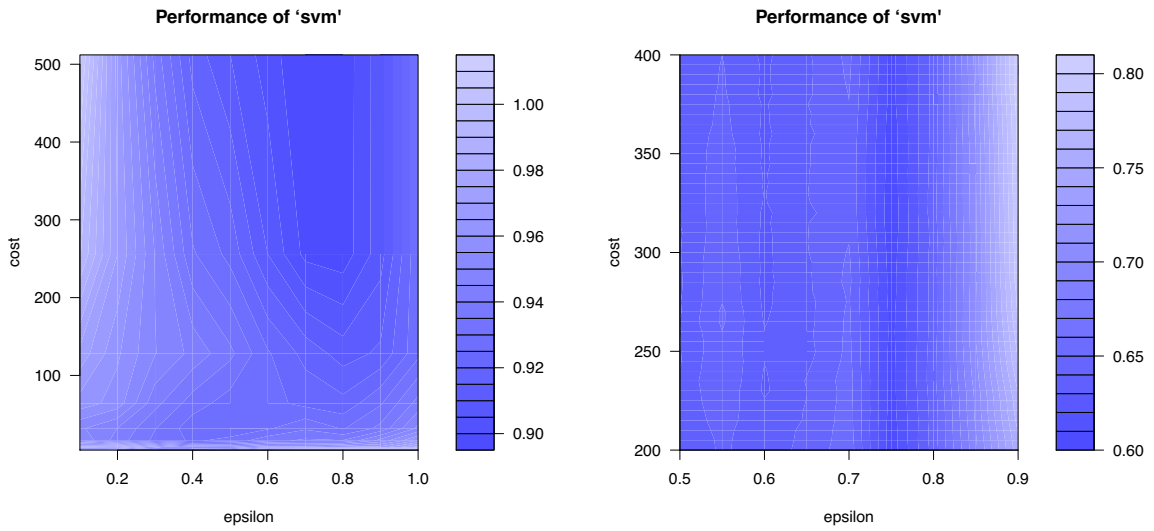
(a) SVR coarse tune (b) SVR fine tune

Figure B.10: Atorvastatin reduced-model degree 5 polynomial SVR tuning



(a) SVR coarse tune (b) SVR fine tune

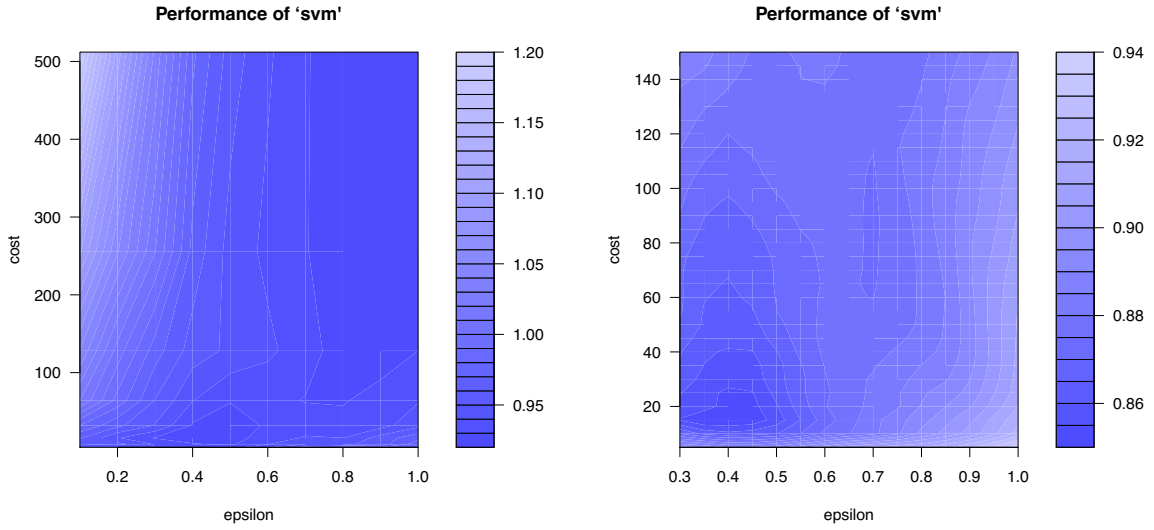
Figure B.11: Atorvastatin reduced-model radial kernel SVR tuning



(a) SVR coarse tune

(b) SVR fine tune

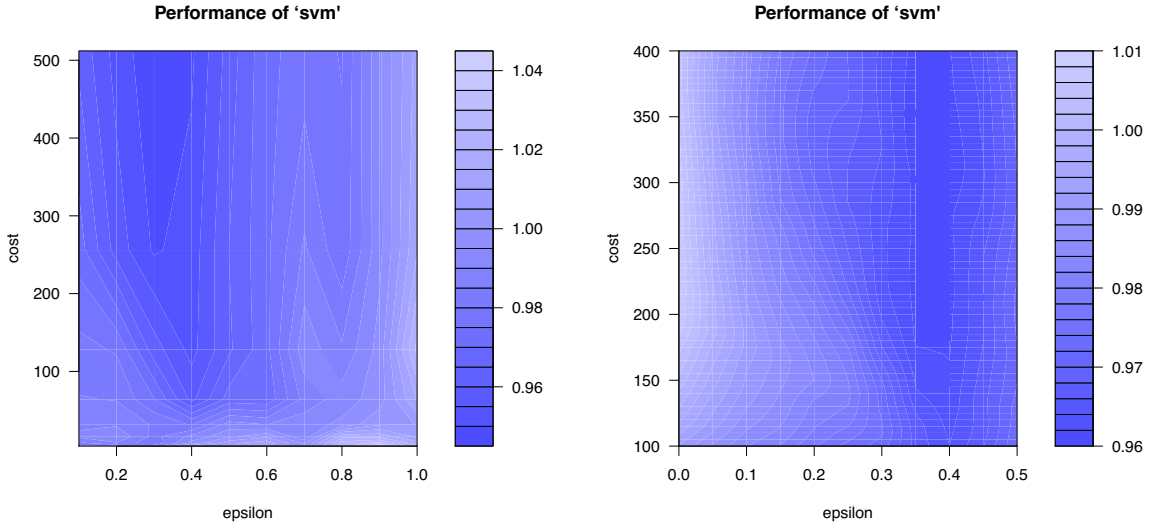
Figure B.12: Atorvastatin full concomitant medication model linear kernel SVR tuning



