A Framework for the Multilevel Integration of Molecular, Clinical, and Population Data in the Context of Breast Cancer: Challenges and Considerations of Socioecological Conditions and Pharmacogenomics

A DISSERTATION
SUBMITTED TO THE FACULTY OF
UNIVERSITY OF MINNESOTA
BY

Matthew K. Breitenstein

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

David S. Pieczkiewicz, PhD – Adviser
Jyotishman Pathak, PhD – Co-adviser

January 2015

Acknowledgements

This work has been partially supported by NIH grant U19-GM61388-13, PGRN Network Resource, Pharmacogenomics of Phase II Drug Metabolizing Enzymes. Mentorship, statistical consultation, and bio-medical expertise that contributed to the success of this work has been provided by Jyotishman Pathak PhD, Liewei Wang MD, PhD, Richard Weinshilboum MD, Gyorgy Simon PhD, Euijung Rui PhD, Sebastian Armasu MS, and Balmiki Ray MBBS. Advisement and committee service has been provided by David Pieczkiewicz PhD, Jyothishman Pathak PhD, Gyorgy Simon PhD, Nathan Shippee PhD, and Terrence Adam MD, PhD. Additional data support services have been provided by the Division of Biomedical Statistics and Informatics at Mayo Clinic, the Minnesota Population Center, and the Resources of the Rochester Epidemiology Project. The National Historical Geographic Information System of the Minnesota Population Center was supported by the National Science Foundation and the National Institutes of Health. The Rochester Epidemiology Project was supported by the National Institute on Aging of the National Institutes of Health under Award Number R01AG034676.

Dedication

To my colleagues, friends, and family.

Table of Contents

List Tables

List of Figures

Chapter 1: Prelude: A Conceptual Framework for Translational Informatics

Background

Despite medicine's rigorous pace of advancement, clinical research remains limited by scalability and portability issues. All too commonly cancer epidemiology relies on manual chart reviews to generate data in a registry-specific or study-specific or manner. Further, methodological approaches for integration of data between data sources warrant improvement. Recently, the National Cancer Institute (NCI) provided a vision for the modernization of cancer epidemiology that highlighted the need for methodological developments where biomedical health informatics would be of tremendous value (Khoury 2013). Specific aims for methodological improvements included 1) incorporation of multilevel analysis into translational research, 2) integration of data across the life course, 3) integration of big data science into translational research, and 4) reliably capture exposure phenomena on a massive scale (Khoury 2013). Specifically, advancement in our understanding and utilization of multilevel modeling is necessary to advance our understanding of the complex, multifactorial causes of cancer (Lynch 2013).

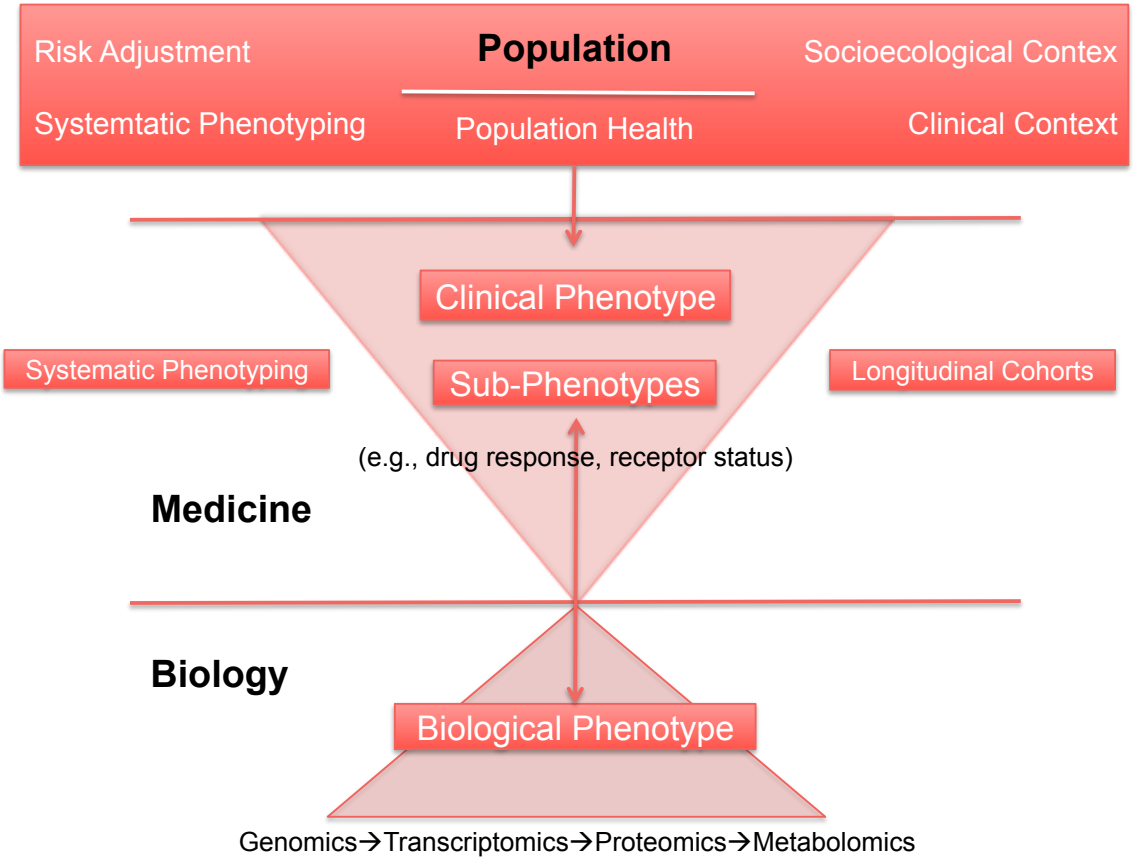
In the past we have witnessed the field of biomedical informatics grow from its infancy in biometry to clinical informatics. However, as we now see the field of biomedical informatics approaching data mining and data science we see both a tantalizing opportunity for advancement and humbling assessment of the infancy of even data standardization and normalization. Further, as we think about the needs of cancer epidemiology, we see the need for multilevel modeling powered by scalable and portable informatics-driven approaches even more.

Conceptual Framework

For the consideration of the committee I present a conceptual framework that guided the integration of biological and population premises across levels of analysis: A Multilevel Framework for Translational Informatics (**Figure 1.1**). Highlighted in this framework is the inherent gap between population-level measures and individual-level measures, while the center of the research question remains on providing insight into a biological phenomenon.

Multilevel modeling remains a computationally and logically difficult endeavor to pursue. Each level in a multilevel model requires precise and specific measures to appropriately attribute the construct of relevance (i.e. systematic phenotyping). However, within each level of the model, constructs will be logically confounded with one another. For example, you do not want to attribute distal effects of social environment to clinical biomarkers of breast cancer. Or conversely, you do not want to attribute the beneficial effect of a drug on all-cause mortality to cancer-specific mortality. While often the confounding cannot be removed, approaches aimed at understanding the latent characteristics around these constructs can provide some level of clarity for the purpose of modeling phenomena across multiple levels, an approach that is utilized directly in the second part and indirectly in the third part.

Figure 1.1: A Multilevel Framework for Translational Informatics



In this dissertation I utilized informatics-driven data acquisition to answer a line of scientific inquiry regarding the pharamacogenomics and pharmacoepidemiology of

metformin in breast cancer and type 2 diabetes mellitus that is centered around this framework. Of particular novelty is a multilevel approach that addresses the NCI's above four specific aims and is powered by testable molecular and population-level premises to elucidate clinical-level questions regarding the potential repurposing of metformin for breast cancer treatment. In what follows I present three topics, consisting of the reproductions of four scientific papers, that as of the time of the writing are currently under review: 1) metformin and insulin pharmacoepidemiology, 2) the modifying impact of socioecological context on breast cancer prevalence, and 3) translational biomedical informatics of metformin pharmacogenomics. While this work elucidates aspects of metformin pharmacogenomics, it primarily aims to demonstrate this framework for utilizing informatics approaches to drive metformin pharmacoepidemiology and translational pharmacogenomics for potential metformin repurposing in breast cancer treatment and similar research in the future.

Overview

Metformin is an oral biguanide and is a widely prescribed anti-diabetic medication (Thompson 2014) that is considered to be a first-line treatment for T2DM (Inzucchi 2012). While metformin if generally well tolerated (Gong 2012) its pharmacogenomics are not clearly understood and approximately 30% of patients do not respond to it. Due to the epidemic growth of T2DM in the US and the accumulating evidence highlighting potential repurposing of metformin for cancer prevention and treatment it is imperative to understand molecular mechanisms and clinical impacts of metformin. Further, in order to appropriately separate effects due to metformin and breast cancer from social stress, a known modifier of breast cancer biology, it is necessary to incorporate this characteristic into the model in a way that does not lead to over fitting. Below I provide a brief abstract of the corresponding part and chapter components.

Part #1: Pharmacoepidemiology of Metformin in Breast Cancer

Metformin use in type 2 diabetes mellitus (T2DM) patients has been shown to reduce the incidence of breast cancer. However, the impact of metformin and insulin on breast cancer specific survival outcomes remains controversial. Further, the impact of

metformin and insulin on breast cancer-specific mortality outcomes beyond those attributable to mortality outcomes due to T2DM severity remains unclear. In this part, I evaluated the effect for metformin and insulin on mortality outcomes using a traditional and alternate analysis approach, which I am going describe in the following two chapters, respectively.

Chapter 2: Association of Metformin Exposure with Breast Cancer Survival

In this chapter I utilized electronic health record (EHR) augmented cancer registry data and standard survival analysis approaches to model the effect of metformin and insulin on all cause and breast cancer-specific mortality outcomes in patients with breast cancer and T2DM. This standard approach left confounded the implications for metformin and insulin that was due to T2DM severity and for breast cancer-specific mortality outcomes.

Chapter 3: Disease-Specific Effects of Metformin in Breast Cancer Patients with Type II Diabetes Mellitus

In this chapter I utilized an alternative analysis approach that aimed to separate the effects of metformin and insulin due to T2DM severity and for breast cancer treatment outcomes using a trained linear offsets for T2DM severity.

Part #2: Accounting for Socioecological Context in Clinical Research

Neighborhoods and social structure are known to impact health. However, population attributes are commonly confounded with each other and latent measures are commonly skewed by limited inter-neighborhood comparison, limiting our ability to quantify causality. While causal estimations of neighborhood effects cannot be reliably attained, translational research endeavors benefit from utilizing robust phenotypes that incorporate socio-ecological context. Further, translational research also stands to benefit from novel approaches that allow for more nuanced stratification between phenotypes and facilities that go beyond the potentially problematic standard practice of simply placing the facility as a random effect or socioeconomic status as an additive effect in clinical research models. The aim of this work was to develop a population-based risk

adjustment index that accounted for socio-ecological context using a generalizable approach that enables future translational research.

Chapter 4: Community-Based Measures of Social Stress on Breast Cancer Prevalence

Due to the confounding relationship between clinical and social factors it is likely that the effect of socio-ecological conditions (SECs) is being inappropriately attributed to clinical variables in retrospective clinical research. Utilizing validated measures of social stress as the SEC of interest, obesity as the clinical indicator of interest, and breast cancer prevalence as the clinical outcome of interest, we demonstrated the impact of isolating SECs from the effect of clinical indicators in clinical research models. I demonstrated the need to account for SECs to more appropriately model clinical indicators that impact clinical outcomes.

Part #3: Leveraging the Electronic Health Record-Linked Biorepository in Translational Biomedical Informatics

Metformin is a first-line antihyperglycemic agent commonly prescribed in type 2 diabetes mellitus (T2DM), but whose pharmacogenomics are not clearly understood. Approximately 30% of patients do not respond to metformin with outcomes ranging from no impact on glycemic control to adverse reactions; it is imperative to understand molecular mechanisms of metformin further. In this part I sought to understand the impact of variants in candidate genes thought to modify glycemic response to metformin treatment using EHR-linked genetic data.

Chapter 5: The Impact of Genomic Variation in Metformin Pharmacogenomic Determinants on Glycemic Response

Seventeen genes with suspected pharmacokinetic or pharmacodynamic implications were selected based on systematic review for study. Our analysis cohort consisted of 258 T2DM patients who had new metformin exposure, existing genomic data, and longitudinal electronic health records. Change in glycemic response to metformin exposure via A1c measures pre and post metformin exposure serve as the outcome of interest. After quality control, gene-level and SNP-level analysis were conducted on 17

candidate genes and 463 SNPs within those genes, controlling for key covariates of sex, age, and body mass index (BMI) at index metformin exposure.

Chapter 6: Leveraging the Electronic Health Record(EHR)-Linked Biorepository: An EHR-Driven Hypothesis for Glycemic Response to Metformin

In this part I explored the potential association of the flavin-containing monooxygenase(FMO)-5 gene, a biologically plausible biotransformer of metformin, and modifying glycemic response to metformin treatment. Using a cohort of 258 T2DM patients who had new metformin exposure, existing genetic data, and longitudinal electronic health records, I compared genetic variation within FMO5 to change in glycemic response. Gene-level and SNP-level analysis identified marginally significant associations for FMO5 variation, representing an EHR-driven pharmacogenetics hypothesis for a potential novel mechanism for metformin biotransformation. However, functional validation of this EHR-based hypothesis is necessary to ascertain its clinical and biological significance.

Part #5: Conclusion

Chapter 7: Conclusion for the Framework on Translational Informatics

Part #1: Pharmacoepidemiology of Metformin in Breast Cancer

Metformin use in type 2 diabetes mellitus (T2DM) patients has been shown to reduce the incidence of breast cancer. However, the impact of metformin and insulin on breast cancer specific survival outcomes remains controversial. Further, the impact of metformin and insulin on breast cancer-specific mortality outcomes beyond those attributable to mortality outcomes due to T2DM severity remains unclear. In this part, I first demonstrate associations for metformin and insulin on breast cancer specific mortality outcomes using survival analysis using EHR data. Finally, I demonstrated a significant protective effect for metformin and a significant detrimental effect for insulin on breast cancer survival outcomes that is beyond their implication T2DM severity using linear offsets to separate effects.

Chapter 2: Association of Metformin and Insulin Exposure with Breast Cancer Survival Outcomes

Abstract

Purpose

While metformin use in diabetic patients has been shown to reduce the incidence of breast cancer, the impact of metformin on survival after breast cancer incidence still remains unclear. Previous studies have been limited by small cohort size and missing key clinical attributes. Using electronic health record (EHR) data we examined metformin exposure after breast cancer diagnosis on mortality outcomes by breast cancer subtype.

Patients and Methods

1,180 patients with unilateral, non-recurrent breast cancer diagnosis between 1998 and 2011 and type II diabetes mellitus were identified using a combination of manual chart review and phenotyping of electronic health records. Median age at breast cancer diagnosis was 67(34-95) and median follow-up time was 83 (6-191) months. Exposure to metformin was defined as ≥6 months of known exposure after breast cancer diagnosis. Univariate, multivariate, and stratified Cox Proportional Hazard Regression was utilized to model metformin exposure and survival outcomes.

Results

Univariate survival analysis identified significant associations for metformin exposure (n=330) after breast cancer diagnosis on decreased all cause mortality (HR=0.680,p=0.0131) and decreased breast cancer-specific mortality (HR=0.433,p=0.0195) outcomes. Multivariate analyses identified a consistent significant protective association between metformin exposure and breast cancer-specific mortality and a consistent detrimental association between insulin exposure and breast cancer-specific mortality during the same period. Potential treatment and severity interaction were identified during iterative model development. After developing an overall drug exposure model controlling for key covariates of patient demographics, severity, and

treatment metformin demonstrated a significant protective association (HR=0.269,p-val=0.0378) with breast cancer specific mortality.

Conclusion

This study demonstrates a consistent overall protective association for metformin exposure after breast cancer diagnosis with decreases in all cause and breast cancer-specific mortality. While potentially valuable treatment and severity interaction were uncovered, increased power is needed to effectively evaluate these interactions.

Background and Introduction

Women with type II diabetes mellitus (T2DM) are at an increased risk of breast cancer incidence[1]. Further, the risk of cancer-related mortality is also increased in cancer patients with T2DM, including breast cancer[2]. While metformin use in diabetic patients appears to decrease the risk of breast cancer incidence[3], studies aimed at understanding the impact of metformin on survival for breast cancer patients have only identified significant positive associations for all cancer metformin exposure and all cause survival[2]. Specifically, metformin exposure before and after breast cancer incidence among patients with T2DM did not exhibit a statistically significant association with all cause survival outcomes[2] and cumulative duration of metformin exposure after T2DM diagnosis did not exhibit an association with all cause and breast cancer-specific survival in an aged cohort[4]. Among studies evaluating specific types of breast cancer, metformin exposure was associated with improved overall survival outcomes for stage ≥2 HER2+[5] but not for triple receptor-negative receptor breast cancer treated with chemotherapy[6].

Consequently, the association of metformin exposure with all cause and breast cancer-specific mortality remains unclear. Previous studies aimed at understanding this relationship have been limited by lacking disease-specific mortality information or by lacking access to data available in an electronic health record (EHR). Our current study aims to use an age representative and typologically diverse cohort of women with breast cancer and T2DM, augmented with EHR data, to understand this association. Specifically, we aim to study if a protective association exists between metformin exposure after diagnosis of breast cancer and all cause and breast cancer-specific

mortality. Our preliminary finding of a potential association between metformin exposure and survival has potentially profound implications for breast cancer therapy. Specifically, this potential association has direct pharmacogenetic implications for inhibiting breast cancer cell growth by targeting components of the adenosine monophosphate-activated protein kinase (AMPK)[7] pathway, the insulin/insulin-like growth factor-1 (IGF1)[8] signaling pathway, and directly and indirectly via additional pathways[9] with specific drug therapy combinations.

Materials and Methods

Cohort inclusion was limited to female patients with a non-recurrent confirmed diagnosis of unilateral breast cancer and T2DM. Inclusion was also limited to eligible new breast cancer cases between the years 1998 and 2011 to allow for a minimum two year follow-up period. Male patients and those with type I diabetes mellitus or ambiguous diabetes type were excluded from this study. Patients with survival or follow-up < 6 months were silenced from the cohort to control for immortal time bias. Metformin and insulin exposure were parameterized to represent >=6 months of drug exposure within the treatment period post breast cancer diagnosis. Patients needed to have multiple recorded mentions of an individual class of antidiabetic medications spanning a period of greater than 30 days to be considered as having exposure to an individual class of antidiabetic medications. Patients without multiple recorded mentions of an individual class of antidiabetic medications spanning more than 30 days post breast cancer diagnosis were parameterized as not having exposure to individual class of antidiabetic medications (n=25 for metformin). The resulting cohort contained female patients (n=1,180) with unilateral, non-recurrent breast cancer diagnosis who had a median age of breast cancer diagnosis of 67 (34-95) years and median follow-up period of 83 (6-191) months. A description of the cohort can be found in **Table 2.1**.

Table 2.1. Descriptive statistics of cohort.

	event	all
Demographics		
all cause mortality	306	1188
breast cancer-specific mortality	71	1053
follow up period	83 (6 , 191)	1188

	age at diagnosis of breast cancer		67 (34 , 95)	1188
	resident of rural zip code		277	1162
	Primary care patient			
	white		1072	1188
Antidiabetic Therapies				
	metformin		355	
	insulin		245	
	tzd		73	
	sulfonylurea		182	1188
Diabetic Severity				
	A1c		6.5 (4.5 , 15.1)	
		uncontrolled diabetes	205	687
	BMI		31.8 (14.5 , 61.7)	1135
		underweight	5	
		normal	149	
		overweight	281	
		moderately obese	324	
		severely obese	376	
Breast Cancer Severity				
	Stage	In Situ	185	1152
		1	596	
		2	14	
		3	240	
		4	25	
	Grade	1	197	1188
		2	424	
		3	273	
		4	294	
Breast Cancer Receptor Status				
	Estrogen Receptor Positive		772	886
	Progesterone Receptor Positive		695	864
	HER2/neu status actionable		527	906
		HER2/neu positive	98	906
	Triple Negative		31	864
Breast Cancer Treatment				
	Surgery		444	979
	Chemotherapy		242	1174
	Radiation Therapy		227	890
	Hormone therapy		503	1168
	Immuno therapy		12	1187
	Fulvestrant		43	1188
	Aromatase Inhibitors		567	1188
	Herceptin		96	1188
	Methotrexate		72	1188
	Ovarian ablation		23	1188
	Selective estrogen response modulators		918	1188
Charlson Comorbidity Index				
	Myocardial Infarction		73	1188

Congestive heart failure	159
peripheral vascular disease	132
cerebrovascular disease	166
Dementia	62
chronic pulmonary disease	263
ulcer	88
mild liver disease	119
Diabetes	1188
Diabetes with organ damage	206
hemiplegia	20
moderate/severe renal disease	135
moderate/severe liver disease	24
metastatic solid tumor	394
Aids	0
Rheumatologic disease	66
Cancer	1188

Diagnosis, medication, patient attributes, and last follow-up data points were extracted from Mayo Clinic's Enterprise Data Trust[10] and electronic death certificate data containing ICD-10-CM codes between the years 2000 and 2013 for underlying and direct cause of death were extracted from Minnesota Electronic Death Certificates. Survival time was defined as the duration in months between first identified diagnosis of breast cancer and censor or death date. If a patient was not identified as deceased from either the Minnesota Death Registry or Mayo Clinic's EHR, patient survival time was censored at the last known follow-up date. Kaplan-Meier analysis was used to calculate unadjusted survival models and Cox Proportional Hazard Regression was utilized to calculate adjusted survival models. All statistical analysis was performed using Linux SAS v9.3 and data visualization was assisted by a SAS macro for Cox Proportional Hazard Regression[11].

Results

Univariate survival analysis demonstrated a protective association of metformin exposure after breast cancer diagnosis with decreases in all cause mortality (HR=0.680,p-val=0.0131) and breast cancer-specific mortality (HR=0.433,p-val=0.0195). Since elucidating the impact of metformin in this cohort of patients requires considerations for

both T2DM **(Table 2.2)** and breast cancer **(Table 2.3)**, a series of sensitivity analyses were utilized to identify parameters appropriate for stratification in the final drug exposure model.

