Lifestyle, Biochemical, and Genetic Risk Factors for Prostate Cancer

Karl Smith Byrne

Brasenose College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2017

Acknowledgements

Primarily, I thank my supervisors Ruth and Tim. The continued encouragement and support I have received from them has made the last three years at the Cancer Epidemiology Unit (CEU) an enjoyable and engaging experience. Further, the conferences they made me aware of and suggested I attend have led directly to future job opportunities, and so for that too I am incredibly grateful.

Within CEU, I am also thankful to Paul Appleby, Georgina Fensom, Isobel Barnes, and Ben Cairnes for their help over the years with questions that arose surrounding methods in this thesis. Similarly, I thank Ellie Watts, Rob Smith, Kuan Ai Seon, Julie Schmidt, and, in particular, Seamus Kent for their support in the DPhil room.

I would also like to thank the two friends, my Brasenose library family, that made sitting each day writing my thesis a genuinely enjoyable experience; Ashleigh Arton and Ragna Eide, you have made what could have been a miserable and tedious experience, one of the best summers of my life. I am profoundly lucky to have met both of you.

Also I thank Anna Carlqvist, Dylan Morris, my brother Finn McGuirk, Jenny Tran, and my uncle Hugh Byrne for proofing my thesis for the inevitable typos and grammatical errors that arose in the process of writing.

Many people will go unnamed in these acknowledgements. However, many many people have made it possible for me to come to Oxford, Cambridge before that, or Dundee before that, and to all these people I am thankful.

I would also like to thank my funders, the Clarendon Fund and Brasenose College for their continued support throughout.

Lifestyle, Biochemical, and Genetic Risk Factors for Prostate Cancer

Karl Smith Byrne, Brasenose College, University of Oxford.
D.Phil., Trinity term, 2017

Abstract

Despite considerable research there are no established risk factors for prostate cancer beyond age, family history, ethnicity, genetic factors, and insulin-like growth factor I. However, evidence from existing studies of risk factors for prostate cancer suggest that many lifestyle and biochemical exposures, and genetic variants may be risk factors for prostate cancer.

Exposures investigated in this thesis are: vasectomy status, microseminoprotein-beta (MSP), human kallikrein 2 (HK2), and the lactase persistence SNP, rs4988235. Additionally, this thesis contains an analysis of how germline polymorphism, including rs4988235, may determine the intake of dairy produce.

Vasectomy has been implicated as a risk factor for total and aggressive prostate cancer. However, vasectomy was not associated with risk of prostate cancer overall (hazards ratio (HR): 1.05; [95% CI 0.96-1.15]), with risk for high grade or advanced stage tumours (1.01; [0.84-1.21] & 0.83; [0.64-1.07], respectively), or for death from prostate cancer (0.88; [0.68-1.12]) in 84,753 men in the European Prospective Investigation into Cancer and Nutrition (EPIC).

A previous prospective study in the Multi-Ethnic Cohort found a significant protective association of MSP with prostate cancer. In a nested case-control study (1,871 cases & 1,871 controls) plasma MSP concentrations in EPIC were inversely associated with prostate cancer risk after adjusting for total prostate-specific antigen (PSA) concentration (odds ratio (OR) for highest versus lowest fourth of MSP = 0.65; [95% CI 0.51-0.84], p-trend = 0.001). No heterogeneity in this association was observed by tumor stage or histological grade. Mendelian randomisation (MR) analyses suggest a causal protective association of MSP with prostate cancer risk (OR per unit increase in MSP for MR: 0.96; [95% CI 0.95-0.97]).

Although HK2 has previously been found to improve discrimination for prostate cancer in predictive models, there has been no large prospective investigation of the association of HK2 with prostate cancer in men without elevated PSA. However, in EPIC plasma HK2 concentrations were not associated with prostate cancer risk independent of circulating concentrations of PSA in a nested case-control study (2,867 cases & 2,867 controls) (OR for highest versus lowest fourth of HK2 before adjustment for PSA = 7.09; [95% CI 5.82-8.65], p-trend = 0.001 vs. after adjustment for PSA = 1.29; [0.98-1.67], p-trend = 0.1). Further, there was no evidence that including HK2 in a model of prostate cancer risk based on PSA and age improved discrimination for prostate cancer overall or for high grade tumours (overall area under the curve (AUC): 0.816 vs. 0.816 or high-grade AUC: 0.752 vs. 0.752).

The lactase persistence SNP, rs4988235, has previously been investigated in the estimation of prostate cancer risk as a marker of dairy intake. In these analyses rs4988235 was not associated with prostate cancer risk overall (per-allele OR: 1.01; [95% CI 0.99-1.04]), with risk for high grade or advanced stage tumours (0.99; [0.95-1.04] and 0.99; [0.93-1.04], respectively), or for death from prostate

cancer (0.96; [0.90-1.03]) in the PRACTICAL consortium, a large (48,471 cases & 29,866 controls) prostate cancer genetics consortium. Further, in UK Biobank (up to 41,514 men) rs4988235 was only associated with the intake of dairy milk (difference between CC and TT: 19.9 g/d; [12.8-26.9]) and not the intake of other dairy products (yogurt, ice cream, and cheese).