(a) SVR coarse tune

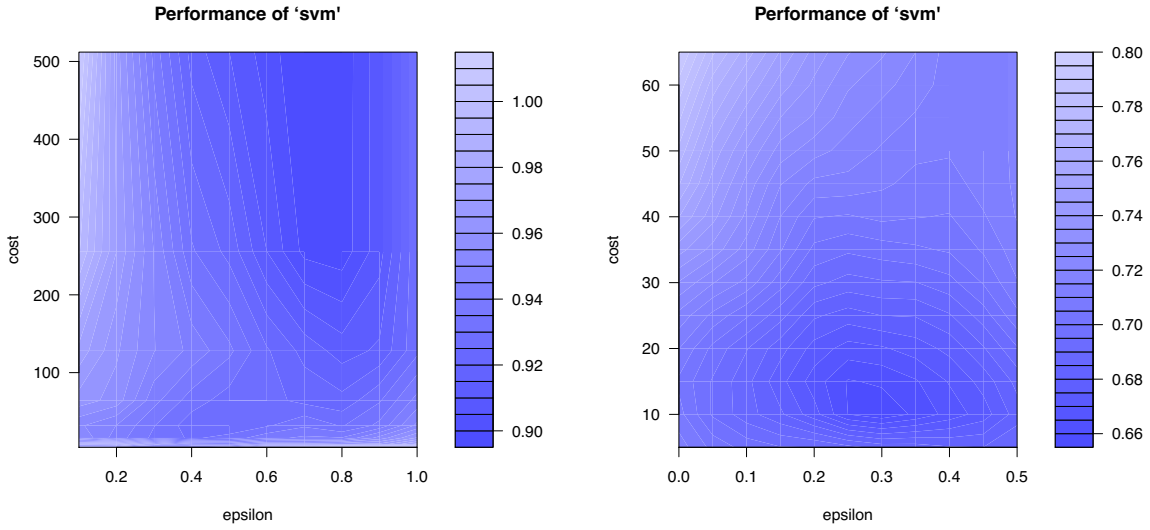
(b) SVR fine tune

Figure B.13: Atorvastatin full concomitant medication model degree 3 polynomial SVR tuning



(a) SVR coarse tune (b) SVR fine tune

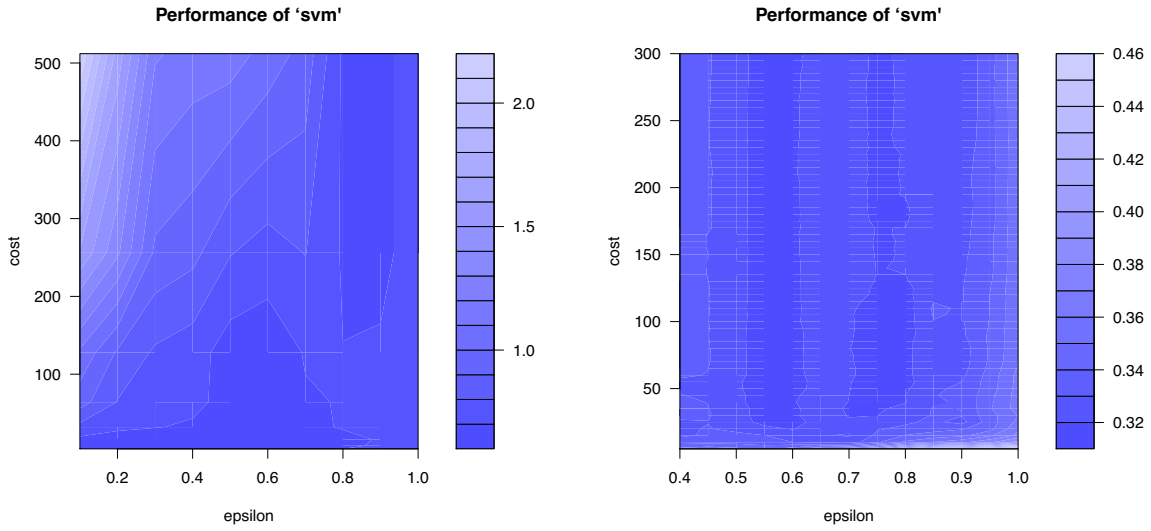
Figure B.14: Atorvastatin full concomitant medication model degree 5 polynomial SVR tuning



(a) SVR coarse tune (b) SVR fine tune

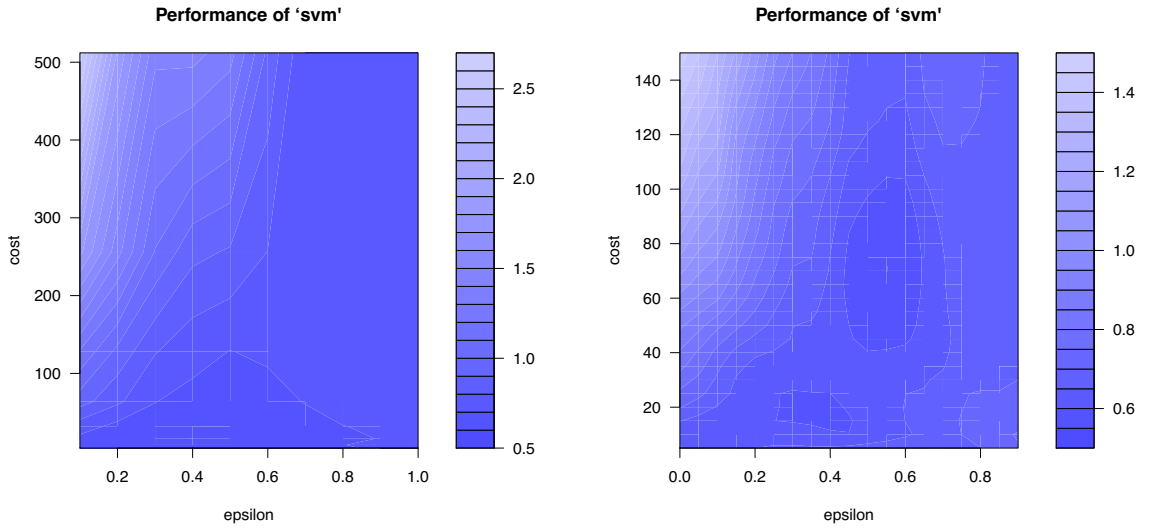
Figure B.15: Atorvastatin full concomitant medication model radial kernel SVR tuning