T2M considerations

Multivariate analysis of common classes of antidiabetic therapies (**Table 2.2**) (metformin, insulin, thiazolidinedione, sulfonylurea) identified metformin as having a strongly significant protective association on all cause mortality (HR=0.500,p-val<0.0001) and breast cancer-specific mortality (HR=0.248,p-val=0.0003) and insulin as having a strongly significant detrimental association on all cause mortality (HR=2.361,p-val<0.0001) and breast cancer-specific mortality (HR=3.190,p-val<0.0001) (**Table 2.2a**). In order to identify potential bias amongst classes of antidiabetic medication exposure by glycemic control of diabetes (A1c < 7), a potentially confounding factor and proxy for diabetes severity, we evaluated these medications in a subgroup of patients with A1c measurements in the EHR (n=687). Glycemic control of diabetes was found not to be associated with significant detrimental associations on breast cancer specific mortality

Table 2.2: Classes of Antidiabetic Medications

Table 2.2a: Antidiabetic Therapies	All Cause Mortality				Breast Cancer-Specific Mortality			
	HR	**p-val**	**CI - low**	**CI - high**	**HR**	**p-val**	**CI - low**	**CI - high**
Metformin	**0.500**	**<0.0001**	0.357	0.702	**0.248**	**0.0003**	0.117	0.527
Insulin	**2.361**	**<0.0001**	1.765	3.159	**3.190**	**<0.0001**	1.848	5.506
Thiazolid-inedione	1.258	0.3594	0.770	2.056	2.163	0.0700	0.939	4.981
Sulfonylurea	1.264	0.2194	0.870	1.835	1.458	0.2987	0.716	2.971

Table 2.2b: Glycemic Control	All Cause Mortality				Breast Cancer-Specific Mortality			
	HR	**p-val**	**CI - low**	**CI - high**	**HR**	**p-val**	**CI - low**	**CI - high**
Metformin	**0.458**	**<0.0001**	0.317	0.662	**0.275**	**0.0016**	0.123	0.612
Insulin	**1.513**	**0.0156**	1.082	2.117	**2.042**	**0.0302**	1.071	3.893
Thiazolid-inedione	0.949	0.8564	0.536	1.678	1.015	0.9780	0.342	3.014
Sulfonylurea	1.010	0.9641	0.667	1.529	1.444	0.3425	0.676	3.085
Uncontrolled diabetes	**1.478**	**0.0117**	1.091	2.003	1.285	0.4475	0.673	2.455

(HR=1.285,p-val=0.4475), and had only a minimal impact on the calculated risk of breast cancer specific mortality for metformin (HR=0.275,p-val=0.0016) and insulin (HR=2.042,p-val=0.0302) (**Table 2.2b**). Since glycemic control of diabetes was demonstrated as having no significant impact on breast cancer specific survival outcomes it was dropped from future multivariate models.

Breast cancer staging and severity considerations

After controlling for specific elements breast cancer staging and severity (**Table 2.3a**), adjusted survival analysis demonstrated a protective association of metformin exposure after breast cancer diagnosis with decreases in all cause mortality (HR=0.544,p-val=0.0005) and breast cancer-specific mortality (HR=0.3000,p-val=0.0020). Parameters for chemotherapy and surgery status appeared to capture breast cancer severity and were included in the final model. Considerations for breast cancer biology **(Table 2.3b)** demonstrated detrimental associations for HER2 negative patients, while still demonstrating similar protective association of metformin exposure after breast cancer diagnosis with decreases in all cause mortality (HR=0.542,p-val=0.0010) and breast cancer-specific mortality (HR=0.343,p-val=0.0078). Since the majority of patients were estrogen (87.1%) or progesterone (80.4%) positive (90.0%) it was expected ER/PR status would not deviate the median fit of the model. HER2 negative status was marked for inclusion in the final model.

Breast cancer treatment considerations

A combined model of breast cancer drug therapy and select diabetes medications to identify potential drug interaction effects and elucidate the potential effect of metformin **(Table 2.3c)**. Fulvestrant (HR=7.045,p-val<0.0001) and methotrexate (HR=2.450,p-val=0.0060) had significant detrimental associations with breast cancer specific mortality **(Table 2.3d)**. SERMS having marginally significant (HR=0.616,p-val=0.0904) protective associations on breast cancer specific mortality **(Table 2.3d)**. Interactions were detected between metformin and fulvestrant (HR=10.025,p-val=0.0002) and insulin and trastuzumab (HR=25.743,p-val<0.0001) using backwards elimination on breast cancer and select diabetes drug therapy interactions **(Table 2.3e)**.

Patients exposed to both metformin and fulvestrant therapy (n=8) suffered a mortality event within 7 years (n=4), 3 of which were known to be breast cancer-specific mortality events. It is important to note that these associations cannot be interpreted as having actual or clinical significance due to potential model over fitting and instability due to a small (n=7) number of interaction events.

Table 2.3: Breast Cancer Staging and Severity

Table 2.3a: Breast Cancer severity	All Cause Mortality				Breast Cancer-Specific Mortality			
	HR	**p-val**	**CI-low**	**CI-high**	**HR**	**p-val**	**CI-low**	**CI-high**
Metformin	**0.544**	**0.0005**	0.386	0.768	**0.300**	**0.0020**	0.139	0.644
Insulin	**2.240**	**<0.0001**	1.642	3.055	**3.254**	**<0.0001**	1.837	5.766
In situ	**0.431**	**0.0004**	0.271	0.685	0.356	0.0547	0.124	1.021
High grade	1.292	0.0658	0.983	1.698	1.613	0.0851	0.936	2.779
Chemotherapy	0.811	0.2277	0.577	1.140	**2.008**	**0.0134**	1.155	3.488
Surgery	**0.478**	**<0.0001**	0.364	0.628	**0.544**	**0.0265**	0.317	0.931
Table 2.3b: Breast Cancer biology	All Cause Mortality				Breast Cancer-Specific Mortality			
	HR	**p-val**	**CI-low**	**CI-high**	**HR**	**p-val**	**CI-low**	**CI-high**
Metformin	**0.542**	**0.0010**	0.038	0.780	**0.343**	**0.0078**	0.156	0.754
Insulin	**2.758**	**<0.0001**	1.974	3.853	**3.907**	**<0.0001**	2.065	7.393
ER positive	1.038	0.8922	0.607	1.774	0.853	0.7535	0.315	2.308
PR positive	0.656	0.0691	0.416	1.034	0.540	0.1527	0.232	1.256
her2 actionable, positive	1.227	0.3971	0.765	1.968	1.941	0.1805	0.735	5.121
her2 actionable, negative	**1.406**	**0.0399**	1.016	1.945	**2.396**	**0.0175**	1.166	4.925
Table 2.3c: Breast Cancer biology	All Cause Mortality				Breast Cancer-Specific Mortality			
	HR	**p-val**	**CI-low**	**CI-high**	**HR**	**p-val**	**CI-low**	**CI-high**
Metformin	**0.658**	**0.0110**	0.477	0.909	**0.389**	**0.0114**	0.188	0.808
Insulin	**2.620**	**<0.0001**	1.955	3.511	**3.281**	**<0.0001**	1.916	5.619
Anti-estrogens	**2.898**	**<0.0001**	1.846	4.548	**7.045**	**<0.0001**	3.508	14.151
Aromatase inhibitors	0.681	0.0045	0.523	0.888	0.902	0.7173	0.518	1.573
Herceptin	1.252	0.2940	0.823	1.906	1.568	0.1948	0.794	3.094
Methotrexate	**1.495**	**0.0511**	0.998	2.238	**2.450**	**0.0060**	1.293	4.644
Ovarian oblation	0.179	0.3257	0.179	1.771	1.108	0.8682	0.330	3.720
SERM	**0.453**	**<0.0001**	0.353	0.581	**0.616**	**0.0904**	0.351	1.079

Table 2.3d: Backwards elimination		Breast Cancer-Specific Mortality			
		Association	Chi-Square	p-val	S.E.
	Metformin	**Protective**	**7.377**	**0.0066**	0.445
	Insulin	**Detrimental**	**7.625**	**0.0058**	0.344
	Anti-estrogens	Detrimental	26.348	<0.0001	0.383
	Metformin*Anti-Estrogens	**Detrimental**	**4.748**	**0.0293**	0.794
	Aromatase inhibitors	Detrimental	0.508	0.4760	0.306
	Herceptin	Detrimental	4.070	0.0437	0.645
	Insulin*Herceptin	Detrimental	3.478	**0.0622**	0.811
	Aromatase inhibitors*Herceptin	**Protective**	**12.632**	**0.0004**	0.768
	Methotrexate	Detrimental	2.282	0.1309	0.464

Table 2.3e: Medication Interaction Model		Breast Cancer-Specific Mortality				
		Association	HR	p-val	L-CI	H-CI
	Metformin	**Protective**	**0.261**	**0.0024**	**0.110**	**0.622**
	Insulin	**Detrimental**	**3.180**	**<0.0001**	**1.778**	**5.686**
	Anti-estrogens	Detrimental	7.614	<0.0001	3.643	15.914
	Metformin*Anti-Estrogens	Detrimental	**10.025**	**0.0002**	**2.967**	**33.878**
	Aromatase inhibitors	Detrimental	1.063	0.8331	0.601	1.880
	Herceptin	Detrimental	4.346	0.0077	1.476	12.795
	Insulin*Herceptin	Detrimental	**25.743**	**<0.0001**	**8.259**	**80.239**
	Aromatase inhibitors*Herceptin	Protective	0.525	0.3295	0.144	1.918
	Methotrexate	Detrimental	1.983	0.0456	1.013	3.879

Stratified drug exposure model

Using a machine learning approach to identify key covariates from the above components a final model of drug exposure was developed **(Table 2.4).** Age at diagnosis of breast cancer, obesity status, rural resident, high-grade breast cancer, in situ breast cancer, chemotherapy, surgical intervention, and Charlson Comorbidity Index >8 were stratified in the final drug exposure model. Metformin was identified to have a protective association with overall (HR=0.545,p-val=0.0059) and disease specific mortality (HR=0.269,p-val=0.0378).

Table 2.4: Overall Model Stratified by Key Covariates

Stratified* Model		All Cause Mortality				Breast Cancer-Specific Mortality			
		HR	**p-val**	**CI-low**	**CI-high**	**HR**	**p-val**	**CI-low**	**CI-high**
	Metformin	**0.545**	**0.0059**	0.354	0.840	**0.269**	**0.0378**	0.078	0.929
	Insulin	1.438	0.0695	0.971	2.129	2.147	0.1321	0.794	5.802
	Anti-estrogens	2.013	0.0316	1.064	3.809	2.806	0.1189	0.767	10.266
	Herceptin	1.117	0.8641	0.314	3.973	2.253	0.5067	0.205	24.771
	Methotrexate	1.073	0.8190	0.601	1.919	2.135	0.2157	0.643	7.096
	Metformin*anti estrogen	4.757	0.0506	0.996	22.714	9.750	0.1067	0.613	155.116
* stratified by: Age at diagnosis of breast cancer, BMI >=35, rural residence, high grade cancer, in situ cancer, chemotherapy, surgical intervention									

In summary, metformin exposure after breast cancer diagnosis appeared to have a significant protective association with all cause survival in both the unadjusted and adjusted models, whereas insulin appeared to have a significant detrimental association with breast cancer-specific survival. However, it is still not clear if the protective and detrimental associations for metformin and insulin, correspondingly, are due truly to a treatment effect during the course of breast cancer treatment or due to their implications for T2DM severity.

Discussion

While existing literature points to a decreased incidence of breast cancer in diabetic patients exposed to metformin[3], existing studies[2,4] have measured metformin exposure occurring after T2DM incidence and have not clearly addressed metformin exposure occurring after breast cancer incidence. Further these studies were limited by lacking either clinical EHR data or cause of death data and may have confounded measures of incidence and survival when studying the impact of metformin. Our study limited potential confounding of incidence with survival by focusing on metformin exposure occurring only after breast cancer diagnosis.

While this analysis remains preliminary, the findings hint that a protective association between long-term metformin exposure after breast cancer diagnosis and all cause and breast cancer-specific survival in a cohort of breast cancer and T2DM patients

may exist. The unadjusted models demonstrated a significant positive association with overall survival, while the adjusted models display a significant protective association between long-term metformin exposure after diagnosis of breast cancer as well as all cause and breast cancer-specific survival. Some confounding relationships were accounted for in the adjusted models, while others remain to be addressed in future work: Drug relationships including metformin exposure prior to breast cancer diagnosis and exposure to other antidiabetic therapies, including insulin, thiazolidinedione, and sulfonylurea concurrent and non-concurrent with metformin exposure. Breast cancer severity, receptor status, grade, stage, surgery status, and treatment regiment need to be utilized. Finally, the most important confounding that remains to be addressed in future work is disambiguating between changes in breast cancer survival outcomes due to metformin and insulin implications for T2DM severity and their implication as a potential treatment effect in breast cancer.

Future research will broaden focus with a larger cohort to understand if metformin exposure both lowers incidence of breast cancer and increases survival outcomes and to what extent. Specifically, metformin exposure in certain individuals may have already lowered the risk of developing breast cancer to a threshold where metformin exposure after breast cancer diagnosis becomes ineffective. Additional analysis will aim to ensure cohort homogeneity between exposure groups and appropriate model weighting. Some specific analytic approaches will include marginal structural modeling and penalized Cox regression.

In conclusion, these findings hint at a potential protective relationship between length of metformin exposure occurring after breast cancer diagnosis and increased survival outcomes. While compelling, adjusted estimates of survival remain preliminary and should be interpreted with caution until fully vetted in future survival research, but provide a foundation to proceed forward with future pharmacogenetic and translational genetic pathway research.

References

1. Boyle P, Boniol M, Koechlin A, et al. Diabetes and breast cancer risk: a meta-analysis. British journal of cancer. 2012;107(9):1608-1617.

2. Currie C, Gale EA, Poole C, Johnson J, Jenkens-Jones S, Morgan C. Mortality After Incident Cancer in People With and Without Type 2 Diabetes. Diabetes care. 2012;35.

3. Chlebowski RT, McTiernan A, Wactawski-Wende J, et al. Diabetes, metformin, and breast cancer in postmenopausal women. Journal of clinical oncology. Aug 10 2012;30(23):2844-2852.

4. Lega IC, Austin PC, Gruneir A, Goodwin PJ, Rochon PA, Lipscombe LL. Association Between Metformin Therapy and Mortality After Breast Cancer: A population-based study. Diabetes care. 2013;36(10):3018-3026.

5. He X, Esteva FJ, Ensor J, Hortobagyi GN, Lee MH, Yeung SC. Metformin and thiazolidinediones are associated with improved breast cancer-specific survival of diabetic women with HER2+ breast cancer. Annals of oncology. 2012;23(7):1771-1780.

6. Bayraktar S, Hernadez-Aya LF, Lei X, et al. Effect of metformin on survival outcomes in diabetic patients with triple receptor-negative breast cancer. Cancer. 2012;118(5):1202-1211.

7. Brown KA, Samarajeewa NU, Simpson ER. Endocrine-related cancers and the role of AMPK. Molecular and Cellular Endocrinology. 2013;366:170-179.

8. Gallagher EJ, LeRoith D. Diabetes, cancer, and metformin: connections of metabolism and cell proliferation. Annals of the New York Academy of Sciences. 2011;1243:54-68.

9. Hardie DG, Ross FA, Hawley SA. AMPK: a nutrient and energy sensor that maintains energy homeostasis. Molecular Cell Biology. 2012;13:251-262.

10. Chute CG, Beck SA, Fisk TB, et al. The Enterprise Data Trust at Mayo Clinic: a semantically integrated warehouse of biomedical data. Journal of the American Medical Informatics Association. 2010;17:131-135.

11. Liu L, Forman S, Barton B. Fitting Cox Model Using PROC PHREG and Beyond in SAS. SAS Global Forum. 2009.

Chapter 3: Disease-Specific Effects of Metformin in Breast Cancer Patients with Type II Diabetes Mellitus

Abstract

Purpose

While metformin use in type 2 diabetes mellitus (T2DM) patients has been shown to reduce the incidence of breast cancer, the impact of metformin and insulin on breast cancer specific survival outcomes remains unclear. Further, due to the implications of metformin and insulin for T2DM severity, the impact of metformin and insulin on mortality outcomes beyond those attributable to T2DM severity remains unclear.

Patients and Methods

1,180 female patients with unilateral, non-recurrent breast cancer diagnosis between 1998 and 2011 and T2DM at breast cancer diagnosis were identified using Mayo Clinic electronic health record (EHR) data. Median age at breast cancer diagnosis was 67 (34-95) and median follow-up time was 83 (6-191) months. Stratified Cox proportional hazard regression was utilized to separate the effects of metformin (n=299) and insulin (n=197) ≥6 months on T2DM severity (breast cancer-specific mortality events removed) and breast cancer treatment outcomes. A linear offset of T2DM severity was utilized to separately study the effect of metformin and insulin in multivariate Cox models of breast cancer treatment and biology.

Results

Significant univariate associations on breast cancer specific mortality were observed for metformin exposure (n=322, HR=0.293 ,p=0.0042) and insulin exposure (n=218, HR=0.2.889, p<0.0001). Significant protective effects for metformin (HR=0.544, p=0.0017) and detrimental effects for insulin (HR=2.144, p<0.0001) were observed due to their implication for T2DM severity. For breast cancer-specific disease impact, significant (p<0.0500) effects in models of breast cancer treatment and biology adjusted for T2DM severity were observed for metformin (HRs=0.388, 0.339) and insulin (HRs=1.956, 2.201).

Conclusion

In this EHR-driven study, we have demonstrated a significant protective effect for metformin and a significant detrimental effect for insulin on breast cancer survival outcomes that is beyond their implication T2DM severity.

Background and Introduction

Metformin is an oral biguanide and is a widely prescribed anti-diabetic medication (Thompson 2014) that is considered to be a first-line treatment for T2DM (Inzucchi 2012) and is well tolerated (Gong 2012). The pharmacological effect of Metformin is marked by inhibition of hepatic glucose production, reduced intestinal glucose absorption, and improved glucose uptake and utilization throughout the body (Gong 2012). Women with type II diabetes mellitus (T2DM) are at an increased risk of developing breast cancer (Boyle 2012). However, a prospective study of metformin for diabetes therapy has showed a decrease in breast cancer incidence below rates for patients without T2DM, with fewer ER and PR positive cases and fewer HER2 negative cases observed (Chlebowski 2012). Further, evidence suggests that insulin may influence cancer prognosis (Pollak 2008).

While preclinical evidence exists to suggest potential repurposing of metformin for breast cancer treatment (Thompson 2014), retrospective and clinical evidence for metformin treatment in breast cancer is insufficient. Specifically, inconclusive evidence exists of the associations between metformin exposure after breast cancer diagnosis and breast cancer specific mortality (Zhang 2014). Eight studies have addressed the impact of metformin on all cause mortality outcomes (He 2012, Hou 2013, Peeters 2013, Xiao 2014, Lega 2013, Bayraktar 2012, Currie 2012), with four of these studies (He 2012, Hou 2013, Peeters 2013, Xiao 2014, Xu 2014) identifying a significant protective association between metformin therapy and all-cause mortality. However, substantial heterogeneity between cohorts and substantially different approaches to control for residual confounding in analysis was observed among these studies (Zhang 2014). Further, only three studies have addressed the impact of metformin on breast-cancer specific mortality (He 2012, Lega 2013, Peeters 2013), with one study reporting a significant protective

association (HR = 0.47,p=0.023) between metformin exposure and risk of breast cancer-specific mortality (He 2012). However, that study was limited to a small number of patients (n=154) with stage >=2 HER2+ breast cancer and T2DM. Heterogeneity among studies may be due to methodological limitations of variable adjustment for differing confounding factors (Lega 2014). A protective association for HER2 positive cases was demonstrated for metformin exposure after breast cancer incidence on breast-cancer specific survival (He 2012). However, associations between metformin exposure and survival were not demonstrated for patients with triple negative breast cancer (Bayraktar 2012).

Metformin use in nondiabetic, newly diagnosed breast cancer patients is subject to active clinical trials. Prospective window of opportunity clinical trials in non-diabetic breast cancer patients have identified associations between metformin exposure prior to surgery and anti-cancer effects (Hadad 2011, Niraula 2012). A phase III clinical trial (NCIC CTG MA.32) evaluating the effect of adjuvant metformin exposure on recurrence and mortality outcomes in non-diabetic, early-stage breast cancer patients is ongoing (Goodwin 2011). A Phase II clinical trial (NCT00909506) testing adjuvant metformin therapy in overweight or pre-diabetic patients is currently underway.

Sufficient evidence does not exist to make inferences regarding the potential impact of metformin exposure after breast cancer diagnosis on breast cancer-specific mortality outcomes. While the ongoing trial has the potential to clarify this issue in early stage patients, further population analysis is needed to elucidate associations by biological subgroups. Specifically, plausible additive or interaction effects of metformin and breast cancer therapies need to be evaluated. More broadly, we need to understand the overall association of metformin exposure with breast cancer-specific mortality outcomes in an age-representative cohort with robust clinical indicators. Using robust clinical indicators contained within the electronic health record (EHR), our study aims to add clarity to this ongoing dialogue by evaluating the impact of metformin use after breast cancer diagnosis and potential drug interactions with survival outcomes in an age-representative and typologically diverse cohort of women with breast cancer and T2DM. This study isolates the effect of metformin exposure (>= 6 months) during the treatment period following

breast cancer diagnosis for breast cancer treatment outcomes beyond its implications for T2DM severity. Additionally, this study follows a similar progression evaluating the effect of insulin due to its potential modifying effect on breast cancer treatment outcomes.

Materials and Methods

Using a combination of manually attributed cancer registry for breast cancer patients, structured queries, and natural language processing (NLP) a cohort of 13,163 eligible women with unilateral, non-recurrent breast cancer was identified. Diagnosis of breast cancer was confirmed by inclusion in the Mayo Clinic Cancer Registry for breast cancer, non-recurrent cancer status was confirmed by using a combination of structured EHR and cancer registry data. Breast cancer receptor status was attributed via an NLP algorithm of clinical notes (under review). 72.2% of patients had complete breast cancer receptor status information available. Manual chart review was performed on 96.4% of patients to confirm breast cancer receptor status. Attribution of a T2DM phenotype was performed using modified methodology developed by eMERGE (Kho 2012), where 1,188 patients were identified as having T2DM and breast cancer. Diagnosis, patient attributes, and last follow-up data points were extracted from the EHR. Medication exposure was extracted from a combination of structured EHR data and NLP of medication reconciliation forms (Pathak 2011). Disease specific mortality was identified from electronic death certificate data from Minnesota Electronic Death Certificates. If a patient was not identified as deceased from either the Minnesota Death Registry or Mayo Clinic's EHR, patient survival time was censored at the last known follow-up date.

