

**INVESTIGATING THE CAUSAL RELATIONSHIP BETWEEN TYPE 2 DIABETES  
MELLITUS AND OVARIAN CANCER USING TWO-SAMPLE MENDELIAN  
RANDOMIZATION**

**LANGAT KIPKIRUI VICTOR**

**A Thesis Submitted to the Board of Graduate Studies in Partial Fulfillment of the  
Requirements for Conferment of the Degree of Master of Science in Applied Statistics of  
the University of Kabianga**

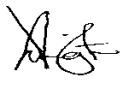
**UNIVERSITY OF KABIANGA**

**FEBRUARY, 2022**

## DECLARATION AND APPROVAL

### Declaration

This thesis is my original work and has not been presented for the award of a diploma or a degree in this or any other University:


Signed:.....  .....Date:.....23<sup>rd</sup> February 2022.....

Langat Kipkirui Victor

PGC/AST/005/17

### Approval


This thesis has been submitted for examination with our approval as the University supervisors:

Signed.....  ..... Date:....24<sup>th</sup> February 2022.....

Dr. Reuben Cheruiyot Lang'at

Department of Mathematics, Actuarial and Physical Sciences

University of Kabianga

Signed.....  ..... Date... 24<sup>th</sup> February 2022 .....

Dr. Ayubu Anapapa Okango

Department of Mathematics, Actuarial and Physical Sciences

Murang'a University of Technology

## **COPYRIGHT**

No part of this thesis may be produced, stored in any retrieval system or transmitted in any form, mechanical, photocopying, recording or otherwise without prior permission from the author or the University of Kabianga.

© Langat Kipkirui Victor, 2022

## **DEDICATION**

This thesis is dedicated to my lovely cherished parents: Richard Bore and Salina Bore, siblings: Kennedy Langat, Janet Chepwogen and Caleb Kiprono, Sister In-law: Caroline Chepngeno and my nephews: Kayron Kipngetich, Kayden Kimutai, and Baraka Kipruto.

## **ACKNOWLEDGMENT**

I would like to thank the Almighty God for the gift of life, good health, knowledge and the strength that has enabled me to reach where I am. Without His grace I would not have been able to accomplish this task which has been demanding and enriching in whole spheres of life.

I would like to acknowledge the members of Department of Mathematics and Computer Science, University of Kabianga for their support and guidance. In particular, my sincere acknowledgements go to my lead supervisor, Dr. Reuben Cheruiyot Lang'at for his tireless academic support and availability throughout this period of research. I also thank my second supervisor Dr. Ayubu Anapapa Okango for instilling statistics skills in me and going through my work.

My wholehearted appreciation goes to my parents and siblings. They have been my source of inspiration and have continuously given me moral, spiritual and financial support up to this far in my life.

## ABSTRACT

Ovarian cancers have registered rising cases of morbidity and mortality over the years. There is an assumption that Type 2 Diabetes Mellitus has a causal relationship with ovarian cancer due to the alarming rising incidence statistics. This research aimed at using a two-sample Mendelian Randomization design to undertake the causal relationship investigation. The specific objectives were to find out whether there exists heterogeneity, horizontal pleiotropy, and the causal relationship between Type 2 Diabetes Mellitus and ovarian cancer. Some of the epidemiologists have applied observational studies for example, cohort and case-control to investigate this causal relationship but ended up with conflicting points of view. The reason for the opposing results is that observational studies are prone to confounding errors and reverse causation, which has led to inaccurate conclusions. This research employed Mendelian randomization technique which uses genetic variants as instrumental variables, which undergo random allocation at conception hence it is therefore not affected by confounding factors. Apart from that, in this method the genetic variants were non-modifiable and thus not altered by reverse causation. The study used the inverse variance weighted technique and Mendelian Randomization-Egger method to obtain the causal estimates and test for the model sensitivity using the R software. The study indicated that there was no evidence of causal relationship between Type 2 Diabetes Mellitus and ovarian cancer (Mendelian Randomization-Egger:  $\beta = -0.0476$ , standard error = 0.0619, p-value = 0.4479, Inverse Variance Weighted:  $\beta = -0.0165$ , standard error = 0.0257, p-value = 0.5217). The odds ratios indicated that the two-sample Mendelian randomization had the power to detect 0.0464 and 0.0164 decrease in variability per 1 standard deviation for Mendelian Randomization-Egger and Inverse Variance Weighted respectively (Mendelian Randomization-Egger: odds ratio = 0.9536, confidence interval: 0.8447, 1.0765, Inverse Variance Weighted: odds ratio = 0.9836, confidence interval: 0.9352, 1.0345). This approach alleviated the usual problem of reverse causation and confounding factors hence depicting clearly that there is no causal relationship between Type 2 Diabetes Mellitus and ovarian cancer.

## TABLE OF CONTENTS

<b>DECLARATION AND APPROVAL .....</b>	<b>ii</b>
<b>COPYRIGHT .....</b>	<b>iii</b>
<b>DEDICATION .....</b>	<b>iv</b>
<b>ACKNOWLEDGMENT .....</b>	<b>v</b>
<b>ABSTRACT.....</b>	<b>vi</b>
<b>TABLE OF CONTENTS .....</b>	<b>vii</b>
<b>LIST OF TABLES .....</b>	<b>xi</b>
<b>LIST OF FIGURES .....</b>	<b>xii</b>
<b>LIST OF ABBREVIATIONS AND ACRONYMS .....</b>	<b>xiii</b>
<b>DEFINITION OF TERMS .....</b>	<b>xiv</b>
<b>CHAPTER ONE .....</b>	<b>1</b>
<b>INTRODUCTION.....</b>	<b>1</b>
1.1 Overview .....	1
1.2 Background of the Study.....	1
1.2.1 Type 2 Diabetes Mellitus.....	1
1.2.2 Ovarian cancer .....	3
1.2.3 Relationship between Type 2 Diabetes Mellitus and ovarian cancer.....	4

1.3 Statement of the Problem .....	5
1.4 General Objective .....	6
1.5 Specific Objectives .....	6
1.6 Research Hypotheses .....	6
1.7 Justification of the Study.....	7
1.8 Significance of the Study .....	7
1.9 Limitation of the Study .....	8
1.10 Scope of the Study .....	8
1.11 Outline of the Thesis .....	8
<b>CHAPTER TWO .....</b>	<b>9</b>
<b>LITERATURE REVIEW .....</b>	<b>9</b>
2.1 Introduction .....	9
2.2 Observational Studies .....	9
2.2.1 Cohort studies .....	10
2.2.2 Case-control studies .....	11
2.3 Mendelian Randomization .....	12
2.3.1 Two-sample Mendelian randomization .....	13
2.4 Horizontal Pleiotropy.....	15
2.5 Heterogeneity .....	16
2.6 Mendelian Randomization Analysis Techniques .....	17



2.6.1 Inverse variance weighted method .....	19
2.6.2 MR-egger method .....	20
2.7 Two-Sample Mendelian Randomization Application .....	22
2.8 Knowledge Gap .....	23
<b>CHAPTER THREE .....</b>	<b>24</b>
<b>RESEARCH METHODOLOGY .....</b>	<b>24</b>
3.1 Introduction .....	24
3.2 Study Design .....	24
3.3 Data Sources.....	24
3.4 Study Population.....	25
3.5 Mendelian Randomization Analysis to Determine the Causal Relationships between T2DM and Ovarian cancer .....	26
3.6 Sensitivity Analysis .....	28
3.7 Results Presentation.....	30
3.8 Ethical Consideration.....	30
<b>CHAPTER FOUR.....</b>	<b>31</b>
<b>RESULTS AND DISCUSSIONS .....</b>	<b>31</b>
4.1 Introduction .....	31
4.2 Data Analysis .....	31

4.3 Study Results .....	33
4.3.1 Homogeneity of the study data .....	33
4.3.2 Horizontal pleiotropy.....	36
4.3.3 Causal relationship between Type 2 Diabetes Mellitus and ovarian cancer .....	37
4.3.4 Sensitivity results .....	40
<b>CHAPTER FIVE.....</b>	<b>45</b>
<b>SUMMARY, CONCLUSIONS, AND RECOMMENDATIONS .....</b>	<b>45</b>
5.1 Introduction .....	45
5.2 Summary .....	45
5.3 Conclusions .....	46
5.4 Recommendations.....	47
5.5 Suggestions for Further Research.....	47
<b>REFERENCE.....</b>	<b>48</b>
<b>APPENDICES .....</b>	<b>55</b>
Appendix 1: Type 2 Diabetes Mellitus and Ovarian Cancer Genetic Data .....	55
Appendix 2: R Codes for the Study.....	56
Appendix 3: Published Paper .....	59

## LIST OF TABLES

<b>Table 4.1.</b> The Homogeneity Results of the Causal Relationship between Type 2 Diabetes Mellitus and Ovarian Cancer .....	34
<b>Table 4.2.</b> The MR-Egger Intercept Output of the Causal Relationship between Type 2 Diabetes Mellitus and Ovarian Cancer .....	36
<b>Table 4.3.</b> The Two-Sample Mendelian Randomization Results of the Causal Relationship between Type 2 Diabetes Mellitus and Ovarian Cancer .....	37
<b>Table 4.4.</b> The Direction of the Causal Relationship between Type 2 Diabetes Mellitus and Ovarian Cancer .....	40
<b>Table 4.5.</b> The Odds Ratio Output of the Causal Relationship between Type 2 Diabetes Mellitus and Ovarian Cancer .....	41

## LIST OF FIGURES

<b>Figure 2.1.</b> Two-Sample MR Conceptual Framework Showing the Causal Relationship between Type 2 Diabetes Mellitus and Ovarian Cancer.....	14
<b>Figure 4.1.</b> Forest Plot Displaying the Results of Single and Multi-SNP Analyses on the Causal Relationship between Type 2 Diabetes Mellitus and Ovarian Cancer .....	35
<b>Figure 4.2.</b> Scatter Plot Representing Two-Sample Mendelian Randomization Results of the Causal Relationship between Type 2 Diabetes Mellitus and Ovarian Cancer. ....	39
<b>Figure 4.3.</b> The Leave-One-Out Graph Displaying the IVW Results of the Causal Relationship between Type 2 Diabetes Mellitus and Ovarian Cancer while Excluding One SNP each Time .....	43
<b>Figure 4.4.</b> The Funnel Plot Displaying the Causal Relationship between Type 2 Diabetes Mellitus and Ovarian Cancer .....	44

## LIST OF ABBREVIATIONS AND ACRONYMS

<b>CI</b>	Confidence Interval
<b>DNA</b>	Deoxyribonucleic acid
<b>GIV</b>	Genetic Instrumental Variables
<b>GLOBOCAN</b>	Global Cancer Incidence, Mortality and Prevalence
<b>GWAS</b>	Genome-Wide Association Studies
<b>IGF-1</b>	Insulin like-Growth Factors 1
<b>InSIDE</b>	Instrument Strength Independent of Direct Effect
<b>IV</b>	Instrumental Variable
<b>IVW</b>	Inverse Variance Weighted
<b>LD</b>	Linkage Disequilibrium
<b>MR</b>	Mendelian Randomization
<b>OCAC</b>	Ovarian Cancer Association Consortium
<b>OR</b>	Odds Ratio
<b>SD</b>	Standard Deviation
<b>SE</b>	Standard Error
<b>SNPs</b>	Single Nucleotide Polymorphisms
<b>T2DM</b>	Type 2 Diabetes Mellitus
<b>WHO</b>	World Health Organization

## DEFINITION OF TERMS

**Causal inference** is the science or process of drawing conclusion on the causal relationship between two or more events basing on a particular data

**Causal relationship** refers to a scenario where one condition (exposure) leads to the occurrence of another disease (outcome)

**Causality** refers to cause and effect; a situation where one phenomenon leads to the occurrence of another event

**Confounding factor** is a variable that has a major role on a particular cause-and-effect relationship but it has not been included in the study

**Epidemiology** is a branch of biology that deals with the distribution and causation of health-related occurrences among a particular population

**Genetics** is a branch of science that deals with the study of genes, their variations, and heredity of organisms

**Genome-wide association studies (GWAS)** is group of researchers that identify the genetic variants that have association with a particular phenotype

**Linkage disequilibrium** is the situation where the alleles of different loci have a non-random relationship.

**Mendelian randomization** is an approach used to investigate the causal relationship between an exposure and outcome

**Pleiotropy** is a situation where the genetic variants affect multiple phenotypes those appear to appear to have no association

**Reverse causation** is a situation where, it is expected that factor A causes factor B but in reality factor B causes factor A

## **CHAPTER ONE**

### **INTRODUCTION**

#### **1.1 Overview**

This chapter gives the insights about Type 2 Diabetes Mellitus and ovarian cancer. It also indicates the statement of the problem, objectives of the study, significance, limitation and the scope of the thesis.

#### **1.2 Background of the Study**

People have noticed the rising morbidity and mortality cases associated with cancers and are now blaming the doctors for the late diagnosis of malignancies and the government for lack of efficient facilities to fight this menace (Pilleron *et al.*, 2021). Many of the people neglect how their health condition may be contributing to the development of cancer (Savard and Morin, 2001; Gibson *et al.*, 2015). Ovarian cancer is one of the deadly malignancies due to its late diagnosis and women with Type 2 Diabetes Mellitus (T2DM) have low survival rates from it (Urpilainen *et al.*, 2018). Therefore, there is a need for causal relationship determination between Type 2 Diabetes Mellitus and ovarian cancer using two-sample Mendelian randomization.

##### **1.2.1 Type 2 Diabetes Mellitus**

Diabetes is one of the metabolic disturbance conditions, which can occur when there is underproduction of insulin, or the body does not effectively use the hormone. Insulin is a hormone that helps significantly in the regulation of blood sugar (Tao *et al.*, 2015). The

World Health Organization (WHO) describes T2DM as a chronic disease distinguished by hyperglycemia, raised blood sugar.