B.2.2 Rosuvastatin



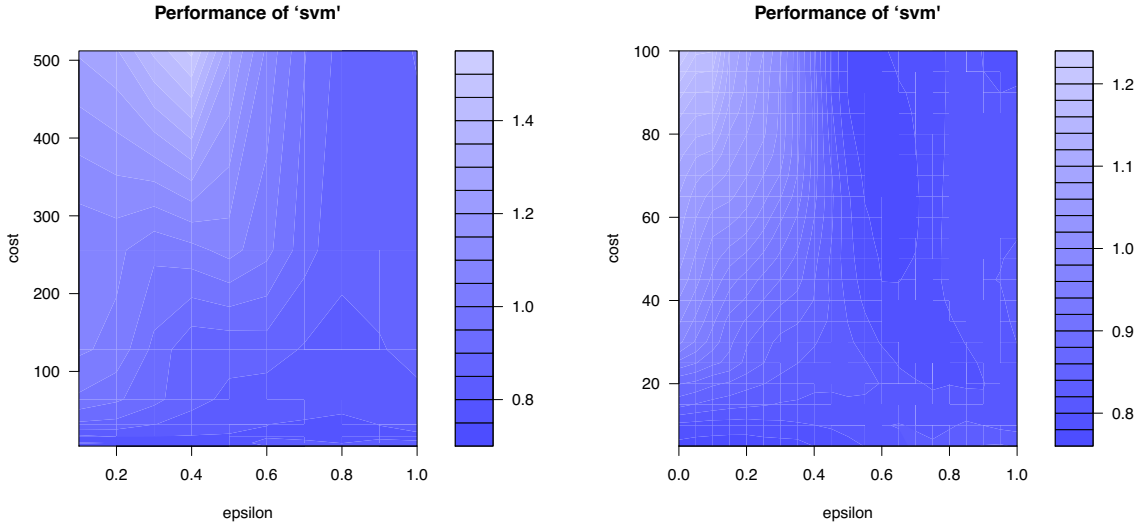
(a) SVR coarse tune (b) SVR fine tune

Figure B.16: Rosuvastatin linear kernel SVR tuning



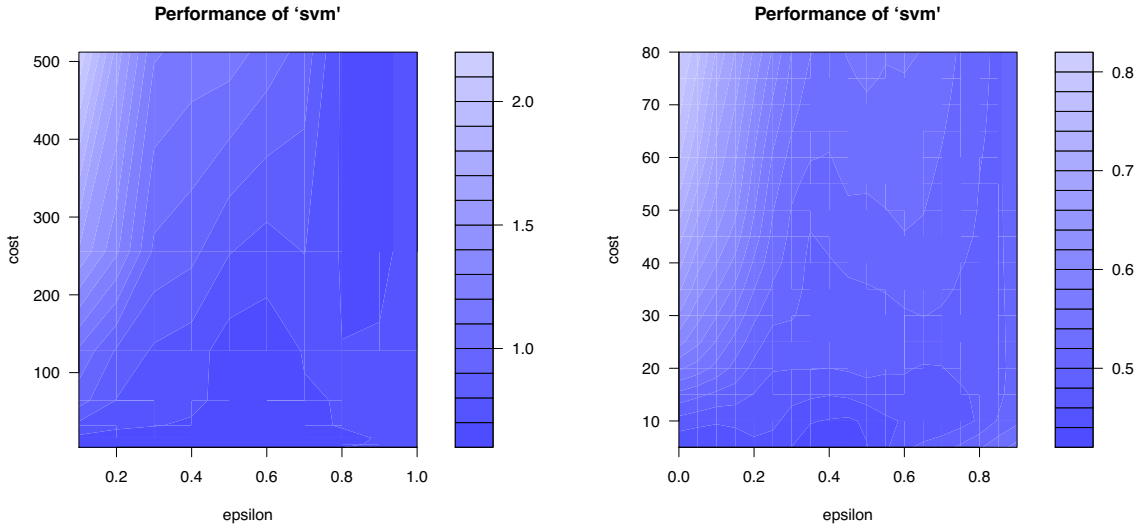
(a) SVR coarse tune (b) SVR fine tune

Figure B.17: Rosuvastatin degree 3 polynomial kernel SVR tuning



(a) SVR coarse tune (b) SVR fine tune

Figure B.18: Rosuvastatin degree 5 polynomial kernel SVR tuning



(a) SVR coarse tune (b) SVR fine tune

Figure B.19: Rosuvastatin radial kernel SVR tuning

Appendix C

Rosuvastatin NGS Novel Variant Identification

C.1 Supplemental Tables: Identified Variant Allele Frequencies

Table C.1: ABCC1 variant allele frequencies

Chromosome	Position	Case MAF (%)	Control MAF (%)
16	16043022	0	1
16	16043174	70	70
16	16101875	2.5	2
16	16108282	35	27
16	16108642	7.5	5
16	16108730	0	1
16	16110244	32.5	42
16	16110253	27.5	41
16	16110848	0	1
16	16126758	20	13
16	16126764	5	26
16	16126986	0	1
16	16127235	22.5	17
16	16130491	0	2
16	16130514	5	16
16	16130524	22.5	14
16	16130701	0	1
16	16138076	7.5	13
16	16138086	2.5	6
16	16138204	0	1
16	16138313	5	3
16	16138322	27.5	24
16	16139714	27.5	24
16	16139878	27.5	24
16	16140041	10	5
16	16141810	12.5	8
16	16141823	25	6
16	16141835	0	1
16	16142224	5	2
16	16142358	10	5
16	16146795	15	11
16	16149759	0	1
16	16149871	10	7
16	16149901	10	4
16	16150208	5	9
16	16150364	2.5	0
16	16150375	0	2
16	16161928	7.5	3

ABCC1 variant minor allele frequencies (continued)

Chromosome	Position	Case MAF (%)	Control MAF (%)
16	16161976	35	40
16	16162019	75	79
16	16162039	0	9
16	16162264	0	6
16	16162338	5	2
16	16165289	5	11
16	16170477	0	1
16	16173221	0	1
16	16173232	0	4
16	16173548	2.5	3
16	16180896	0	1
16	16184232	32.5	42
16	16184235	0	1
16	16184623	40	44
16	16192565	5	5
16	16196309	7.5	9
16	16196833	0	1
16	16196839	0	1
16	16200756	40	52
16	16200908	40	52
16	16205141	15	27
16	16205143	15	27
16	16205161	15	27
16	16205501	7.5	21
16	16208928	0	1
16	16216139	0	1
16	16218641	0	1
16	16225538	5	0
16	16225971	0	2
16	16228242	20	17
16	16228282	2.5	0
16	16228482	0	1
16	16228548	0	1
16	16230069	5	1
16	16230290	25	16
16	16230427	0	1
16	16232433	5	10
16	16232607	0	10