Cohort inclusion was limited to female patients with a non-recurrent confirmed diagnosis of unilateral breast cancer and T2DM. Inclusion was also limited to eligible new breast cancer cases between the years 1998 and 2011 to allow for a minimum two year follow-up period. Male patients and those with type I diabetes mellitus or ambiguous diabetes type were excluded from this study. Patients with survival or follow-up < 6 months were silenced from the cohort to control for immortal time bias. Metformin and insulin exposure were parameterized to represent >=6 months of drug exposure within the treatment period post breast cancer diagnosis. Patients needed to have multiple

recorded mentions of an individual class of antidiabetic medications spanning a period of greater than 30 days to be considered as having exposure to an individual class of antidiabetic medications. Patients without multiple recorded mentions of an individual class of antidiabetic medications spanning more than 30 days post breast cancer diagnosis were parameterized as not having exposure to individual class of antidiabetic medications (n=25 for metformin). The resulting cohort contained female patients (n=1,180) with a median age of breast cancer diagnosis of 67 (34-95) years and median follow-up period of 83 (6-191) months. A detailed diagram of cohort development can be found in **Figure 3.1** and a description of the cohort can be found in **Table 3.1**. Patients with metformin and fulvestrant exposure were silenced from analysis due to a small sample size (n=8) and unexpected effect on breast cancer specific mortality. Primary care patients represented 34.1% of the patients in the study cohort. Median age was similar between the primary care (68) and non-primary care (67) cohort, and similar number of metformin exposure events, 32.4% vs. 25.4% respectively.

Figure 3.1: Cohort Development

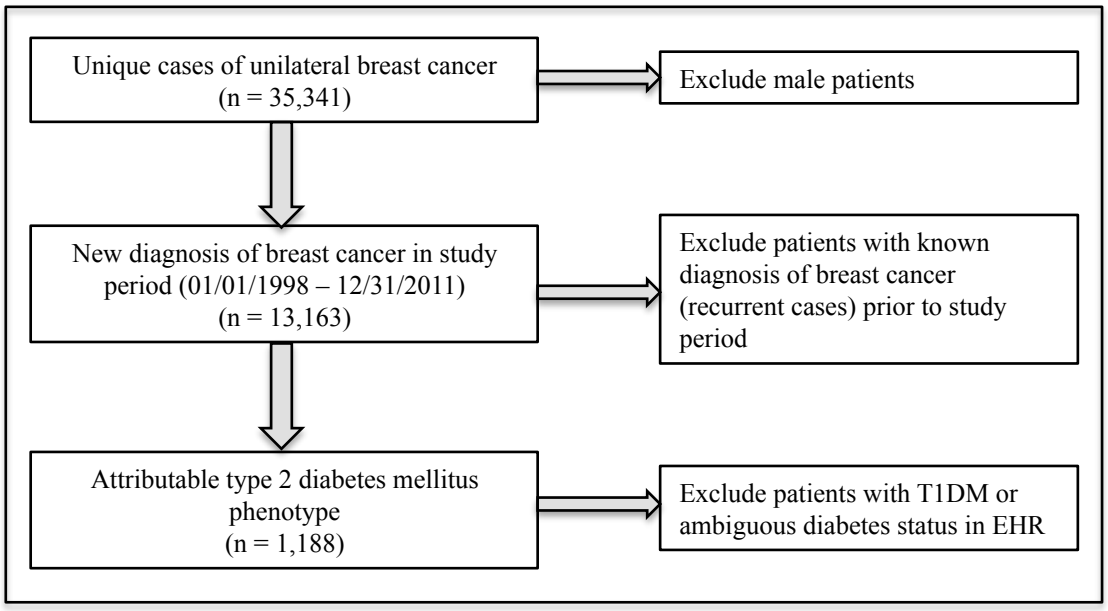


Table 3.1: Description of Cohort

		event	all
Demographics			
	all cause mortality	306	1188
	breast cancer-specific mortality	71	1053

	follow up period		83 (6 , 191)	1188
	age at diagnosis of breast cancer		67 (34 , 95)	
		Age < 50	98	
		50>=Age<67	477	
		67>=Age<80	504	
		Age>=80	109	1188
	resident of rural zip code		277	1162
	primary care patient		405	1188
	white		1072	1188
Antidiabetic Therapies				
	metformin		355	
	insulin		245	
	tzd		73	
	sulfonylurea		182	1188
Diabetic Severity				
	A1c		6.5 (4.5 , 15.1)	
		uncontrolled diabetes	205	687
	BMI		31.8 (14.5 , 61.7)	
		Underweight (BMI<18.5)	5	1135
		Normal (18.5>=BMI>25)	149	
		Overweight (25>=BMI>30)	281	
		Moderately obese (30>=BMI>35)	324	
		Severely obese BMI>=35	376	
Breast Cancer Severity				
	Stage	In Situ	185	1152
		1	596	
		2	14	
		3	240	
		4	25	
	Grade	1	197	1188
		2	424	
		3	273	
		4	294	
Breast Cancer Receptor Status				
	Estrogen Receptor Positive		795	919
	Progesterone Receptor Positive		728	893
	HER2/neu status actionable		925	925
		HER2/neu positive	106	925
	Triple Negative		31	857
Breast Cancer Treatment				
	Surgery		444	979
	Chemotherapy		242	1174
	Radiation Therapy		227	890
	Hormone therapy		503	1168
	Immune therapy		12	1187
	Fulvestrant		43	1188

	Aromatase inhibitors	567	1188
	Trastuzumab	96	1188
	Methotrexate	72	1188
	Ovarian oblation	23	1188
	Selective estrogen receptor modulators	918	1188
Charlson Comorbidity Index			
	Myocardial Infarction	73	1188
	Congestive heart failure	159	
	peripheral vascular disease	132	
	cerebrovascular disease	166	
	Dementia	62	
	chronic pulmonary disease	263	
	ulcer	88	
	mild liver disease	119	
	diabetes	1188	
	Diabetes with organ damage	206	
	hemiplegia	20	
	moderate/severe renal disease	135	
	moderate/severe liver disease	24	
	metastatic solid tumor	394	
	Aids	0	
	Rheumatologic disease	66	
	Cancer	1188	

Since clear clinical treatment progressions exist for T2DM (Inzucchi 2012) anti-diabetic therapies carry implications for T2DM severity In order to isolate the effects of metformin and insulin beyond their impact on survival due to T2DM severity we engaged in a series of Cox Proportional Hazard Regression models. First, to identify separate baseline associations of metformin and insulin exposure on morality outcomes we developed separate univariate models. Second, to assess baseline additive associations for anti-diabetic therapies with morality outcomes we developed a multivariate model. Third, a stratified model was utilized to independently model the separate effects of anti-diabetic therapies on mortality outcomes (stratum 1) and breast cancer specific mortality (stratum 2). Specifically, this stratified approach separately attributed the impact of anti-diabetic drug therapy on T2DM severity to stratum 1 and separately attributed the additional breast cancer-specific mortality burden on stratum 2. Stratification was utilized to control for covariates as opposed to a paired model to minimize potential unbalance in subgroup cohort size during the matching process. Since stratum 1 contained the impact of anti-diabetic therapies on mortality outcomes we utilized its linear coefficient as an offset to

adjust for anti-diabetic drug therapy exposure associated with T2DM severity. Finally, we modeled the effect of metformin and insulin exposure, adjusted for T2DM severity on breast cancer specific mortality with breast cancer treatment and biology considerations. Exploratory interaction analysis was performed for both considerations. Backwards elimination was utilized to develop a final model of metformin and insulin exposure, adjusted for T2DM severity, for both considerations. Statistical analysis was performed using Linux SAS v9.3 and R 3.1.0.

Results

Baseline associations

Univariate survival analysis (**Table 3.2a**) demonstrated a protective baseline association of metformin exposure after breast cancer diagnosis with decreases in all cause mortality (HR=0.635,p-val=0.0049) and breast cancer-specific mortality (HR=0.293,p-val=0.0042). A detrimental baseline association for insulin exposure after breast cancer diagnosis with increases in all cause mortality (HR=2.124,p-val<0.0001) and breast cancer-specific mortality (HR=2.889,p-val<0.0001). Additive associations for metformin, insulin, and other anti-diabetic therapies (i.e. thiazolidinedione, sulfonylurea) demonstrated similar associations with morality outcomes (**Table 3.2b**).

T2DM severity

The stratified model (**Table 3.2c**) identified a significant protective impact for metformin (HR=0.544 ,p-val=0.0017) and significant detrimental impact for insulin (HR=2.144, p-val<0.0001), controlling for other anti-diabetic therapies, on mortality outcomes in stratum 1. Due to the T2DM severity implications associated with T2DM therapy progression, these baseline associations were anticipated. The linear coefficients of stratum 1 were utilized as offsets in additional modeling on breast cancer outcomes Separately, the stratified model identified a significant protective association for metformin (HR=0.352, p-val=0.0273) and marginally significant detrimental association for insulin (HR=1.775, p-val=0.0772) on breast cancer specific mortality in stratum 2 that is beyond the impact of metformin and insulin on morality outcomes due to T2DM severity. Since stratum 2 did not control for confounding additive effects related to breast

cancer the impact of metformin and insulin on breast cancer outcomes was not representative. However, the directionality of these associations indicated decreases in baseline hazard ratios for both metformin and insulin.

Table 3.2: Diabetes Medications

Table 2a: Univariate Analysis of Metformin and Insulin										
		All Cause Mortality					Breast Cancer-Specific Mortality			
	N	**HR**	**p-val**	**CI-low**	**CI-high**	**N**	**HR**	**p-val**	**CI-low**	**CI-high**
Metformin	330	**0.635**	**0.0049**	0.463	0.871	299	**0.293**	**0.0042**	0.127	0.679
Insulin	219	**2.124**	**<0.0001**	1.608	2.806	197	**2.889**	**<0.0001**	1.712	4.876
Table 2b: Multivariate model of T2DM medications										
		All Cause Mortality					Breast Cancer-Specific Mortality			
	N	**HR**	**p-val**	**CI-low**	**CI-high**		**HR**	**p-val**	**CI-low**	**CI-high**
Metformin	330	**0.628**	**0.0160**	0.430	0.917	299	**0.248**	**0.0003**	0.117	0.527
Insulin	219	**2.119**	**<0.0001**	1.499	2.996	197	**3.190**	**<0.0001**	1.848	5.506
Thiazolidinedione	66	0.956	0.8866	0.514	1.777	58	2.163	0.0700	0.939	4.981
Sulfonylurea	170	1.150	0.5388	0.737	1.794	157	1.458	0.2987	0.716	2.971

Table 2c: Stratified T2DM Severity Model (Two baseline hazard ratios)					
	N	**HR**	**p-val**	**CI-low**	**CI-high**
Strata 1: All Cause and Breast Cancer Specific Mortality					
Metformin	330	**0.544**	**0.0017**	0.372	0.795
Insulin	219	**2.144**	**<0.0001**	1.515	3.034
Thiazolidinedione	66	1.297	0.3139	0.782	2.150
Sulfonylurea	170	1.283	0.1971	0.879	1.872
Strata 2: Breast Cancer Specific Survival Endpoint					
Metformin	299	**0.352**	**0.0273**	0.139	0.889
Insulin	197	1.775	0.0772	0.939	3.352

Breast cancer treatment perspective

A full and reduced multivariate model of metformin, insulin, and breast cancer treatment, adjusting for T2DM severity can be found in **Table 3.3**. A backwards elimination model (not shown) evaluation potential interaction effects between metformin, insulin, and breast cancer treatments revealed no evidence of treatment interactions. The reduced model (**Table 3.3b**) identified a significant protective effect for metformin (HR=0.388, p-val=0.0353) and detrimental effect for insulin (HR=1.956, p-val=0.0170) on breast cancer-specific mortality.

Table 3.3: Breast Cancer Treatment Perspective, Adjusted for T2DM Severity

Table 3.3a: Full Model	Breast Cancer Specific Mortality				
	N	**HR**	**p-val**	**CI-low**	**CI-high**
Age	1046	**0.552**	**0.0005**	0.394	0.773
Metformin	299	**0.388**	**0.0356**	0.161	0.938
Insulin	197	**1.944**	**0.0191**	0.115	3.389
Fulvestrant	24	**5.177**	**<0.0001**	2.456	10.910
Aromatase inhibitors	513	0.815	0.4861	0.459	1.448
Trastuzumab	83	1.427	0.3236	0.704	2.893
Selective estrogen receptor modulators	838	0.616	0.0947	0.349	1.087
Methotrexate	58	**2.729**	**0.0022**	1.437	5.184
Chemotherapy	216	**1.878**	**0.0298**	1.064	3.316
Table 3.3b: Reduced Model		**HR**	**p-val**	**CI-low**	**CI-high**
Age	1046	**0.559**	**0.0008**	0.398	0.786
Metformin	299	**0.388**	**0.0353**	0.161	0.937
Insulin	197	**1.956**	**0.0170**	1.127	3.394
Fulvestrant	24	**4.973**	**<0.0001**	2.537	9.747
Selective estrogen receptor modulators	838	0.578	**0.0461**	0.337	0.991
Methotrexate	58	**2.800**	**0.0017**	1.475	5.315
Chemotherapy	216	**1.951**	**0.0158**	1.133	3.358

Breast cancer biology perspective

A full and reduced multivariate model of metformin, insulin, and breast cancer biology, adjusting for T2DM severity can be found in **Table 3.4**. A backwards elimination model (not shown) testing interaction effects between metformin, insulin, and breast cancer biology revealed no evidence of interactions by receptor status. However, it is important to note that clinical guidelines for testing overexpression of the relevant receptors has varied throughout the study period, having potential to bias these associations towards the null due to limited sample size and misattribution of receptor impact. The reduced model (**Table 3.4b**) identified a significant protective effect for metformin (HR=0.339, p-val=0. 0159), a detrimental effect for insulin (HR=2.201, p-val=0.0131), and significant additive associations for estrogen receptor and HER2 on breast cancer-specific mortality.

Table 3.4: Breast Cancer Biology Perspective, Adjusted for T2DM Severity

Table 3.4a: Full Model	Breast Cancer Specific Mortality				
	N	**HR**	**p-val**	**CI-low**	**CI-high**
Age	1046	**0.493**	**0.0003**	0.337	0.723
Metformin	299	**0.344**	**0.0182**	0.142	0.834
Insulin	197	**2.019**	**0.0310**	1.066	3.824
Progesterone receptor - positive	636	**0.784**	**0.5879**	0.325	1.891
Estrogen receptor - positive	706	0.543	0.2114	0.209	1.414
HER2 actionable - negative	394	**2.217**	**0.0279**	1.090	4.509
HER2 actionable - positive	88	2.442	0.0535	0.987	6.045
Table 3.4b: Reduced Model		**HR**	**p-val**	**CI-low**	**CI-high**
Age	1046	**0.489**	**0.0001**	0.339	0.703
Metformin	299	**0.339**	**0.0159**	0.141	0.817
Insulin	197	**2.201**	**0.0131**	1.181	4.101
Estrogen receptor - positive	706	**0.476**	**0.0366**	0.237	0.955
HER2 actionable - negative	394	**2.425**	**0.0107**	1.231	4.886
HER2 actionable - positive	88	**2.434**	**0.0451**	1.020	5.811

Patient demographic considerations

Table 3.5: Patient Demographics

Table 3.5: Demographics	All Cause Mortality				Breast Cancer-Specific Mortality			
	HR	**p-val**	**CI - low**	**CI - high**	**HR**	**p-val**	**CI - low**	**CI - high**
Metformin	**0.625**	**0.0039**	0.454	0.860	**0.433**	**0.0218**	0.212	0.885
Insulin	**1.807**	**0.0001**	1.338	2.441	**2.587**	**0.0009**	1.475	4.538
Age < 50	*ref*	*ref*	*ref*	*ref*	*ref*	*ref*	*ref*	*ref*
50>=Age<67	0.879	0.6181	0.530	1.459	**0.368**	**0.0107**	0.171	0.793
67>=Age<80	1.586	0.0630	0.975	2.579	0.612	0.1699	0.304	1.234
Age>=80	**3.983**	**<0.0001**	2.324	6.828	0.951	0.9167	0.371	2.435
White	0.756	0.1954	0.495	1.154	0.796	0.5679	0.363	1.744
Rural residence	**1.307**	**0.0415**	1.010	1.690	0.966	0.9048	0.548	1.703
BMI (>= 35)	1.043	0.7547	0.799	1.362	0.743	0.2915	0.427	1.291
Charlson Score > 8	**2.475**	**<0.0001**	1.952	3.139	**8.713**	**<0.0001**	4.773	15.906

Postmenopausal women with a diagnosis of breast cancer at < 67 years old were associated with decreased breast cancer specific mortality (HR=0.368, p-val=0.0107) when compared to premenopausal women. Surprisingly, obesity (BMI>=35) was not

significantly associated with either all cause or breast cancer specific mortality outcomes (**Table 3.5**) since obesity is associated with both increased diabetes and breast cancer incidence. However, the majority (61.7%) of patients in this study cohort were obese (**Table 3.1**). Rural residence was selected for inclusion in the final due to a significant overall association (HR=1.307, p-val=0.0415) between rural patient residence and overall survival.

Discussion

Existing literature evaluating the associations of metformin exposure occurring after breast cancer incidence with mortality outcomes have yielded inconclusive and inconsistent results (Zhang 2014, Lega 2014). Cohorts have been heterogeneous (Lega 2014), definitions of metformin exposure have varied (Lega 2014), and methodological approaches to approach for confounding have varied (Lega 2014). Further, studies not utilizing breast cancer specific morality as an endpoint (Hou 2013, Xiao 2014, Bayraktar 2012, Currie 2012, Xu 2014) are critically limited in their ability to elucidate the impact of metformin and insulin on mortality that is beyond that which is due to T2DM severity. Our study added clarity to these studies by evaluating metformin exposure occurring after breast cancer diagnosis and addressing robust indicators including breast cancer receptor status and breast cancer-specific mortality not available in previous studies. Further, our analysis efforts independently isolated the effects of metformin and insulin due to implications for T2DM severity and their impact on breast cancer mortality outcomes. In this study, we consistently identified a significant protective effect for metformin and significant detrimental effect for insulin on breast cancer-specific mortality that is beyond their impact on T2DM severity.

This study is uniquely poised to elucidate the impact of metformin and insulin exposure on morality outcomes that is due to their effect on breast cancer mortality outcomes during breast cancer treatment and not due to the effect T2DM severity associated with the original purpose of metformin and insulin as a T2DM treatment regiment. We are further advantaged in our ability to include robust structured (Chute 2010), semi-structured (Pathak 2011), and unstructured (Savona 2010) clinical and treatment indicators not commonly available in population or claims-based studies. These

approaches afford independence from manual chart review for cohort ascertainment and are highly granular but also highly portable and scalable. Finally, we are uniquely poised to perform future translation biomedical research on parallel subcohorts due to the large percentage of patients who have donated, plasma, serum, and tumor samples for research to biobanks.

A limitation for this study includes lacking insight into duration of metformin and insulin treatment prior to breast cancer treatment for referral patients (n=783) due to incomplete historical clinical data. While, the efforts of isolating the confounding relationship between T2DM severity and T2DM treatment remove this limitation for exposure occurring during breast cancer treatment, we are unable to evaluate the potential modifying impact of metformin and insulin duration of exposure occurring prior to breast cancer treatment on breast cancer incidence and severity. Further, an estimated approximately 30% of patients do not respond to metformin (Cook 2007), necessitating alternative anti-diabetic therapy. Due to incomplete historical data we were not able to account for a potential difference between patients that did not respond to metformin treatment prior to breast cancer treatment. Finally, disease-specific cause of death could not be ascertained for 44.1% (n=135) of eligible patients with known mortality events (n=306) due to deaths occurring outside of MN, which has potential to skew interpretation of disease-specific survival analysis towards the null. Median follow-up period was similar amongst patients with known mortality (61.5 months) vs. patients with known mortality but missing cause of death (65.0 months).

As research on the potential repurposing of metformin in breast cancer proceeds forward it is important to note that the biological function and mechanism of metformin remains unclear (Todd 2014). The pharmacokinetics (PK) of metformin, the transportation throughout the body, are moderately understood (Gong 2012). Metformin is not metabolized, with absorption of metformin known to occur in the small and large intestine (Graham 2011). Uptake of metformin from the blood is known to occur in the kidneys and liver (Todd 2014), but can be reasonably assumed to occur in any tissue with abundance of organic cation transporters (OCT)(Viollet 2012). Eventually metformin is excreted unchanged in the urine(Graham 2011). However, the pharmacodynamics (PD)

of metformin, the physiological and biochemical impact of metformin in the body, are not clearly understood(Gong 2012). Metformin works primarily by inhibiting hepatic glucose production by reducing gluconeogenesis in the liver (Hundal 2000) and is also known to reduce intestinal glucose absorption(Sakar 2010). Further, metformin appears to improve glucose uptake and utilization systemically (Gong 2012). However, there is considerable variation in glycemic response to metformin(Gong 2012) and serious adverse reactions to metformin can occur (Bailey 1996). Further, an important confounding factor to consider based on in vitro models is that 'metformin is probably unable to exert cytotoxic or cytostatic effects on breast cancer subtypes at pharmacological concentrations and normal plasma glucose levels' (Sadighi 2014), suggesting that the effect of metformin on breast cancer is more complex than cytotoxicity. Further research integrating multi-omics approaches to elucidate the biological function and mechanism of metformin in breast cancer is needed.

Conclusion

Metformin exposure during the treatment period after breast cancer diagnosis demonstrated a significant protective effect on breast cancer-specific mortality, while insulin during the same period demonstrated a significant detrimental effect. Our study demonstrates that the strength of these effects is beyond their implications for T2DM severity.

References

Bayraktar S, Hernadez-Aya LF, Lei X, et.al. Effect of metformin on survival outcomes in diabetic patients with triple-negative breast cancer. Cancer 2012;118(5):1202-1211.

Bonanni B, Puntoni M, Cazzaniga M, et al. Dual effect of metformin on breast cancer proliferation in a randomized presurgical trial. Journal of clinical oncology: official journal of the American Society of Clinical Oncology. Jul 20 2012;30(21):2593-2600.

Boyle P, Boniol M, Koechlin A, et.al. Diabetes and breast cancer risk: a meta-analysis. British Journal of Cancer. 2012;107(9):1608-1617.

Chlebowski RT, McTiernan A, J. W-W, al. e. Diabetes, metformin, and breast cancer in

postmenopausal women. Journal of Clinical Oncology. 2012;30(23):2844-2852.

Chute CG, Beck SA, Fisk TB, Mohr DN. The Enterprise Data Trust at Mayo Clinic: a semantically integrated warehouse of biomedical data. Journal of the American Medical Informatics Association: JAMIA. Mar-Apr 2010;17(2):131-135.