It is possible that alternative germline polymorphisms may determine differences in the intake of dairy milk, which may be useful for future MR studies into prostate cancer risk. However, no SNPs were genome-wide significant in relation to the intake of dairy milk in a discovery GWAS in the UK Biobank (22,041 men). Further, putative associations for SNPs significant at the lower tentative genome-wide significance threshold ($p < 10^{-5}$) were not replicated when investigated in an independent sample of men from the UK Biobank (50,701 men).

In conclusion, there is no strong evidence that vasectomy, HK2, or the lactase persistence SNP as a marker of dairy intake, are associated with prostate cancer risk. Given the null GWAS for dairy milk intake, rs4988235 remains the strongest known genetic determinant of dairy milk intake. Lastly, observational and MR results provide compelling evidence that MSP is a potentially causal protective factor for prostate cancer risk. Future research should focus on replicating the putative causal association of MSP with prostate cancer and better understanding the purported role of MSP in tumour suppression or pathogen defense.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Rationale

Prostate cancer is the second most commonly diagnosed cancer in men and between 1990 and 2012 rates of prostate cancer diagnosis rose in both developed and developing nations [1, 2]. However, prostate cancer incidence is highly heterogeneous globally, varying by up to 30-fold between regions. Prostate cancer mortality accounts for approximately 7% of cancer deaths globally. Similar to prostate cancer incidence, mortality varies highly between regions (up to 10-fold globally). However, in contrast to incidence, the majority of prostate cancer deaths occur in less developed nations [3, 4].

The only established risk factors for prostate cancer are age [5, 6], family history of prostate cancer [7, 8], ethnicity [9, 10], select germline genetic polymorphisms [11, 12, 13, 14], and circulating concentrations of insulin-like growth factor I (IGF-I) [15]. None of these, with the potential exception of IGF-I through its association with the intake of dairy produce [16, 17], are modifiable risk factors. Previous research has investigated numerous exposures, many of which may be modifiable, as potential risk factors for prostate cancer, such as: the consumption of dairy milk [18], processed meat [19], or poultry [20]; hip or waist circumference, body fat, or body mass index [21, 22]; vasectomy [23, 24] or sexually transmitted diseases [25, 26]; or microseminoprotein-β [27] or the Kallkrein proteins [28, 29]. However, at the time of writing, none of these have been confirmed as risk factors for prostate cancer.

The absence of consensus for many putative risk factors stems, at least in part, from inconsistency in the epidemiological evidence, its sparsity, and the methodological limitations of previous studies, which are discussed briefly below. Differences in study design (retrospective and prospective) have likely led to inconsistencies in the state of evidence due to difficulties that these designs may have in overcoming common epidemiological biases, such as recall, selection, or detection biases. Further, small sample size for previous studies used to generate evidence for risk factors has contributed to inconsistencies in evidence. Notably, small sample size is often discussed with regard to statistical power concerns surrounding null results. However, it is also important to note that for a given significance threshold ($p < 0.05$, for example) that the probability of overestimating an association or misestimating its direction increases as a function of the noise in the exposure, and so noisy measures in

small samples have also likely produced false positives [30, 31]. Difficulties of measurement for dietary, lifestyle or anthropometric factors, or for novel biomarkers, which are likely subject to measurement error and within person variation over time, have also likely contributed to the current state of the literature [32].

To help resolve the potential over- and underestimated associations of putative risk factors with prostate cancer risk in the literature, there is a necessity for evidence from large-scale prospective epidemiological studies with access to repeat questionnaires, and biomarker and genetic data that may facilitate novel statistical methods. Such large studies can: take advantage of greater power to look into associations of potential risk factors with high risk subtypes of prostate cancer, which may be less affected by concerns attributed to PSA testing and overdiagnsis [3, 4]; investigate the potential effect of misclassification or measurement error of exposures during follow-up through repeat questionnaires; use available genetic data to generate robust instruments for novel statistical methods that may address the potential causal nature of risk factors, such as Mendelian randomisation. As such, this thesis will use available data from two large prospective epidemiological cohorts (the European Prospective Investigation into Cancer and Nutrition [EPIC] and UK Biobank) and a large prostate cancer genetics consortium (PRACTICAL) to investigate the association of a selection of lifestyle, biochemical, and genetic factors with prostate cancer risk.