ABCC1 variant minor allele frequencies (continued)

Chromosome	Position	Case MAF (%)	Control MAF (%)
16	16235366	2.5	1
16	16235515	0	1
16	16235681	10	16
16	16235939	37.5	42
16	16236004	22.5	37
16	16236138	0	1
16	16236431	0	3
16	16236483	2.5	0
16	16236523	90	92
16	16236650	12.5	14
16	16237379	5	5
16	16237456	0	1

Table C.2: NR1I2 variant minor allele frequencies

Chromosome	Position	Case MAF (%)	Control MAF (%)
3	119498808	5	15
3	119499015	90	92
3	119499507	52.5	46
3	119499608	90	91
3	119499856	0	1
3	119500035	25	25
3	119500664	90	91
3	119501039	50	46
3	119501263	0	1
3	119501307	17.5	16
3	119501327	2.5	0
3	119501780	40	51
3	119501798	40	52
3	119526203	5	2
3	119526349	20	34
3	119526372	20	34
3	119526654	0	2
3	119529113	40	50
3	119529605	0	1
3	119529689	2.5	6
3	119530027	97.5	87
3	119530141	2.5	2
3	119530858	2.5	10
3	119532652	5	5
3	119532980	0	1
3	119533733	17.5	28
3	119533773	37.5	50
3	119533910	100	100
3	119534097	0	1
3	119534153	10	7
3	119535780	5	2
3	119535795	2.5	8
3	119536429	85	80
3	119536559	10	9
3	119536575	0	1
3	119536581	70	76
3	119536718	5	6
3	119536817	100	97
3	119536897	10	9
3	119536926	100	96
3	119537254	2.5	10
3	119537291	2.5	10
3	119537353	40	52
3	119537625	0	1

Table C.3: SLCO1B3 variant minor allele frequencies

Chromosome	Position	Case MAF (%)	Control MAF (%)
12	20963135	5	6
12	20963307	2.5	0
12	20966330	7.5	5
12	20966451	5	1
12	20966548	47.5	54
12	20966590	0	4
12	20966681	2.5	2
12	20966722	7.5	3
12	20968527	0	1
12	20968828	40	46
12	20968982	7.5	5
12	20975338	7.5	5
12	20975408	7.5	22
12	20975801	0	1
12	21008356	0	4
12	21011235	20	42
12	21011296	20	42
12	21011310	20	42
12	21011480	77.5	74
12	21011581	17.5	29
12	21013641	0	2
12	21013678	5	6
12	21013948	77.5	74
12	21014025	2.5	0
12	21014062	2.5	0
12	21014139	20	42
12	21014163	20	42
12	21014178	5	5
12	21014269	15	22
12	21014343	0	2
12	21015046	0	15
12	21015075	27.5	24
12	21015119	2.5	1
12	21015139	27.5	38
12	21015205	20	42
12	21015243	20	42
12	21015526	0	1
12	21015610	20	42
12	21015760	77.5	74

SLCO1B3 variant minor allele frequencies (continued)

Chromosome	Position	Case MAF (%)	Control MAF (%)
12	21015815	5	1
12	21015864	5	1
12	21015906	10	14
12	21028093	17.5	41
12	21028208	10	14
12	21030454	0	2
12	21030582	20	42
12	21030584	2.5	1
12	21030590	17.5	29
12	21030672	0	1
12	21031020	20	42
12	21031076	17.5	14
12	21031153	5	4
12	21032173	5	4
12	21032242	25	44
12	21034017	2.5	1
12	21034198	0	1
12	21036102	0	14
12	21036168	20	43
12	21036270	20	42
12	21036300	2.5	0
12	21036411	77.5	74
12	21036502	0	1
12	21036634	0	1
12	21036683	35	34
12	21036686	2.5	0
12	21051169	0	1
12	21051489	30	31
12	21054369	80	78
12	21068699	10	17
12	21069049	7.5	1
12	21069690	0	1
12	21069803	0	1
12	21070135	5	0
12	21070137	100	92
12	21070243	0	1

Appendix D

Machine Learning Clinic (MLC)

D.1 Overall Development Goal

Throughout the completion of this doctoral thesis, I have gained an immeasurable wealth of experience from collaborating with many clinical and pharmacological researchers. These researchers are experts in their field, and make meaningful substantive research possible for biostatisticians and other researchers interested in interdisciplinary modelling. A barrier that I have noticed in many of these interactions has been the unavailability and inaccessibility of advanced modelling techniques to researchers who are experts in their substantive field, but not necessarily experts in statistical modelling. There are many opportunities for interdisciplinary knowledge transfer going forward, particularly in the area of medical research, as statisticians also need to know the substantive questions that are meaningful and of interest to clinicians. In order to facilitate easier access to statistical modelling that does not involve knowing a statistical programming language (like R, SAS, or Stata), I began developing an interactive statistical platform called Machine Learning Clinic (MLC). Initially, the goal of development was to pro-

vide a platform to perform the analysis needed for the different research questions addressed in this thesis, and that could be easily adaptable to other research domains within clinical pharmacology, epidemiology, and other substantive areas. Eventually, the goal of the platform became helping researchers more broadly to make informed decisions about regression and modelling parameter selection for the specific context of clinical pharmacology, and providing extensive documentation allowing the user to see exactly what model parameters are being used in each model, and why. MLC integrates information about statistical interpretation for use as an educational tool and to improve statistical interpretation in journal publications about the statistics used for different analyses. Because the eventual dissertation did not include quality testing and research about the efficacy of the platform, I have included it in the appendix as supplemental material. The remainder of the appendix will document the development process, and provide screenshots of the application to show the user interface and statistical outputs.

D.2 MLC Software Development

The software was developed in Shiny, a web application framework for R[1]. RStudio was used as a development environment for writing this application[2].

D.2.1 Data Import Design

The data loading interface provides a file browser for the user to select a comma separated values (.csv) file containing their data. The first row of the file is expected to contain variable names for each column in the data set. The file may contain an arbitrary number of rows, and having more covariates than observations is permissible. After the file loads, the user is

provided with a preview table, and a warning message detailing the number of missing observations removed from the data set, as MLC does not yet include data imputation functionality (Figure D.1).