Cook MN, Girman CJ, Stein PP, Alexander CM. Initial monotherapy with either metformin or sulphonylureas often fails to achieve or maintain current glycaemic goals in patients with Type 2 diabetes in UK primary care. Diabetic Medicine. 2007;24:350-358.

Currie CJ, Gale EAM, Poole CD, Johnson JA, Jenkins-Jones S, Morgan CL. Mortality After Incident Cancer in People With and Without Type 2 Diabetes. Diabetes Care. 2012;35:299-304.

Ferro A, Goyal S, Kim S, et al. Evaluation of Diabetic Patients with Breast Cancer Treated with Metformin during Adjuvant Radiotherapy. International journal of breast cancer. 2013;2013:659723.

Gong L, Goswami S, Giacomini KM, Altman RB, Klein TE. Metformin pathways: pharmacokinetics and pharmacodynamics. Pharmacogenetics and genomics. Nov 2012;22(11):820-827.

Goodwin PJ, Stambolic V, Lemieux J, et al. Evaluation of metformin in early breast cancer: a modification of the traditional paradigm for clinical testing of anti-cancer agents. Breast cancer research and treatment. Feb 2011;126(1):215-220.

Graham G.C., Punt J, Arora M, et al. Clinical Pharmacokinetics of Metformin. Clinical Pharmacokinetics. 2011;50(2):81-98.

Hadad S, Iwamoto T, Jordan L, et al. Evidence for biological effects of metformin in operable breast cancer: a pre-operative, window-of-opportunity, randomized trial. Breast cancer research and treatment. Aug 2011;128(3):783-794.

Hadad SM, Hardie DG, Appleyard V, Thompson AM. Effects of metformin on breast cancer cell proliferation, the AMPK pathway and the cell cycle. Clinical Translational Oncology. Dec 2013.

He X, Esteva FJ, Ensor J, Hortobagyi GN, Lee MH, Yeung SC. Metformin and

thiazolidinediones are associated with improved breast cancer-specific survival of diabetic women with HER2+ breast cancer. Annals of oncology: official journal of the European Society for Medical Oncology / ESMO. Jul 2012;23(7):1771-1780.

Hou G, Zhang S, Zhang X, Wang P, Hao X, Zhang J. Clinical pathological characteristics and prognostic analysis of 1,013 breast cancer patient with diabetes. Breast cancer research and treatment. 2013;137:807-816.

Hundal RS, Krssak M, Dufour S, et al. Mechanism by Which Metformin Reduces Glucose Production in Type 2 Diabetes. Diabetes. 2000;49.

Inzucchi SE, Bergenstal RM, Buse JB, et al. Management of Hyperglycemia in Type 2 Diabetes: A Patient-Centered Approach. Diabetes Care. 2012;35:1364-1379.

Kho AN, Hayes MG, Rasmussen-Torvik L, et al. Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study. Journal of the American Medical Informatics Association : JAMIA. Mar-Apr 2012;19(2):212-218.

Kim J, Lim W, Kim EK, et al. Phase II randomized trial of neoadjuvant metformin plus letrozole versus placebo plus letrozole for estrogen receptor positive postmenopausal breast cancer (METEOR). BMC cancer. 2014;14:170.

Lega IC, Austin PC, Gruneir A, Goodwin PJ, Rochon PA, Lipscombe LL. Association Between Metformin Therapy and Mortality After Breast Cancer: A population-based study. Diabetes Care. 2013;36(10):3018-3026.

Lega IC, Shah PS, Margel D, Beyene J, Rochon PA, Lipscombe LL. The effect of metformin on mortality following cancer among patients with diabetes. Cancer Epidemiology Biomarkers & Prevention. 2014.

Lin HC, Hsu YT, Kachingwe BH, Hsu CY, Uang YS, Wang LH. Dose effect of thiazolidinedione on cancer risk in type 2 diabetes mellitus patients: a six-year population-based cohort study. Journal of clinical pharmacy and therapeutics. Mar 24 2014.

Ma J, Guo Y, Chen S, et al. Metformin enhances tamoxifen-mediated tumor growth

inhibition in ER-positive breast carcinoma. BMC cancer. 2014;14:172.

Niraula S, Dowling RJ, Ennis M, et al. Metformin in early breast cancer: a prospective window of opportunity neoadjuvant study. Breast cancer research and treatment. Oct 2012;135(3):821-830.

Pathak J, Murphy SP, Willaert BN, et al. Using RxNorm and NDF-RT to Classify Medication Data Extracted from Electronic Health Records: Experiences from the Rochester Epidemiology Project. American Medical Informatics Association Annual Symposium. 2011:1089-1098.

Peeters PJ, Bazelier MT, Vestergaard P, et al. Use of metformin and survival of diabetic women with breast cancer. Curr Drug Saf. 2013;8:357-363.

Pollak M. Insulin and insulin-like growth factor signaling in neoplasia. Nature Reviews Cancer. 2008;8: 915-928.

Sadighi S, Amanpour S, Behrouzi B, Khorgami Z, Muhammadnejad S. Lack of Metformin Effects on Different Molecular Subtypes of Breast Cancer under Normoglycemic Conditions: An in vitro Study. Asian Pacific Journal of Cancer Prevention. 2014;15(5):2287-2290.

Savova GK, Masanz JJ, Ogren PV, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. Journal of the American Medical Informatics Association : JAMIA. Sep-Oct 2010;17(5):507-513.

Sakar Y, Meddah B, Faouzi MYA, Cherrah Y, Bado A, Ducroc R. Metformin-Induced Regulation of Intestinal D-Glucose Transporters. Journal of Physiology and Pharmacology. 2010;61(3):301-307.

Thompson AM. Molecular pathways: preclinical models and clinical trials with metformin in breast cancer. Clinical cancer research : an official journal of the American Association for Cancer Research. May 15 2014;20(10):2508-2515.

Todd JN, Florez JC. An update on the pharmacogenomics of metformin: progress, problems and potential. Pharmacogenomics. 2014;15(4):529-539.

Vazquez-Martin A, Oliveras-Ferraros C, Menendez JA. The antidiabetic drug metformin suppresses HER2 (erbB-2) oncoprotein overexpression via inhibition of the mTOR effector p70S6K1 in human breast carcinoma cells. Cell Cycle. 2009;8(1):88-96.

Viollet B, Guigas B, Sanz Garcia N, Leclerc J, Foretz M, Andreelli F. Cellular and molecular mechanisms of metformin: an overview. Clinical science. Mar 2012;122(6):253-270.

Xiao Y, Zhang S, Hou G, Zhang X, Hao X, Zhang J. Clinical pathological characteristics and prognostic analysis of diabetic women with luminal subtype breast cancer. Tumour Biol. 2014.

Xu H, Aldrich MC, Chen Q, et al. Validating drug repurposing signals using electronic health records: a case study of metformin associated with reduced cancer mortality. Journal of the American Medical Informatics Association. 2014;0:1-10.

Zhang ZJ, Li S. The prognostic value of metformin for cancer patients with concurrent diabetes: a systematic review and meta-analysis. Diabetes , Obesity and Metabolism. 2014.

Zhou K, Donnelly L, Yang J, et al. Heritability of variation in glycaemic response to metformin: a genome-wide complex trait analysis. The Lancet Diabetes & Endocrinology. 2014;2(6):481-487.

Zhu P, Davis M, Blackwelder AJ, et al. Metformin selectively targets tumor-initiating cells in ErbB2-overexpressing breast cancer models. Cancer prevention research. Feb 2014;7(2):199-210.

Zordoky BN, Bark D, Soltys CL, Sung MM, Dyck JR. The anti-proliferative effect of metformin in triple-negative MDA-MB-231 breast cancer cells is highly dependent on glucose concentration: implications for cancer therapy and prevention. Biochimica et biophysica acta. Jun 2014;1840(6):1943-1957.

Part #2: Accounting for Socioecological Context in Translational Research

Due to the confounding relationship between clinical and social factors it is likely that the effect of socio-ecological conditions (SECs) is being inappropriately attributed to clinical variables in retrospective clinical research. Utilizing validated measures of social stress as the SEC of interest and Charlson Comorbidity Index as the training outcome I created a SEC Index that aimed to capture direct and latent community effects surrounding social stress. While not a perfect solution, I demonstrated a potential approach to incorporate SEC into clinical research models that minimizes power loss and model over fitting.

Chapter 4: Studying the Confounding Effects of Socio-Ecological Conditions in Retrospective Clinical Research: A Use Case of Social Stress

Abstract

Socio-ecological Conditions (SECs) are important to include in clinical research models as they have been known to impact the health of patients. However, current clinical research models account for these factors only in an unsatisfyingly rudimentary way. In this study, I developed an SEC Index that captured the latent and direct effects of social stress, one of the many kinds of SEC, on patients' general health as measured by the Charlson Comorbidity Index. I demonstrated that the above SEC Index had a significant effect in a clinical model, a patient-level model with the specific clinical outcome of breast cancer prevalence. Further, I demonstrated that including the SEC Index of social stress into the clinical models significantly increased their performance. This study demonstrated a viable approach that is interchangeable to include any SEC of interest, to more appropriately account for SECs in clinical research models.

Introduction

Socio-ecological Conditions (SECs) have been known to impact the health of patients, but current clinical research models account for these factors only in an unsatisfyingly rudimentary way. Phenomena such as access to health care and social support networks are examples of SEC; factors that can profoundly impact a patient's health and prognosis once health deteriorates. SECs not only influence outcomes, but they confound patient characteristics and clinical factors, inhibiting the ability of clinical models to estimate the effect of clinical factors independently of SECs. Specifically, without adjusting for SECs we can only estimate the combined effect of SEC and the clinical factors.

In what follows, we describe SECs in more details, present our methodology and demonstrate its utility on a large tertiary care provider in the Midwest U.S. In this study, we set out to develop an SEC Index, a summary of socio-ecological measures that quantifies the effect of SECs on patients' general health. Then we utilize this SEC Index in a clinical risk prediction model with a specific end point (e.g. breast cancer prevalence)

and show how the inclusion of the SEC Index results in a statistically significantly better model.

Background

Socio-ecological conditions (SECs) are the embodiment of social and ecological population factors that exist within a defined geographical region (i.e. community) and are known to impact health of an individual patient[1] and enhance stability of retrospective clinical research[2]. A community is an amalgamation of social interactions and geographical proximity exhibiting many confounded and latent characteristics. These characteristics cannot be effectively measured as independent metrics, and are known to vary across geographic regions[1]. While person-level measurements of socioeconomic status are commonly included in statistical analysis in an attempt to control for alternative confounding factors, they do not adequately represent the underlying phenomena at the root of an SEC[4]. Further, placing measures of SEC directly in models can lead to model overfitting and potentially reduces power. In our study we chose to focus on the SEC mechanism social stress because of its hypothesized relevance to breast cancer and because it has established, validated population measures. However, it is most important to emphasize that any SEC phenomenon of interest, such as factors measuring social contagions, demographic change, and social capital, can be readily substituted within the study design.

Social stress is a phenomenon experienced by individuals when they do not have the resources to address an acute situation[4]. Further, social stress is known to be highly confounded with socioeconomic status (SES)[4]. Social stress has been associated with negative impacts on health, including increased prevalence of asthma, diabetes, gastrointestinal disorders, myocardial infarction, cancer, and rheumatoid arthritis[5]. Social stressors are commonly socially patterned, and can manifest at both the individual and community level[6]. Further, mouse models of social stress have identified correlations between localized (i.e. non-systemic) mammary adipose-specific metabolic changes and increased mammary tumor growth[7].

We chose to use validated measures of social stress: Index of Dissimilarity(D-score)[8], Townsend Index of Socioeconomic Derivation Index(T-score)[9], and poverty[4]. For brevity, we omit the exact definitions, but at a high level, D-score quantifies the homogenization of racial distribution across geographic areas in relation to the entire geographic region of study; and T-score, which is composed of four components, measures unemployment, non-car ownership, household overcrowding, and non-home ownership. Poverty is a measure collected and provided by the US Census Bureau and provides a direct measures of socioeconomic deprivation.

Study Aim

In this work, we seek to understand if variation in established measures of socio-ecological conditions (SEC) for social stress are associated with breast cancer prevalence. Specifically, we developed an SEC Index using the above validated SEC measures of social stress; and later used this index as a covariate in addition to patient-level clinical covariates in a model predicting breast cancer prevalence (clinical model)

Materials

This study utilizes a combination of clinical and population-based data sources. A cohort consisting of primary care patients (n=228,069) with longitudinal clinical data were aggregated from Mayo Clinic's EHR and Enterprise Data Trust[10] using a combination of structured queries. Validated population measures of social stress were calculated using 2010 American Community Survey and US Census datasets. All population data was clustered at the census block group level. A SAS address–census block group crosswalk *(proc geocode)* was used to assign patients to their Census Block Group (CBG) of residence. CBGs (n=278) corresponding to Mayo Clinic's (in Rochester, Minnesota, USA) primary care coverage area (Dodge, Fillmore, Goodhue, Houston, Mower, Olmsted, Wabasha, and Winona Counties in Minnesota, USA) were included in the analysis. Patients who had designated residence in more than one CBG corresponding to Mayo Clinic's primary care coverage area were excluded from this study. Further, to eliminate estimation bias in low coverage areas patients who resided in census block groups with <50 patients were excluded from this study. CBGs were assigned a random

identifier and patient-level data was de-identified prior to analysis to ensure patient privacy and confidentiality. The final analysis cohort contained deidentified data for a relatively homogeneous cohort of 94,561 patients and 237 CBGs. A detailed diagram of cohort demographics can be found in **Table 4.1**.

Table 4.1: Cohort Demographics (n=94,561)		
Age		
>=18 to ≤25	14,583	15.4%
>25 to ≤35	16,761	17.7%
>35 to ≤45	16,262	17.2%
>45 to ≤55	18,160	19.2%
>55 to ≤65	11,793	12.5%
>65 to ≤75	8,393	8.9%
>75 to ≤85	6,143	6.5%
>85	2,466	2.6%
Caucasian	85,238	90.1%
Male	43,833	46.4%
Coverage		
Low (<33%)	17,939	19.0%
Medium (33 to 50%)	18,893	20.0%
High (50 to 100%)	57,729	61.0%
Poverty	6,720	7.1%
T-score		
High Unemployment	15,440	16.3%
High Non-Car Ownership	6,720	7.1%
D-Score High	9,323	9.9%

Methods

Our proposed method has two steps. First, we develop the SEC Index using a generic endpoint to quantify the effect of the SEC measures on health in general. For this step, we utilize 30% of the data set. In the second step, we utilize the remaining 70% of the data and build the specific clinical model, which includes the SEC Index as an independent variable, with breast cancer prevalence as the clinical end point. Note that to avoid model overfitting, the portion of the data (30%) on which the SEC Index is

developed has no overlap with the portion of the data (70%) that the clinical model is constructed on.

SEC Index Construction

We construct an SEC Index to capture the effect of a number of known measures of social stress on the patients' general health. We quantify patients' general health through the Charlson Comorbidity Index[11], with index value in excess of 3 indicating high risk of mortality (poor health). The independent variables include the D-score[8] (averaged over each census block group and dichotomized[12] into high and low at .5), components of the T-score[9], poverty[13], rurality[14], and coverage. The D-score quantifies homogenization of racial distribution across geographic areas in relation to the entire geographic region of study. CBG poverty is defined as absolute poverty thresholds, a measured and defined by the US Census Bureau[13]. In our study, D-score was calculated for individual patients (stratified at 90^{th} percentile) and then averaged within each census block group. T-score was measured by unemployment (90^{th} percentile), non-car ownership (20^{th} percentile), household overcrowding (10^{th} percentile), and non-home ownership (90^{th} percentile). Poverty was stratified at the 90^{th} percentile. Coverage was defined as the percentage of the population in the CBG who received their *primary care* at Mayo Clinic. Coverage needs to be adjusted for, as patients in certain CBGs only receive specialty care (such as breast cancer treatment) from Mayo Clinic, falsely suggesting that those regions have people who are disproportionally sick or with better access. The independent variables in the SEC Index are not patient-level variables; they are aggregated to CBG-level as they are aimed to capture CBG-level effects. The SEC Index itself is a binomial propensity score model and the index score is the link-space (linear) prediction from this model.

Applying the SEC Index

To show the effectiveness of the SEC Index, we develop two breast cancer prevalence models on the remaining 70% of the cohort. The first model, our baseline, contained patient measures of age and gender and did not utilize the SEC Index; the second model contained age, gender, and our trained SEC Index. Both models were logistic regression models.

Evaluation

We used concordance as the metric of model performance. For a randomly selected pair of patients, with exactly one of the two having breast cancer, concordance is the probability that the predicted risk of breast cancer is higher for the breast cancer patient than for the one without breast cancer. Concordance is also known as C-statistic or Area under the ROC curve (AUC). 100 replications of bootstrap simulation were used to estimate the model performance for the two clinical models.

In bootstrapping, a simulated data set of the same size as the original is created by sampling the patients of the original data set with replacement. As a result of sampling with replacement, some of the original patients are excluded from the simulated data set and others are included multiple times. The patients excluded are referred to as "out-of-bag" patients and are set aside for validation. In each of the 100 replications, the SEC Index model was constructed (on 30% of the simulated data set) and the two clinical models (one with and one without the SEC Index) were developed on the remaining 70% of the simulated data set as described above. The concordance for the two clinical models was calculated on the out-of-bag (validation) patients. After the 100 replications, we had 100 SEC Index models and 100 pairs of clinical models with 100 pairs of concordance values. A paired t-test was utilized to compare the 100 pairs of concordance measures.

Results

We first present the overall results of the bootstrap simulation. Finally, to offer further insight into our methodology, we also present the SEC Index model and the clinical model of a specific replication.

Bootstrap Simulation

Overall, the model for breast cancer prevalence that included the SEC Index in addition clinical characteristics for age and gender in bootstrap replications (n=100) performed strongly significantly (**$t=5.8457, p=6.489 \times 10^{-8}$**) better than the model without the SEC Index, indicating that the inclusion of the SEC Index is significantly beneficial to the model performance. Further, in 22% of the individual bootstrap replications, the

SEC variable was designated as a statistically significant predictor of breast cancer prevalence.

Table 4.2: Propensity Training Model					
Variable		**Estimate**	**Standard Error**	**Z Test**	**P-value**
Intercept		-1.8026	0.0225	-79.975	< 2e-16
Poverty		0.0943	0.0381	2.478	0.0132
T-score					
	V1	-0.1303	0.0284	-4.595	4.32e-06
	V2	0.1914	0.0643	2.978	0.0029
Coverage					
	33 to 50% (vs. <33%)	-0.1230	0.0305	-4.034	5.48e-05
	50 to 100% (vs. <33%)	-0.2765	0.0254	10.883	< 2e-16
D-Score		0.1971	0.0316	6.248	4.16E-10

Specific Example

We randomly chose 1 of the 22 bootstrap models where the trained SEC Index demonstrated a significant impact on breast cancer prevalence measures **(Table 4.2)**. This model included measures for poverty, coverage, T-Score (unemployment and non-car ownership), and D-Score. Census block groups with high social stress SEC demonstrated a significant (p=0.0168) detrimental (Beta=0.2948) effect **(Table 4.3)**.

Table 4.3: Demonstration of SEC				
Variable	**Estimate**	**Standard Error**	**Z Test**	**P-value**
Intercept	-5.7700	0.3562	-16.197	< 2e-16
Age (years), baseline: ≥18 to ≤25				
>25 to ≤35	1.6063	0.2809	5.719	1.07e-08
>35 to ≤45	2.7252	0.2665	10.225	< 2e-16
>45 to ≤55	3.4421	0.2622	13.130	< 2e-16
>55 to ≤65	4.0885	0.2617	15.624	< 2e-16
>65 to ≤75	4.5772	0.2615	17.501	< 2e-16
>75 to ≤85	4.5463	0.2626	17.316	< 2e-16
>85	4.5198	0.2676	16.889	< 2e-16
Male	-2.3522	0.0676	-34.786	< 2e-16
SEC	**0.2948**	0.1233	2.391	**0.0168**

Discussion

We have demonstrated that the use of an SEC Index for social stress significantly increased the performance for prediction of breast cancer prevalence even in our study region located in the Upper Midwest, where differences among CBGs in terms of SEC are relatively modest. We expect the impact of using the SEC Index to amplify when applied to a region where SEC differences among regions are more pronounced. We wish to emphasize that the purpose of this study is neither to recommend the use of specific SEC measures nor to quantify a neighborhood effect, which has been proven theoretically impossible[15]. Rather, our intent is to better identify clinical effect, which our method successfully accomplished.

Strengths and Limitations

Our proposed method offers significant benefits. The most important benefit is that it helps separate the effect of SECs from the effect of clinical variables: without accounting for SEC, we would have only been able to measure the combined effect of the clinical variables and SEC; accounting for SEC helped elucidate the true effect of the clinical variables. SEC measures are so highly correlated with each other that efforts to separate their effect has been deemed fruitless[15]. Including the individual SEC measures into a clinical model would make overfitting inevitable and would limit degrees of freedom, while including the SEC Index contains the collinearity problem of the SEC measures in the SEC Index model. Further, being able to capture the effect of SEC through the patients' generic health (as measured by the Charlson Comorbidity Index) and being able to use it for a specific clinical end point (breast cancer prevalence) enables large-scale generalizability across organizations and coverage areas. The (arguably imperfect) separation between SECs and the clinical variables that the SEC Index affords helps capture the differences in SECs between organizations and coverage areas, leading to more accurate estimates for the clinical effects. Finally, the proposed methodology also allows us to incorporate additional validated or new measures of SEC that may help better separate the impact of SECs and patient characteristics. The SEC measures used in this study are merely a sample of the measures in existence. Alternative measures of interest, for example social contagions, demographic change, and social capital, can be

incorporated into the SEC Index in a straightforward way without causing the clinical model to overfit the data.

Future Work

Despite medicine's rigorous pace of advancement, appropriately capturing SEC patterning of disease remains an important topic. With the advent of harnessing social media data and focus on consumer health informatics it is important to consider the lingering issue of how we can quantify the effect of SEC using relatively stable population-based data through validated measures. Advancing our understanding and utilization of SECs is necessary to advance our understanding of the complex, multifactorial causes of cancer[16].

Conclusion

This study demonstrates a viable approach to account for SECs, including social stress, in retrospective clinical research. An important distinction exists between the utilization of SECs to control for confounding effects in retrospective research and utilizations in population health or clinical decision support, critical considerations remains in how to accurately address the long-standing concerns of social epidemiology in consumer health informatics.