The WHO reported that Diabetes cases have exponentially risen from about 108 million in 1980 to approximately 422 million in 2014 (Roglic, 2016). This situation reflects that the global prevalence had also changed from about 4.7% in 1980 to about 8.5% in 2014. The WHO statistics show that in 2012, about 2.2 million deaths had an association with high blood glucose. This insight depicts that T2DM is one of the menaces in the societies affecting approximately 3.0% to 4.0% of the adults (Tsilidis *et al.*, 2015). WHO has made some steps to reduce the prevalence of T2DM, for instance, it has designated 14<sup>th</sup> November every year as awareness day on the global epidemic of Diabetes. Apart from that, it has developed the standard and required norms for Diabetes diagnosis and care.

Kibirige *et al.* (2019) indicated that Africa has a Diabetes Mellitus disease prevalence of approximately 3.1%. This prevalence indicates that about 15.9 million adults in Africa are battling with Diabetes Mellitus. According to Kibirige *et al.* (2019), most people in the African continent are undiagnosed with the disease hence there is an expectation that the incidences will increase by about 156% by 2045. Mercer *et al.* (2019) noted that most African countries are trying to improve the Diabetes care programs, which will ensure accessibility, quality, and safety of medications.

The WHO estimated that the Diabetes prevalence in Kenya stands at around 3.3% and it is expected to rise to about 4.5% by 2025. Jones (2013) stated that in 2010, Diabetes Mellitus led to 2% of the total deaths. The Kenyan government has stepped up in helping the people with Diabetes by subsidizing the prices of insulin. However, the insulin supply usually runs out and there is mismanagement of funds directed to fighting this menace (Jones, 2013).



### **1.2.2 Ovarian cancer**

Cancer refers to a collection of associated conditions that lead to abnormal cell growth that has the possibility of spreading (Greenstein, 2016). Many types of cancers exist, for example, breast, prostate, colorectal, ovarian malignancies, among others.

According to 2018 global cancer statistics, there were about 300,000 new ovarian cancer incidences recorded (Bray *et al.*, 2018). The 2018 Global Cancer Incidence, Mortality and Prevalence (GLOBOCAN) estimates indicated that ovarian cancer is the eighth most prevalent malignancy among women globally. Momenimovahed *et al.* (2019) noted that ovarian cancer accounts for about 3.4% of all malignancies in women using GLOBOCAN 295,414 cases. The research further stated that approximately 184,799 deaths had an association with ovarian cancer, accounting for about 4.4% of cancer-related demise in 2018. Most of the diagnoses of ovarian cancer usually occur in the advanced stages. The late diagnoses account for about two-thirds of the cases; hence the survival rates tend to be low due to lack of effective screening strategies (Wang *et al.*, 2017). These low survival rates necessitate the identification of the predisposing factors to reduce the chances of this type of cancer.

Ovarian cancer is ranked second in Africa among the gynecological malignancies. (Akinfolarin, 2020). The major obstacle of the management of this disease in Africa is lack of sufficient screening facilities. This situation hence leads to the late diagnosis of the ovarian cancer. Most African countries are now combating ovarian cancer through escalation of public awareness and making sure that the machines for detection and diagnosis of the diseases are available. They also try to ensure that there is affordability of the treatment for all cancers.

In Kenya, ovarian cancer is ranked third among the major causes of deaths from gynecologic tumors (Cheserem *et al.*, 2013). Like other countries, Kenya experiences a challenge in the diagnosis of the ovarian cancer because of the non-specific characteristics and symptoms at its onset. Due to this reason, more than half of the women with ovarian cancer come to know of their status at the advanced stages (Cheserem *et al.*, 2013). The Kenyan government is currently trying to invest on the screening machines and training of the medical personnel with an aim of curbing the cancer menace as a whole.

### **1.2.3 Relationship between Type 2 Diabetes Mellitus and ovarian cancer**

Many studies indicate that T2DM is one of the significant predisposing factors for most types of malignancies (Harding *et al.*, 2015; Tsilidis *et al.*, 2015). The reason for this is that T2DM has a relationship with insulin resistance, hyperinsulinemia, and chronic inflammation, which contribute to the development of cancers (Juong *et al.*, 2015). Women with T2DM have ovarian steroid hormone, which alters the levels of estrogen, androgen, and progesterone. For instance, ovarian steroid hormone leads to an increase of estrogen and androgen levels while resulting in the reduction of progesterone. This situation, therefore, creates the potential carcinogenic conditions for the ovaries. T2DM tends to increase insulin or insulin like growth-factors 1 (IGF-1) levels, which have a relationship with the development of ovarian cancer (Joung *et al.*, 2015). The reason for this action is that higher levels of insulin and IGF-1 intensifies proliferation and slows down apoptosis in the affected cells, hence leading to the advancement of ovarian malignancy.

Several scholars have tried to investigate the causal relationship between T2DM and ovarian cancer using classical epidemiological methods like case-control and cohort studies (Wang *et al.*, 2017; Urpilainen *et al.*, 2018). Some of the reviews suggested that women with T2DM have a high probability of contracting ovarian cancer than their counterparts.

In contrast, other studies indicated that there was no sufficient evidence supporting the relationship. For instance, Wang *et al.* (2017) concluded that women with Diabetes Mellitus have a high probability of becoming victims of ovarian cancer, especially Asians. On the other hand, Urpilainen *et al.* (2018) demonstrated that there is no proof of an association between T2DM and ovarian cancers among women using metformin or oral anti-diabetic medicines.

These conflicting points of view indicate that classical epidemiological studies may be giving unreliable results due to biases arising from confounding and reverse causality. Therefore, the study used a two-sample Mendelian randomization (MR) method to investigate whether there is a causal relationship between T2DM and ovarian cancer.

### **1.3 Statement of the Problem**

An ideal research methodology would be expected to carry out causal inference without being affected by confounding factors and reverse causation. However, many studies have given conflicting points of view regarding the causal relationship between T2DM and ovarian cancer (Craig *et al.*, 2016.) Many of these studies use observational epidemiological methods that do not take into account biases from confounding factors like the degree of glycemic control; and reverse causality. Observational studies have been found to be affected by the possible confounding factors like familial history, the mutation of genes, menstrual periods, and the oral contraceptive usage in the analysis (Bashir and Litonjua, 2018). Therefore, there is need to adopt a better technique that do not suffer the setbacks as the observational methods. MR method is such a technique that can be utilized in the determination of the causal relationship between T2DM and ovarian cancer. Sekula *et al.* (2016) noted that a MR is an approach that uses genetic variants (SNPs) as instrumental variables to test the causal relationship between the exposure (T2DM) and the

outcome (ovarian cancer). The genetic alleles undergo randomized allocation at conception; hence they are free from confounding factors and reverse causality. Notably, there are limited researches in literature that explains the causal relationship between T2DM and ovarian cancer, and in particular those that have used the two-sample MR approach. There is therefore a need to use a robust method to determine whether there is a causal relationship between T2DM and ovarian cancer.

#### **1.4 General Objective**

To investigate if there is a causal relationship between T2DM and ovarian cancer using two-sample Mendelian randomization

#### **1.5 Specific Objectives**

The specific objectives of the study were;

1. To examine the homogeneity of the study data using Cochran's Q and I-square statistics.
2. To establish whether or not the selected Single Nucleotide Polymorphisms (SNPs) affect the outcome (ovarian cancer) through a biological pathway that is independent of the exposure (T2DM) in the study using MR Egger intercept.
3. To determine whether or not the exposure (T2DM) causes the outcome (ovarian cancer) using Inverse Variance Weighted (IVW) technique and MR-Egger method.

#### **1.6 Research Hypotheses**

The null hypothesis ( $H_0$ ): There is no causal relationship between Type 2 Diabetes Mellitus (T2DM) and ovarian cancer

The alternative hypothesis ( $H_1$ ): There is causal relationship between Type 2 Diabetes Mellitus (T2DM) and ovarian cancer

## **1.7 Justification of the Study**

Many epidemiological studies indicate that there is a high development of ovarian cancer in women patients because of the daily doses glargine, which is long-acting insulin (Joung *et al.*, 2015). It was also deduced that about 40-80% of the T2DM patients get the recommendation from the doctors to use insulin therapy, which will assist them in controlling glycemic levels. However, insulin, IGF-1, and the ovarian steroid hormone can act synergistically and increase the chances of contracting ovarian cancer (Joung *et al.*, 2015). Therefore, this study aims at determining if there is sufficient suggestion of the causal relationship between T2DM and ovarian cancer.

## **1.8 Significance of the Study**

Most of the epidemiologists and scholars have been using observational methods while carrying out the causal inference between two diseases. However, some of the scientists have indicated that these observational studies face drawbacks from confounding factors and reverse causation. Therefore, this study is significant because it gives the scholars and epidemiologist the alternative option of carrying out the causal inference using a method that thrives against confounding factors and reverse causation because it uses genetic variants as instrumental variables.

This study aims at finding out the causal relationship between Type 2 Diabetes Mellitus and ovarian cancer. Therefore, this study is significant to the stakeholders since with the appropriate results, scientists and medical practitioners will be better placed in coming up and prescribing medication accordingly

## **1.9 Limitation of the Study**

This research used a MR technique to determine the causal relationship between T2DM and ovarian cancer. This method is recommendable since it uses genetic variants as instrumental variables, hence able to establish whether there is causal relationship between an exposure and the outcome. This model has different types of randomization, like two-sample, bidirectional, multivariable, factorial, and more (Zheng *et al.*, 2017). Therefore, this study limits itself to using two-sample Mendelian randomization since this method utilizes datasets from different population hence able to widen the scope of the study unlike the other techniques.

## **1.10 Scope of the Study**

Mendelian randomization uses genotypes to evaluate causality between two diseases since they are not susceptible to reverse causation and confounding factors (Koellinger and De Vlaming, 2019). The study used the secondary data. Therefore, the summary statistics were retrieved from Ovarian Cancer Association Consortium (OCAC) and DIAGRAMplusMetaboChip consortium.

## **1.11 Outline of the Thesis**

This thesis is organized into five chapters. Chapter one outlines the background of the study leading to the identification of the problem, the objectives of the study, justification, significance, limitation and the scope of the study. The literature review of the various observational methods and the Mendelian randomization are presented in chapter two. Chapter three gives the step by step research methodology of the study. The results of the study are widely discussed in chapter four. Finally, chapter five gives the summary, discussions and recommendations of the study.

## CHAPTER TWO

### LITERATURE REVIEW

#### 2.1 Introduction

This chapter gives an overview of observational studies and the challenges that necessitate the use of the Mendelian randomization technique. It has also discussed the Mendelian randomization method and its advantages as well as identifying the knowledge gap.

#### 2.2 Observational Studies

The main aim of some epidemiological researches is to investigate the relationship between an exposure variable (environmental or genetic) and a medical condition outcome (Lucas and McMichael, 2005). Most of the epidemiological researches are non-experimental; hence they are prone to errors. The two researchers suggested that the investigators should proceed cautiously in making causal inferences because of the random, systematic, and confounding errors that may occur. The researchers should understand that having correlated variables does not necessarily indicate that there is a causal relationship because confounding factors may be the prime reason for the close relationship. Dickers *et al.* (2006) note that various observational epidemiological methods help in causal inference, for example, cohort studies and case-control studies. They indicated that the researchers simply observe the association that the exposure has on the outcome (disease) status of each participating individual. This situation indicates that there is a high probability of encountering confounding factors and reverse-causation thus there is a need to adopt a more sophisticated epidemiological technique like Mendelian randomization.

### 2.2.1 Cohort studies

In modern epidemiological studies, a cohort refers to a group of individuals with specific characteristics who are under investigation to establish the incidence of an occurrence of a disease outcome (Song and Chung, 2010). This method requires a researcher to first select a disease-free group by the exposure. The researcher then has to follow up on the cohort on schedule until the outcome of interest occurs (Song and Chung, 2010). Dekkers *et al.* (2012) noted that there are two types of cohort methodology; prospective and retrospective. According to them, prospective studies start from the present into the future and use specified data collection techniques. This feature makes it advantageous in that it can acquire specific exposure data and has a high probability of getting complete information (Dekkers *et al.*, 2012). However, this method is vulnerable to high loss of data if the event of interest takes long to occur. Retrospective cohort studies, on the other hand, use the past acquired data to analyze the exposure-outcome link in the present (Lucas and McMichael, 2005). This method is less costly and takes a shorter time compared to the prospective cohort study. However, Song and Chung (2010) deduce that the researcher has limited control over data collection and thus the information put into use may be incomplete.

Many researchers have used cohort studies to determine a relationship between exposure and outcome. For example, Huxley *et al.*, (2006) investigated the relationship between fatal coronary heart disease and diabetes in men and women using 37 prospective cohort studies. The researchers found out that diabetic patients had a high probability of contracting fatal coronary heart disease than non-diabetic individuals (5.4% vs. 1.6%). Apart from that, the researchers concluded that women had a 50% higher chance of fatal coronary heart disease than men (Huxley *et al.*, 2006). However, they noted that the study had some limitations. For instance, the researchers indicated that they did not exclude potential confounding



effects from the menopausal situation and hormone replacement therapy because the information was not readily accessible. Besides that, the researchers did not address the issue of reverse causation in their investigation hence weakening the prospective cohort design.

### **2.2.2 Case-control studies**

This technique got its recognition first in 1926 when Janet Lane-Claypon used the method to show that the low fertility rate increases the chances of contracting breast cancer (Song and Chung, 2010). This research design requires the researcher to start by selecting a group of people with the disease of interest, case-patients. The investigator should then enroll the individuals without the disease to act as a control group (Dicker *et al.*, 2006). The two groups of individuals should come from the same population. The examiner then compares the previous exposure data between the two selected categories. The main reason for having a control group is to act as the baseline or to give the expected degree of exposure in the population (Dicker *et al.*, 2006). Case-control studies are efficient, especially, when investigating rare diseases or incidences with long latency (Song and Chung, 2010). Besides that, Dicker *et al.*, (2006) indicated that there are suitable in studying dynamic populations where follow-ups are difficult. However, the method is vulnerable to bias, especially, in the selection stage.