1.2 Research Aims

The research presented in this thesis aims to investigate risk factors as aetiological markers for prostate cancer using prospectively measured lifestyle factors (vasectomy status), biochemical markers (MSP and Human Kallikrein 2 [HK2]), and genetic markers (rs10993994 and rs4988235). As such, the thesis will primary deal with prostate cancer aetiology and not risk prediction/early detection. The thesis is also concerned with investigating how germline polymorphism determines the intake of dairy products; given the role for the intake of various dairy produce as putative risk factors for prostate cancer, a better understanding of their genetic determinants may allow for more sophisticated analyses of dairy produce and prostate cancer risk through Mendelian randomisation methods. The main objectives are to:

- Investigate the association of vasectomy with prostate cancer risk in the EPIC cohort, with a focus on high grade and advanced stage tumours, and death from prostate cancer
- Estimate the association of prediagnostic circulating MSP concentrations with prostate cancer risk in an EPIC prostate cancer nested case-control study, and assess whether this association varies by tumour characteristics using observational epidemiology. Additionally, I assess the potential causal nature of the association of MSP with prostate cancer risk using Mendelian randomisation
- Investigate the association of prediagnostic circulating HK2 concentrations with prostate cancer risk overall, and by prostate cancer tumour characteristics in an EPIC prostate cancer nested case-control study
- Investigate the association of the lactase persistence SNP, rs4988235, as a marker of dairy intake, with prostate cancer risk overall, and by prostate cancer tumour characteristics, and with death from prostate cancer in the PRACTICAL consortium
- Assess how the intake of dairy produce varies by rs4988235 in the UK Biobank, and whether additional information on rs4988235, as a marker for the intake of dairy produce, can aid in the interpretation of the association of rs4988235 with prostate cancer risk
- Conduct a genome-wide association study for the intake of dairy milk, and assess whether any genetic variation, in addition to the lactase persistence SNP, may be used as a genetic instrument to further our understanding of the association of dairy intake with prostate cancer risk

1.3 Outline of the Thesis

- **Chapter 1** is an introduction to the thesis, the rationale and general aims, and an outline of the key research objectives that will be explored in chapters 4 to 8.

- **Chapter 2** is an overview of prostate cancer epidemiology, which includes a discussion of temporal and geographical trends in incidence and mortality, and established and putative risk factors for prostate cancer.

- **Chapter 3** is a description of the data used in this thesis; this includes a brief description of the design and methods for the European Prospective Investigation into Cancer and Nutrition (EPIC), UK Biobank, and the PRACTICAL consortium.

- **Chapter 4** investigates the association of vasectomy with prostate cancer risk in EPIC. Vasectomy is a method of birth control for men in which the vas deferens is cut, blocked, or sealed to prevent sperm from traveling from the testes to the seminal fluid. Much of the evidence in favour of vasectomy as a risk factor for prostate cancer is from retrospective case-control studies with modest samples sizes, which may be subject to a variety of biases [33]. In contrast, many of the large prospective studies have not supported a significant association of vasectomy with prostate cancer [34, 24], and moreover, there is no established biological rationale for an association of vasectomy with prostate cancer [35]. Nevertheless, a recent investigation in the Health Professionals Follow-Up Study found an approximately 20% increase risk for aggressive and fatal prostate cancer [23]. The study described in this chapter uses a prospective cohort design with 4,377 cases, of which 641 had had a vasectomy (approximately 15%). The work from this chapter has been published in the Journal of Clinical Oncology and is available at: $http://ascopubs.org/toc/jco/current$

- **Chapter 5** estimates the association of circulating concentrations of MSP with prostate cancer risk. I used both a nested case-control design in EPIC, and Mendelian randomisation, which uses published estimates for the association of rs10993994 with prostate cancer from the iCOGS

genotyping project [36] and EPIC SNP data from the OncoArray chip [37] and BPC3 genotyping [38]. MSP is an abundant protein in the immunoglobulin factor family secreted by the prostate epithelium into the seminal fluid [39], which has previously been associated with a 2% decrease in prostate cancer risk per one unit (ng/ml) increase in circulating MSP concentration [27]. The study described in this chapter is based on data from 1,871 cases and 1,871 matched controls.

- **Chapter 6** investigates the association of circulating concentrations of the prostate protein HK2 with prostate cancer risk in EPIC using a nested case-control design. There are limited data on the association of HK2 with prostate cancer risk from prospective cohorts. However, there is evidence that expression and circulating concentration of HK2 may be higher in cases than in controls [40, 41, 42, 43, 44, 45, 46], and that HK2 concentration may be higher in more severe prostate cancer subtypes [47, 48, 49, 50, 51, 52, 53]. Nonetheless, it is unclear whether this is due to the co-localisation of HK2 with PSA, or instead due to an independent association of HK2 with prostate cancer risk (either as a risk factor or as a marker of disease). Indeed, there is minimal evidence that a model of prostate cancer risk calculated using PSA and HK2 combined compared to total PSA alone improves discrimination for both prostate cancer overall (1% to 4%) and for high grade disease (1% to 6%) [40, 41, 42, 43, 44, 45, 46]. The study described in this chapter contains 2,867 cases with 2,867 matched controls.