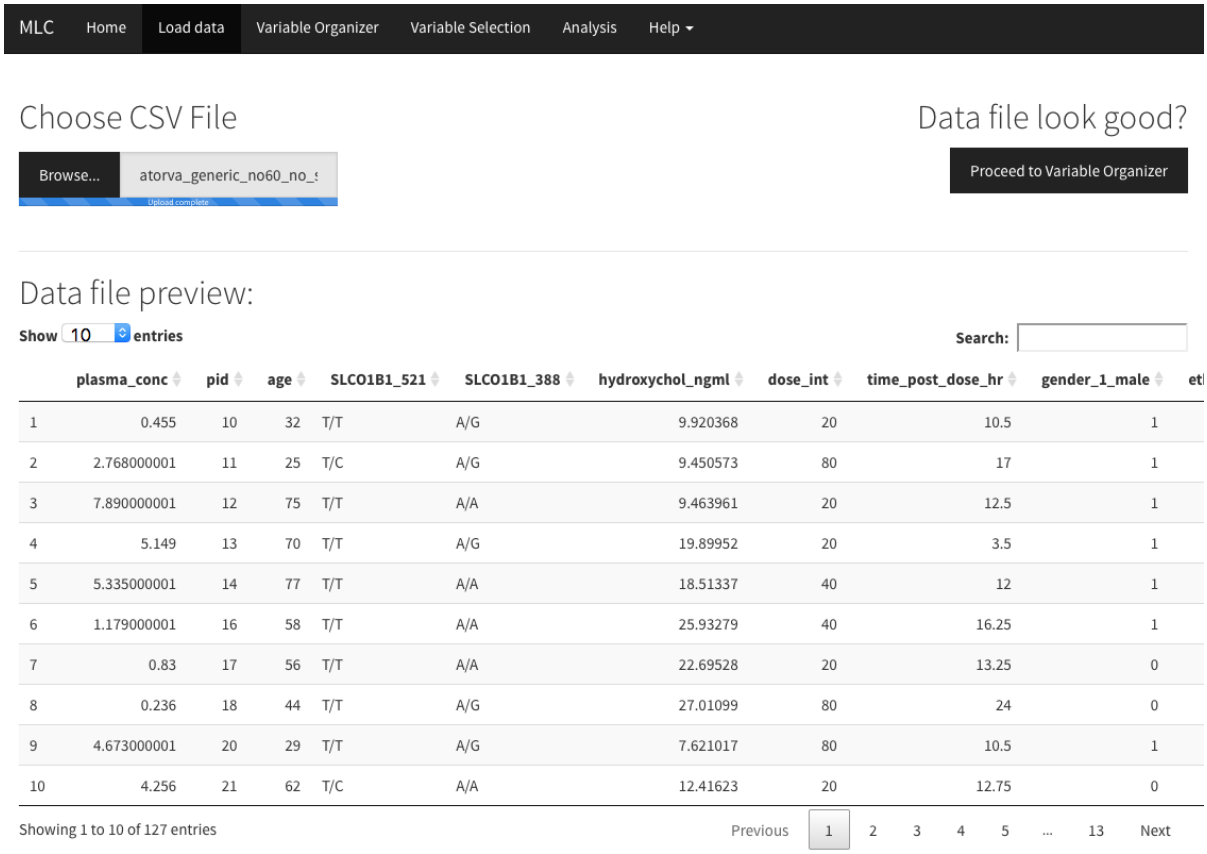


Figure D.1: MLC: Data loading interface

After reviewing the previewed data, an interface is available to specify the structure of the data for analysis. The user is given the option to center and scale their data, which would be of use in cases where interpreting variables in the original scale is different. The user may then choose the outcome variable, and whether this should be log transformed; this is particularly important for outcomes such as plasma concentration of drugs with a shorter half life that display non-linear characteristics over time. A primary use case of the application is modelling clinical information and and possibly a column of patient identifiers; the user is

required to specify whether there is a patient ID variable, and if so, which variable contains this information before selecting options for the remaining variables in the data set so as to not accidentally include patient IDs as regressors in any model.

Because a main intended application of the platform is for use with clinical pharmacological data sets, specific options are available for encoding genetic variables such as single nucleotide polymorphisms (SNPs). It is expected that the format of any genetic variables contained in the data file will take the form of a pair of gene bases (ie. 'A/T', 'C/G'). Variables containing this pattern of data are automatically detected, and after selecting the base of interest (A, T, C or G), the user is given the following encoding options for each genetic variable detected: "Count (0,1,2)", a raw measure of the number of times the base appears in the pair; "Present/absent (0,1)", whether or not the base appears in the pair given; "Homozygous (0,1)", whether or not both members of the pair are the base of interest; and "Exclude", to remove the variable from any analyses (Figure D.2).

Finally, the user has the option to exclude any dichotomous variables, which are automatically detected by the software, and to exclude any continuous variables or indicate whether any of the variables should be treated as factors (Figure D.3).

D.2.2 Logistic and Linear Regression Implementation

Linear regression results are obtained using a linear model (`lm()`) or generalized linear model (`glm()`). For data sets with a continuous outcome variable, the family is specified as Gaussian; for dichotomous outcomes, the family is specified as Binomial. Model performance is calculated as per the user's choice of 5- or 10-fold CV. This process is repeated 30 times to