Many epidemiologists have put into use the case-control research design in establishing the causal link between the exposure and the outcome. For example, Ness *et al.*, (2002) conducted research trying to establish if there is a causal relationship between impotence and reproductive drug use and ovarian cancer using this method. The researchers used 8 case-control studies carried out between 1989 and 1999. The study had 5,207 case-patients

and 7,705 controls (Ness *et al.*, 2002). The researchers found out that some specific biological causes of impotence and not the use of fertility medicines increases the chances of developing ovarian cancer (Ness *et al.*, 2002). However, the researchers did not explain the possibility of confounding factors and reverse causation affecting the study. This situation indicates the limitation of case-control studies and thus there is a need to adopt a better method. One such a sophisticated epidemiological technique is Mendelian Randomization. In section 2.2 we discuss this method.

### **2.3 Mendelian Randomization**

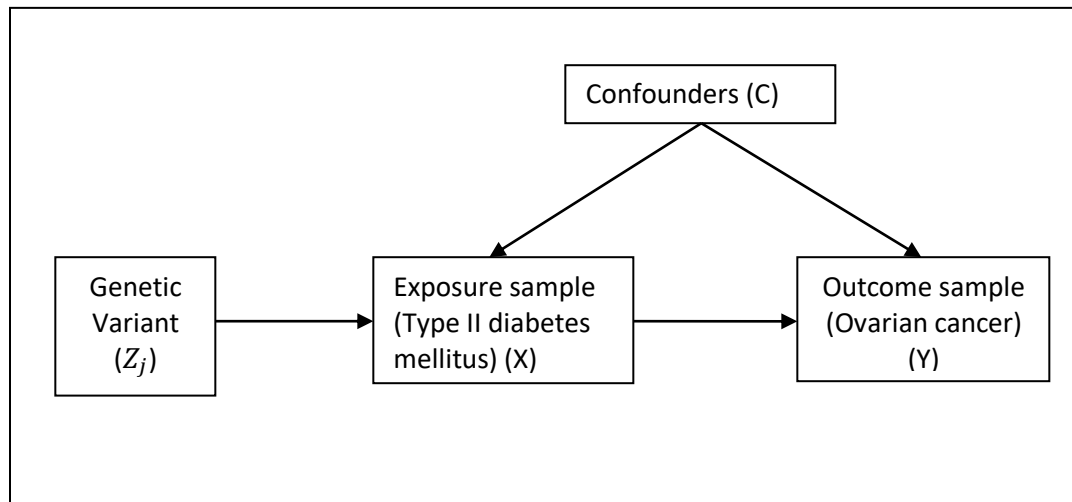
MR is a research model that assists in establishing the causal relationship between a modifiable predisposing factor (exposure) and the outcome. Sheehan *et al.*, (2008) noted that it uses instrumental variables, which makes Mendelian randomization paradigm to be recommendable because they mimic the random allocation of genetic variants to the risk factors. This situation, therefore, ensures that confounding factors and reverse causation does not alter the causal analysis. Burgess (2012), on the other hand, noted that this method got its basis from Grigor Mendel's two laws of inheritance of 1866 (law of segregation and independent assortment). He defines the instrumental variables as exogenous components with endogenous exposure, and therefore, helps in estimating the causal effect of altering the predisposing factor while maintaining other variants constant. Sekula *et al.*, (2016) added that the use of genetic variants in the method makes it not to be prone to reverse causation because they are non-modifiable. Grover *et al.*, (2017) indicate that the most genetic variants applied in this research design in single-nucleotide polymorphisms (SNPs) because their allele's allocations occur randomly at conception before any exposure or outcome.

Different study designs for Mendelian randomization exist, for instance, bi-directional MR, two-step MR, and two-sample MR. Bi-directional MR enables the researchers to investigate if the chosen exposure causes the outcome or vice versa. This method is advantageous because it can determine if latent confounding may be causing the correlation. However, the interpretation of its results may be difficult due to the complexities of biology such as feedback loops (Zheng *et al.*, 2017). The two-step MR method helps in assessing if an intermediate trait plays a role of causal mediation between a selected exposure and the outcome, hence enabling the estimation of direct and indirect effects. However, this technique does not consider linearity and homogeneity assumptions. This study uses a two-sample MR method in the determination of the existence or non-existence of the causal relationship between T2DM and ovarian cancer. Its discussion is in section 2.2.1. These MR models have enabled the researchers to determine the causal link between the exposures and outcomes in different circumstances using the most suitable design.

### **2.3.1 Two-sample Mendelian randomization**

Zheng *et al.*, (2017) noted that two-sample MR enables the researchers to estimate the causal effects in a case where exposure and outcome data are from different samples. Lawlor (2016) stated that it is not necessary to obtain the genetic data from the same population. The researcher indicated that using two independent populations will help to sideline ‘winners’ curse’ that would have led to underestimation of the true causal effects if one group of individuals was used. Apart from that, Lawlor (2016) indicated that using two samples in the causal relationship analysis reduces the effects of weak instruments hence increasing the probability of obtaining true causal estimates. This concept makes this model suitable for the study because there are two samples; T2DM and ovarian cancer.

Zheng *et al.*, (2017) indicated that this method is also advantageous because it greatly increases the scope of Mendelian randomization analysis.



**Figure 2.1. Two-Sample MR Conceptual Framework Showing the Causal Relationship between Type 2 Diabetes Mellitus and Ovarian Cancer**

The conceptual framework indicates the causal association between the risk factor (X), T2DM and the outcome (Y), ovarian cancer using the instrumental variants (SNPs). This model indicates that the confounders (C) do not alter the genetic variant. Some of the confounders in this study may include familial history, gene mutation, menstrual periods, and the use of oral contraceptives. Mendelian randomization method has three core assumptions that need fulfillment for it to give unbiased results (Grover *et al.*, 2017). The assumptions are as follows:

1. The genetic variant should have a strong relationship with the exposure.
2. The genetic variant should be independent of the confounding factors.
3. The genetic variant should only have a relationship with the outcome via exposure.

Walker *et al.*, (2019) noted that SNPs are credible instruments for determining the causal association between an exposure and a disease outcome because of their random allocation

at conception and thus free from subsequent alteration from environmental factors. This insight deduces that if the three assumptions hold, the resulting MR effect estimates are not due to confounding and reverse causation. Assumption one needs biological support indicating that the gene that encodes the exposure biomarker has the selected genetic variant (Sekula *et al.*, 2016). This assumption is empirically verifiable since the researcher can use the F statistic, odds ratio, risk ratio, and regression coefficient ( $r^2$ ) to estimate the relationship. Assumptions two and three are not empirically verifiable, however, they are testable to some extent. Koellinger and De Vlaming (2019) noted that large-scale Genome-Wide Association Studies (GWASs) have led to the discovering of various genetic loci for many risk factors. This discovery has led to the use of many SNPs in determining the causal link between the exposure and the outcome hence addressing assumption two.

#### **2.4 Horizontal Pleiotropy**

Swerdlow *et al.*, (2016) define pleiotropy as a case where the genetic variant has a relationship with more than one phenotype. There are two categories of pleiotropy; horizontal and vertical. Horizontal pleiotropy refers to an instance when the SNP has a relationship with the outcome through the pathway that is independent of the risk factor of interest (Type 2 Diabetes Mellitus). This manifestation invalidates assumption two of Mendelian randomization. This issue, therefore, led to the formulation of objective two of the study which aimed at determining whether or not there exist horizontal pleiotropy in the data used. In contrast, vertical pleiotropy occurs when the SNP has a relationship with other phenotypes that may present themselves in between the risk factor and the outcome (Walker *et al.*, 2019). The adoption of stratified analysis or the use of allelic scores to exclude vertically pleiotropic variants helps in avoiding the occurrence (Sekula *et al.*, 2016).

Mendelian randomization technique has some threats and epidemiological researchers need to be keen on them. The researchers should ensure that they do not use weak genetic instrumental variants since it may give a biased effect estimate of the relationship between the exposure and the outcome (Sekula *et al.*, 2016). One of the ways to solve this threat is to stick to the Genome-wide significance threshold (P-value  $<5 \times 10^{-8}$ ). Davies *et al.*, (2018) explained that this threshold helps in reducing the number of false-positive relationships emanating from the statistical tests. Besides that, the researchers should ensure that they use large sample sizes, which will help in increasing the model effect size. Some of the indicators of effect size are the beta-coefficient, regression coefficient ( $r^2$ ), and the F-statistic. Swerdlow *et al.*, (2016) suggested that SNPs with  $F > 10$  are suitable for the Mendelian randomization analysis. Linkage disequilibrium (LD) is another problem encountered in MR. Linkage disequilibrium refers to an instance where the alleles have a statistical relationship at different loci (Sheehan *et al.*, 2008). This situation may lead to the occurrence of confounding in the study and thus the researchers need to select only independent genetic variants as instruments (Sekula *et al.*, 2016).

## **2.5 Heterogeneity**

Another problem that may occur is genetic heterogeneity. This threat arises when several alleles can influence the selected phenotype at different genetic loci (Sheehan *et al.*, 2008). Heterogeneity in the study leads to inconsistencies, and thus the researchers need to test for this issue using Cochran's Q statistic (Higgins *et al.*, 2003). Bowden *et al.*, (2019) stated that having extreme heterogeneity shows that either there is model assumption violation or some of the genetic variants used are weak instruments. Apart from that, the researchers can employ  $I^2$  statistics to estimate the percentage of the total discrepancy in the study that is due to heterogeneity and not chances. Higgins *et al.* (2003) suggested that percentage

values of 25%, 50%, and 75% represent low, moderate and high heterogeneity in the study, respectively. This issue led to the development of objective one which necessitated the study to investigate whether or not the data used in the analysis is homogenous.

## **2.6 Mendelian Randomization Analysis Techniques**

Mendelian randomization design has various methods for obtaining the causal estimates, for instance, Wald Ratio, Maximum Likelihood, Median-based techniques, Model-based systems, Inverse Variance Weighted, MR-Egger. Walker *et al.*, (2019) indicate that the Wald ratio is suitable when one IV is available. This method divides the regression coefficients of the outcome and the exposure. The Maximum Likelihood method utilizes the odds of the model, which focuses on the exposure-outcome relationship and the spread of the genetic relationship estimates. The median-based methods (simple, weighted, penalized-weighted) obtains the Wald ratios for each instrumental variable and picks the median value as specified by the technique to act as the required estimate (Walker *et al.*, 2019). On the other hand, model-based methods for example, simple model-based, weighted model-based techniques generate clusters from the causal effects of each instrumental variable. The largest cluster of the IVs will then provide the causal effect estimate for the study. Walker *et al.*, (2019) state that model-based methods require zero-mode bias terms for each IV. Inverse Variance Weighted technique also generates the Wald ratios and integrates the output using the meta-analysis method. This approach uses either fixed or multiplicative random effect models. The slope of the model gives the required causal estimates of the study. MR-Egger also utilizes Wald ratios for each SNP and uses an adapted Egger regression technique to amalgamate the results (Walker *et al.*, 2019). The Egger regression slope coefficient indicates the causal effect while the intercept of the Egger regression provides the average pleiotropic effect of the genetic instrumental

variables (GIV) under study. This research study focused on Inverse Variance Weighted (IVW) and MR-Egger techniques since they can work well when using summarized data and are capable of optimizing the likelihood function i.e. reducing estimator variance.

Assume that all the associations between variables in Fig 2.1 in section (2.2.1) are linear and that there is no effect adjustment. Besides that, assume that all the genetic variants  $Z_j$  ( $j = 1, 2, \dots, N$ ) have no linkage disequilibrium (LD). Let  $\beta_{Y_j}$  denote the association between the genetic instrument variable  $Z_j$  ( $j = 1, 2, \dots, N$ ) and the outcome  $Y$  (ovarian cancer). Also, let  $\beta_{X_j}$  represent the relationship between the genetic variant  $Z_j$  ( $j = 1, 2, \dots, N$ ) and the exposure  $X$  (type 2 diabetes). Burgess and Thompson (2017) suggested that it is possible to breakdown the relationship between the genetic instrument variable  $Z_j$  ( $j = 1, 2, \dots, N$ ) and the outcome  $Y$  (ovarian cancer) into the sum of pleiotropic (direct) and causal (indirect) effects as follows:

$$\beta_{Y_j} = \alpha_j + \theta \beta_{X_j} \tag{2.1}$$

where,

$\beta_{Y_j}$  Indicates the relationship between the genetic instrument  $Z_j$  ( $j = 1, 2, \dots, N$ ) and the outcome  $Y$ , ovarian cancer

$\alpha_j$  represents the effect of the GIV on the outcome that is not through the exposure (pleiotropy)

$\theta$  represents the causal effect of the risk factor  $X$ , type 2 diabetes on the outcome  $Y$ , ovarian cancer.