- **Chapter 7** looks into the association of the lactase persistence SNP, rs4988235, as a marker of the intake of dairy produce, with prostate cancer risk in the PRACTICAL consortium. The ability to digest the lactose sugar in many dairy products depends on lactase enzyme activity. The predominant genetic polymorphism associated with lactase enzyme activity in European populations is rs4988235 (2:135851076_C_T) [54]. Indeed, as much as a 50% increase in the consumption of milk and milk beverages has been observed in men homozygous for the T allele compared to men homozygous for the C allele in European populations [55, 56, 57, 58, 59]. A secondary aim of this chapter is to investigate how the intake of dairy produce varies by rs4988235 in the UK Biobank, and how these findings may aid the interpretation of any

association of rs4988235 with prostate cancer risk. The study described in this chapter contains 48,471 cases and 29,866 controls for the investigation of rs4988235 with prostate cancer risk, and up to 41,514 men in the cross-sectional association of the intake of dairy produce with rs4988235.

- **Chapter 8** aims to find a genetic instrument for the intake of dairy milk to advance our understanding of the association of dairy intake and prostate cancer risk. To that end, this chapter is a genome-wide association study (GWAS) for the intake of dairy milk in the UK Biobank. The Family Food Survey suggests the British population consumed approximately two litres of milk-based dairy products per person per week in 2015, which is estimated as 11.3% of daily energy intake [60]. Dairy intake has principally been associated with rs4988235 [55], however, research to date indicates that the behavioural differences in the intake of dairy produce by rs4988235 are largely due to the differences in the consumption of dairy milk [55]. A recent meta-analysis of dairy products and calcium intake and prostate cancer risk found that high intake of milk and low-fat milk may increase the risk of prostate cancer (RR per 200 g/d was 1.03 [95% CI: 1.00-1.06] and 1.06 [95% CI: 1.01-1.11], respectively) [18]. The study described in this chapter uses data from 22,041 men to conduct a discovery GWAS for the intake of dairy milk from the interim genetic data release in the UK Biobank. Putative genetic variants associated with dairy milk intake were subsequently tested for replication in an indepedent sample of men from the UK Biobank.

- **Chapter 9** reviews the main findings of this thesis, discusses the strengths and limitations of these studies, and provides recommendations for future research.

- **Appendix A** contains ancillary tables with analysis results from the MR-Base platform for rs10993994 [61] and detailed results for germline polymorphism association with the intake of dairy milk in UK Biobank.

- **Appendix B** contains a copy of the baseline questionnaires for EPIC UK (Oxford) and UK Biobank.

I anticipate that the research from this thesis will further our understanding of prostate cancer epidemiology and aetiology through the analysis of lifestyle, biochemical, and genetic factors in large prospective cohorts (EPIC and UK Biobank) and in the largest prostate cancer genetics dataset (the PRACTICAL consortium).

1.4 The role of the author in the thesis

All work presented is my own unless otherwise stated, and conducted under the supervision of Dr. Ruth Travis and Prof. Timothy Key. The Literature review, statistical analyses, and authorship of chapters was my own. Funding for the maintenance of EPIC Oxford samples came from a Cancer Research UK grant secured by Tim and Ruth. My doctoral funding came from the Clarendon Fund and Brasenose College.

The idea to investigate the association of vasectomy, MSP, HK2, and rs4988235 with prostate cancer risk were developed in consultation with my supervisors, Ruth and Tim - for MSP, I had additional consultation with Prof. Hans Lilja. These factors were chosen as, in each case, previous evidence had suggested an association with prostate cancer risk but, as yet, there was a lack of sufficient large prospective evidence for the estimation of associations with prostate cancer (vasectomy, MSP, and HK2), or large enough samples to ensure analyses were adequately powered to detect an association (rs4988235). In the case of the cross-sectional association of rs4988235 with the intake of dairy produce and the GWAS for the intake of dairy milk, concepts were also developed in consultation with Dr. Travis and Prof. Key.

All assays for total PSA, MSP, and HK2 were conducted by laboratory technicians at the Wallenberg Research Laboratories, Department of Translational Medicine, Lund University. All hormone assays were performed by the laboratory of the Hormones and Cancer Team at IARC. Data for chapters that use the EPIC cohort were provided from the EPIC-database by Paul Appleby. Data for rs4988235 from the PRACTICAL consortium were provided as part of the OncoArray project, which was facilitated by Dr. Sara Benlloch.

1.5 Publications resulting from this thesis

At the time of writing, one manuscript has been published from work described in this thesis with the title "Vasectomy and Prostate Cancer Risk in the European Prospective Investigation Into Cancer and Nutrition (EPIC)". A second manuscript is ready for submission entitled "The role of plasma microseminoprotein-beta in prostate cancer: an observational and Mendelian randomization nested case-control study in the European Prospective Investigation into Cancer Nutrition"; this has been reviewed by local and international co-authors and will be submitted to the Journel of Clinical Oncology once the main publication from the GAME-ON OncoArray on prostate cancer risk is accepted for publication. My co-authors for these papers read and provided advice on the content and methodology and, to that extent, these papers may not be considered entirely my own work.