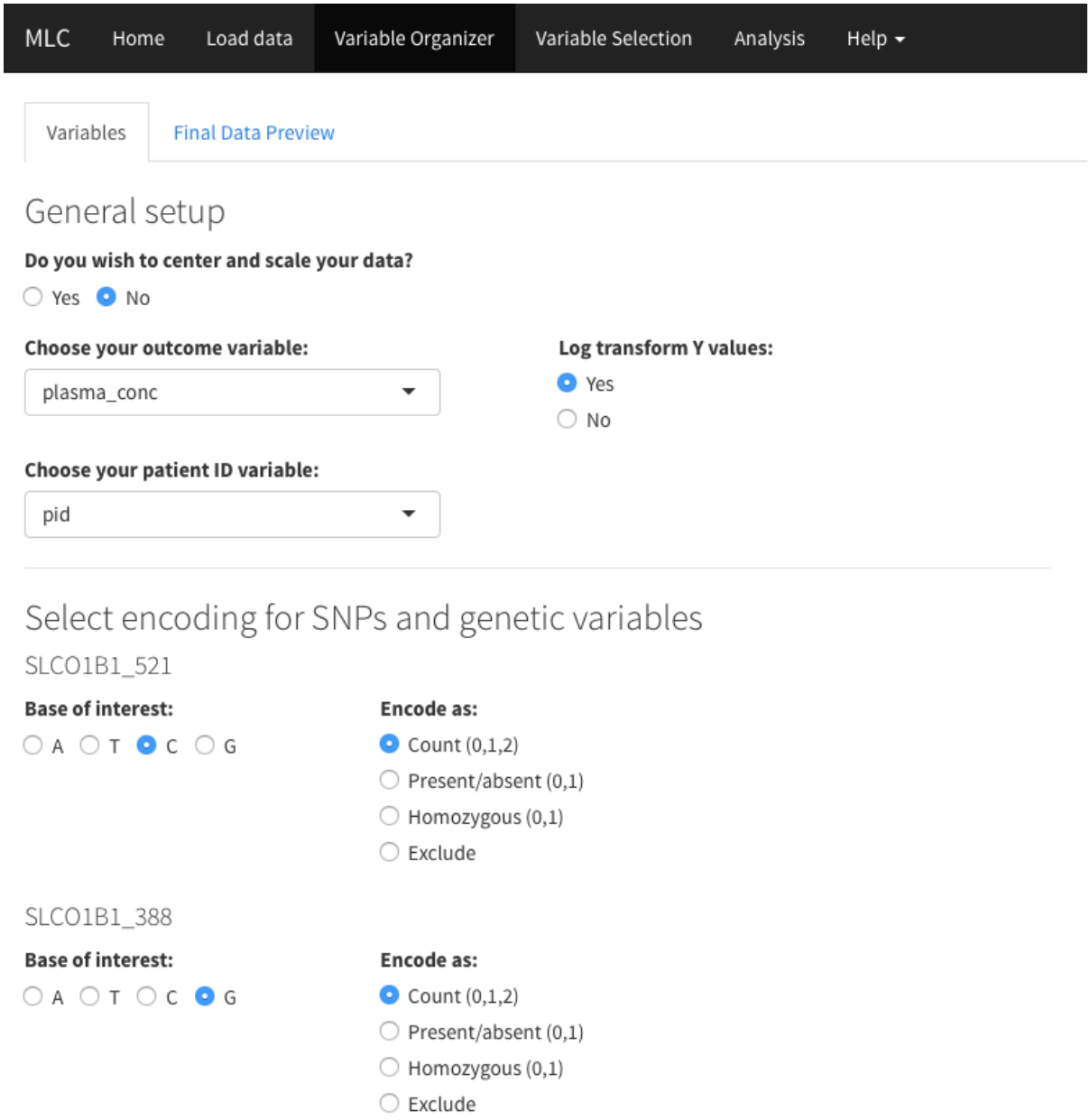


Figure D.2: MLC: Initial variable setup

obtain more stable performance estimates for smaller data sets, and performance measures are averaged over all resultant folds. Calculated CV measures include AIC, RMSE, Adjusted R^2 for linear models, and classification performance for logistic models.

For both linear and logistic regression, the outcome variable and number of observations present in the analysis are displayed, followed by a table of covariates, coefficient estimates,

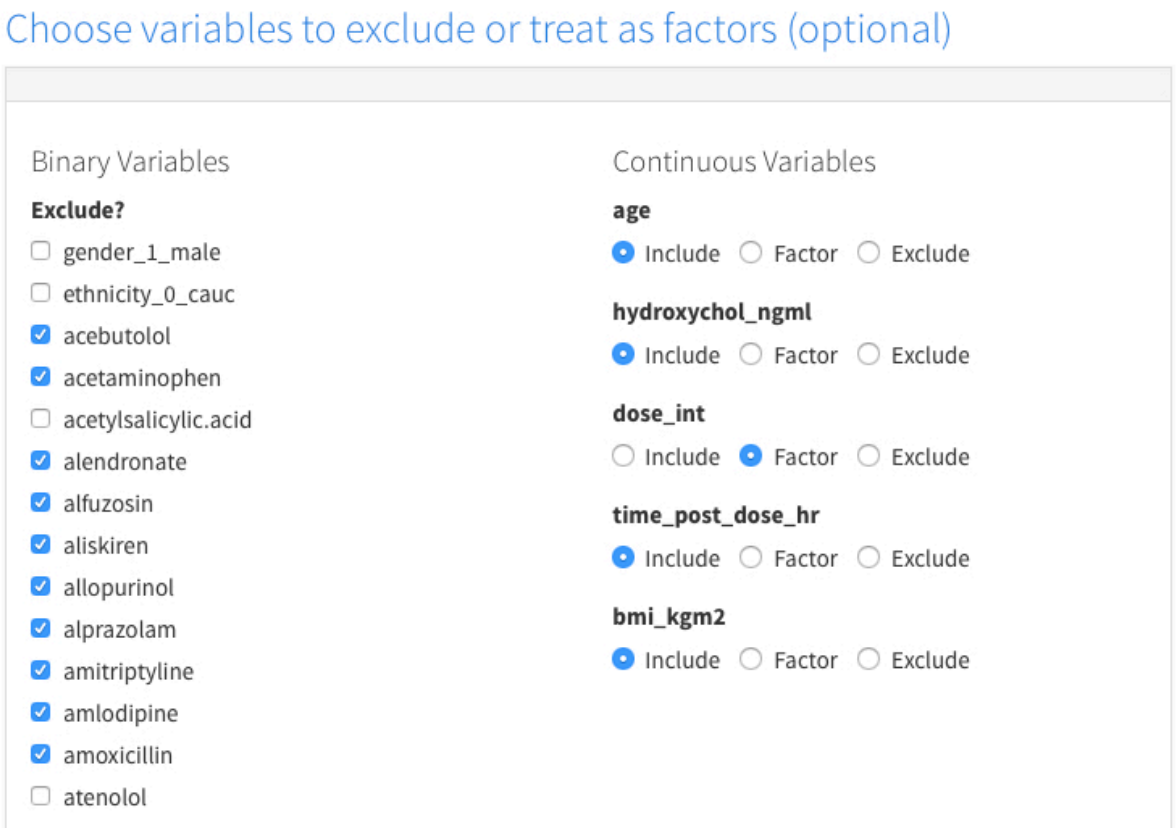


Figure D.3: MLC: Variable exclusion and factor representation

a 95% confidence interval, and corresponding P-value, indicated with stars for those whose CI's do not cross zero (P-value<0.05). Below the table is a slider to give the user the option of adjusting the number of significant digits to appear in the table. For logistic regression models, the coefficient estimates are converted to odds ratios for ease of interpretation. Alongside the regression summary table is an interface for interpreting regression coefficients, including a possible interpretation for the intercept (Figure D.4). The user also has the option to output the model fit of the linear regression to an .RData file in order to conduct comparisons with other similar models using ANOVA etc.