References

1. Ed. Kawachi, I., Berkman, L.F., Neighborhoods and Health. Oxford University Press, 2003.
2. Adkins, D.E., Vaisey, S., Toward a Unified Stratification Theory: Structure, Genome, and Status Across Human Societies. Sociological Theory, 2009. 27(2): p.99-121.
3. Oakes, J.M., The measurement of SES in health research: current practice and steps toward a new approach. Social Science & Medicine, 2003. 56(4): p.769-784.
4. Aneshensel, C.S., Social Stress: Theory and Research. Annual Review of Sociology, 1992. 18: p.15-38
5. McEwen, B.S., Stellar, E., Stress and the Individual. Archives of Internal Medicine, 1993. 153: p.2093-2101.

6. DuBois, D.L., Felner, R.D., Brand, S., Adan, A.M., Evans, E.G., A Prospective Study of Life Stress, Social Support, and Adaptation in Early Adolescence. Child Development, 1992. 63(3): p.542-557.
7. Volden PA, Wonder EL, Skor MN. Chronic Social Isolation is Associated with Metabolic Gene Expression Changes Specific to Mammary Adipose Tissue. Cancer Prevention Research, 2013. 6(7):634-45.
8. Friedman, S., Index of dissimilarity. In V. Parrillo (Ed.), *Encyclopedia of social problems.* Thousand Oaks, CA: SAGE Publications, Inc., 2008. p. 489.
9. Townsend, P., Phillimore, P., Beattie, A. Health and Deprivation: Inequality and the North. Croom Helm, 1988.
10. Chute CG, Beck SA, Fisk TB, Mohr DN. The Enterprise Data Trust at Mayo Clinic: a semantically integrated warehouse of biomedical data. JAMIA. Mar-Apr 2010;17(2):131-135.
11. Charlson, M.E., Pompei, P., Ales, K.L., MacKenzie, C.R., A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. Journal of Chronic Disease, 1987. 40(5): p. 373-383.
12. Gilthrope, M.S., The importance of normalisation in the construction of deprivation indices. Journal of Epidemiology and Community Health, 1995. 49: p.S45-S50.
13. US Census Bureau 2009 Poverty Thresholds. shttp://www.census.gov/hhes/www/poverty/data/threshld/index.html
14. 2010 Census Urban and Rural Classification and Urban Area Criteria. https://www.census.gov/geo/reference/ua/urban-rural-2010.hf

JM. The (mis)estimation of neighborhood effects: causal inference for a practicable social epidemiology. Social Science & Medicine, 2004. 58: p.1929-1952.

15. Lynch, S.M., Rebbeck, T.R., Bridging the Gap between Biologic, Individual, and Macroenvironmental Factors in Cancer: A Multilevel Approach. Cancer Epidemiology, Biomarkers & Prevention, 2013. 22(4): p.485-495.

Part #3: Leveraging the Electronic Health Record(EHR)-Linked Biorepository in Translational Bioinformatics

While metformin is generally well tolerated approximately 30% of patients do not respond to metformin with some patients having adverse reactions. The pharmacogenomics of metformin are not clearly understood; it is imperative to understand molecular mechanisms of metformin further. In this part I sought to explore genetic variation that is responsible for glycemic response to metformin.

Chapter 5: The Impact of Genomic Variation on Glycemic Response to Metformin Using EHR-Linked Biorepository Data.

Abstract

Metformin is a first-line antihyperglycemic agent commonly prescribed in type 2 diabetes mellitus (T2DM), but whose pharmacogenomics are not clearly understood. Due to the epidemic growth of T2DM in the US and the accumulating evidence highlighting potential repurposing of metformin for cancer prevention and treatment it is imperative to understand molecular mechanisms of metformin further. In this pharmacogenomics study we seek to identify potential personalized medicine targets that are associated with response to metformin treatment. Specifically, we seek to elucidate the impact of key genetic variants that modify glycemic response to metformin treatment. Seventeen genes with suspected pharmacokinetic or pharmacodynamic implications were selected based on systematic review for study. Our analysis cohort consisted of 258 T2DM patients who had new metformin exposure, existing genomic data, and longitudinal electronic health records. Change in glycemic response to metformin exposure via A1c measures pre and post metformin exposure serve as the outcome of interest. After quality control, gene-level and SNP-level analysis were conducted on 17 candidate genes and 463 SNPs within those genes was performed, controlling for key covariates of sex, age, and body mass index (BMI) at index metformin exposure. PRKAB2, the gene encoding the beta subunit 2 of adenosine monophosphate-activated protein kinase complex, is associated with significant (p=0.0194) change in glycemic response after exposure to metformin. SLC29A4, the gene encoding the plasma membrane monoamine transporter expressed in the intestine, the next most significant gene (p=0.0614), has potential to also be associated with glycemic response after exposure to metformin. This study serves as an example of how EHR-linked biorepositories can be used to test pharmacogenomics hypotheses using intermediary clinical phenotypes.

1. Background

Metformin is recommended as a first-line therapy for type 2 diabetes mellitus (T2DM)[1] and is believed to be the most prescribed drug worldwide[2]. Evidence is also

accumulating that highlights the potential repurposing of metformin for cancer prevention and treatment[3]. However, the details underlying the molecular mechanism of action for metformin are not fully understood[2]. It is imperative to understand the molecular mechanisms of metformin further, particularly genetic variation in clinically relevant targets of metformin [4].

Metformin is in the biguanides class of medications and is primarily utilized to regain glycemic control in diabetic or pre-diabetic patients. Metformin is a relatively safe antidiabetic therapy[5]. However, serious adverse reactions can occur[6]. The pharmacokinetics (PK) of metformin, the transportation throughout the body, are moderately understood[7]. Metformin is not metabolized, with absorption of metformin known to occur in the small and large intestines[5]. Uptake of metformin from the blood is known to occur in the kidneys and liver[2], but can be reasonably assumed to occur in any tissue with abundance of organic cation transporters (OCT)[8]. Eventually metformin is excreted unchanged in the urine[5]. However, the pharmacodynamics (PD) of metformin, the physiological and biochemical impact of metformin in the body, are not clearly understood[7]. Metformin works primarily by inhibiting hepatic glucose production by reducing gluconeogenesis in the liver[9] and is also known to reduce intestinal glucose absorption[10]. Further, metformin appears to improve glucose uptake and utilization systemically[7]. However, there is considerable variation in glycemic response to metformin[7]. While genetic factors may partially explain glycemic response to metformin, further studies are needed to understand the impact of variation in key transporter genes on glycemic response in clinical populations[2].

Our study aims to add clarity to metformin pharmacogenomics by understanding the impact of common variants in metformin candidate genes (n=17) on altered glycemic response in a clinical population. Candidate genes selected for inclusion in this study are suspected metformin PK or PD determinants as designated in systematic reviews of metformin pharmacogenomics[2,5,7,8,11]. Gene-level and SNP-level analyses were performed in this study to identify genes significantly associated with change in glycemic response after exposure to metformin and directionality of the effect of corresponding SNPs.

2. Materials and Methods

In order to elucidate a pharmacogenomic perspective on glycemic response to metformin, all patients selected for inclusion into this study have known exposure to metformin and T2DM (n=258) with corresponding glycemic cross sections.

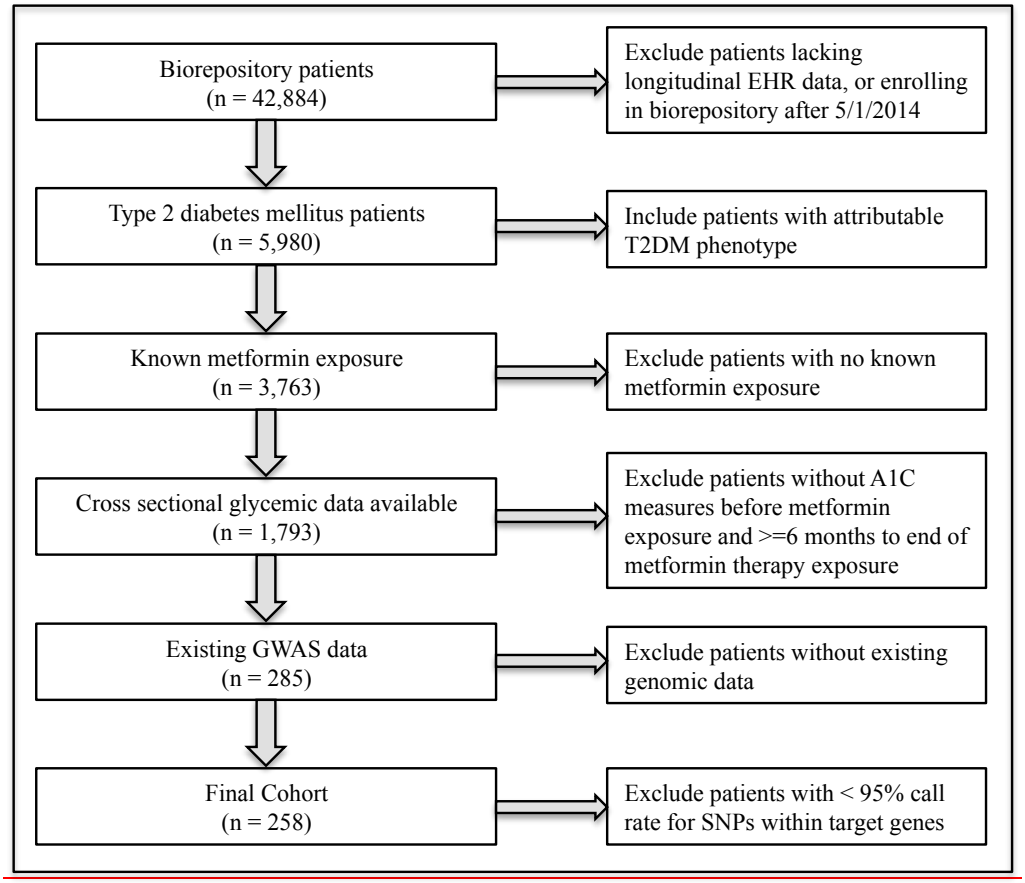
2.1. Materials

All genomic data utilized in this study was a part of a local biorepository with linked longitudinal electronic health record data[12]. Patients with existing GWAS data, T2DM, known metformin exposure, and longitudinal glycemic indicators were included in this study. Patients without known metformin exposure >= 6 months, without A1c measures >=6 months apart prior to and concurrent with the metformin exposure, and without genomic data with >=95% call rate within candidate genes were excluded from this study. Clinical phenotypes were developed using EHR-based algorithms and data. A final cohort of Caucasian patients (n=258) with known metformin exposure, T2DM, longitudinal A1c measures, and genomic data with >= 95% call rate was utilized in this study. A description of the cohort is found in **Table 5.1** and a detailed diagram of cohort development is found in **Figure 5.1**.

Table 5.1: Cohort Demographics

	n	**(%)**	
Female	89	(34.5)	
Male	169	(64.5)	
BMI < 30	64	(24.8)	
BMI >= 30 to < 35	100	(38.8)	
BMI >=35	93	(36.1)	
Median A1c > 7.0	101	(39.1)	
	Median	**Range**	
Change in A1c	0.0733	-6.45	3.51
Age	64	30	84

Figure 5.1: Cohort Development



2.2. Methods

17 candidate genes suspected to be determinants of metformin PK or PD were selected for inclusion in this study based on suspected relevance in recent systematic reviews[2,5,7,8,11] (**Table 5.2**).

Attribution of a T2DM phenotype was performed using modified methodology developed by eMERGE [13]. Metformin exposure was ascertained using a combination of validated structured[14] and semi-structured[15] EHR data collection methodologies. To compare the genetic modification of glycemic response to metformin, measures of A1c were compared prior to metformin exposure and during the period of metformin exposure following a 6-month period of delay. In this study, A1c was calculated as the difference between the average of A1c measures within 6 months prior to metformin exposure and the average of A1c measures between >= 6 months after exposure to

metformin and the end of metformin exposure. This approach minimizes the impact of any one A1c measure and biases change in A1c measures towards the null. Age, gender, and morbid obesity (BMI >= 35), a known modifier of T2DM state[1], were selected for inclusion as covariates in the model. Age and BMI measures were calculated at first recorded exposure to metformin.

Table 5.2: Candidate Genes

PGX	Gene Name	Chromo-some	SNP counts	Protein Name	Protein Full Name
PK	SLC22A1	6	38	OCT1	organic cation transporter 1
PK	SLC22A2	6	31	OCT2	organic cation transporter 2
PK	SLC22A3	6	33	OCT3	organic cation transporter 3
PK	SLC29A4	7	13	PMAT	plasma membrane monoamine transporter
PK	SLC47A1	17	19	MATE1	multidrug and toxin extrusion protein 1
PK	SLC47A2	17	20	MATE2-K	multidrug and toxin extrusion protein 2
PD	Adenosine monophosphate-activated protein kinase (AMPK) complex				
	PRKAA1	5	19	AMPK-A1	AMPK-alpha subunit 1
	PRKAA2	1	24	AMPK-A2	AMPK-alpha subunit 2
	PRKAB1	12	13	AMPK-B1	AMPK-beta subunit 1
	PRKAB2	1	35	AMPK-B2	AMPK-beta subunit 2
	PRKAG1	12	8	AMPK-G1	AMPK-gamma subunit 1
	PRKAG2	7	132	AMPK-G2	AMPK-gamma subunit 2
	PRKAG3	2	9	AMPK-G3	AMPK-gamma subunit 3
PD	STK11	19	10	LKB1	liver kinase B
PD	PPARG	3	42	PPARG	peroxisome proliferator-activate receptor gamma
PD	ATM	11	13	ATM	ataxia telangiectasia mutated serine/threonine protein kinase
PD	GCKR	2	14	GCKR	glucokinase regulatory protein

PGX = pharmacogenomic implication, PK = Pharmacokinetics,
PD = Pharmacodynamics

2.2.1. Quality Control

For the 17 candidate genes, we selected SNPs 50 kb upstream and downstream of each gene using 1000 genomes project variants and NCBI build 37 as the reference genome. By this mapping rule, a total of 8440 SNPs were mapped to the 17 genes, but only 1065 SNPs were available in the genotype data. For the remaining SNPs, two main quality control filters were applied: (i) SNPs with unacceptable high rates of missing

genotype calls (>10%); and (ii) monomorphic SNPs were excluded. The quality control of the genotype data was performed by PLINK v 1.07[16]. The call rate was < 0.90 for 601 SNPs and 1 SNP was monomorphic, leaving 463 SNPs for the single SNP and gene-level analyses. From the total of 285 samples with available genotype, we excluded 27 samples with call rate < 0.95, leaving 258 samples available for analysis in the final cohort.

2.2.2. Gene-Level Analysis Method

We analyzed the association of each gene with the change in a1c using Van der Waerden rank, or rank based inverse Gaussian, transformed change in A1c. Gene level tests were performed using principal component analysis (PCA) as described previously[17]. For each gene, principal components were created using linear combinations of ordinally scaled SNPs (i.e., 0, 1, 2 copies of minor allele) and the smallest set of resulting principal components that explained at least 90% of the SNP variance was included in linear regression models. Instead of including the entire set of SNPs for each gene, the principal component approach reduces the degrees of freedom, avoids model fitting issues due to multi-collinearity of the SNPs from linkage disequilibrium (LD) and potentially improves the statistical power. Finally, to assess overall significance of a gene, we computed the likelihood ratio test (LRT) by comparing the null model containing only the covariates with the full model containing covariates and the set of resulting principal components. At the gene-level, results of the 17 simultaneous hypothesis tests were utilized for filtering. Plots of LD displaying r2 for each gene based on 258 Caucasian samples were created using Haploview v 4.2. The statistical package R 2.15.0 was used for the gene-level analysis[18].

2.2.3. SNP-Level Analysis Methods

We tested the association between each SNP and Van der Waerden rank transformed change in A1c using a linear regression model, adjusting for age, gender and morbid obesity. Coefficient estimates were calculated per minor allele, that is, with each minor allele, the A1c level changes by ‘beta’. SNP-level results were not corrected for multiple testing.

3. Results

A series of analyses were performed to elucidate the potential modifying impact of common variants in candidate genes on glycemic response to metformin. First, a gene-level analysis was performed to identify genes significantly associated with change in glycemic response after exposure to metformin (Section 3.1). Since this gene-level analysis was based on principal components having undetermined signs they are unable to indicate directionality, requiring additional SNP-level analysis. The SNP-level analysis was performed on significant and marginally significant candidate genes identified in Section 3.1 to confirm and determine directionality of these associations (Section 3.2). Finally, an exploratory SNP-level analysis was conducted to identify potential SNP-level associations in non-significant candidate genes (Section 3.3).

3.1. Gene-Level Results

The estimates for male (Coef=0.04,P-val=0.737), age (Coef=-0.0009,P-val=0.881), morbid obesity (Coef=0.22,P-val=0.083) were not significantly associated with change in A1c at alpha=0.05 significance level. In the multivariate model, none of the covariates were associated with change in A1c at alpha=0.05 significance level.

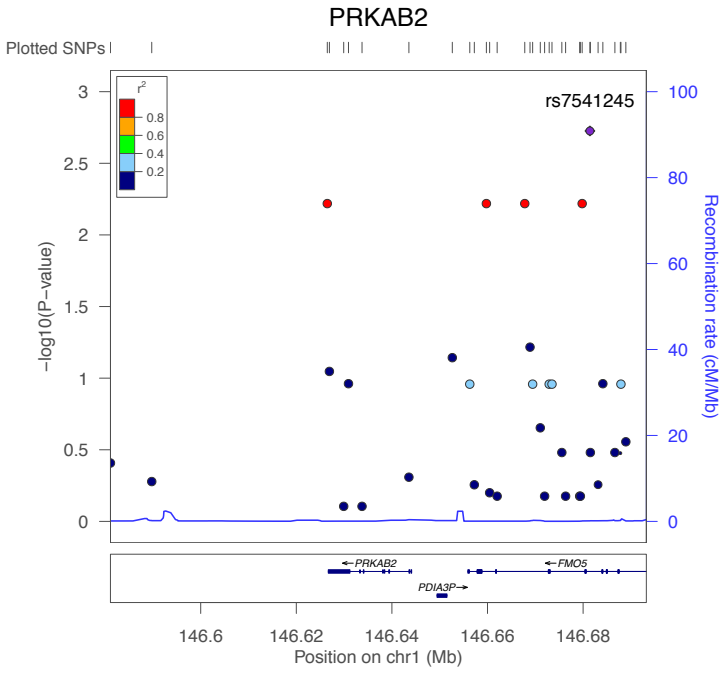
Table 5.3: Principal Component (PC) Analysis

Gene Name	nSNPs	nPCs	P-value
PRKAB2	**35**	**5**	**0.0194**
SLC29A4	13	9	0.0614
PRKAG1	8	4	0.1548
SLC47A1	19	7	0.2121
GCKR	14	4	0.2365
SLC47A2	20	5	0.3237
STK11	10	8	0.4450
PRKAA1	19	6	0.5104
ATM	13	5	0.5256
PRKAA2	24	12	0.5439
SLC22A3	33	6	0.6017
PPARG	42	13	0.7499
PRKAB1	13	4	0.7665
PRKAG3	9	6	0.8113
PRKAG2	132	46	0.8292
SLC22A1	38	12	0.8487
SLC22A2	31	8	0.9086

After controlling for age, gender, and morbid obesity the PCs for the 17 candidate genes capturing 90% of the summed variation in SNPs identified 1 candidate gene as being significantly associated with glycemic response. PRKAB2, the beta subunit 2 of adenosine monophosphate-activated protein kinase complex, represented by 5 PCs of 35 SNPs was marginally significantly associated (**p=0.0194**) with change in A1c (**Table 5.3**). The next most significant (p=0.0614) gene was SLC29A4, the gene coding for plasma membrane monoamine transporter (PMAT). The Locus Zoom plots (**Figure 5.2**) identified some linkage disequilibrium for 4 SNPS in PRKAB2 but were far from the most significant SNP (rs7541245,p=0.0019), the SNP of interest, based on the 1000 Genomes European reference population from March 2012 release. Blocks of linkage disequilibrium (**Figure 5.3**) were designated within the PRKAB2 and SLC29A4 genes. Gene-level analysis does not give a direction of the associations; SNPs within a gene can have both negative and positive associations with the outcome. An alternative SNP-level analysis was performed to gain insight in to the directionality of these associations.

Figure 5.2: Locus Zoom plots for select candidate genes.