Chapter 2

Prostate Cancer Epidemiology

2.1 Introduction

Prostate cancer is the second most highly incident cancer among men, globally, and between 1990 and 2012 rates of prostate cancer incidence were found to have risen for many nations in both the developed and developing world. Nonetheless, relatively little is known about prostate cancer aetiology or how best to make clinical predictions for risk stratification at a population level.

Much of the temporal and geographical variation in prostate cancer incidence may be attributed to differences in diagnostic practices, such as the use of prostate-specific antigen (PSA) testing. However, there is also evidence that the variance in prostate cancer incidence may be, at least in part, explained by demographic, hormonal, or genetic factors - there is less evidence in favor of lifestyle, anthropometric, or dietary factors.

Risk stratification of prostate cancer has mostly identified high risk groups of men based on age, family history, and ethnicity. To date, these strategies have not demonstrated an ability to discriminate prostate cancer by stage or grade of tumour. In contrast, there is evidence that a baseline PSA value, measurements of a protein called human kallikrein 2, or a risk score generated from a panel of Kallikrein proteins may predict high grade and advanced prostate cancer.

What follows is a select summary of prostate cancer epidemiology that includes a discussion of recent trends in incidence and mortality rates, PSA testing, current established and putative risk factors, and risk prediction.

2.2 Overview of prostate cancer incidence and mortality

In low and middle income countries there is often a paucity of high quality regional and national cancer registry data for geographic and temporal trends; estimates are typically aggregates from available regional registries or, in their absence, averages from neighbouring countries. Although this may indicate a bias for the following description of prostate cancer incidence and mortality, a recent publication suggested that reported statistics are a reasonable estimation of underlying cancer rates and temporal trends [62].

2.2.1 Geographical Trends

Prostate cancer is the second most commonly diagnosed cancer among men, globally [1] (see Figure 2.1), and in 2012, accounted for approximately 1.1 million (15%) new cancer cases among men [63]. However, incidence varies by as much as 30-fold between selected registries, evidenced by two-thirds (759,000) of prostate cancer diagnoses occurring in more developed regions. Highest incidence rates are in traditionally high-income western regions such as North America and Australia/New Zealand with age-standardised rates (ASR) of 97.2 and 111.6 per 100,000 per year, respectively. In contrast, incidence rates are low among men in less developed regions, such as Northern African (8.1), or in Eastern and South-Central Asian (ASR: 10.5 and 4.5 per 100,000, respectively) [63].

Prostate cancer mortality varies by 10-fold, globally (approximately 3 to 30 per 100,000), and accounted for approximately 6.6% of cancer deaths among men (307,000) in 2012 [63] (see Figure 2.2). Although two-thirds of diagnosed prostate cancer cases occur in more highly developed nations (165,000 or 54% of total deaths from prostate cancer), age-standardised prostate cancer mortality is highest among less developed nations. Specifically, mortality rates are low for more highly developed nations, such as Western Europe and Northern America (10.9 and 9.8 per 100,000, respectively), and very low in countries with a high proportion of men of Asian ethnicity such as South-Central Asia (ASR: 2.9 per 100,000) [63].

2.2.2 Temporal Trends

A recent study of prostate cancer incidence between 1990 and 2012 for 36 countries found that incidence was increasing in the majority (N=24) of countries included; although many of these countries were from high income and developed regions such as Japan, Israel, and Western, Southern, and Eastern Europe incidence also increased in the Caribbean, China, and Brazil. In contrast, incidence has decreased in the USA among both Black and Caucasian populations, Finland, Sweden, and Iceland between 2002 and 2012 [2]. Geographic variation in the incidence of prostate cancer has typically been attributed to changes in screening and diagnostic practices [64, 65, 66]. After the introduction of prostate-specific antigen (PSA) testing in the USA in the 1980s prostate cancer incidence increased by approximately 100% [4], before its subsequent decline. However, these increases may also be explained by the number of countries with increased incidence that are either less developed or developing nations, improved access to health care and treatment, or accuracy of regional and national cancer registries.

Prostate cancer mortality rates between 1990 and 2012 have fallen for the majority (N=24) of countries in North America, Oceania, Asian, and Europe; in Latvia, Lithuania, Estonia, Iceland, Slovenia, Slovakia, Malta, Croatia no change in mortality rates during this time period was observed, and in Bulgaria, Belarus, Russia, the Philippines, and Singapore prostate cancer mortality has increased [2]. As will be discussed below, the role of PSA testing in the reduction of mortality is controversial. However, some variation in mortality is likely attributable to screening and diagnostic practices[67, 68]. In the case of Singapore, it is possible that, despite being a high income nation, access to better diagnostic methods has not been present for long enough to reflect a reduction in prostate cancer mortality. It is also likely that access to healthcare and treatment, and death certificate accuracy and reporting, have driven differences in prostate cancer mortality [69, 70].