$\beta_{X_j}$  shows the association between the genetic variant  $Z_j$  ( $j = 1, 2, \dots, N$ ) and the exposure  $X$ , type 2 diabetes mellitus



### 2.6.1 Inverse variance weighted method

Assume that there is one genetic variable  $Z_j$  that fulfills the Mendelian randomization assumptions given in section 2.2.1. If that is the case, Burgess and Thompson (2017) indicates that simple ratio of association can help to estimate the causal effect of the exposure (T2DM) on the outcome (ovarian cancer) as follows:

$$\hat{\theta}_j = \frac{\hat{\beta}_{Y_j}}{\hat{\beta}_{X_j}} \quad (2.2)$$

where,

$\hat{\theta}_j$  represents the estimated causal effect of the exposure  $X$ , T2DM, on the outcome  $Y$ , ovarian cancer

When the research study uses multiple genetic instrumental variables, each simple ratio estimates need averaging using a meta-analysis approach to give inverse-variance weighted (IVW) estimates (Burgess and Thompson, 2017). This approach assumes that all the genetic instrumental variables have no linkage disequilibrium (LD) hence giving independent evidence on the causal effect. The delta expansion method provides the variance of the simple ratio estimate, which is as follows:

$$\frac{se(\hat{\beta}_{Y_j})^2}{\hat{\beta}_{X_j}^2} \quad (2.3)$$

where,

$se$  is the standard error

Therefore, assuming the fixed-effect model then the IVW estimate is:

$$\hat{\theta}_{IVW} = \frac{\sum_j \hat{\beta}_{Y_j} \hat{\beta}_{X_j} se(\hat{\beta}_{Y_j})^{-2}}{\sum_j \hat{\beta}_{X_j} se(\hat{\beta}_{Y_j})^{-2}} \quad (2.4)$$

where,

$\hat{\theta}_{IVW}$  is inverse variance weighted estimate

Weighted linear regression can also give the same results using inverse-variance weights

$$se\left(\hat{\beta}_{Y_j}\right)^{-2} \quad (2.5)$$

The condition for weighted linear regression is that there is no intercept in the model, for instance:

$$\hat{\beta}_{Y_j} = \hat{\theta}_{IVW}\hat{\beta}_{X_j} + \epsilon_{I_j} \quad (2.6)$$

where,

$$\epsilon_{I_j} \sim \mathcal{N}\left(0, \sigma^2 se\left(\hat{B}_{Y_j}\right)^{-2}\right) \quad (2.7)$$

where,

$\epsilon_{I_j}$  represents the error term

Burgess and Thompson (2017) suggested that in a scenario where there are no pleiotropic effects ( $\alpha_j = 0$ ), then  $\hat{\theta}_j$  will give a consistent estimate of the causal effect. This situation indicates that  $\theta_{IVW}$  also gives a consistent estimate of the causal effect when pleiotropy is absent. However, the inverse-variance weighted method is not suitable when at least one violation of the Mendelian randomization assumptions.

### 2.6.2 MR-egger method

This method is a technique that helps in analyzing summarized genetic data. This method can determine the existence of directional pleiotropy, test for the causal effect, and can also be an estimate for the causal relationship between the exposure and the outcome (Sekula *et al.*, 2016). Burgess and Thompson (2017) noted that MR-Egger is a powerful Mendelian randomization method because it can establish if the genetic instrumental variables have a

directional pleiotropic effect on the outcome (ovarian cancer) and give a consistent measure under weaker assumptions called Instrument Strength Independent of Direct Effect (InSIDE). InSIDE, in this case, means that the pleiotropic effects  $\alpha_j$  have an independent distribution from the genetic relationships with the exposure  $\hat{\beta}_{X_j}$  (type 2 diabetes mellitus).

MR-Egger is a modified weighted linear regression (Equation 2.4) since it takes the intercept to be a non-zero. Therefore, its equation is as follows:

$$\hat{\beta}_{Y_j} = \theta_{0E} + \theta_{1E}\hat{\beta}_{X_j} + \epsilon_{E_j} \quad (2.8)$$

where,

$$\epsilon_{E_j} \sim \mathcal{N}\left(0, \sigma^2 se\left(\hat{\beta}_{Y_j}\right)^{-2}\right) \quad (2.9)$$

where,

$\theta_{0E}$  represents the intercept,

$\theta_{1E}$  indicates the slope

MR-Egger estimate will be equal to the IVW measure if the intercept is zero. Under the weaker assumption (InSIDE), if the sample sizes and SNPs number increase, the MR-Egger method will give a consistent causal estimate. In a scenario where there is a fixed number of instrumental variables, as the sample size increases the MR-Egger estimate is consistent provided that the inverse-variance weights in (2.5) are equal to zero. Burgess and Thomson (2017) indicate that MR-Egger is a significant sensitivity method but it may give biased estimates and inflate Type 1 error rate due to the impacts of the outliers and violation of InSIDE assumptions.

## 2.7 Two-Sample Mendelian Randomization Application

Most epidemiologists have applied two-sample MR to determine the causal relationship between a particular predisposing factor and the outcome. For example, Libuda *et al.*, (2019) carried a research study aiming to establish if vitamin D and depression have a causal relationship. The researchers used six genome-wide significant genetic variants from GWAS. This selection ensured the satisfaction of Mendelian randomization assumptions. The MR analysis depicted that vitamin D and broad depression had no causal relationship (IVW;  $b=0.025$ ,  $SE= 0.038$ ,  $P = 0.52$ ). This conclusion is against the results from observational studies suggesting that they might be the victims of confounding and reverse causation (Libuda *et al.*, 2019).

Seddighi *et al.*, (2019) also employed two-sample MR to test the causal relationship between cancer and Alzheimer's disease. The researchers selected the SNPs that attained the standard threshold ( $P\text{-value} < 5 \times 10^{-8}$ ) and ensured that there was no linkage disequilibrium to satisfy the Mendelian randomization assumptions. The researchers concluded that cancers have a relationship with lower odds of incident Alzheimer's disease (Seddighi *et al.*, 2019). This study ensured that there was no confounding or reverse causation.

Gage *et al.*, (2018), on the other hand, employed two-sample MR to investigate the causal relationship between education and smoking. He noted most people believe that lower educational attainment has a relationship with rising cases of smoking. However, ascertaining the causality was a challenge, hence led to the use of two-sample MR. The researchers used summary statistics from GWAS and applied complementary MR

paradigms like IVW, MR-Egger, and weighted median regression. He stated that the results were consistently indicating the causal relationship.

## **2.8 Knowledge Gap**

Most researchers have acknowledged that observational studies like cohort and case-control studies may give biased results when carrying out a causal relationship investigation due to the issues of confounding and reverse causation (Wang *et al.*, 2017; Urpilainen *et al.*, 2018.) Two-sample MR is an epidemiological method that takes into consideration the issues of confounding and reverse causation while undertaking the causal relationship between an exposure and the outcome since it utilizes genetic variants that undergo random allocation at conception and are non-modifiable as instrumental variables (Sekula *et al.*, 2016). Therefore, the application of two-sample MR to investigate the causal relationship between T2DM and ovarian cancer is expected to give better results.

## **CHAPTER THREE**

### **RESEARCH METHODOLOGY**

#### **3.1 Introduction**

This chapter describes the data sources for the study. It also discusses the two-sample Mendelian randomization analysis techniques employed in the investigation. Apart from that, it outlines the data analysis and sensitivity tests procedures.

#### **3.2 Study Design**

In this study, the data, Type 2 Diabetes Mellitus and ovarian cancer SNPs were ensured that they were homogeneous using the Cochran's Q and I-squared statistics. The study also ensured that T2DM and ovarian cancer SNPs data did not exhibit horizontal pleiotropy. This was done using the MR Egger Intercept. The next step used two-sample MR design that is, IVW and MR-Egger to investigate the causal relationship between T2DM (exposure) and ovarian cancer (outcome).

#### **3.3 Data Sources**

The summary statistics data was retrieved from two different GWAS consortiums. The exposure genetic (T2DM) data was obtained from the DIAGRAMplusMetaboChip consortium. The outcome genetic (ovarian cancer) data was obtained from Ovarian Cancer Association Studies (OCAC).

One of the important reasons for understanding the concepts of human genetics is that it enables the scientists to come up with improved methods of disease diagnosis and treatment. Waddington (2016) defined a gene as unit that contains the hereditary information about a particular individual. On the other hand, deoxyribonucleic acid (DNA)

is a complex molecule that has the genetic information coding that helps in transmitting the hereditary traits. According to Crow (2017), DNA in every individual is composed of the same chemical units, which are adenine (A), thymine (T), guanine (G), and cytosine (C). Usually, strands A matches with T while C go with G. Alliance (2009), indicated that the strands forms the genetic sequences that assist in the transmission of the hereditary traits called the alleles.

Some of the studies show that all individuals are about 99.9% genetically similar (Crow, 2017; Waddington, 2016; Alliance, 2009). However, they indicated that the differences come about because of the genetic variations such as mutation and polymorphisms that makes one to differ in form of physical traits, and the level of risk for certain diseases. DNA sequence variation occurs when a single nucleotide is tempered, which leads to the formation of SNP. According to the National Institute of health (2019), SNPs are biological markers that assist in predicting how an individual response to certain drugs, susceptibility to environmental chances and the risk to developing a particular disease. This suggestion indicates that the SNPs help in linking the two datasets from different consortium.

### **3.4 Study Population**

The Type 2 Diabetes Mellitus (exposure) data was based on a study done by Morris *et al.*, (2012) using DIAGRAMplusMetaboChip consortium. They used a sample size of 149,821 (ncase=34,840, ncontrol=114,981) of mixed population. The population had both males and females. The ovarian cancer (outcome) data was based on a study done by Phelan *et al.*, (2017) using the summary statistics from Ovarian Cancer Association Consortium (OCAC). This study used a sample size of 66450 (ncase=25,509, ncontrol=40,941) of European population. This thesis required harmonization and clumping of the data which

were used by Morris *et al.* (2012) and that was used by Phenal *et al.*, (2017) using the MR Base platform. From then on, Cochran's Q statistic helped to determine whether or not the data were homogenous. The next step was to establish whether or not horizontal pleiotropy exists using the MR-Egger method. Finally, the study used the two-sample Mendelian randomization technique to establish the causal relationship between T2DM and ovarian cancer.

MR-Base assisted in the implementation of the Two-sample Mendelian randomization technique because it is a database that can access summary statistics from Genome-Wide Association Studies. R statistical software version 3.5.1 was used in the analysis which led to finding the causal estimates and performs sensitivity analysis of the Two-sample MR causal model.

### **3.5 Mendelian Randomization Analysis to Determine the Causal Relationships between T2DM and Ovarian cancer**

The study started with the extraction of the SNPs associated with the risk factor (T2DM) and clumping the data to ensure the independence of the exposure instrumental variables. This process involves ensuring sufficient fulfillment of the standard threshold for Genome-wide significance ( $p - value < 5 \times 10^{-8}$ ). Apart from that, it was ensured that the genetic variants extracted have no linkage disequilibrium by defining the regression coefficient to be less than two ( $r^2 < 0.2$ .)

The next step was to retrieve the exposure SNPs from the outcome trait (ovarian cancer). In this process, if the required SNP is not available in the outcome (ovarian cancer) GWAS, the MR-Base will provide an SNP proxy that is in LD with the wanted instrumental variable. Therefore, it was necessary to ensure that the regression coefficient  $r^2 > 0.8$  to



give a probability of finding a particular proxy. The study also allowed palindromic SNP, which indicates that the alleles on the forward strand are the same as those on the reverse strand, hence enabling the study to use the allele frequency information.

The next phase was to harmonize the extracted data. This process ensured that the exposure and the outcome SNPs' effects must correspond to the same allele hence giving a new data frame that has combined both the variables. There are three ways of harmonizing the data. The first option was to assume that all the alleles follow the forward strand. The other one is to analyze the forward strand alleles by understanding their allele frequency information. The last option was to rectify the strands for non-palindromic SNPs and eliminate all the palindromic genetic instruments (Burgess and Thompson, 2017). The study made use of the allele frequency information to infer the forward strand, which is option two.

This kind of study requires using the data that is homogenous and has no horizontal pleiotropy so as to give reliable and consistent results. Therefore, this study had to check the homogeneity of the data using the Cochran's Q statistic which gives the level of the heterogeneity. Apart from that, the  $I^2$  statistic was employed to indicate the percentage of the variation in the data. This process of checking the homogeneity of the data using Cochran's Q and  $I^2$  statistics formed the first objective of the study. The second objective which revolved around the horizontal pleiotropy was analyzed using the MR-Egger intercept. After achieving the two constraints checks of this study, it was possible to move to objective number three, which concerned determining whether or not the causal relationship between Type 2 Mellitus and ovarian cancer using IVW and MR-Egger techniques.

Burgess and Thompson (2017) stated that the slope coefficients of IVW and MR-Egger models give the causal relationship estimates. The scatter plot enabled the researcher to determine the direction of the causal relationship between T2DM and ovarian cancer. Besides that, the MR-Egger regression intercept helped in determining the pleiotropy in the study (Burgess and Thompson, 2017.)

The output included the MR results table and the method comparison graph (scatter plot). According to Walker *et al.*, (2019), the MR result's table gives the causal estimates from each MR method (IVW and MR-Egger). The method comparison plot, on the other hand, depicts the effect of the IVs on the exposure (Type 2 Diabetes Mellitus) against the effect of the ovarian cancer SNPs.

### **3.6 Sensitivity Analysis**

Mendelian randomization method is one of the recommendable epidemiological methods for testing the causal relationship between the predisposing factor and the disease outcome as long as the model fulfills its instrumental variables assumptions. Heterogeneity refers to a case where the research data are inconsistent hence may give non-reliable estimates. Therefore, it is crucial to test for heterogeneity when performing Mendelian randomization. This ensures the fulfillment of the instrumental variable assumptions. Bowden *et al.*, (2018) state that Cochran's Q statistic can help in estimating heterogeneity among the ratio estimates. This is a method that assists in determining whether or not heterogeneity exists in the data where the outcome is binary. According to Hoaglin (2016), Cochran's Q statistic has some assumptions that need to be fulfilled. Let  $k$  be binary measurements and  $N$  to be subject which may be a set of matched variables. Let also  $Y_{i,j}$  be the binary response from the subject  $i$  in category  $j$  ( $i = 1$  to  $N$ ,  $j = 1$  to  $k$ ). Then, one of them states that the responses should be binary and should come from  $k$  matched samples. The other

assumption indicates that the variables should be independent and that they were selected randomly from a population considered to be large. The last assumption states that the sample size should be sufficiently large, that is,  $n \geq 4$  and  $nk \geq 24$  (Hoaglin, 2016).