PRKAB2-top, SLC29A4-lower



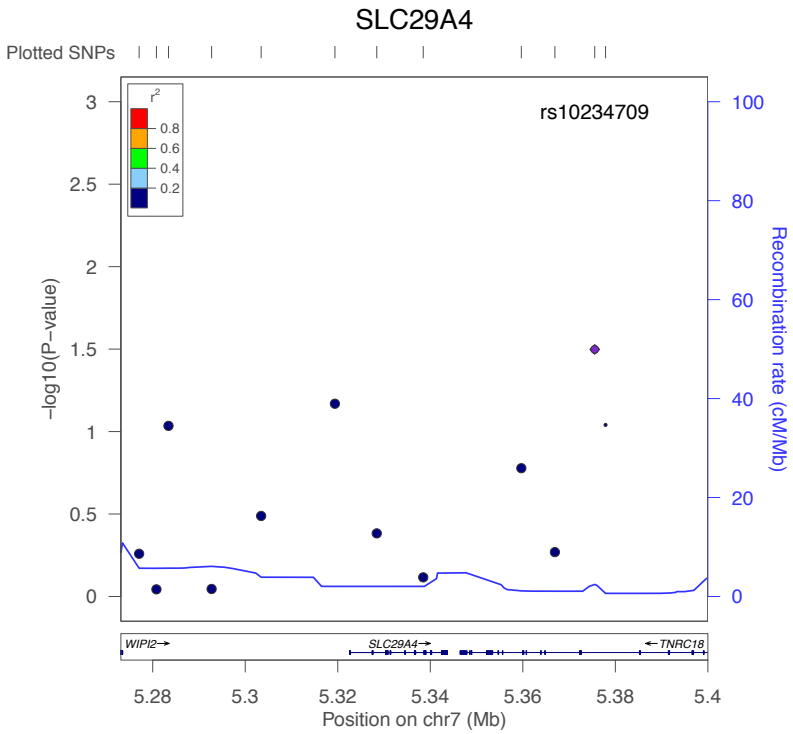
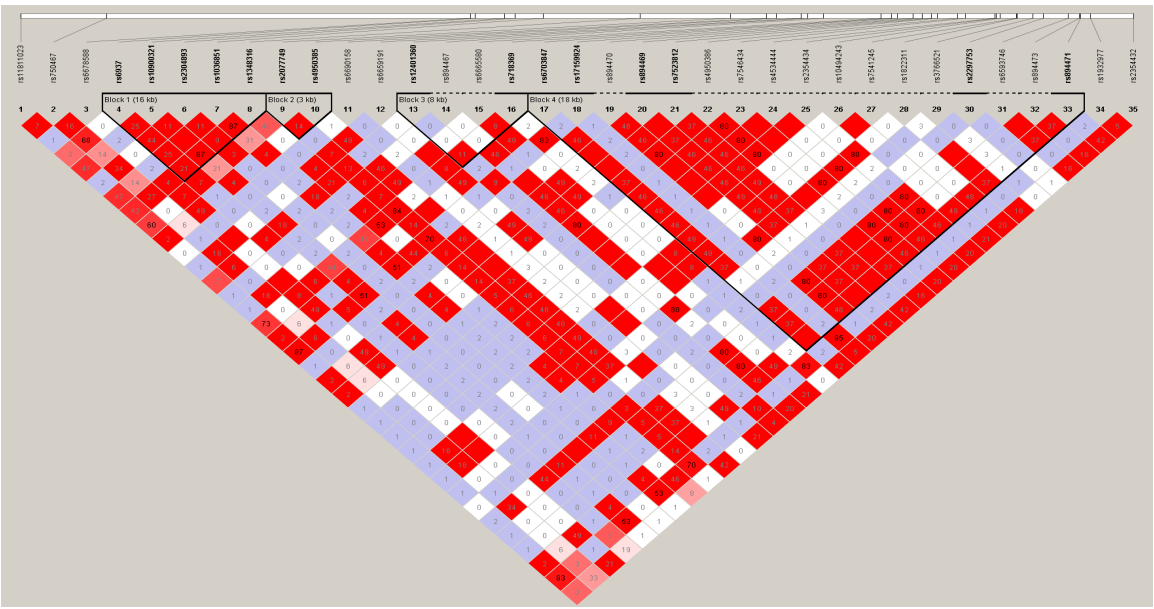
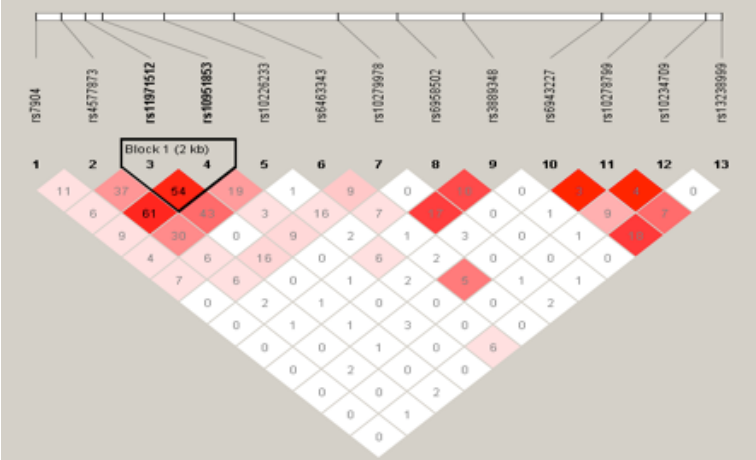


Figure 5.3: Linkage Disequilibrium (LD) blocks. The LD values as measured using r2 are given by numbers and the LD values as measured by D' are shown by color intensity (red squares indicate strong LD, pink squares indicate intermediate LD, and white squares indicate low LD, with evidence for ancestral recombination; blue indicates limited data).

PRKAB2



SLC29A4



3.2. SNP-Level Results of Two Most Significant Candidate Genes

48 SNPs in 2 significant or marginally significant candidate genes were evaluated for directionality. 5 SNPs (rs6665580, rs6659191, rs6678588, rs7541245, rs10494243) in high LD within PRKAB2, a PD determinant, were found to be significantly associated with a decrease in glycemic response after metformin exposure. 1 SNP (rs10234709) in SLC29A4, a PK determinant, was found to be significantly associated with a decrease in glycemic response after metformin exposure. Detailed SNP-level associations can be found in **Table 5.4**.

3.3. SNP-Level Results of Remaining Candidate Genes

Adjusted SNP-level analysis (**Table 5.5**) of non-significant candidate genes found variation in SNPs (n=425) for the genes PRKAA1 (1), PRKAG2 (3), SLC22A3 (2), SLC47A1 (4), SLC47A2 (2), and STK11 (1) to be significantly associated with change in glycemic response after metformin exposure. However, gene-level analysis did not identify significant variation within these genes and after applying Bonferroni correction for multiple testing no SNPs could be considered significant with the lowered significance threshold ($p<1.1765E-4$).

Table 5.4: SNP-level Analysis of Significant or Marginally Significant Candidate Genes

PRKAB2							PRKAB2 (cont.)						
SNP	**Minor Allele**	**Major Allele**	**MAF**	**BETA**	**95% CIs**	**P-value**	**SNP**	**Minor Allele**	**Major Allele**	**MAF**	**BETA**	**95% CIs**	**P-value**
rs7541245	**A**	**C**	**0.0311**	**-0.7885**	**(-1.28;-0.297)**	**0.0019**	rs2354432	C	T	0.1376	-0.0685	(-0.317;0.180)	0.5894
rs6678588	C	A	0.0330	-0.6809	(-1.163;-0.199)	0.0060	rs12401360	G	A	0.2287	0.0485	(-0.149;0.246)	0.6310
rs6659191	C	T	0.0330	-0.6809	(-1.163;-0.199)	0.0060	rs894467	C	T	0.0310	0.1036	(-0.368;0.575)	0.6667
rs6665580	**G**	**A**	**0.0330**	**-0.6809**	**(-1.163;-0.199)**	**0.0060**	rs894470	A	G	0.0310	0.1036	(-0.368;0.575)	0.6667
rs10494243	**T**	**C**	**0.0330**	**-0.6809**	**(-1.163;-0.199)**	**0.0060**	rs7546434	A	G	0.0310	0.1036	(-0.368;0.575)	0.6667
rs718369	C	T	0.2810	-0.1834	(-0.374;0.007)	0.0607	rs4534444	T	C	0.0310	0.1036	(-0.368;0.575)	0.6667
rs2077749	A	G	0.3159	-0.1631	(-0.34;0.014)	0.0719	rs2354434	A	G	0.0310	0.1036	(-0.368;0.575)	0.6667
rs6937	C	T	0.1376	0.2096	(-0.032;0.451)	0.0897	rs10900321	T	C	0.3895	0.0245	(-0.151;0.2)	0.7849
rs2304893	C	T	0.0659	0.2813	(-0.062;0.624)	0.1093	rs1036851	C	T	0.3895	0.0245	(-0.151;0.2)	0.7849
rs2297753	A	G	0.0659	0.2813	(-0.062;0.624)	0.1093	**SLC29A4**						
rs4950385	A	G	0.0640	-0.2787	(-0.619;0.062)	0.1101	**SNP**	**Minor Allele**	**Major Allele**	**MAF**	**BETA**	**95% CIs**	**P-value**
rs6703847	G	A	0.0640	-0.2787	(-0.619;0.062)								
rs894469	C	T	0.0640	-0.2787	(-0.619;0.062)	0.1101	**rs10234709**	**A**	**C**	**0.3852**	**-0.1987**	**(-0.379;-0.018)**	**0.0318**
rs7523812	G	A	0.0640	-0.2787	(-0.619;0.062)	0.1101	rs10279978	A	G	0.2965	0.1878	(-0.013;0.388)	0.0678
rs894471	G	A	0.0640	-0.2787	(-0.619;0.062)	0.1101	rs13238999	A	C	0.3839	-0.1473	(-0.317;0.023)	0.0911
rs17159924	G	A	0.2461	-0.1234	(-0.321;0.074)	0.1101	rs10951853	T	C	0.4593	-0.1406	(-0.304;0.022)	0.0923
rs1932977	C	T	0.2539	-0.1064	(-0.298;0.085)	0.2222	rs6943227	G	T	0.2829	0.1341	(-0.055;0.324)	0.1668
rs4950386	A	G	0.0252	0.2541	(-0.257;0.765)	0.2781	rs7904	T	C	0.3547	-0.1124	(-0.292;0.067)	0.2207
rs1822311	T	C	0.0252	0.2541	(-0.257;0.765)	0.3305	rs6463343	A	G	0.4360	-0.0896	(-0.268;0.088)	0.3249
rs6593746	G	A	0.0252	0.2541	(-0.257;0.765)	0.3305	rs6958502	A	G	0.2074	-0.0888	(-0.302;0.124)	0.4143
rs894473	A	G	0.0261	0.2553	(-0.262;0.772)	0.3305	rs10278799	T	C	0.0717	0.1049	(-0.229;0.438)	0.5383
rs11811023	C	T	0.2713	-0.0843	(-0.277;0.108)	0.3339	rs4577873	A	G	0.4554	0.0507	(-0.116;0.217)	0.5508
rs1348316	G	A	0.4225	-0.0602	(-0.231;0.111)	0.3912	rs3889348	A	G	0.3643	0.0271	(-0.151;0.205)	0.7656
rs750467	T	C	0.1686	0.0709	(-0.148;0.29)	0.4913	rs10226233	T	C	0.3488	-0.0109	(-0.181;0.159)	0.9003
rs3766521	C	A	0.0039	0.4161	(-0.960;1.792)	0.5270	rs11971512	G	A	0.4147	0.0103	(-0.159;0.180)	0.9052
rs6690158	T	C	0.0039	0.4143	(-0.96;1.788)	0.5550	MAF = minor allele frequency						

Table 5.5: SNP-level Analysis of Non-significant Candidate Genes

SNP	MA	BETA	P-value	Gene Name
rs4725434	T	0.2211	0.0136	PRKAG2
rs3127602	T	0.2307	0.0147	SLC22A3
rs3123629	A	0.2266	0.0168	SLC22A3
rs12539356	T	0.2312	0.0211	PRKAG2
rs2120274	A	0.1938	0.031	SLC47A1
rs1534665	C	-0.1942	0.0318	PRKAG2
rs2301759	C	0.2203	0.0345	STK11
rs2453586	A	0.1849	0.0385	SLC47A1
rs4621031	C	-0.1908	0.041	SLC47A2
rs962801	T	-0.1909	0.0414	SLC47A2
rs11749180	A	-0.208	0.0464	PRKAA1
rs2245639	C	0.1755	0.0493	SLC47A1
rs2165895	A	0.1755	0.0493	SLC47A1
MA = Minor Allele				

4. Discussion

The primary purpose of this study was to add clarity to metformin pharmacogenomics by understanding the impact of common variants in metformin candidate genes (n=17) on altered glycemic response in a clinical population derived from an EHR-linked biorepository. All candidate genes were selected due to a suspected role in metformin PK/PD. Gene-level and SNP-level variants were found to be associated with decreased glycemic response to metformin.

4.1. Interpretation of associations

Gene-level analysis found variation in the PD candidate PRKAB2 gene to be significantly associated with change in glycemic response after exposure to metformin. SLC29A4 was found marginally significant (P-value=0.0614) in gene-level analysis. SNP-level analysis confirmed variations in both PRKAB2 and SLC29A4 genes to be associated with decreased glycemic response. While the additional SNP-level analysis for non-significant gene-level associations identified significant SNPs within the PRKAA1, PRKAG2, SLC22A3, SLC47A1, SLC47A2, and STK11, these SNPs are no longer significant after adjusting for multiple testing and cannot be meaningfully interpreted.

4.2. Molecular significance of associations

AMPK is a heterotrimeric enzyme composed of alpha, beta, and gamma subunits, encoded by 5 genes, each of which uniquely determines protein stability and activity. AMPK acts as a metabolic master switch regulating several intracellular systems and plays an important role in cellular energy homeostasis, the maintenance of cellular ATP levels[19]. Metformin is a known AMPK activator, with genetic variations in AMPK suspected to impact the response to metformin[11]. In this study we found variations in PRKAB2 (AMPK-B2), a PD determinant, to be associated with decreased change in glycemic response to metformin. While some SNPs in PRKAA1 and PRKAG2 subunit coding genes were potentially implicated in impacting change in glycemic response, an association was not identified in the gene-level analysis. Despite suspected AMPK involvement in response to metformin, the specific targets of metformin are not clearly understood. In this study we identified 3 LD blocks within the PRKAB2 gene that

deserve further investigation for developing personalized metformin therapy considerations.

The SLC29A4 gene, encoding for the PK determinant PMAT, is the primary source of metformin absorption in the intestine[2], having systemic implications for metformin availability via the bloodstream. While gene-level analysis identified a marginally significant association, the locus zoom (Figure 5.2) and LD plots (Figure 5.3) identified low linkage disequilibrium across the gene. SNP-level analysis points to the identified variant (rs10234709) as being associated with decreased glycemic response to metformin. Of note is the exploratory SNP-level analysis that identified covariate adjusted significant associations for SNP variants (rs3127602 and rs3123629) in the SLC22A3 gene with increases in glycemic response to metformin. SLC22A3 is the gene that encodes for OCT3, which is also known to be responsible for intestinal absorption of metformin and is also implicated in metformin uptake in the liver from the blood. Finally OCT1, encoded by SLC22A1, is responsible for uptake of metformin into the blood from the intestine and has been previously found to be associated with glycemic response[20]; our study found no such association. While we find some evidence pointing towards the importance of intestinal absorption of metformin in modifying glycemic response, the evidence is not clear.

In addition to the filtered gene-level analysis, our exploratory SNP-level analysis of non-significant candidate genes identified some findings of note. SNP-level variation in SLC47A1, a gene that encodes for MATE1, a PK determinant that is expressed in the liver, appears to be associated with increased glycemic response to metformin. While this evidence only hints at a potential association, variation in this transporter is clinically relevant due to potential drug-drug interactions that can occur in liver with transporter variation[7], further study is needed. Regarding PD determinants, kinases such as ATM[21] or STK11[22] are thought modulate AMPK activity by metformin[23]. While we able to identify rs2301759 as a significant (beta=0.2203,p=0.0345) covariate adjusted SNP within the STK11 gene, it is important to note that Bonferroni correction negates significance of this association. We were unable to identify a significant association between ATM variation and glycemic response to metformin with the lowest covariate

adjusted SNP (rs1800058) having a p-value of 0.1546.

4.3. Strengths and Limitations

In this study, we leverage EHR-linked biorepository data and EHR-based phenotyping methods to study variants associated with Metformin PD and PK. While our cohort is modestly powered (N=258), we posit that utilizing a clinical endpoint that is sensitive to PD and potentially PK determinants strengthens our study. However, relying solely on the endpoint of glycemic response has potential to bias PK associations towards the null as PK alterations might only indirectly impact glycemic response. In particular, due to the wide distribution and systemic impact of metformin the distinction between PK and PD determinants may not necessarily be distinct at the clinical level. Further, while the genetic contribution to variation in metformin renal clearance in the kidneys is estimated to be approximately 90%[24], bioavailability and concentration of metformin likely will very widely based on renal function, topics that will be addressed in our future work.

Like most existing pharmacogenomic studies of metformin we have utilized a candidate gene approach that focuses PK implications on transporters[2]. However, our study includes PD determinants, which are not as thoroughly understood, to increase our understanding of metformin pharmacogenomics. Further, not all SNPs within candidate genes were available for analysis due to GWAS sequencing being originally performed for other studies. The secondary nature of the GWAS data has potential to bias findings either due to original patient selection criteria or sequencing criteria. Finally, while an agnostic training step was not utilized to identify candidate genes, our selection of candidate genes draws strength from the large body of literature on which it is based. Our work seeks to clarify the associations and importance of genes speculated in the literature as being PK or PD determinants of metformin.

4.4. Implications and prospects for personalized medicine

Advances in pharmacogenomics are needed to understand the relationship between genetic variation in key proteins, like those included in this study, and PD implications[11]. Further, with the potential repurposing of metformin in cancer

prevention and treatment it is important to understand the pharmacogenomics of metformin in both T2DM and cancer. While the development of personalized metformin therapy benefits from understanding pharmacogenomic associations such as these, further insight into the mechanism of metformin action is needed. Further, the utilization of multi-omics approaches might be considered to identify additional metformin targets [4,23]. Additional information clarifying further metformin pharmacogenomics will make it possible to develop personalized metformin therapies. In this study we identified potential biomarkers that alter the clinically relevant outcome of glycemic response to metformin.

5. Conclusion

PRKAB2, the gene encoding the beta subunit 2 of adenosine monophosphate-activated protein kinase complex, appears to be associated with decreases in glycemic response after exposure to metformin, with rs7541245 having the strongest SNP association. SLC29A4, the next most significant gene, with rs10234709 having the strongest SNP association, has potential to also be associated with decreases in glycemic response after exposure to metformin. In this study we were able to replicate a metformin PD determinant and potentially a metformin PK determinant using an intermediary phenotype powered by EHR-linked biorepository data.

References

1. Inzucchi SE, Bergenstal RM, Buse JB, et al. Management of Hyperglycemia in Type 2 Diabetes: A Patient-Centered Approach. Diabetes Care. 2012;35:1364-1379.

2. Todd JN, Florez JC. An update on the pharmacogenomics of metformin: progress, problems and potential. Pharmacogenomics. 2014;15(4):529-539.

3. Franciosi M, Lucisano G, Lapice E, Strippoli GF, Pellegrini F, Nicolucci A. Metformin therapy and risk of cancer in patients with type 2 diabetes: systematic review. PloS one. 2013;8(8):e71583.

4. Wang L, Weinshilboum R. Metformin pharmacogenomics: biomarkers to mechanisms. Diabetes. Aug 2014;63(8):2609-2610.

5. Graham G.C., Punt J, Arora M, et al. Clinical Pharmacokinetics of Metformin.

Clinical Pharmacokinetics. 2011;50(2):81-98.

6. Bailey CJ, Path MRC, Turner RC. Metformin. The New England Journal of Medicine. 1996;334(9):574-579.

7. Gong L, Goswami S, Giacomini KM, Altman RB, Klein TE. Metformin pathways: pharmacokinetics and pharmacodynamics. Pharmacogenetics and genomics. Nov 2012;22(11):820-827.

8. Viollet B, Guigas B, Sanz Garcia N, Leclerc J, Foretz M, Andreelli F. Cellular and molecular mechanisms of metformin: an overview. Clinical science. Mar 2012;122(6):253-270.

9. Hundal RS, Krssak M, Dufour S, et al. Mechanism by Which Metformin Reduces Glucose

Production in Type 2 Diabetes. Diabetes. 2000;49.

10. Sakar Y, Meddah B, Faouzi MYA, Cherrah Y, Bado A, Ducroc R. Metformin-Induced Regulation of Intestinal D-Glucose Transporters. Journal of Physiology and Pharmacology. 2010;61(3):301-307.

11. Chen S, Zhou J, Xi M, et al. Pharmacogenetic Variation and Metformin Response. Current Drug Metabolism. 2013;14:1070-1082.

12. Bielinski SJ, Chai HS, Pathak J, et al. Mayo Genome Consortia: A Genotype-Phenotype Resource for Genome-Wide Association Studies With an Application to the Analysis of Circulating Bilirubin Levels. Mayo Clinic proceedings. 2011;86(7):606-614.

13. Kho AN, Hayes MG, Rasmussen-Torvik L, et al. Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study. Journal of the American Medical Informatics Association : JAMIA. Mar-Apr 2012;19(2):212-218.

14. Chute CG, Beck SA, Fisk TB, Mohr DN. The Enterprise Data Trust at Mayo Clinic: a semantically integrated warehouse of biomedical data. Journal of the American Medical Informatics Association : JAMIA. Mar-Apr 2010;17(2):131-135.

15. Pathak J, Murphy SP, Willaert BN, et al. Using RxNorm and NDF-RT to Classify Medication Data Extracted from Electronic Health Records: Experiences from the Rochester Epidemiology Project. American Medical Informatics Association Annual

Symposium. 2011:1089-1098.

16. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. American journal of human genetics. Sep 2007;81(3):559-575.

17. Gauderman WJ, Murcray C, Gilliland F, Conti DV. Testing association between disease and multiple SNPs in a candidate gene. Genetic epidemiology. Jul 2007;31(5):383-395.

18. Team RDC. R: A language and environment for statistical computing. . Vienna, Austria: R Foundation for Statistical Computing; 2012.

19. Winder WW, Hardie DG. AMP-activated protein kinase, a metabolic master switch: possible roles in Type 2 diabetes. American Journal of Physiology. 1999;277(1).

20. Christensen MM, Brasch-Andersen C, Green H, et al. The pharmacogenetics of metformin and its impact on plasma metformin steady-state levels and glycosylated hemoglobin A1c. Pharmacogenetics and genomics. Dec 2011;21(12):837-850.

21. Zhou K, Bellenguez C, Spencer CC, et al. Common variants near ATM are associated with glycemic response to metformin in type 2 diabetes. Nature Genetics. 2011;43(2):117-120.

22. Jablonski KA, McAteer JB, de Bakker PIW, et al. Common Variants in 40 Genes Assessed for Diabetes Incidence and Response to Metformin and Lifestyle Intervention in the Diabetes Prevention Program. Diabetes. 2010;59:2672-2681.

23. Pawlyk AC, Giacomini KM, McKeon C, Shuldiner AR, Florez JC. Metformin Phramacogenomics: Current Status and Future Directions. Diabetes. 2014;63:2590-2599.

24. Leabman MK, Giacomini KM. Estimating the contribution of genes and environment to variation in renal drug clearance. Pharmacogenetics. Sep 2003;13(9):581-584.

Chapter 6: Leveraging an Electronic Health Record-Linked Biorepository to Generate a Metformin Pharmacogenomics Hypothesis

Abstract

Metformin is a first-line antihyperglycemic agent commonly prescribed in type 2 diabetes mellitus (T2DM), but whose pharmacogenomics are not clearly understood. Further, due to accumulating evidence highlighting the potential for metformin in cancer prevention and treatment efforts it is imperative to understand molecular mechanisms of metformin. In this electronic health record(EHR)-based study we explore the potential association of the flavin-containing monooxygenase(FMO)-5 gene, a biologically plausible biotransformer of metformin, and modifying glycemic response to metformin treatment. Using a cohort of 258 T2DM patients who had new metformin exposure, existing genetic data, and longitudinal electronic health records, we compared genetic variation within FMO5 to change in glycemic response. Gene-level and SNP-level analysis identified marginally significant associations for FMO5 variation, representing an EHR-driven pharmacogenetics hypothesis for a potential novel mechanism for metformin biotransformation. However, functional validation of this EHR-based hypothesis is necessary to ascertain its clinical and biological significance.