2.3 Prostate-specific antigen testing

PSA is a protein highly abundant in the seminal fluid that is secreted by the prostate epithelial cells [39]. One estimate suggests 21% of men age ≥ 45 in Britain in 2010 had a previous PSA test within the past 12 months [71] have had at least one PSA test [3], a common blood test of circulating concentrations of PSA introduced in the 1980s [4]. High circulating concentrations of PSA have been associated with more aggressive forms of prostate cancer [72], and in several studies PSA among men aged between 50 and 60 was associated with death from prostate cancer at up to 30 yrs follow-up with receiver operating curves between 0.80 and 0.90 [73, 74]. In addition, there is evidence that PSA improves upon other routinely used decision-aids for biopsy, such as a digital rectal examination [75]. As such, in the last 25 years PSA has been used in many countries as either a screening tool for prostate cancer or as a decision aid for clinicians deciding whether men should be referred for biopsy.

PSA testing has been shown to significantly reduce prostate cancer mortality in the European Randomized Study of Prostate Cancer Screening trial (ERSPC) by up 27% at 13 years of follow-up [76] and by up to 56% in the Göteborg randomised population-based prostate-cancer screening trial after 14 years [77]. Although Prostate, Lung, Colorectal, and Ovarian cancer screening trial (PLCO) [78] did not report a significant reduction in prostate cancer mortality, there was significant PSA testing contamination the control arm, with up to 50% of men having recieved a PSA test within the last year [79]. Thus, it is unclear whether the PLCO trial was able to estimate the efficacy of PSA testing for the reduction of prostate cancer mortality.

A common criticism of PSA testing is that it lacks specificity and has poor positive predictive value; in ERSPC [76] and PLCO [78] the positive biopsy rates were 24% and 35%, respectively, in men with elevated PSA (≥ 3 or 4 ng/ml, respectively). Furthermore, of men who receive a positive biopsy after a PSA test, there is evidence that less than a third go on to suffer from a clinically aggressive tumour [80]. As such, the widespread use of PSA is implicated in the substantial overdiagnosis of prostate cancer, and it is not currently recommend as a screening tool by Urological Associations in Europe [81] or the USA[82].

2.4 Risk Factors

2.4.1 Established risk factors for prostate cancer

The only established risk factors for prostate cancer, to date, are age, family history of prostate cancer, ethnicity, select germline genetic polymorphisms, and circulating concentrations of insulin-like growth factor I (IGF-I).

Age

Only 1.6% of prostate cancer cases and 0.7% of deaths from prostate cancer occur in men below the age of 50 years, globally. However, both incidence and mortality rates rise exponentially among men aged 50 years and greater; 2012 worldwide age-specific incidence rates were 6.3 per 100,000 for men aged 45 to 49 years, which rises to 26.9, 67.0, 136.4, 227.9, and 305.4 per 100,000 for each successive five-year age bracket, and 386.9 per 100,000 for men aged $\geq$ 75 years. A similar pattern is observed for prostate cancer mortality with rates of 0.4 per 100,000 among men aged 45 to 49, which rises to 1.1, 4.4, 12.6, 33.1, 77.3 per 100,000 for each successive five-year age bracket, and 223.3 per 100,000 for men aged $\geq$ 75 years in 2012 [5, 6].

Family history

A recent meta-analysis [83] and a subsequent analysis of 635,443 men with ancestral genealogy data in the Utah Population Database (UPD) [7] both found an increased risk of prostate cancer among men who have at least one first-degree relative with prostate cancer (RR: 2.46 [95% CI: 2.39-2.53]) compared to men without a family history of prostate cancer. Further, the UPD study reported a dose-dependence of number of first-degree relatives on prostate cancer risk (RR for $\geq$ 4 first-degree relatives: 7.65 [95% CI: 6.28-9.23]). In the UPD study, prostate cancer risk also attenuated with decreased relatedness (RR for $\geq$ 1 second-degree relatives: 1.51 [95% CI: 1.47-1.56] & RR for $\geq$ 1 third-degree relatives: 1.15 [95% CI: 1.12-1.19]) [7]. Such evidence is the basis of clinical guidelines in Europe and the USA to consider men with a family history of prostate cancer at higher risk for prostate cancer [84, 81].

However, the previous two studies did not present prostate cancer risk associated with family history by tumour characteristics or with respect to age at diagnosis. A high frequency of PSA testing has been previously associated

with the over diagnosis of low grade or localised stage disease, particularly among men older at diagnosis [85]. As such, it is also possible that the familial aggregation of prostate cancer is at least partly due to the frequency of PSA testing among related men, and not to a truly shared genetic predisposition to prostate cancer.