Following the regression summary table are the raw output of the regression model as given by `summary(lm)`, and the explicit regression equation, should users need to reference it. An

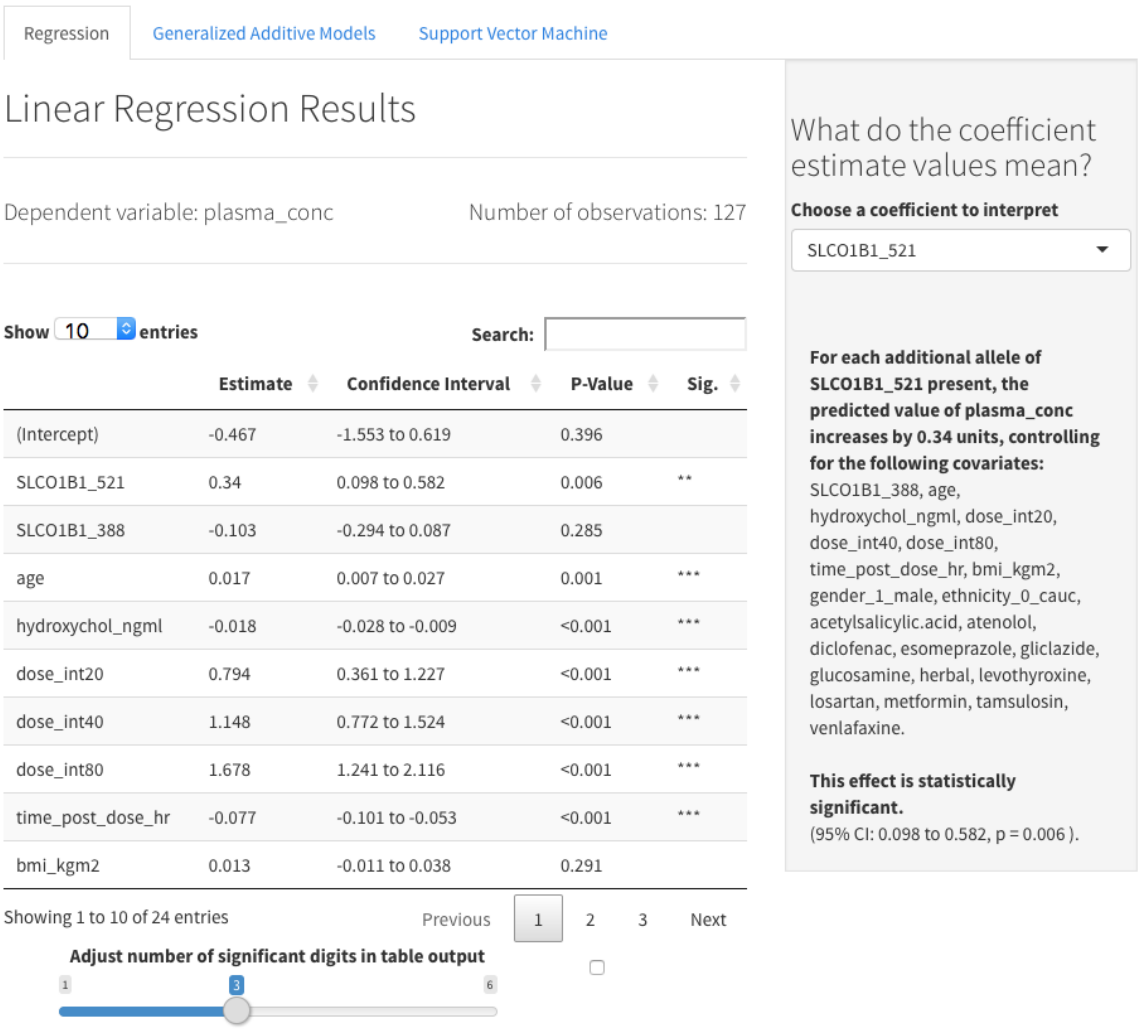


Figure D.4: MLC: Linear regression results table and coefficient interpretation

evaluation of model performance follows; performance across folds and a table containing performance per fold are shown, as well as options for the user to choose between 5- or 10-fold CV and a panel for more information on CV topics. In both the 5- and 10-fold CV protocols, 100 random splits of the data are taken and the CV is repeated on each. The final summary CV numbers are the average of all of the folds from all of the splits.

For linear regression, the additional visual display to help with model understanding is an array of diagnostic regression plots, with corresponding explanations for how to interpret each

of the four plots: residuals vs. fitted, normal Q-Q, scale-location, and residuals vs. leverage. At the bottom of the plots is a link to the resource used to help formulate the explanations, which contains additional specific examples of what the plots might show if the regression model assumptions are violated[3].

The additional visual display for logistic regressions has not yet been implemented, but will contain an interactive ROC curve with sliders to change the cutoff probability for classifying a case as 0 or 1, as classical regression diagnostic plots can be misleading to interpret when generated from logistic models[3]. This is meant to help the user explore the relationship between sensitivity, specificity and overall accuracy with regard to how one chooses to define the probability of a positive or negative classification outcome.

D.3 Future Work

Because of time constraints associated with finishing the analyses outlined in the research goals for this thesis, the GAM interface was not fully developed and the CV procedure was not incorporated into the application. Similarly, the tuning process and statistical output for the SVR analysis has yet to be transferred from regular R code to code appropriate for use with Shiny. Once these statistical functions are implemented, development will focus on implementing tools for visual analytics and interactive visualization, in order to help the user have a more intuitive understanding of model fit quality, and the relationship between the output and the statistical parameters chosen.

References

[1] Winston Chang, Joe Cheng, JJ Allaire, Yihui Xie, and Jonathan McPherson. shiny: Web Application Framework for R, 2017. R package version 1.0.0.

[2] R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2014. Available online at `http://www.R-project.org/`.

[3] Bommae Kim. Understanding diagnostic plots for linear regression analysis, 2015.

RHIANNON V. **ROSE**

EDUCATION

2014-2018: UNIVERSITY OF WESTERN ONTARIO

PHD, EPIDEMIOLOGY AND BIOSTATISTICS

Improving Prediction of Systemic Statin Exposure Using Concomitant Medications, Non-Linear Modelling and Novel SNP Discovery | Improved predictive modelling of statin systemic exposure for the purpose of minimizing the risk of future adverse drug events. Developed a selection algorithm for selecting relevant concomitant medications for prediction of atorvastatin plasma concentration, applied nonlinear modelling techniques, and identified novel genes that could play a role in moderating rosuvastatin plasma concentration.