Introduction

Metformin is a first-line antihyperglycemic agent commonly prescribed for type 2 diabetes mellitus (T2DM) patients[1], whose pharmacogenomics are not clearly understood[2], but are thought to be absent of biotransformation[3]. Further, glycemic response to metformin is variable[3] and serious adverse reactions to metformin have been known to occur[4]. Due to increasing evidence highlighting the potential for metformin in cancer prevention and treatment, it is imperative to understand molecular mechanisms of metformin further.

Background

Metformin is primarily utilized to regain glycemic control in diabetic or pre-diabetic patients. Metformin is a relatively safe antidiabetic therapy[5]. However, serious

adverse reactions can occur[4] and there is considerable variation in glycemic response to metformin, with ~30% of patients unable to achieve glycemic control with metformin[3]. While genetic factors may partially explain clinical glycemic response to metformin due to pharmacokinetic(PK) determinants[3], the transportation throughout the body variation, the identification and impact of metformin pharmacodynamic(PD) determinants, the physiological and biochemical impact of metformin in the body, remains uncertain[2]. Regarding PKs, Metformin is thought to not be metabolized[3], with absorption of metformin known to occur in the small and large intestines[5]. Uptake of metformin from the blood is known to occur in the kidneys and liver[2], but can be reasonably assumed to occur in any tissue with abundance of organic cation transporters (OCT). Eventually metformin is excreted unchanged in the urine[5]. Regarding PDs, metformin works primarily by inhibiting hepatic glucose production by reducing gluconeogenesis in the liver[6] and is also known to reduce intestinal glucose absorption[7]. Further, metformin appears to improve glucose uptake and utilization systemically[3].

Metformin is a nitrogen-rich biguanide. Flavin-containing monooxygenases(FMO)-5 has demonstrated narrow substrate specificity, but has been known to catalyze oxygenation of nitrogen-containing drugs[8]. FMO5 is expressed in the kidneys and liver[8]. The FMO5 gene exists near PRKAB2, a known PD regulator of metformin response, away from the single gene cluster for the remaining FMOs in chromosome 1q23-q25 region. Metformin is excreted unchanged in the urine[5], hinting that metformin does not undergo biotransformation. However, studies such as these do not produce 100% yield, hinting at room for deviation from this paradigm. While metformin is thought to be absent of biotransformation[3], it is biologically plausible that FMO5 might carry out N-oxygenation of metformin.

FMOs show overlapping substrate specificity among family members[8]; a signal corresponding to FMO5 might also correspond to an additional FMO gene. All FMOs contain eight coding exons that share 50 to 80% sequence identity, with mutant FMOs are known to react to alternative chemical sites[9]. FMOs are localized in the endoplasmic reticulum of the cell whose expression is tissue-specific[8]. The extent of which reactions are catalyzed by FMOs in vivo cannot be determined by measuring end products excreted

in bile or urine[10].

The primary purpose of this study was to add clarity to metformin pharmacogenomics by understanding the impact of common variants in the FMO5 gene on altered glycemic response in a clinical population derived from an EHR-linked biorepository. Due to some shared functional similarity among genes in the FMO gene family, we selected the remaining FMO genes (FMO1 – FMO4) as exploratory gene candidates as our secondary hypothesis.

Methods

In this EHR-linked genetic study, both the approaches for obtaining clinical phenotypes and genotypes had important considerations for both study design and study interpretation. Our primary hypothesis of interest holds that genetic variation within FMO5 has potential to modify glycemic response to metformin monotherapy. Secondary to the primary hypothesis is an exploratory hypothesis that posits similar potential associations for FMO1 – FMO4 due to functional similarity[8]. However, their function is not identical. Further, due to the close proximity of the FMO1 – FMO4 to each other and their relative distance from FMO5 on chromosome 1q21 our secondary hypothesis is considerably weaker than our primary hypothesis for FMO5. In this study, we utilized the longitudinal EHR at Mayo Clinic and genome-wide association study (GWAS) data from the subjects enrolled in the Mayo Genome Consortia (Mayo GC)[11].

Clinical Phenotypes

The application of EHR-based phenotypes dramatically impacts study design and interpretability of findings. In this study we had 4 key phenotype aspects to consider: 1) T2DM phenotype, 2) metformin exposure phenotype, and 3) change in A1c. First, attribution of a T2DM phenotype was performed using a modified methodology developed by eMERGE[12]. A key point of differentiation is that our T2DM phenotype relied on diagnosis codes and did not initially consider laboratory values or medication. However, our second and third considerations relied on lab values and medication exposure events that were more specific than the criteria for the eMERGE T2DM phenotype algorithm. Second, our metformin exposure period was designated as a new

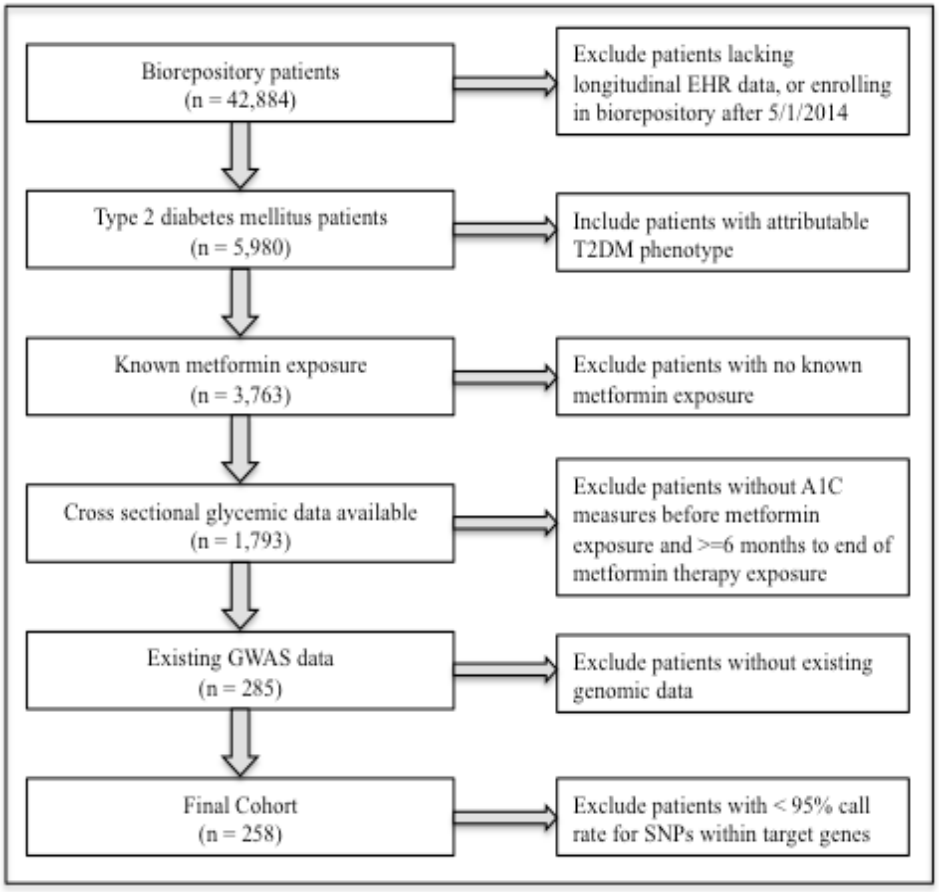
prescription of metformin that extended ≥6 months to ensure adequate primary care visits, multiple A1c measures, and maintenance dose achievement. Since our study aimed to understand genomic variation in relation to patients who respond or do not respond to metformin, maintenance dose was not a consideration. To accurately populate this metformin exposure phenotype our study design required longitudinal data access from primary care patients. Specifically, study inclusion criteria required ≥1 year of patient history and ≥2 primary care visits to ensure accurate capture of the first date of metformin exposure, which aimed to exclude patients that were false positives for a new recorded exposure to metformin due to medication reconciliation that occurred at transfer of primary care. Metformin exposure events were ascertained using a combination of validated structured and semi-structured EHR data collection methodologies that leveraged our prior work[13,14] where a total of 1 generic name (metformin) and 4 brand name medications (Fortamet®, Glucophage®, Glumetza®, and Riomet®) were queried. Patients with <6 months of metformin exposure or on combination drugs that included metformin or other prescribed antidiabetic drugs during the ≥6 month exposure period were excluded from the study. Third, to compare the association of genetic modification to glycemic response to metformin, measures of A1c were compared prior to metformin exposure and during the period of metformin exposure following a 6-month period of delay to allow for the achievement of maintenance dosage. A1c measures were required ≤6 months prior to metformin exposure and ≥6 months after metformin exposure. A1c measures were averaged across sections that occurred before and up to the date of metformin exposure. A1c measures were averaged across the period occurring ≥6 months after initial metformin exposure and until either metformin exposure ceased or anti-diabetic combination therapy was initiated. This approach minimizes the impact of any one A1c measure and biases change in A1c measures towards the null.

Genotyping and Quality Control

The Mayo GC stores existing GWAS data generated from multiple studies. These data were harmonized to the forward strand mapped to become on the same strand as the 1000 genome cosmopolitan reference population. Genotypes for unmappable or ambiguous SNPs were excluded. We selected SNPs 20 kb upstream and downstream of

each gene using 1000 genomes project variants and NCBI build 37 as the reference genome. By this mapping rule, a total of 1,381 SNPs were mapped to the 5 genes, but only 205 SNPs were available in the genotype data. Further, due to their proximity the FMO1, FMO2, FMO3, and FMO4 genes some SNPs belong to multiple genes. For the remaining SNPs, two main quality control filters were applied:(i) SNPs with unacceptable high rates of missing genotype calls (>10%); and (ii) monomorphic SNPs were excluded. The quality control of the genotype data was performed by PLINK v1.07[15]. A detailed diagram of cohort development is found in **Figure 6.1**.

Figure 6.1: Study Cohort Development Process



Analysis

The SNP-level and gene-level analyses were performed on the final analysis cohort where 258 Caucasian subjects had metformin exposure, complete EHR data, and 90 SNPs after quality control. In the analysis, we adjusted for age, gender, and morbid

obesity (BMI ≥35), a known modifier of T2DM state[1] as fixed covariates in our model. Age and BMI measures were calculated at first recorded exposure to metformin. The endpoint of change in A1c was transformed using Van der Waerden rank, otherwise known as rank based inverse Gaussian, to normalize and accommodate linear regression modeling. Batch adjustment did not change the results of GWAS data (data not shown) and was not adjusted in the displayed results. SNP-level and gene-level results were described, but not displayed, after application of Bonferroni correction.

SNP-Level Analysis

SNP-level analyses were performed on each SNP in FMO genes pertaining to both our primary and secondary hypothesis to identify top SNPs and determine directionality of their associations. Using Van der Waerden rank transformation on change in A1c, linear regression models were applied adjusting for age, gender and morbid obesity. Coefficient estimates were calculated per minor allele, that is, with each minor allele, the A1c level changes by 'beta'. SNP-level results are displayed as unadjusted for multiple testing. Finally, similar analysis adjusting for the top SNP was performed to identify potentially independent SNPs in each gene. Locus Zoom plots were also created for better visualization using the LD in the 1000 Genomes European reference population from March 2012 release.

Gene-Level Analysis

Gene-level tests were performed using principal component analysis (PCA)[16]. For each gene, principal components (PC) were created using linear combinations of ordinally scaled SNPs (i.e., 0, 1, 2 copies of minor allele) and the smallest set of resulting principal components that explained at least 90% of the SNP variance within the gene was included in linear regression models. Instead of including the entire set of SNPs for each gene, the PC approach reduces the degrees of freedom, avoids model fitting issues due to multi-collinearity of the SNPs from linkage disequilibrium (LD) and potentially improves the statistical power. Finally, we computed the likelihood ratio test (LRT) to assess overall significance of a gene by comparing the null model containing only the covariates with the full model containing covariates and the set of resulting principal

components. The statistical package R 2.15.0 was utilized for the gene-level analysis. Plots of LD displaying r^2 for FMO5 gene was created using Haploview v 4.2.

Results

Our EHR-based phenotyping algorithm identified 1,793 T2DM subjects (**Figure 6.1**). Among those, 258 subjects had 90 SNP data that passed quality control criteria. Cohort demographics can be found in **Table 6.1**. The estimates for male (Coefficient=0.0435, P-value=0.737), age (Coefficient=-0.0009, P-value=0.881), morbid obesity (Coefficient=0.2214, P-value=0.083) were not significantly associated with change in A1c at alpha=0.05 significance level in the univariate analysis. Further, none of the covariates were associated with change in A1c at alpha=0.05 significance level in a multivariate model.

Table 6.1: Demographics (n=258)	
Variable	**n (%)**
Female, N(%)	89 (34.5)
Male, N (%)	169 (64.5)
BMI <30, N (%)	64 (24.8)
BMI(≥30 to <35 kg/m^2)	100 (38.8)
BMI ≥35 (kg/m^2)	93 (36.1)
Median A1c >7.0 (DCCT %), N (%)	101 (39.1)
Change in A1c (DCCT %), median (range)	0.07 (-6.45, 3.51)
Age (years), median (range)	64 (30, 84)

SNP-Level Results

Among 31 genotyped SNPs within FMO5 gene, 4 SNPs had p-values less than 0.05 for the association with a decrease in glycemic response during metformin exposure, with rs7541245 having the most significant signal. While after adjusting for multiple testing rs7541245 was marginally significant, this signal is vey close to passing correction (0.00188-observed vs. 0.00161-Bonferonni threshold), and was still appropriate for consideration. None of the SNPs in FMO1-FMO4 gene cluster were found to be significant. The FMO5 linkage disequilibrium (LD) plot (**Figure 6.2**) contained 4 LD blocks and appeared to show 9 independent SNPs. The conditional analysis that adjusted

Figure 6.2: FMO5 Linkage Disequilibrium Blocks

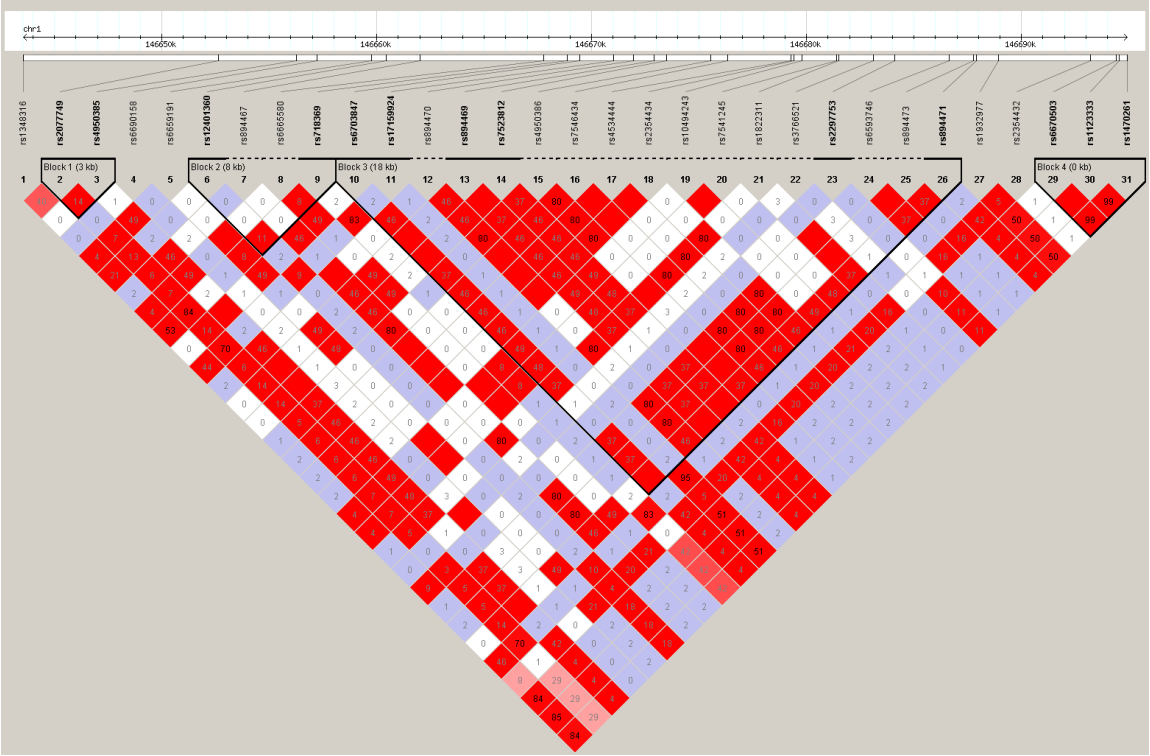
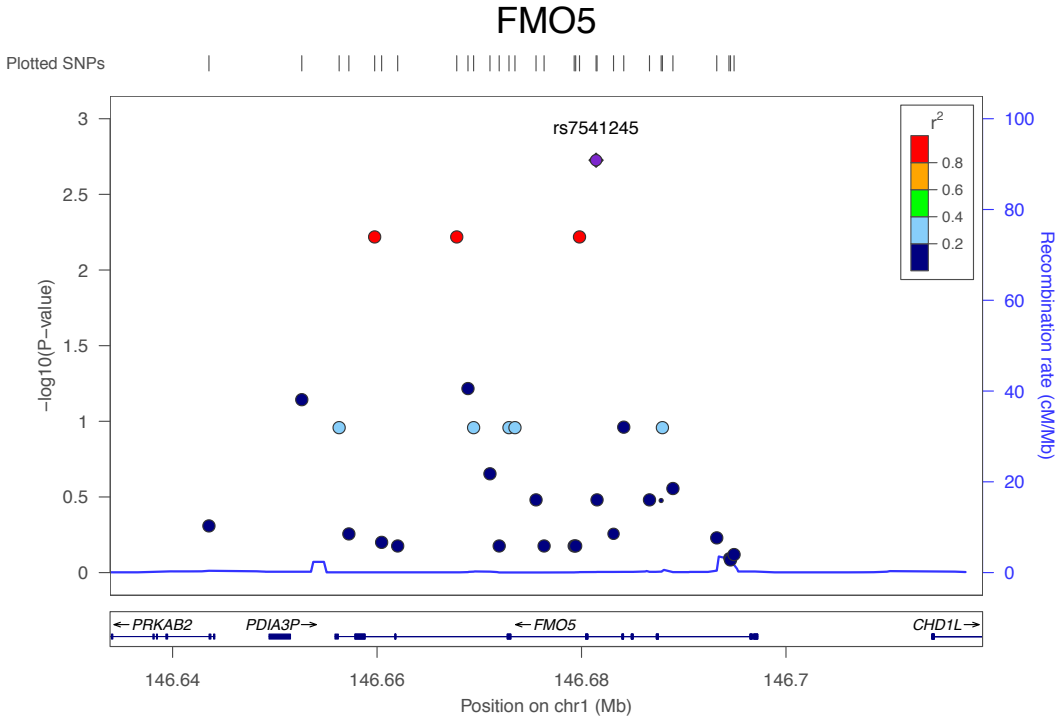
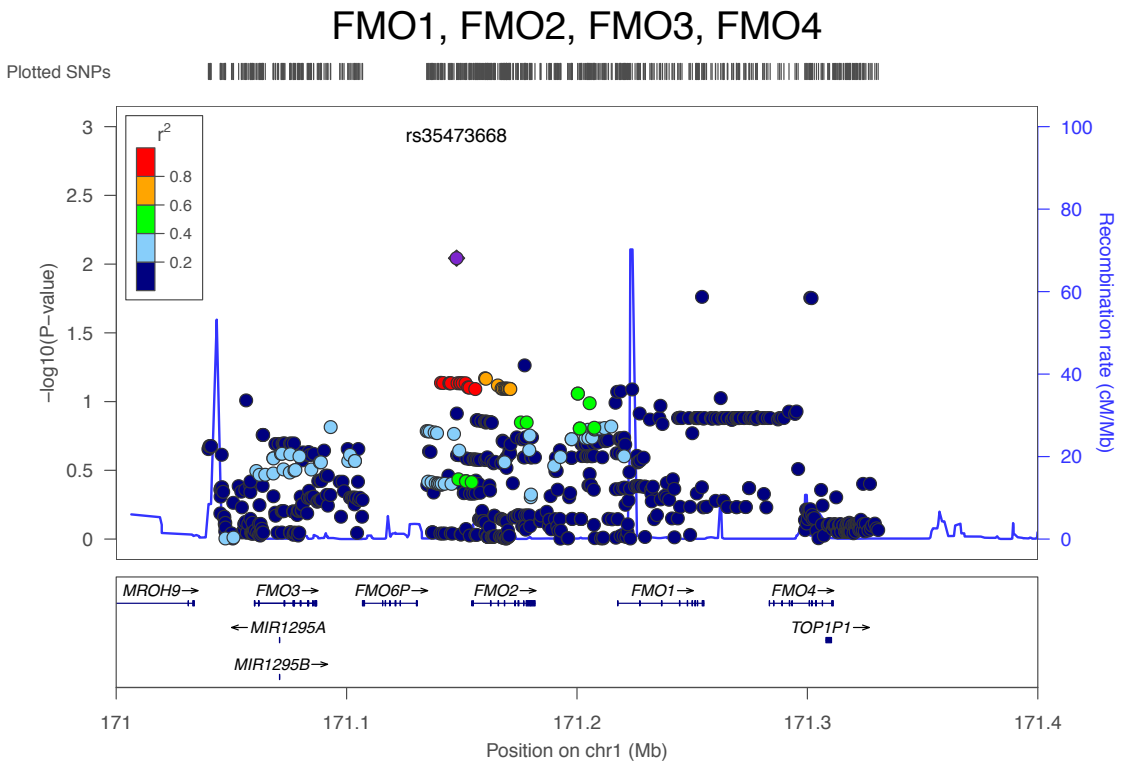


Figure 6.3: FMO5 Locus Zoom Plot



for the top most significant SNP in each gene and clinical covariates was performed. FMO5 rs7541245 was the main signal on FMO5 gene as no SNPs reached p-values less than 0.05 which pointed to the remaining SNPs within FMO5 being in high LD with rs7541245 and hence, not independent. The Locus Zoom plot for FMO5 can be found in **Figure 6.3**. For reference purposes only, a Locus Zoom plot containing FMO1 – FMO4 can be found in **Figure 6.4.**

Figure 6.4: Locus Zoom plot for FMO1 – FMO4



Gene-Level Results

Our primary hypothesis for the FMO5 gene, represented by 5 PCs and 31 genotyped SNPs, was marginally significantly associated (**p=0.0185**) with glycemic response (**Table 6.2**) after controlling for age, gender and morbid obesity. No significant associations were identified for our secondary hypothesis tests of the remaining FMO genes.