A nationwide, population-based registry study from Sweden [8] investigated the association of family history with high risk prostate cancer (Gleason score $\geq$ 8, T3-4, PSA $\geq$ 20 ng/mL, N1 and/or M1). The absolute risk of a diagnosis with high risk prostate cancer by age 65 approximately doubled from 1.4% (95% CI: 1.3-1.4) for men with no family history of prostate cancer to 3.0% (95% CI: 2.6-3.4) for men with a brother previously diagnosed with high risk prostate cancer. Further, there was strong evidence that the reported association of family history with prostate cancer was inflated due to PSA testing practices; while only 2.0% of men in the general population were diagnosed with low risk prostate cancer (T1-2, Gleason score $\leq$ 6, PSA $\leq$ 10ng/mL, Nx/N0, Mx/M0) by 65 years, 7.6% of men with one brother and 20.1% of men with two brothers also previously diagnosed with prostate cancer.

Thus, although the association of family history with prostate cancer may be inflated, the approximate doubling of high risk disease for men with a family history supports a true genetic predisposition to prostate cancer.

Ethnicity

The incidence of prostate cancer varies by approximately three-fold between Black American men and Asian American men; age-adjusted prostate cancer incidence from Surveillance, Epidemiology, and End Results Program (SEER) were 203.5 (Black), 121.9 (Caucasian), and 68.9 (Asian) per 100,000 [9] - a similar pattern of age-adjusted incidence was found from a study in the UK [10]. There is also evidence that tumour stage and histological grade may be more severe at diagnosis among Black than Caucasian American men [86, 87]. However, a recent study observed a reduction in the percent difference in Black and Caucasian men in the USA being diagnosed with high risk prostate cancer. Among men aged 65 to 74 year Black men were 7% more likely to be diagnosed with high risk prostate cancer compared to Caucasian men, which was reduced to 4% for men aged 25 to 54 [88]. In contrast, the Kaiser Permanente cohort found that, relative to Caucasian men,

ethnically Asian men were at a 30% reduced risk of being diagnosed with prostate cancer, and further reported that there was a higher proportion of men diagnosed with high risk prostate cancer among Asian-American men compared to Caucasian American men [89].

The ethnic disparities in prostate cancer incidence and tumour severity at diagnosis are likely multifactorial and represent a combination of diagnostic practices, access to healthcare, attitudes toward treatment, socioeconomic factors, as well as the disease biology.

Much focus has been given to case ascertainment due to the differences in the frequency of PSA testing between ethnic groups; fewer prostate cancer diagnoses are preceded by a PSA test among Black men when compared to Caucasian men [90] using Medicare records between 1991 and 1998. However, a subsequent study including additional years of information, and a detailed analysis including age at PSA test, found that differences in PSA testing rates between Black and Caucasian men attenuated by the year 2000 among men aged $\geq$ 50 years. Further, among men aged <50 years by the year 2000 PSA testing was at a higher frequency among Black men than Caucasian men in the USA [91], which has also been reported in a large UK-based prospective study that finds Black men are 30% more likely to receive a PSA test compared to Caucasian men [92]. Data on the frequency of PSA testing among Asian men is not widely available, however, there is evidence that they may be up to 40% less likely to get a PSA when compared to Caucasian men [92].

SEER estimates for prostate cancer mortality vary four-fold in the USA by ethnicity: 44.2 (Black), 19.1 (Caucasian), 9.1 (Asian) per 100,000. Mortality is also high among men from countries with predominantly Black populations, such as the Caribbean and sub-Saharan Africa (ASR: 29 and 19-24 per 100,000, respectively). In contrast, mortality rates are lower for highly developed nations, such as Western Europe and Northern America (10.9 and 9.8 per 100,000, respectively), and very low among men in countries with a high proportion of men of Asian ethnicity such as South-Central Asia (ASR: 2.9 per 100,000) [63]. Further, the CONCORD study of 628,000 men from 31 countries found consistently lower five-year survival for men diagnosed with prostate cancer between 1990 and 1994 (Black: 85.8%, Caucasian: 92.4%) [93], although it is notable that the difference is minimal and overall survival is high. Although, the disparity in prostate cancer mortality may be

accounted for by access to diagnoses and/or treatment among Black communities or in countries with a predominantly Black population, the prospective Multiethnic-Cohort report differences in risk of death from prostate cancer after adjustment for lifestyle and socioeconomic factor, and history of PSA testing. As such, it remains probable that at least a proportion of the variation in prostate cancer mortality is due to differences in the underlying risk due to inherited genetic susceptibility [94].

Genetic polymorphisms

Early studies of germline risk for prostate cancer investigated rare, high risk loci due to their potential value as a tool for clinicians. However, by virtue of their rarity, it is unlikely they account for a large proportion of prostate cancer heritability. Recent large-scale GWAS have identified 103 more common single nucleotide polymorphisms (SNP) associated with prostate cancer [11, 12, 13], which together explain 33% of familial risk for prostate cancer [14]. Many of these variants occur in intronic or intergenic regions that were not previously suspected to be clinically relevant. However, a SNP enrichment study for the 103 known risk variants for prostate cancer suggests that the majority of high risk variants for prostate cancer can be mapped to androgen receptor and FoxA1 binding regions (74 of 103 variants) [95].