Therefore, the Cochran's Q statistic is given as:

$$Q = \frac{(k-1)[kC - T^2]}{kT - R} \quad (3.1)$$

Where,

$$C = \sum_{j=1}^k (\sum_{i=1}^N Y_{i,j})^2 \quad (3.2)$$

$$T = \sum_{i=1}^N (\sum_{j=1}^k Y_{i,j}) \quad (3.3)$$

$$R = \sum_{i=1}^N (\sum_{j=1}^k Y_{i,j})^2 \quad (3.4)$$

It is also noted that if the instrumental variables under study are valid that is, fulfills the model assumptions, then Cochran's Q statistic will follow the Chi-square distribution asymptotically with N-1 degrees of freedom where N is the total number of SNPs used. Besides that,  $I^2$  helped in estimating the variation percentage given by the following equation;

$$I^2 = \left( \frac{Q - df}{Q} \right) \times 100 \quad (3.5)$$

where,

Q represents the Cochran's Q statistic

df indicates the degree of freedom

Walker *et al.*, (2019) stated that R gives a heterogeneity statistics table containing the variations in the causal estimate across the instrumental variables used. Apart from that, the odds ratios (exponent of beta coefficients) were used to give the power of the analysis. The odds ratios will give a detection of increase or decrease in variability per 1 standard deviation (SD.)

### **3.7 Results Presentation**

The results were presented in form of tables (MR results, heterogeneity statistic, and horizontal pleiotropy) and plots (method comparison graph, leave-one-out, forest and funnel plots).

### **3.8 Ethical Consideration**

The research acknowledged all the works of other scholars and cited them as is appropriate to avoid plagiarism.

## CHAPTER FOUR

### RESULTS AND DISCUSSIONS

#### 4.1 Introduction

This chapter gives an overview of data analysis procedure, the outcome of the process and the discussions.

#### 4.2 Data Analysis

“R” statistical software version 4.0.3 (2020-10-10) aided in running the codes for this study. Besides that, MR-Base was used as an online platform, which provides an interface allowing MR analyses and sensitivity tests to be performed (Walker *et al.*, 2019).

The exposure variable in this study was T2DM whose GWAS ID is “ieu-a-24”. It was retrieved from DIAGRAMplusMetaboChip consortium. On the other hand, the outcome variable was ovarian cancer whose GWAS ID is “ieu-a-1120”. It was retrieved from Ovarian Cancer Association Consortium (OCAC). Other packages required in R during the analysis include: “devtools”, “TwoSampleMR”, “digest”, “githubinstall”, and “googleAuthR”.

The first process was to extract the exposure data, Type 2 Diabetes Mellitus and clump them. This move assisted in identifying the independent alleles among the correlated SNPs (Walker *et al.*, 2019). This was fulfilled by sticking to the GWAS standard threshold significance level ( $p - value < 5 \times 10^{-8}$ ) and defining the regression coefficient to be  $r^2 < 0.2$  which ensured that there was no linkage disequilibrium.

The next step was to list all the available outcomes in the MR-Base platform and extract the exposure data, Type 2 Diabetes Mellitus. In this case, the use of SNP proxy was allowed, which is in LD with the targeted SNP. This was achieved by defining the minimum r-square to find the SNP proxy to be 0.8, that is  $r^2 > 0.8$ . On the other hand, it was assumed that all the alleles are aligned in the forward strand. According to Walker *et al.*, (2019), palindromic SNPs refer to a situation where the pair of alleles on the forward-strand are the same as those on the reverse strand. This study infers the palindromic SNPs and the maximum minor allele frequency acceptable threshold are defined by 0.3.

The next step was to harmonize the exposure and the outcome data. Walker *et al.*, (2019) defined harmonization as a way of specifying the effect and other alleles in the same way in both the exposure and outcome data. In this study, interpretation of the forward strand was by use of the allele frequency information. From this action, a new data frame that had a combination the Type 2 Diabetes Mellitus and ovarian cancer was obtained.

After obtaining the required data frame that contains the targeted exposure and outcome variables, it was necessary to ensure that the data was homogenous and that there was no horizontal pleiotropy. The homogeneity of the data was checked using the Cochran's Q statistic which indicated the level of heterogeneity. The  $I^2$ , on the other hand, assisted in indicating the percentage of the variation in the data. This action helped in answering the objective number one of the study. The MR-Egger intercept helped in determining whether or not the horizontal pleiotropy existed. This step assisted in ensuring that objective number two of the study is achieved.

At this point, it is possible to perform the two-sample Mendelian randomization analyses as highlighted as objective number three of the study. This gave the results of five different

Mendelian randomization methods; MR-Egger, weighted median, IVW, simple mode, and weighted mode. Since this study focused mainly on MR-Egger and IVW methods, the scatter plot was restricted to only depict these two techniques.

### **4.3 Study Results**

Thirty nine (39) variants of the exposure variable (T2DM) were retrieved after clumping the data. However, the process found the proxies for 3 SNPs in the outcome data (ovarian cancer.) After harmonizing the study data, 3 SNPs (rs10830963, rs1801282, rs243088) were found to be palindromic with intermediate allele frequencies hence eliminated. Therefore, the study used 33 SNPs to determine whether or not T2DM and ovarian cancer have a causal relationship.

The results showed that the exposure p-values were less than 0.05, which suggested that T2DM was strongly associated with the targeted SNPs. Contrary to that, the outcome p-values were greater than 0.05 indicating that ovarian cancer was only associated with the targeted SNPs via the exposure. This situation shows that this study fulfills the assumptions of Mendelian randomization. The F-statistic ( $F= 65.269$ ,  $P = 0.000$ ) for this research study was greater than 10, the GWAS standard threshold. This reflects that the SNPs used in this study were considered to be strong instrumental variables as suggested by Swerdlow *et al.*, (2016).

#### **4.3.1 Homogeneity of the study data**

Heterogeneity is one of the factors that were analyzed when carrying out two-sample Mendelian randomization. The data used should be homogenous so that the results can be reliable and give consistent results. Therefore, it was necessary to test whether or not the

data used in the study was homogenous as per the objective one. The results are shown in Table 4.1 and Figure 1.

**Table 4.1.**

**The Homogeneity Results of the Causal Relationship between Type 2 Diabetes Mellitus and Ovarian Cancer**

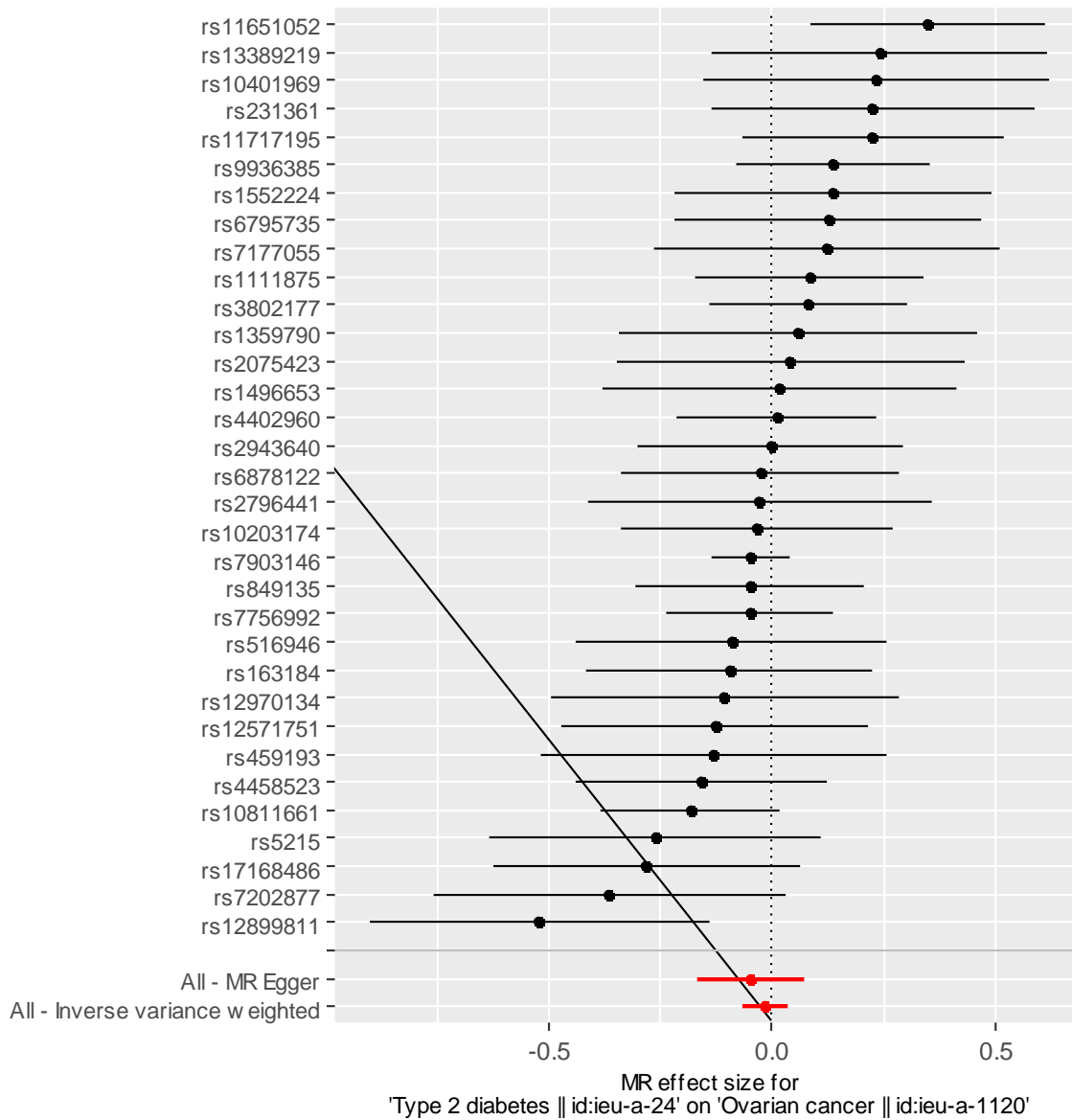
Method	Q	Q_df	Q_pval	I-sq
MR Egger	39.387	31	0.143	21.293
Inverse variance weighted	39.776	32	0.162	19.550

*Q: Cochran's Q statistic, df: degree of freedom, pval: p-value, I-sq: I-square statistic*

According to this study, the first objective ought to find out whether or not the study data was homogenous. The results indicated that all 33 genetic variants were not heterogeneous (MR-Egger:  $Q=39.387$ ,  $p\text{-value}=0.144$ , IVW:  $Q=39.778$ ,  $p\text{-value}=0.162$ ) as shown in Table 4.1. The corresponding p-values are 0.144 and 0.162, which are greater than 0.05. This indicates that there was sufficient evidence to conclude that the data used was homogenous.

The I-square statistics (MR-Egger=21.293, IVW=19.776) also indicated low heterogeneity in the study since they were less than 25%. This indicates that about 21.29% and 19.54% variance in this study is due to heterogeneity rather than chance. According to Higgins *et al.*, (2003), the I-square statistics, which are less than 25% indicate low variances due to heterogeneity hence negligible.





**Figure 4.1. Forest Plot Displaying the Results of Single and Multi-SNP Analyses on the Causal Relationship between Type 2 Diabetes Mellitus and Ovarian Cancer**

The forest plot in Figure 4.1 displays the causal estimates of each SNP utilized in this study. This was generated using the Wald ratio. Apart from that, the graph indicates the multi-SNP causal estimates using the MR-Egger and IVW techniques. This forest plot has no discrepancy in the causal estimates displayed. Therefore, this indicates that the data

(T2DM and ovarian cancer SNPs) used in this study were homogeneous hence the results achieved will be considered reliable.

#### 4.3.2 Horizontal pleiotropy

One of the assumptions of the Mendelian randomization states that the outcome variable should only be associated with the genetic variants through the study exposure factor. The presence of horizontal pleiotropy invalidates this assumption. Therefore, it was necessary to test for this occurrence by interpreting the MR-Egger intercept as per the objective two of the study. The results are shown in Table 4.2

**Table 4.2.**

**The MR-Egger Intercept Output of the Causal Relationship between Type 2 Diabetes Mellitus and Ovarian Cancer**

Egger_intercept	Se	pval
0.004	0.004	0.584

*se: standard error, pval: p-value*

The second objective of the research aimed at finding out if there was presence of the horizontal pleiotropy in the data. The results showed that there was no or minimal horizontal pleiotropy in this study since the MR-Egger intercept was 0.0038 (p-value=0.5839). The p-value (0.5839) was greater than 0.05 hence there was sufficient evidence to conclude that there was no horizontal pleiotropy in the data used in the study. This indicated that the research study fulfilled the Mendelian randomization assumption that states that the selected SNPs should be associated with the outcome via the exposure.

### 4.3.3 Causal relationship between Type 2 Diabetes Mellitus and ovarian cancer

The main aim of this research study was to investigate the causal relationship between T2DM and ovarian cancer. The results of this objective were obtained by interpreting the beta coefficients of the methods employed which gave the causal effects. Five Mendelian randomization techniques were displayed in this section, however, this study focused on MR-Egger and IVW methods as shown in Table 4.3 and Figure 4.2.