2012-2014: UNIVERSITY OF WATERLOO

MMATH, COMPUTER SCIENCE

gLOP: A Cleaner Dirty Model for Multitask Learning | Development of a novel penalized regression technique for health research. Leverages information from multiple patients for predictive modelling in the setting where the number of covariates is much greater than the available patient observations**.**

2007-2012: UNIVERSITY OF WATERLOO

BA, HONOURS PSYCHOLOGY

EXPERIENCE

2018 - PRESENT

PHARMACOEPIDEMIOLOGIST, SANOFI PASTEUR

Worked within a dynamic team to provide guidance and leadership for the development and implementation of pharmacoepidemiological and observational post-authorization safety studies for the seasonal influenza vaccine. Conducted literature reviews and analysis for safety signal evaluation and interpretation; acted as a point of contact and expert for pharmacoepidemiological support within Sanofi Pasteur, particularly with respect to vaccines for influenza, meningitis, typhoid, hepatitis A, tuberculosis and products in development. Participated in working groups for process improvement and methods development.

2015 – 2018

RESOURCE BIOSTATISTICIAN, PERSONALIZED MEDICINE LABORATORY (LHSC)

Improved the quality of research in the lab by providing guidance on study design, sample size, statistical analysis and interpretation of results for studies in clinical pharmacology. Aided in the preparation of manuscripts for publication and grant proposals for research funding.

2017 – 2018

CO-INSTRUCTOR/TEACHING ASSISTANT, EPIDEMIOLOGY OF MAJOR DISEASES

Engaged experts in various areas of clinical research to instruct students about major chronic and infectious illnesses. Taught a weekly tutorial session; created quizzes and exams and graded all student submissions.

2017

TEACHING ASSISTANT, PUBLIC HEALTH

Assessed quality of student submissions including four proposed public health protocols; instructed students on fundamental principles and methods in public health such as disease surveillance and outbreak investigations when the instructor was unavailable.

2016

TEACHING ASSISTANT, ANALYTIC EPIDEMIOLOGY

Critically appraised research-oriented student submissions, including directed acyclic graphs for causal modelling, and a CIHR-style operating grant proposal. Answered student questions on study design, strength of evidence, confounding, and scientific writing.

SKILLS AND SOFTWARE

- Epidemiology and biostatistics
- Statistical modelling and analysis
- Critical appraisal of study methodology
- Machine learning and prediction
- R and R-Shiny; Latex
- SAS, STATA, Matlab
- Microsoft Office

SELECTED PUBLICATIONS

1. Borrie AE, Rose RV, Choi YH, Perera FE, Read N, Sexton T, Lock M, Vandenberg TA, Hahn K, Dinniwell R, Younus J, Logan D, Potvin K, Yaremko B, Yu E, Lenehan JG, Welch S, Tyndale R, Teft WA, Kim RB. **Letrozole concentration is associated with CYP2A6 variation but not with arthralgia in patients with breast cancer**. **Breast Cancer Research and Treatment**. 2018. (Accepted)
2. Jansen LE, Teft WA, Rose RV, Lizotte DJ, Kim RB. **CYP2D6 genotype and endoxifen plasma concentration do not predict hot flash severity during tamoxifen therapy**. **Breast Cancer Research and Treatment**. 2018 Jul 6:1-8.
3. Wilson A, Jansen LE, Rose RV, et al. **HLA-DQA1-HLA-DRB1 polymorphism is a major predictor of azathioprine-induced pancreatitis in patients with inflammatory bowel disease**. **Alimentary Pharmacology and Therapeutics**, 00:1–6, 2017.
4. Markus Gulilat, Anthony Tang, Steven Gryn, Peter Leong-Sit, Allan Skanes, Jeffrey Alfonsi, George Dresser, Sara Henderson, Rhiannon Rose, Daniel Lizotte, et al. **Marked interpatient and sex-dependent variation in rivaroxaban and apixaban plasma levels in routine care**. **Journal of Pharmacological and Toxicological Methods**, 88:171, 2017.
5. Markus Gulilat, Anthony Tang, Steven E Gryn, Peter Leong-Sit, Allan C Skanes, Jeffrey E Alfonsi, George K Dresser, Sara L Henderson, Rhiannon V Rose, Daniel J Lizotte, et al. **Interpatient**

variation in rivaroxaban and apixaban plasma concentrations in routine care. **Canadian Journal of Cardiology**, 2017.

6. Rhiannon Rose, Daniel Lizotte, et al. **glop: the global and local penalty for capturing predictive heterogeneity**. In **Machine Learning For Healthcare (MLHC)**, pages 134–149, 2016.
7. James Yungjen Tung, Rhiannon Victoria Rose, Emnet Gammada, Isabel Lam, Eric Alexander Roy, Sandra E Black, and Pascal Poupart. **Measuring life space in older adults with mild-to-moderate Alzheimer's disease using mobile phone GPS**. **Gerontology**, 60(2):154– 162, 2014.
8. Rhiannon Rose. **glop: A cleaner dirty model for multitask learning**. Master's thesis, **University of Waterloo**, 2014.
9. Rhiannon Rose, Sandra Black, Eric Roy, Pascal Poupart, and James Tung. **Ambulatory assessment beyond the actigraph: The voice, activity and location monitoring in Alzheimer's disease (VALMA) project**. **Alzheimer's & Dementia**, 8(4):P230, 2012.
10. James Y Tung, Rhiannon V Rose, Emnet Gammada, Isabel Lam, Sandra E Black, Eric A Roy, and Pascal Poupart. **Measuring lifespace in older adults with mild Alzheimer's disease using smartphone GPS**. **Journal of Exercise, Movement, and Sport**, 44(1):66, 2012.
11. Rhiannon V Rose, Amanda J Clark, Dave A Gonzalez, and Eric A Roy. **Still making attentional errors: A modified slip induction task**. **Journal of Exercise, Movement, and Sport (SCAPPS refereed abstracts repository)**, 43(1), 2011.
12. Rhiannon V Rose, Amanda J Clark, and Eric A Roy. **Measuring action slips in older and younger adults**. **Journal of Exercise, Movement, and Sport (SCAPPS refereed abstracts repository)**, 42(1), 2010.

AWARDS

• **NSERC Alexander Graham Bell Canada**	2014-2017
• **Western University Doctoral Excellence Research Award**	2016-2017
• **NSERC Alexander Graham Bell Canada Graduate Scholarship - Masters**	2012
• **University of Waterloo President's Graduate Scholarship**	2012
• **University of Waterloo Special Graduate Scholarship**	2011
• **Toronto Rehabilitation Institute Research Day – Best Poster Award**	2009
• **University of Waterloo Merit Scholarship**	2007