Table 6.2: Principal Component Analysis

Gene	Genotyped SNPs (n)	nPCs	P-value
FMO5	31	5	**0.0185**
FMO4	12	4	0.5623
FMO3	14	5	0.5464
FMO2	19	4	0.3581
FMO1	15	6	0.5479

Discussion

In this study, we leverage EHR-linked biorepository data and EHR-based phenotyping methods to study common variants within FMO5, our gene of primary interest. While the FMO5 gene appeared to be of marginal significance in relation to glycemic response to metformin, our secondary hypothesis for the remaining FMO genes demonstrated no significance. Given the study design and execution of phenotypes, results of this study can be interpreted most accurately as pharmacogenetics hypothesis generating. However, this hypothesis could represent a novel mechanism for the biotransformation of metformin and mechanism of metformin action that has been previously unidentified. Functional studies are indeed warranted.

In our study not all SNPs within candidate genes were available for analysis due to GWAS sequencing being originally performed for other studies. No effect difference was observed between cohort batches which indicated that our findings were not biased due to original patient selection criteria or sequencing criteria. Having all patients with T2DM and metformin allowed for us to identify genetic variation as the consideration of interest. However, the limited sample size paired with a relatively weak clinical outcome had potential to bias associations towards the null. While utilizing a clinical endpoint enabled us to engage in exploratory research, our signal strength was limited by modest cohort size (n=258) and the study criteria design. Specifically, by removing patients with <6 months of metformin exposure during metformin exposure we potentially removed patients who were complete non-responders to metformin or who experienced an adverse reactions to metformin. By study design these would not have been able to attain glycemic control with metformin, biasing our outcome phenotype towards positive glycemic response (i.e. decreased A1c) to metformin.

Alterations in FMO genes are known to induce differential biotransformation of nitrogen-rich compounds, such as metformin[10]. In this study, it appeared that the utility of metformin (i.e. glycemic response) is impaired by alterations in the FMO5 gene, hinting that potential biotransformation of metformin might be occurring in the normal FMO5 gene product. Our finding hints that metformin conjugates resulting from metformin biotransformation via FMO5 might be responsible for the anti-diabetic effects of

metformin. Should these findings be confirmed by functional studies, this hypothesis could represent a novel mechanism for the biotransformation of metformin and mechanism of metformin action that has been previously unidentified.

Conclusion

FMO5 appears to be marginally significantly associated with decreases in glycemic response after exposure to metformin, representing an EHR-driven pharmacogenetics hypothesis that could represent a novel mechanism for the biotransformation of metformin that has been previously unidentified. Functional validation of this hypothesis is warranted to ascertain its clinical and biological significance.

References

1. Inzucchi SE, Bergenstal RM, Buse JB, et al. Management of Hyperglycemia in Type 2 Diabetes: A Patient-Centered Approach. Diabetes Care. 2012;35:1364-1379.

2. Todd JN, Florez JC. An update on the pharmacogenomics of metformin: progress, problems and potential. Pharmacogenomics. 2014;15(4):529-539.

3. Gong L, Goswami S, Giacomini KM, Altman RB, Klein TE. Metformin pathways: pharmacokinetics and pharmacodynamics. Pharmacogenetics and genomics. Nov 2012;22(11):820-827.

4. Bailey CJ, Path MRC, Turner RC. Metformin. The New England Journal of Medicine. 1996;334(9):574-579.

5. Graham G.C., Punt J, Arora M, et al. Clinical Pharmacokinetics of Metformin. Clinical Pharmacokinetics. 2011;50(2):81-98.

6. Hundal RS, Krssak M, Dufour S, et al. Mechanism by Which Metformin Reduces Glucose Production in Type 2 Diabetes. Diabetes. 2000;49.

7. Sakar Y, Meddah B, Faouzi MYA, Cherrah Y, Bado A, Ducroc R. Metformin-Induced Regulation of Intestinal D-Glucose Transporters. Journal of Physiology and Pharmacology. 2010;61(3):301-307.

8. Lattard V, Zhang J, Cashman JR. Alternative Processing Events in Human FMO Genes. Molecular Pharmacology. 2004;65(6):1517-1525.

9. Joosten V, van Berkel WJH. Flavoenzymes. Current Opinion in Chemical Biology.

2007;11:195-202.

10. Ziegler DM. Flavin-Containing Monooxygenases: Catalytic Mechanism and Substrate Specificities. Drug Metabolism Reviews. 1988;19(1):1-33.

11. Bielinski SJ, Chai HS, Pathak J, et al. Mayo Genome Consortia: A Genotype-Phenotype Resource for Genome-Wide Association Studies With an Application to the Analysis of Circulating Bilirubin Levels. Mayo Clinic proceedings. 2011;86(7):606-614.

12. Kho AN, Hayes MG, Rasmussen-Torvik L, et al. Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study. Journal of the American Medical Informatics Association : JAMIA. Mar-Apr 2012;19(2):212-218.

13. Chute CG, Beck SA, Fisk TB, Mohr DN. The Enterprise Data Trust at Mayo Clinic: a semantically integrated warehouse of biomedical data. JAMIA. Mar-Apr 2010;17(2):131-135.

14. Pathak J, Murphy SP, Willaert BN, et al. Using RxNorm and NDF-RT to Classify Medication Data Extracted from Electronic Health Records: Experiences from the Rochester Epidemiology Project. American Medical Informatics Association Annual Symposium. 2011:1089-1098.

15. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. American journal of human genetics. Sep 2007;81(3):559-575.

16. Gauderman WJ, Murcray C, Gilliland F, Conti DV. Testing association between disease and multiple SNPs in a candidate gene. Genetic epidemiology. Jul 2007;31(5):383-395.

V. Conclusion

Chapter 7: Conclusion for the Framework on Translational Informatics

In this dissertation I utilized informatics-driven data acquisition to answer a line of scientific inquiry regarding the pharmacogenomics and pharmacoepidemiology of metformin in breast cancer and T2DM that was centered on the Multilevel Framework for Translational Informatics (**Figure 1.1**). In this work I addressed three topics of relevance: 1) pharmacoepidemiological modeling using EHR-data, 2) incorporating population-level socioecological context into clinical research, and 3) leveraging an EHR-linked biorepository to test and develop pharmacogenetic hypotheses. Together, these parts highlighted some of the nuances in utilizing data from different levels to elucidate clinical questions. Individually, I was able to 1) demonstrate that metformin has a protective effect and insulin a detrimental effect on breast cancer treatment outcomes that is beyond their implication for T2DM severity, 2) develop an Socioecological Conditions (SEC) Index that was significant and improved model performance for predicting breast cancer performance, and 3) develop an EHR-driven molecular hypothesis for the potential biotransformation of metformin via FMO5, a novel mechanism. While this work elucidated aspects of breast cancer pharmacogenomics, it primarily aimed to demonstrate this framework for utilizing informatics approaches to drive metformin pharmacoepidemiology and translational pharmacogenomics for potential metformin repurposing in breast cancer treatment and similar research in the future.

From an informatics perspective the contributions of this dissertation include: (i) multilevel integration of data from a molecular level to socio-ecological level and (ii) capturing exposure phenomena from large-scale (EHR-scale) data through the SEC Index, which is an enabling technology for integrating SECs with clinical data. These contributions are very well aligned with the NCI's strategic goals for biomedical informatics that I briefly described in the Prelude. As the field of biomedical informatics proceeds forward towards data science, relying on reference points that appropriately frame the starting points or boundaries (i.e. premises) become increasingly important. Specifically, since biomedical scientists do not have the luxury to gain knowledge from

near-unlimited data access for algorithm training that is available on internet, robust premises become a critical tool to focusing data power to analysis-specific computationally and scientifically relevant calculations. Further, as we think about the needs of cancer epidemiology the more we see the need for multilevel modeling powered by scalable and portable informatics-driven approaches.

VI. Bibliography

Adkins, D.E., Vaisey, S., Toward a Unified Stratification Theory: Structure, Genome, and Status Across Human Societies. Sociological Theory, 2009. 27(2): p.99-121.

Aneshensel, C.S., Social Stress: Theory and Research. Annual Review of Sociology, 1992. 18: p.15-38

Bailey CJ, Path MRC, Turner RC. Metformin. The New England Journal of Medicine. 1996;334(9):574-579.

Bayraktar S, Hernadez-Aya LF, Lei X, et al. Effect of metformin on survival outcomes in diabetic patients with triple receptor-negative breast cancer. Cancer. 2012;118(5):1202-1211.

Bielinski SJ, Chai HS, Pathak J, et al. Mayo Genome Consortia: A Genotype-Phenotype Resource for Genome-Wide Association Studies With an Application to the Analysis of Circulating Bilirubin Levels. Mayo Clinic proceedings. 2011;86(7):606-614.

Bonanni B, Puntoni M, Cazzaniga M, et al. Dual effect of metformin on breast cancer proliferation in a randomized presurgical trial. Journal of clinical oncology: official journal of the American Society of Clinical Oncology. Jul 20 2012;30(21):2593-2600.

Boyle P, Boniol M, Koechlin A, et al. Diabetes and breast cancer risk: a meta-analysis. British journal of cancer. 2012;107(9):1608-1617.

Brown KA, Samarajeewa NU, Simpson ER. Endocrine-related cancers and the role of AMPK. Molecular and Cellular Endocrinology. 2013;366:170-179.

Charlson, M.E., Pompei, P., Ales, K.L., MacKenzie, C.R., A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. Journal of Chronic Disease, 1987. 40(5): p. 373-383.

Chen S, Zhou J, Xi M, et al. Pharmacogenetic Variation and Metformin Response. Current Drug Metabolism. 2013;14:1070-1082.

Chlebowski RT, McTiernan A, Wactawski-Wende J, et al. Diabetes, metformin, and breast cancer in postmenopausal women. Journal of clinical oncology. Aug 10 2012;30(23):2844-2852.

Christensen MM, Brasch-Andersen C, Green H, et al. The pharmacogenetics of metformin and its impact on plasma metformin steady-state levels and glycosylated hemoglobin A1c. Pharmacogenetics and genomics. Dec 2011;21(12):837-850.

Chute CG, Beck SA, Fisk TB, et al. The Enterprise Data Trust at Mayo Clinic: a semantically integrated warehouse of biomedical data. Journal of the American Medical Informatics Association. 2010;17:131-135.

Cook MN, Girman CJ, Stein PP, Alexander CM. Initial monotherapy with either metformin or sulphonylureas often fails to achieve or maintain current glycaemic goals in patients with Type 2 diabetes in UK primary care. Diabetic Medicine. 2007;24:350-358.

Currie C, Gale EA, Poole C, Johnson J, Jenkens-Jones S, Morgan C. Mortality After Incident Cancer in People With and Without Type 2 Diabetes. Diabetes care. 2012;35.

DuBois, D.L., Felner, R.D., Brand, S., Adan, A.M., Evans, E.G., A Prospective Study of Life Stress, Social Support, and Adaptation in Early Adolescence. Child Development, 1992. 63(3): p.542-557.

Ferro A, Goyal S, Kim S, et al. Evaluation of Diabetic Patients with Breast Cancer Treated with Metformin during Adjuvant Radiotherapy. International journal of breast cancer. 2013;2013:659723.

Franciosi M, Lucisano G, Lapice E, Strippoli GF, Pellegrini F, Nicolucci A. Metformin therapy and risk of cancer in patients with type 2 diabetes: systematic review. PloS one. 2013;8(8):e71583.

Friedman, S., Index of dissimilarity. In V. Parrillo (Ed.), Encyclopedia of social problems. Thousand Oaks, CA: SAGE Publications, Inc., 2008. p. 489.

Gallagher EJ, LeRoith D. Diabetes, cancer, and metformin: connections of metabolism and cell proliferation. Annals of the New York Academy of Sciences. 2011;1243:54-68.

Gauderman WJ, Murcray C, Gilliland F, Conti DV. Testing association between disease and multiple SNPs in a candidate gene. Genetic epidemiology. Jul 2007;31(5):383-395.

Gilthrope, M.S., The importance of normalisation in the construction of deprivation indices. Journal of Epidemiology and Community Health, 1995. 49: p.S45-S50.

Gong L, Goswami S, Giacomini KM, Altman RB, Klein TE. Metformin pathways: pharmacokinetics and pharmacodynamics. Pharmacogenetics and genomics. Nov 2012;22(11):820-827.

Goodwin PJ, Stambolic V, Lemieux J, et al. Evaluation of metformin in early breast cancer: a modification of the traditional paradigm for clinical testing of anti-cancer agents. Breast cancer research and treatment. Feb 2011;126(1):215-220.

Graham G.C., Punt J, Arora M, et al. Clinical Pharmacokinetics of Metformin. Clinical Pharmacokinetics. 2011;50(2):81-98.

Graham G.C., Punt J, Arora M, et al. Clinical Pharmacokinetics of Metformin. Clinical Pharmacokinetics. 2011;50(2):81-98.

Hadad S, Iwamoto T, Jordan L, et al. Evidence for biological effects of metformin in operable breast cancer: a pre-operative, window-of-opportunity, randomized trial. Breast cancer research and treatment. Aug 2011;128(3):783-794.

Hadad SM, Hardie DG, Appleyard V, Thompson AM. Effects of metformin on breast cancer cell proliferation, the AMPK pathway and the cell cycle. Clinical Translational Oncology. Dec 2013.

Hardie DG, Ross FA, Hawley SA. AMPK: a nutrient and energy sensor that maintains energy homeostasis. Molecular Cell Biology. 2012;13:251-262.

He X, Esteva FJ, Ensor J, Hortobagyi GN, Lee MH, Yeung SC. Metformin and thiazolidinediones are associated with improved breast cancer-specific survival of diabetic women with HER2+ breast cancer. Annals of oncology. 2012;23(7):1771-1780.

Hou G, Zhang S, Zhang X, Wang P, Hao X, Zhang J. Clinical pathological characteristics and prognostic analysis of 1,013 breast cancer patient with diabetes. Breast cancer research and treatment. 2013;137:807-816.

Hundal RS, Krssak M, Dufour S, et al. Mechanism by Which Metformin Reduces Glucose Production in Type 2 Diabetes. Diabetes. 2000;49.

Inzucchi SE, Bergenstal RM, Buse JB, et al. Management of Hyperglycemia in Type 2 Diabetes: A Patient-Centered Approach. Diabetes Care. 2012;35:1364-1379.

Jablonski KA, McAteer JB, de Bakker PIW, et al. Common Variants in 40 Genes Assessed for Diabetes Incidence and Response to Metformin and Lifestyle Intervention in the Diabetes Prevention Program. Diabetes. 2010;59:2672-2681.

Joosten V, van Berkel WJH. Flavoenzymes. Current Opinion in Chemical Biology. 2007;11:195-202.

Ed. Kawachi, I., Berkman, L.F., Neighborhoods and Health. Oxford University Press, 2003.

Kho AN, Hayes MG, Rasmussen-Torvik L, et al. Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study. Journal of the American Medical Informatics Association : JAMIA. Mar-Apr 2012;19(2):212-218.

Kim J, Lim W, Kim EK, et al. Phase II randomized trial of neoadjuvant metformin plus letrozole versus placebo plus letrozole for estrogen receptor positive postmenopausal breast cancer (METEOR). BMC cancer. 2014;14:170.

Lattard V, Zhang J, Cashman JR. Alternative Processing Events in Human FMO Genes. Molecular Pharmacology. 2004;65(6):1517-1525.

Leabman MK, Giacomini KM. Estimating the contribution of genes and environment to variation in renal drug clearance. Pharmacogenetics. Sep 2003;13(9):581-584.

Lega IC, Austin PC, Gruneir A, Goodwin PJ, Rochon PA, Lipscombe LL. Association Between Metformin Therapy and Mortality After Breast Cancer: A population-based study. Diabetes care. 2013;36(10):3018-3026.

Lega IC, Shah PS, Margel D, Beyene J, Rochon PA, Lipscombe LL. The effect of metformin on mortality following cancer among patients with diabetes. Cancer Epidemiology Biomarkers & Prevention. 2014.

Lin HC, Hsu YT, Kachingwe BH, Hsu CY, Uang YS, Wang LH. Dose effect of thiazolidinedione on cancer risk in type 2 diabetes mellitus patients: a six-year population-based cohort study. Journal of clinical pharmacy and therapeutics. Mar 24 2014.

Liu L, Forman S, Barton B. Fitting Cox Model Using PROC PHREG and Beyond in SAS. SAS Global Forum. 2009.

Lynch, S.M., Rebbeck, T.R., Bridging the Gap between Biologic, Individual, and Macroenvironmental Factors in Cancer: A Multilevel Approach. Cancer Epidemiology, Biomarkers & Prevention, 2013. 22(4): p.485-495.

Ma J, Guo Y, Chen S, et al. Metformin enhances tamoxifen-mediated tumor growth inhibition in ER-positive breast carcinoma. BMC cancer. 2014;14:172.

McEwen, B.S., Stellar, E., Stress and the Individual. Archives of Internal Medicine, 1993. 153: p.2093-2101.

Niraula S, Dowling RJ, Ennis M, et al. Metformin in early breast cancer: a prospective window of opportunity neoadjuvant study. Breast cancer research and treatment. Oct 2012;135(3):821-830.

Oakes, J.M. The (mis)estimation of neighborhood effects: causal inference for a practicable social epidemiology. Social Science & Medicine, 2004. 58: p.1929-1952.

Oakes, J.M., The measurement of SES in health research: current practice and steps toward a new approach. Social Science & Medicine, 2003. 56(4): p.769-784.

Pathak J, Murphy SP, Willaert BN, et al. Using RxNorm and NDF-RT to Classify Medication Data Extracted from Electronic Health Records: Experiences from the Rochester Epidemiology Project. American Medical Informatics Association Annual Symposium. 2011:1089-1098.

Pawlyk AC, Giacomini KM, McKeon C, Shuldiner AR, Florez JC. Metformin Phramacogenomics: Current Status and Future Directions. Diabetes. 2014;63:2590-2599.

Peeters PJ, Bazelier MT, Vestergaard P, et al. Use of metformin and survival of diabetic women with breast cancer. Curr Drug Saf. 2013;8:357-363.

Pollak M. Insulin and insulin-like growth factor signaling in neoplasia. Nature Reviews Cancer. 2008;8: 915-928.

Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. American journal of human genetics. Sep 2007;81(3):559-575.

Sadighi S, Amanpour S, Behrouzi B, Khorgami Z, Muhammadnejad S. Lack of Metformin Effects on Different Molecular Subtypes of Breast Cancer under Normoglycemic Conditions: An in vitro Study. Asian Pacific Journal of Cancer Prevention. 2014;15(5):2287-2290.

Sakar Y, Meddah B, Faouzi MYA, Cherrah Y, Bado A, Ducroc R. Metformin-Induced Regulation of Intestinal D-Glucose Transporters. Journal of Physiology and Pharmacology. 2010;61(3):301-307.

Sakar Y, Meddah B, Faouzi MYA, Cherrah Y, Bado A, Ducroc R. Metformin-Induced Regulation of Intestinal D-Glucose Transporters. Journal of Physiology and Pharmacology. 2010;61(3):301-307.

Savova GK, Masanz JJ, Ogren PV, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. Journal of the American Medical Informatics Association : JAMIA. Sep-Oct 2010;17(5):507-513.

Team RDC. R: A language and environment for statistical computing. . Vienna, Austria: R Foundation for Statistical Computing; 2012.

Thompson AM. Molecular pathways: preclinical models and clinical trials with metformin in breast cancer. Clinical cancer research : an official journal of the American Association for Cancer Research. May 15 2014;20(10):2508-2515.

Todd JN, Florez JC. An update on the pharmacogenomics of metformin: progress, problems and potential. Pharmacogenomics. 2014;15(4):529-539.

Townsend, P., Phillimore, P., Beattie, A. Health and Deprivation: Inequality and the North. Croom Helm, 1988.

Vazquez-Martin A, Oliveras-Ferraros C, Menendez JA. The antidiabetic drug metformin suppresses HER2 (erbB-2) oncoprotein overexpression via inhibition of the mTOR effector p70S6K1 in human breast carcinoma cells. Cell Cycle. 2009;8(1):88-96.

Viollet B, Guigas B, Sanz Garcia N, Leclerc J, Foretz M, Andreelli F. Cellular and molecular mechanisms of metformin: an overview. Clinical science. Mar 2012;122(6):253-270.

Volden PA, Wonder EL, Skor MN. Chronic Social Isolation is Associated with Metabolic Gene Expression Changes Specific to Mammary Adipose Tissue. Cancer Prevention Research, 2013. 6(7):634-45.

Wang L, Weinshilboum R. Metformin pharmacogenomics: biomarkers to mechanisms. Diabetes. Aug 2014;63(8):2609-2610.

Winder WW, Hardie DG. AMP-activated protein kinase, a metabolic master switch: possible roles in Type 2 diabetes. American Journal of Physiology. 1999;277(1).

Xiao Y, Zhang S, Hou G, Zhang X, Hao X, Zhang J. Clinical pathological characteristics and prognostic analysis of diabetic women with luminal subtype breast cancer. Tumour Biol. 2014.

Xu H, Aldrich MC, Chen Q, et al. Validating drug repurposing signals using electronic health records: a case study of metformin associated with reduced cancer mortality. Journal of the American Medical Informatics Association. 2014;0:1-10.

Zhang ZJ, Li S. The prognostic value of metformin for cancer patients with concurrent diabetes: a systematic review and meta-analysis. Diabetes, Obesity and Metabolism. 2014.

Zhou K, Bellenguez C, Spencer CC, et al. Common variants near ATM are associated with glycemic response to metformin in type 2 diabetes. Nature Genetics. 2011;43(2):117-120.

Zhou K, Donnelly L, Yang J, et al. Heritability of variation in glycaemic response to metformin: a genome-wide complex trait analysis. The Lancet Diabetes & Endocrinology. 2014;2(6):481-487.

Zhu P, Davis M, Blackwelder AJ, et al. Metformin selectively targets tumor-initiating cells in ErbB2-overexpressing breast cancer models. Cancer prevention research. Feb 2014;7(2):199-210.

Ziegler DM. Flavin-Containing Monooxygenases: Catalytic Mechanism and Substrate Specificities. Drug Metabolism Reviews. 1988;19(1):1-33.

Zordoky BN, Bark D, Soltys CL, Sung MM, Dyck JR. The anti-proliferative effect of metformin in triple-negative MDA-MB-231 breast cancer cells is highly dependent on glucose concentration: implications for cancer therapy and prevention. Biochimica et biophysica acta. Jun 2014;1840(6):1943-1957.