**Table 4.3.**

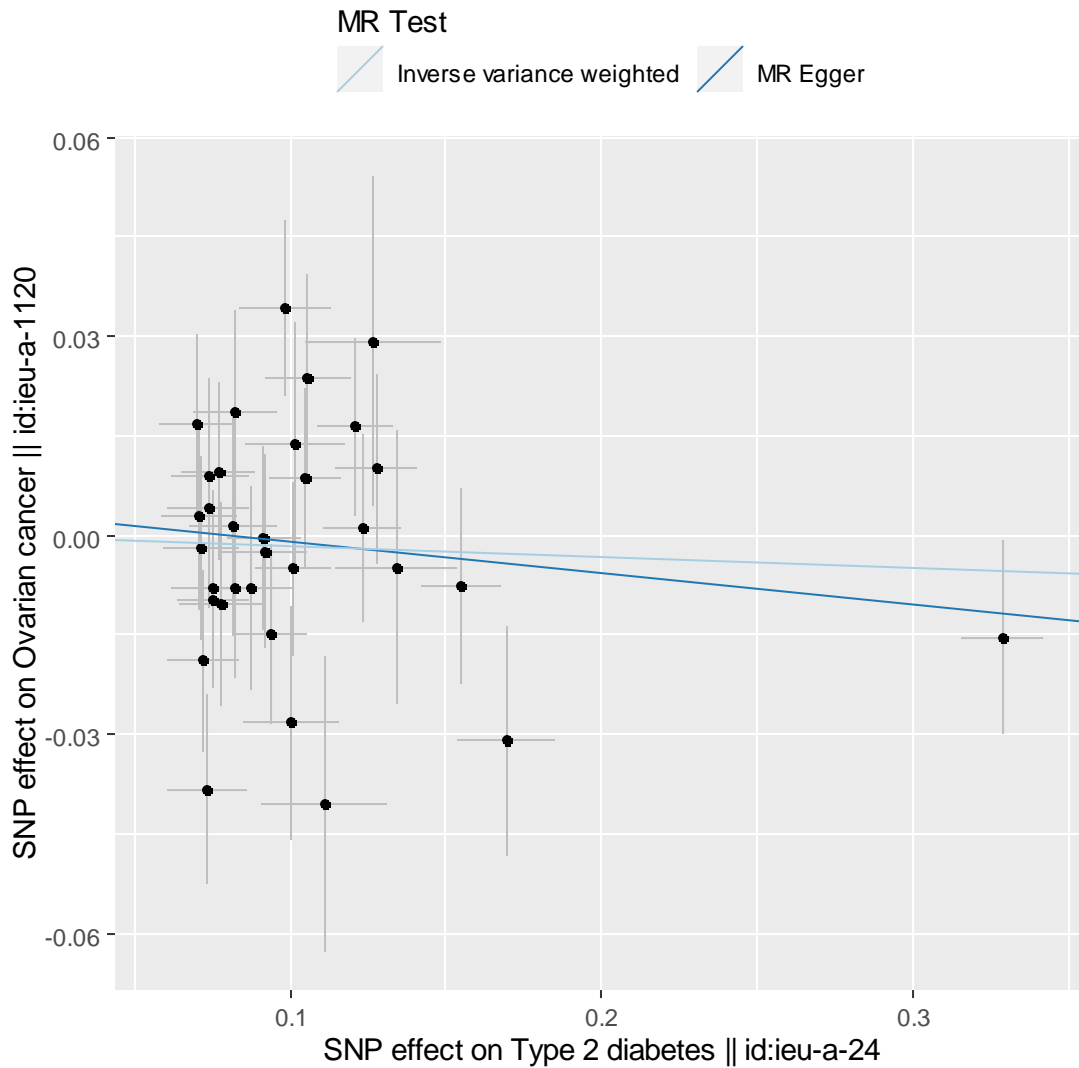
#### The Two-Sample Mendelian Randomization Results of the Causal Relationship between Type 2 Diabetes Mellitus and Ovarian Cancer

Method	n SNP	Beta	exp(beta)	Se	Pval
MR Egger	33	-0.048	0.953	0.062	0.448
Weighted median	33	-0.042	0.959	0.036	0.243
Inverse variance weighted	33	-0.016	0.984	0.026	0.522
Simple mode	33	-0.031	0.969	0.072	0.674
Weighted mode	33	-0.041	0.960	0.04	0.323

*n SNP: number of SNPs, b: beta coefficient, exp(beta): exponential of beta coefficient, se: standard error, pval: p-value*

The third objective of the study tried to determine whether there is causal relationship between T2DM and ovarian cancer. The study revealed in Table 4.3 that there is no causal relationship between T2DM and ovarian cancer (MR-Egger:  $b=-0.048$ ,  $se=0.062$ ,  $p\text{-value}=0.448$ , IVW:  $b=-0.017$ ,  $se=0.026$ ,  $p\text{-value}=0.522$ ). This is because the p-values for these Mendelian randomization methods were greater than 0.05 indicating that there was no evidence of the causal relationship between T2DM and ovarian cancer. The other three

methods (weighted median, simple mode, and weighted mode) gave consistent results, which showed that MR-Egger and IVW techniques are capable of giving reliable results.



**Figure 4.2. Scatter Plot Representing Two-Sample Mendelian Randomization Results of the Causal Relationship between Type 2 Diabetes Mellitus and Ovarian Cancer**

Figure 4.2 graphically represented the effects of the SNPs on exposure (T2DM) against the effects of the SNPs on the outcome (ovarian cancer) as suggested by Walker *et al.*, (2019). The black dots represent each of the SNPs associated with T2DM while the horizontal and the vertical lines depict the standard error of the relationship between type 2 diabetes mellitus and ovarian cancer respectively. It also indicates that there was no causal relationship between T2DM and ovarian cancer.

It was also necessary to check the directionality of the causal relationship between T2DM and ovarian cancer as shown in Table 4.4.

**Table 4.4.**

**The Direction of the Causal Relationship between Type 2 Diabetes Mellitus and Ovarian Cancer**

exposure(T2DM)	outcome(o.c)	correct_causal_direction	steiger_pval
0.030	0.001	TRUE	0.000

*T2DM: type 2 diabetes mellitus, o.c: ovarian cancer, pval: p-value*

Hemani *et al.*, (20118) stated that it is important to ensure that the direction of causality is correct. This will prevent a situation where the researcher maybe thinking that the exposure is causing the outcome while in reality it is the vice versa (Hemani *et al.*, 2018). Therefore, it was necessary to test the causality direction using the estimated variance explained by the exposure (T2DM) and the outcome (ovarian cancer). Figure 4.4, showed that the T2DM (exposure) estimated variance (0.030) was greater that the r-square of the outcome, ovarian cancer, (0.001). The p-value (0.000), on the other hand, was less than 0.05 indicating that the causality direction taken in this study is actually possible. However, Hemani *et al.*, (2018) indicated that it is crucial to note that the causality direction does not show whether or not the causal relation exists.

**4.3.4 Sensitivity results**

This study used multiple genetic variants to carry out the two-sample Mendelian randomization. Therefore, it was highly plausible to assume that all the SNPs, instrumental variables, satisfy the Mendelian randomization assumptions. Hence, it was necessary to perform some of the sensitivity analyses that will either support or question the validity of the research. The interpretation of the odds ratios (exponents of beta coefficients) of

Mendelian randomization assisted in determining the validity of the causal inference from Mendelian randomization. The results are shown in Table 4.5.

**Table 4.5.**

**The Odds Ratio Output of the Causal Relationship between Type 2 Diabetes Mellitus and Ovarian Cancer**

Method	nsnp	B	Se	pval	Lower CL	Upper CL	OR	OR	OR
								Lower CL	Upper CL
MR Egger	33	-0.048	0.062	0.448	-0.169	0.074	0.954	0.845	1.076
Weighted median	33	-0.042	0.037	0.251	-0.115	0.030	0.959	0.892	1.030
Inverse variance weighted	33	-0.016	0.026	0.522	-0.067	0.034	0.984	0.935	1.035
Simple mode	33	-0.031	0.070	0.665	-0.167	0.106	0.970	0.846	1.112
Weighted mode	33	-0.041	0.038	0.299	-0.116	0.035	0.960	0.891	1.035

*nsnp: number of SNPs, b: beta coefficient, se: standard error, pval: p-value, CL:*

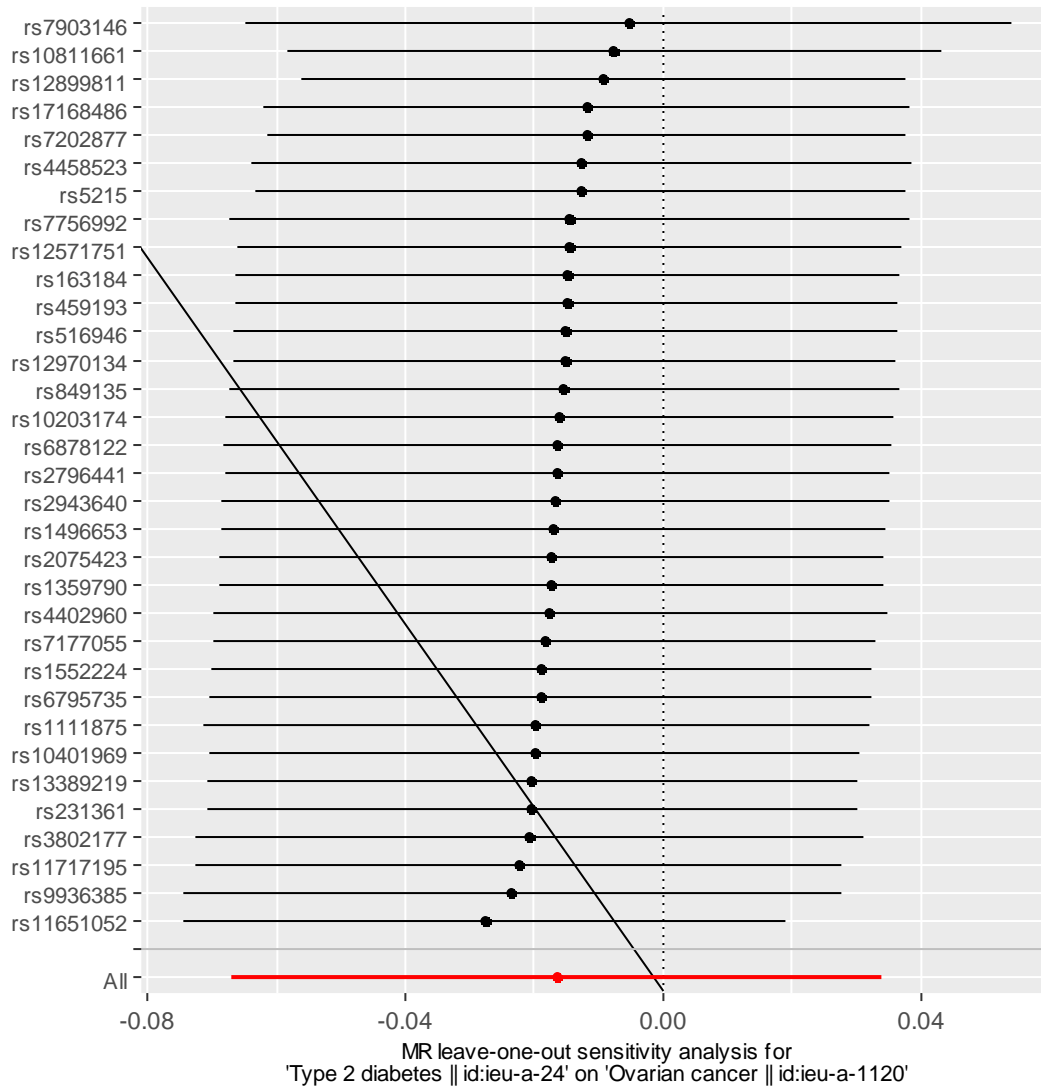
*confidence level, OR: odds ratio*

First, it is necessary to note that taking the exponent of beta coefficient is the same as the odds ratio (OR). Table 4.5 shows that the OR for MR-Egger method was 0.954 with confidence interval (CI) (0.845, 1.077). This indicates that the model, MR-Egger, was capable of detecting 0.046 decrease of variability per 1 standard deviation (SD). On the other hand, IVW technique has the OR of 0.984 with CI (0.935, 1.035). This reflects that IVW was able to detect 0.016 decrease of variability per 1 SD. The other methods

(weighted median, simple mode, and weighted mode) gave the OR which were in the same range as those of MR-Egger and IVW. These results show that two-sample Mendelian randomization gives consistent causal inference.

It was then necessary to investigate if there were potentially influential SNPs, which could be due to horizontal pleiotropy using leave-one-out analysis. The results are given in Figure 4.3



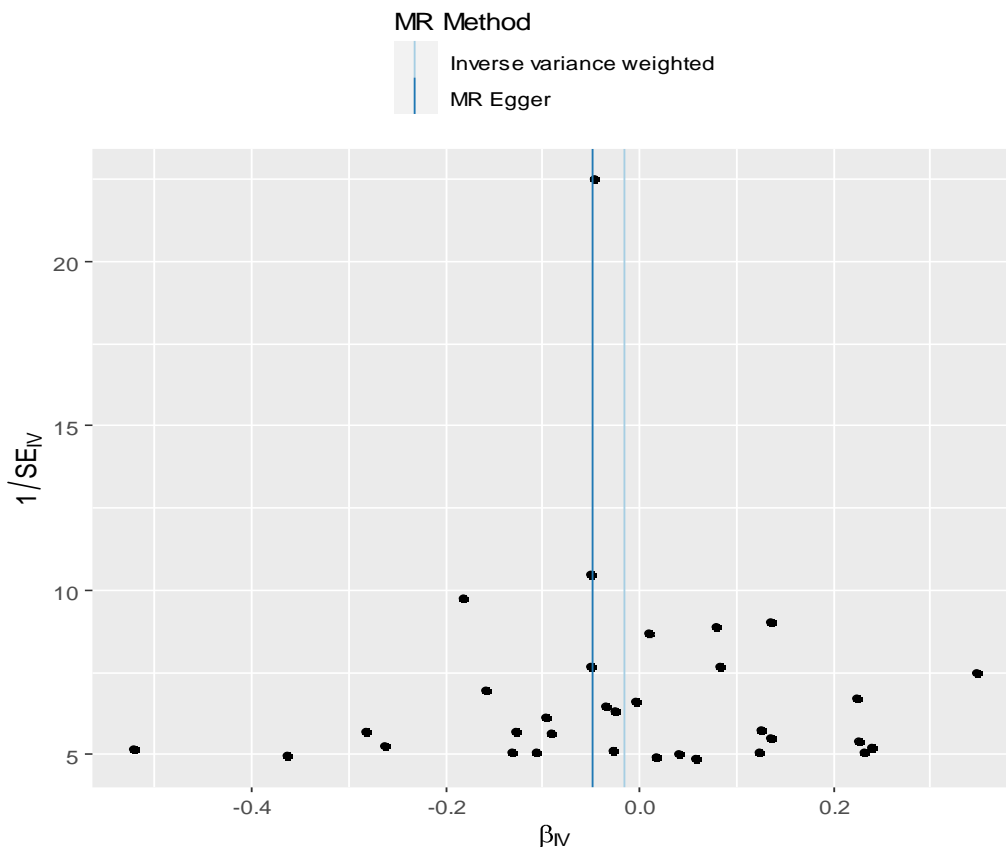


**Figure 4.3. The Leave-One-Out Graph Displaying the IVW Results of the Causal Relationship between Type 2 Diabetes Mellitus and Ovarian Cancer while Excluding One SNP each Time**

One of the assumptions of Mendelian randomization states that the genetic instruments, SNPs, should influence the outcome through the exposure. This situation reflects that there should be minimal or no pleiotropy. Figure 4.3 indicates that all the selected SNPs in this study were consistent. Therefore, it was reasonable to conclude that the results of this study were not influenced by a single outlying SNP. Apart from that, Figure 4.4, is symmetric

indicating that there was no or minimal directional pleiotropy in the study. This information supports the results of the MR-Egger in section 4.2.1.

Directional horizontal pleiotropy is an occurrence that one should be keen to check while carrying out Mendelian randomization. In this study, it was necessary to assess if there was directional horizontal pleiotropy using a funnel plot. The results are given in Figure 4.4 which is symmetric indicating that there was no or minimal directional pleiotropy in the study. This information supports the results of the MR-Egger in section 4.2.1.



**Figure 4.4. The Funnel Plot Displaying the Causal Relationship between Type 2 Diabetes Mellitus and Ovarian Cancer**

## CHAPTER FIVE

### SUMMARY, CONCLUSIONS, AND RECOMMENDATIONS

#### 5.1 Introduction

This chapter highlights the summary, conclusions and the recommendations of the study.

#### 5.2 Summary

The prime objective of this study was to investigate the causal relationship between T2DM and ovarian cancer using two-sample Mendelian randomization. This investigation was necessary because there are studies that have identified that some of the hormones associated with high blood sugars tend to create carcinogenic conditions hence accelerating the growth of cancers (Joung *et al.*, 2019). Apart from that, other scholars have used the observational models like cohort and case control to determine the relationship between T2DM and ovarian cancer but have ended up with conflicting points of view (Wang *et al.*, 2019, Urpilainen *et al.*, 2018). For example, Wang *et al.*, (2019) indicated that women with Type 2 Diabetes Mellitus have higher odds of contracting ovarian cancer, especially, those of Asian origin. On the other hand, Urpilainen *et al.*, (2018) reached to a conclusion that there is no sufficient evidence of the causal relationship between Type 2 Diabetes Mellitus and ovarian cancer. Therefore, this research used two-sample Mendelian randomization, a technique which is not prone to confounding factors and reverse causation unlike the observational studies. Before carrying out the Mendelian randomization analysis, the data were to be put into the check constraints that will ensure that the results obtained are reliable and consistent. Therefore, the first objective of this study sought to find out whether or not the data retrieved were homogenous. The second objective of the study tried to determine whether or not the genetic data had horizontal pleiotropy.

### 5.3 Conclusions

The first objective of this study aimed at determining whether or not the genetic data utilized was homogeneous. This determination was necessary because it aided in checking the fulfilment of the Mendelian randomization model assumption and whether or not the genetic instruments used are weak. The Cochran's Q statistic (MR-Egger:  $Q=39.387$ ,  $p\text{-value}=0.144$ , IVW:  $Q=39.778$ ,  $p\text{-value}=0.162$ ) indicated that there was no sufficient evidence of heterogeneity in the data. This shows that the genetic instruments used were considered strong and that the model assumptions were fulfilled. Apart from that, the  $I^2$  statistics (MR-Egger= $21.293$ , IVW= $19.776$ ) showed that the heterogeneity level was less than 25% for the two techniques hence the variability in the data was negligible.

The second objective of this study revolved around the horizontal pleiotropy. This is a situation where a particular genetic variant (SNP) may be having a causal relationship with the outcome through a biological pathway that is independent of the exposure variable under study. From this study, the MR-Egger intercept was 0.0038 indicating that there was no sufficient evidence of the horizontal pleiotropy in the data. Therefore, it can be concluded that the Mendelian randomization assumptions were fulfilled.

From this study, it can be concluded that Mendelian randomization technique is a robust method. This is so because it uses genetic variants (SNPs) that undergo random allocation at conception and are non-modifiable. These properties of MR technique enable it not to be prone to confounding factors and reverse causation. These problems are common in the other techniques such as observational methods. Two-sample Mendelian randomization increases the scope of the study since it uses the genetic data from samples obtained from totally different populations. This research study has found out that there is no causal

relationship between T2DM and ovarian cancer (MR-Egger:  $b=-0.048$ ,  $se=0.062$ ,  $p\text{-value}=0.448$ , IVW:  $b=-0.017$ ,  $se=0.026$ ,  $p\text{-value}=0.522$ ) and thus other causalities may therefore be of interest in investigation.

#### **5.4 Recommendations**

One of the previous studies had suggested that the diabetes medications like metformin, statin, and oral anti-diabetic may be having some sort of causal relationship with the ovarian cancer. Therefore, there is need for a thorough investigation on the diabetes medications as a follow up to this study. Apart from that, it is recommended that the researchers should investigate the genetic architecture of the ovarian cancer. The research may give some insights to the causes of ovarian cancer and other malignancies. Besides that, the researcher recommends the research institutions to invest in getting the genome data from all the regions of the world. This will increase the scope of the genome analysis and improve precision medicine.

#### **5.5 Suggestions for Further Research**

This study used two-sample Mendelian randomization technique to determine whether or not there exist the causal relationship between Type 2 Diabetes Mellitus and ovarian cancer. Further research can be done using bi-directional Mendelian randomization. This is a technique that uses the genetic instruments for both the exposure and the outcome. One can try to determine whether the exposure causes the outcome or the outcome leads to the occurrence of the exposure.

## REFERENCE

- Akinfolarin, A. C. (2020). Ovarian cancer: An undertreated and understudied entity in sub Saharan Africa. *Tropical Journal of Obstetrics and Gynaecology*, 37(1), 1-2.
- Alliance, G., & New York-Mid-Atlantic Consortium for Genetic and Newborn Screening Services. (2009). Understanding genetics: a New York, mid-Atlantic guide for patients and health professionals.
- Bashir, A., and Litonjua, A. A. (2018). Observational studies of vitamin D associations with asthma: Problems and pitfalls. *Pediatric pulmonology*, 53(10), 1338-1339.
- Bowden, J., Del Greco M, F., Minelli, C., Zhao, Q., Lawlor, D. A., Sheehan, N. A., Thompson, J., and Davey Smith, G. (2019). Improving the Accuracy of Two-Sample Summary-Data Mendelian Randomization: Moving Beyond the NOME Assumption. *International journal of epidemiology*, 48(3), 728–742. <https://doi.org/10.1093/ije/dyy258>
- Bowden, J., Spiller, W., Del Greco M, F., Sheehan, N., Thompson, J., Minelli, C., and Davey Smith, G. (2018). Improving the Visualization, Interpretation, and Analysis of Two-Sample Summary Data Mendelian Randomization via the Radial Plot and Radial Regression. *International journal of epidemiology*, 47(4), 1264-1278.
- Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). Global Cancer Statistics 2018: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: a cancer journal for clinicians*, 68(6), 394-424.
- Burgess, S. (2012). *Statistical Issues in Mendelian Randomization: Use of Genetic Instrumental Variables for Assessing Causal Associations* (Doctoral dissertation, University of Cambridge).

- Burgess, S., and Thompson, S. G. (2017). Interpreting Findings from Mendelian Randomization Using the MR-Egger method. *European journal of epidemiology*, 32(5), 377-389.
- Cheserem, E. J., Kihara, A. B., Kosgei, R. J., Gathara, D., and Gichuhi, S. (2013). Ovarian cancer in Kenyatta National Hospital in Kenya: Characteristics and management.
- Choudhury, A., Aron, S., Botigué, L. R., Sengupta, D., Botha, G., Bensellak, T., and Hanchard, N. A. (2020). High-depth African genomes inform human migration and health. *Nature*, 586(7831), 741-748.
- Craig, E. R., Londoño, A. I., Norian, L. A., and Arend, R. C. (2016). Metabolic Risk Factors and Mechanisms of Disease in Epithelial Ovarian Cancer: A review. *Gynecologic oncology*, 143(3), 674-683.
- Crow, J. F. (2017). *An introduction to population genetics theory*. Scientific Publishers.
- Davies, N. M., Holmes, M. V., and Smith, G. D. (2018). Reading Mendelian Randomization Studies: a guide, glossary, and a checklist for clinicians. *Bmj*, 362, k601.
- Dekkers, O. M., Egger, M., Altman, D. G., and Vandembroucke, J. P. (2012). Distinguishing Case Series from Cohort Studies. *Annals of internal medicine*, 156(1\_Part\_1), 37-40.
- Dicker, R. C., Coronado, F., Koo, D., and Parrish, R. G. (2006). Principles of Epidemiology in Public Health Practice; an Introduction to Applied Epidemiology and Biostatistics.
- Gage, S. H., Bowden, J., Davey Smith, G., and Munafò, M. R. (2018). Investigating Causality in Associations between Education and Smoking: A Two-Sample

- Mendelian Randomization Study. *International journal of epidemiology*, 47(4), 1131-1140.
- Gibson, A. F., Lee, C., and Crabb, S. (2015). 'Take ownership of your condition': Australian women's health and risk talk in relation to their experiences of breast cancer. *Health, Risk & Society*, 17(2), 132-148.
- Greenstein, J. P. (2016). *Biochemistry of cancer*. Elsevier.
- Grover, S., Fabiola Del Greco, M., Stein, C. M., and Ziegler, A. (2017). Mendelian Randomization. In *Statistical Human Genetics* (pp. 581-628). Humana Press, New York, NY.
- Harding, J. L., Shaw, J. E., Peeters, A., Cartensen, B., and Magliano, D. J. (2015). Cancer risk among people with type 1 and type 2 diabetes: disentangling true associations, detection bias, and reverse causation. *Diabetes care*, 38(2), 264-270.
- Hemani, G., Zheng, J., Elsworth, B., Wade, K. H., Haberland, V., Baird, D., and Haycock, P. C. (2018). The MR-Base platform supports systematic causal inference across the human phenome. *elife*, 7, e34408.
- Higgins, J. P., Thompson, S. G., Deeks, J. J., and Altman, D. G. (2003). Measuring Inconsistency in Meta-Analyses. *Bmj*, 327(7414), 557-560.
- Hoaglin, D. C. (2016). Misunderstandings about Q and 'Cochran's Q test'in meta-analysis. *Statistics in medicine*, 35(4), 485-495.
- <https://elifesciences.org/articles/34408>
- Huxley, R., Barzi, F., and Woodward, M. (2006). Excess Risk Of Fatal Coronary Heart Disease Associated With Diabetes In Men And Women: Meta-Analysis Of 37 Prospective Cohort Studies. *Bmj*, 332(7533), 73-78.



- Jones, T. L. (2013). Diabetes Mellitus: the increasing burden of disease in Kenya. *South Sudan Medical Journal*, 6(3), 60-64.
- Joung, K. H., Jeong, J. W., and Ku, B. J. (2015). The Association between Type 2 Diabetes Mellitus and Women Cancer: The Epidemiological Evidences and Putative Mechanisms. *BioMed research international*, 2015.
- Kibirige, D., Lumu, W., Jones, A. G., Smeeth, L., Hattersley, A. T., and Nyirenda, M. J. (2019). Understanding the manifestation of diabetes in sub Saharan Africa to inform therapeutic approaches and preventive strategies: a narrative review. *Clinical diabetes and endocrinology*, 5(1), 1-8.
- Koellinger, P. D., and De Vlaming, R. (2019). Mendelian Randomization: The Challenge Of Unobserved Environmental Confounds.
- Lawlor, D. A. (2016). Commentary: Two-Sample Mendelian Randomization: Opportunities And Challenges. *International journal of epidemiology*, 45(3), 908.
- Libuda, L., Laabs, B. H., Ludwig, C., Bühlmeier, J., Antel, J., Hinney, A., and Peters, T. (2019). Vitamin D and the Risk of Depression: A Causal Relationship? Findings from a Mendelian Randomization Study. *Nutrients*, 11(5), 1085.
- Lucas, R. M., and McMichael, A. J. (2005). Association Or Causation: Evaluating Links Between" Environment And Disease". *Bulletin of the World Health Organization*, 83, 792-795.
- Mercer, T., Chang, A. C., Fischer, L., Gardner, A., Kerubo, I., Tran, D. N., and Pastakia, S. (2019). Mitigating the burden of diabetes in sub-saharan Africa through an integrated diagonal health systems approach. *Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy*, 12, 2261.

- Momenimovahed, Z., Tiznobaik, A., Taheri, S., and Salehiniya, H. (2019). Ovarian Cancer In The World: Epidemiology And Risk Factors. *International journal of women's health, 11*, 287.
- Morris, A. P., Voight, B. F., Teslovich, T. M., Ferreira, T., Segre, A. V., Steinthorsdottir, V., and Prokopenko, I. (2012). Large-Scale Association Analysis Provides Insights Into The Genetic Architecture And Pathophysiology Of Type 2 Diabetes. *Nature genetics, 44*(9), 981.
- National Institutes of Health. (2019). What are single nucleotide polymorphisms (SNPs). *Genetics Home Reference-NIH. US National Library of Medicine. URL: <https://ghr.nlm.nih.gov/primer/genomicresearch/snp>*.
- Ness, R. B., Cramer, D. W., Goodman, M. T., Kjaer, S. K., Mallin, K., Mosgaard, B. J., and Wu, A. H. (2002). Infertility, Fertility Drugs, and Ovarian Cancer: A Pooled Analysis of Case-Control Studies. *American journal of epidemiology, 155*(3), 217-224.
- Phelan, C. M., Kuchenbaecker, K. B., Tyrer, J. P., Kar, S. P., Lawrenson, K., Winham, S. J., and Hansen, T. V. (2017). Identification of 12 new susceptibility loci for different histotypes of epithelial ovarian cancer. *Nature genetics, 49*(5), 680-691.
- Pilleron, S., Soto-Perez-de-Celis, E., Vignat, J., Ferlay, J., Soerjomataram, I., Bray, F., and Sarfati, D. (2021). Estimated global cancer incidence in the oldest adults in 2018 and projections to 2050. *International journal of cancer, 148*(3), 601-608.
- Roglic, G. (2016). WHO Global Report On Diabetes: A summary. *International Journal of Noncommunicable Diseases, 1*(1), 3.
- Savard, J., and Morin, C. M. (2001). Insomnia in the context of cancer: a review of a neglected problem. *Journal of clinical oncology, 19*(3), 895-908.

- Seddighi, S., Houck, A. L., Rowe, J. B., and Pharoah, P. D. (2019). Evidence of a Causal Association between Cancer and Alzheimer's disease: A Mendelian Randomization Analysis. *Scientific reports*, 9(1), 1-12.
- Sekula, P., Fabiola Del Greco, M., Pattaro, C., and Köttgen, A. (2016). Mendelian Randomization as an Approach to Assess Causality Using Observational Data. *Journal of the American Society of Nephrology*, 27(11), 3253-3265.
- Sheehan, N. A., Didelez, V., Burton, P. R., and Tobin, M. D. (2008). Mendelian Randomization and Causal Inference in Observational Epidemiology. *PLoS medicine*, 5(8).
- Song, J. W., and Chung, K. C. (2010). Observational Studies: Cohort and Case-Control Studies. *Plastic and reconstructive surgery*, 126(6), 2234.
- Swerdlow, D. I., Kuchenbaecker, K. B., Shah, S., Sofat, R., Holmes, M. V., White, J., and Casas, J. P. (2016). Selecting Instruments for Mendelian Randomization in the Wake of Genome-Wide Association Studies. *International journal of epidemiology*, 45(5), 1600-1616.
- Tao, Z., Shi, A., and Zhao, J. (2015). Epidemiological perspectives of diabetes. *Cell biochemistry and biophysics*, 73(1), 181-185.
- Tsilidis, K. K., Kasimis, J. C., Lopez, D. S., Ntzani, E. E., and Ioannidis, J. P. (2015). Type 2 Diabetes and Cancer: Umbrella Review of Meta-Analyses of Observational Studies. *Bmj*, 350, g7607.
- Urpilainen, E., Marttila, M., Hautakoski, A., Arffman, M., Sund, R., Ilanne-Parikka, P., and Hinkula, M. (2018). The Role of Metformin and Statins in the Incidence of Epithelial Ovarian Cancer in Type 2 Diabetes: A Cohort and Nested Case–Control

Study. *BJOG: An International Journal of Obstetrics & Gynaecology*, 125(8), 1001-1008.

Urpilainen, E., Marttila, M., Hautakoski, A., Arffman, M., Sund, R., Ilanne-Parikka, P., and Läärä, E. (2018). Prognosis of ovarian cancer in women with type 2 diabetes using metformin and other forms of antidiabetic medication or statins: a retrospective cohort study. *BMC cancer*, 18(1), 1-9.

vanVliet, N. A., Noordam, R., van Klinken, J. B., Westendorp, R. G., Bassett, J. D., Williams, G. R., and van Heemst, D. (2018). Thyroid Stimulating Hormone and Bone Mineral Density: Evidence from a Two-Sample Mendelian Randomization Study and a Candidate Gene Association Study. *Journal of Bone and Mineral Research*, 33(7), 1318-1325.

Waddington, C. H. (2016). *An introduction to modern genetics*. Routledge.

Walker, V. M., Davies, N. M., Hemani, G., Zheng, J., Haycock, P. C., Gaunt, T. R., and Martin, R. M. (2019). Using the MR-Base Platform to Investigate Risk Factors and Drug Targets for Thousands of Phenotypes. *Wellcome open research*, 4.

Wang, L., Wang, L., Zhang, J., Wang, B., and Liu, H. (2017). Association between Diabetes Mellitus and Subsequent Ovarian Cancer in Women: A Systematic Review and Meta-Analysis of Cohort Studies. *Medicine*, 96(16).

Zheng, J., Baird, D., Borges, M. C., Bowden, J., Hemani, G., Haycock, P., and Smith, G. D. (2017). Recent Developments in Mendelian Randomization Studies. *Current epidemiology reports*, 4(4), 330-34

## APPENDICES

### Appendix 1: Type 2 Diabetes Mellitus and Ovarian Cancer Genetic Data

S/N	SNP	ea.exp	oa.exp	ea.out	oa.out	beta.exp	beta.out	palind	chr.out	se.out	ss.out	pval.out	ss.exp	se.exp	pval.exp	chr.exp
1	rs10203174	T	C	T	C	-0.13415	0.004775	FALSE	2	0.02077	66450	0.8182	86197	0.019673	9.50E-12	2
2	rs10401969	C	T	C	T	0.126478	0.02928	FALSE	19	0.02495	66450	0.2406	86196	0.021828	7.04E-09	19
3	rs10811661	C	T	C	T	-0.16943	0.0309	FALSE	9	0.01735	66450	0.074831	86149	0.015689	3.72E-27	9
4	rs10830963	G	C	G	C	0.096396	-0.00757	TRUE	11	0.0151	66450	0.6161	80713	0.013347	5.32E-13	11
5	rs11111875	T	C	T	C	-0.10473	-0.00873	FALSE	10	0.01361	66450	0.5211	86178	0.01161	1.98E-19	10
6	rs11651052	G	A	G	A	-0.09863	-0.03421	FALSE	17	0.01319	66450	0.095139	54567	0.014692	1.97E-11	17
7	rs11717195	C	T	C	T	-0.10574	-0.02369	FALSE	3	0.0158	66450	0.1338	86183	0.014091	6.47E-14	3
8	rs12571751	G	A	G	A	-0.07515	0.009615	FALSE	10	0.01319	66450	0.466	86184	0.011617	1.02E-10	10
9	rs12899811	G	A	G	A	0.073461	-0.03829	FALSE	15	0.01428	66450	0.073261	80656	0.01264	6.34E-09	15
10	rs12970134	A	G	A	G	0.075302	-0.00796	FALSE	18	0.0149	66450	0.5931	83545	0.013198	1.19E-08	18
11	rs13389219	T	C	T	C	-0.07032	-0.01689	FALSE	2	0.01346	66450	0.2095	80649	0.012263	1.00E-08	2
12	rs1359790	A	G	A	G	-0.07373	-0.00426	FALSE	13	0.01506	66450	0.777301	86198	0.012984	1.39E-08	13
13	rs1496653	G	A	G	A	-0.08175	-0.00137	FALSE	3	0.01656	66450	0.9343	86196	0.013838	3.56E-09	3
14	rs1552224	C	A	C	A	-0.10168	-0.01376	FALSE	11	0.01843	66450	0.4553	80642	0.015929	1.79E-10	11
15	rs163184	G	T	G	T	0.082584	-0.00795	FALSE	11	0.01349	66450	0.555801	79357	0.012167	1.18E-11	11
16	rs17168486	T	C	T	C	0.100235	-0.02822	FALSE	7	0.01755	66450	0.1079	80650	0.015302	5.94E-11	7
17	rs1801282	G	C	G	C	-0.12254	-0.0072	TRUE	3	0.01999	66450	0.7188	86188	0.01719	1.05E-12	3
18	rs2075423	T	G	T	G	-0.07089	-0.00291	FALSE	1	0.0141	66450	0.8367	86185	0.012285	8.10E-09	1
19	rs231361	A	G	A	G	0.082508	0.01859	FALSE	11	0.01523	66450	0.2224	80647	0.013562	1.21E-09	11
20	rs243088	T	A	T	A	0.066896	-0.01158	TRUE	2	0.01323	66450	0.3813	80628	0.011876	1.81E-08	2
21	rs2796441	A	G	A	G	-0.07114	0.001974	FALSE	9	0.01393	66450	0.8873	84099	0.012184	5.39E-09	9
22	rs2943640	C	A	C	A	0.091615	-0.00039	FALSE	2	0.01385	66450	0.9778	86189	0.012026	2.69E-14	2
23	rs3802177	A	G	A	G	-0.12757	-0.0101	FALSE	8	0.01436	66450	0.4818	79271	0.013345	1.26E-21	8
24	rs4402960	T	G	T	G	0.123231	0.001173	FALSE	3	0.01418	66450	0.934	84537	0.01237	2.39E-23	3
25	rs4458523	G	T	G	T	0.093859	-0.01492	FALSE	4	0.01354	66450	0.2705	85051	0.011812	2.02E-15	4
26	rs459193	G	A	G	A	0.077951	-0.01022	FALSE	5	0.01539	66450	0.5067	80651	0.01339	5.99E-09	5
27	rs516946	C	T	C	T	0.087361	-0.00801	FALSE	8	0.01543	66450	0.6037	86191	0.013795	2.49E-10	8
28	rs5215	T	C	T	C	-0.07214	0.01889	FALSE	11	0.01368	66450	0.1672	86193	0.011749	8.50E-10	11
29	rs6795735	T	C	T	C	-0.07694	-0.00966	FALSE	3	0.01344	66450	0.4723	86194	0.011805	7.39E-11	3
30	rs6878122	A	G	A	G	-0.09171	0.002406	FALSE	5	0.01454	66450	0.8686	78523	0.013948	5.04E-11	5
31	rs7177055	A	G	A	G	0.074189	0.009084	FALSE	15	0.01469	66450	0.5364	86196	0.012648	4.60E-09	15
32	rs7202877	G	T	G	T	-0.11097	0.04041	FALSE	16	0.02225	66450	0.069419	80654	0.020108	3.50E-08	16
33	rs7756992	G	A	G	A	0.15481	-0.00773	FALSE	6	0.01476	66450	0.6002	86198	0.012557	6.95E-35	6
34	rs7903146	T	C	T	C	0.328476	-0.01534	FALSE	10	0.01459	66450	0.2933	80647	0.013054	1.20E-139	10
35	rs849135	A	G	A	G	-0.10094	0.005036	FALSE	7	0.01315	66450	0.701699	80653	0.011945	3.06E-17	7
36	rs9936385	C	T	C	T	0.120896	0.0164	FALSE	16	0.01338	66450	0.2202	80653	0.012146	2.61E-23	16

## Appendix 2: R Codes for the Study

```
#Packages installed

library(devtools)

library(TwoSampleMR)

library(digest)

library(githubinstall)

library(googleAuthR)

#listing the outcomes available in MR-Base

ao<-available_outcomes()

#let Type 2 Diabetes Mellitus be the exposure and ovarian cancer be the outcome

#extracting the instruments from the Type 2 Diabetes Mellitus GWAS (ID: "ieu-a-24")

exposure_T2DM <- extract_instruments(c('ieu-a-24'))

#clumping the exposure data

exposure_T2DM <- clump_data(exposure_T2DM)

#extracting the instruments from the ovarian cancer GWAS (ID: "ieu-a-1120")

outcome_OC <- extract_outcome_data(exposure_T2DM$SNP, c('ieu-a-1120'), proxies = 1,
rsq = 0.8, align_alleles = 1, palindromes = 1, maf_threshold = 0.3)

#harmonizing the exposure (Type 2 Diabetes Mellitus) and the outcome (ovarian cancer)

data <- harmonise_data(exposure_T2DM, outcome_OC, action = 2)
```

```
#objective 1: determining the homogeneity of the study data
```

```
hetero<-mr_heterogeneity(data)
```

```
#displaying the forest plot
```

```
singleSNP<-mr_singlesnp(data)
```

```
forestplot<-mr_forest_plot(singleSNP)
```

```
#objective 2: the horizontal pleiotropy
```

```
pleio<-mr_pleiotropy_test(data)
```

```
#objective 3: determining the causal relationship between Type 2 Diabetes Mellitus and  
ovarian cancer
```

```
T2DMOV_results<-mr(data, method_list=c("mr_egger_regression", "mr_ivw"))
```

```
#displaying the Mendelian randomization scatter plot
```

```
Scatter_plot<- mr_scatter_plot(T2DMOV_results, data)
```

```
#finding directionality of the causal relationship between Type 2 Diabetes Mellitus and  
ovarian cancer
```

```
dir<-directionality_test(data)
```

```
#displaying the odds ratio for sensitivity analysis
```

```
or<-generate_odds_ratios(T2DMOV_results)
```

```
#displaying leave-one-out graph
```

```
lou<-mr_leaveoneout(data)
```

```
louplot<-mr_leaveoneout_plot(lou)
```

```
#displaying the funnel plot
```

```
forestplot<-mr_forest_plot(singleSNP)
```



### **Appendix 3: Published Paper**