To date, the strongest candidate risk SNP associated with prostate cancer is rs10993994 in the promoter region of the MSMB gene on chromosome 10; T homozygotes had a 57% elevated risk of prostate cancer risk compared to C homozygotes - for heterozygotes, a 21% increased risk was observed [36]. The MSMB gene encodes for microseminoprotein-β, one of the three most abundantly secreted proteins by the prostate epithelium into the seminal fluid [39], which has previously been inversely associated with prostate cancer risk [27].

Insulin-like growth factors and their associated binding proteins

Insulin-like growth factors (IGFs) and their associated binding proteins (IGFBPs) are involved in cellular differentiation, proliferation and apoptosis [96]. A large meta-analysis of individual participant data (Up to $N^{cases} = 10{,}554$, $N^{controls} = 13{,}618$) using almost all published prospective data worldwide (>98%) in the Endogenous Hormones and Prostate Cancer Collaborative Group found a significant, modest positive association between circulating

concentrations of prediagnostic IGF-I, IGF-II, and IGFBP-2, and prostate cancer risk (OR per 80% increase (95% CI): 1.22 (1.12-1.31), 1.30 (1.16-1.45), and 1.17 (1.05-1.31), respectively)[15].

Experimental studies suggest IGF-I has mitotic and anti-apoptotic effects [97], and a recent agnostic pathway analysis of cancer susceptibility polymorphism found the IGF-I pathway was associated with prostate cancer risk [12]. Further, circulating concentrations of IGF-I have been positively associated with the intake of animal protein [16, 17]. Given evidence showing an association of the intake of protein from dairy sources with prostate cancer (discussed below) [55, 98], to date, IGF-I is the strongest candidate modifiable risk factor for prostate cancer.

There is less evidence for the association of other IGF-II and IGFBP-2 with prostate cancer risk. A common prostate cancer susceptibility SNP has been found in the gene that codes for IGF-II [11], and increased IGF-II has previously been found associated with disease progression, which has led to the suggestion that it may be a tumour marker and not an aetiological risk factor [99]. Similarly, IGFBP-2 has been suggested as a marker of tumour progression, as circulating concentrations have been found to increase commensurately with tumour progression [100]. However, IGFPB-2 can also inactivate PTEN [101], and has been suggested to have role in the glucose metabolism following several studies that reported IGFBP-2 as a mediator in the association between adipisoty and aggressive prostate cancer [102, 103]. As such, its biological function with respect to prostate cancer is yet unclear.

2.4.2 Putative risk factors for prostate cancer

There is evidence in favor of a number of dietary, anthropometric, metabolic, or other lifestyle factors such as sexual activity or vasectomy. Below is review of a selection of these putative risk factors.

2.4.3 Dietary factors

Interest in diet as a risk factor for prostate cancer emerged in response to the observation that prostate cancer incidence was higher among men in many Western countries than in Southeast Asian countries [1, 104]. Further, numerous studies found that men who migrate from countries with traditionally low to high prostate cancer incidence adopt the prostate cancer risk profiles

of men born in the high incidence county [105, 106, 107, 108]. These studies suggested that at least a proportion of the global variation in prostate cancer incidence may be due to differences in lifestyle factors such as diet, and that the effect of modifying such factors may be observed within one generation [109, 110].

To that end, Doll and Peto (1981) [109] suggested that as much 35% of cancer risk could be attributed to diet. However, subsequent research from large prospective studies have yet to conclusively ascertain the influence of diet on prostate cancer risk [111, 112]. To date, strong candidates for dietary risk factors for prostate cancer are folate and B_{12}, and dairy products.

Folate and B_{12}

A recent large meta-analysis of individual participant data (Up to N^{cases} = 6,875, $N^{controls}$ = 8,104) of circulating folate and B_{12} concentrations, and prostate cancer found modest positive significant associations (ORs for the top vs bottom fifths were 1.13 [95% CI, 1.02-1.26], ptrend = 0.018, for folate and 1.12 [95% CI, 1.01-1.25], ptrend = 0.017, for vitamin B12). Further, there was evidence that circulating folate concentrations were significantly associated with risk for high grade tumours [113]. Folate is obtained largely from green leafy vegetables and B_{12} is found most commonly in red meat and offal. Folate and B_{12} may influence prostate cancer risk via their role in deoxyribonucleic acid (DNA) synthesis, methylation, and repair [114].

Intake of protein from dairy

There is some evidence from the European Investigation into Cancer and Nutrition cohort (EPIC) that the intake of protein from dairy is associated with a modest increase in prostate cancer risk (OR for the top vs bottom fifth were 1.22 [95% CI: 1.07-1.41, Ptrend=0.02]) [98]. Additionally, whole milk has been strongly positively associated with risk of death from prostate cancer in the Physicians' Health Study (HR for 1 serving/d vs non-consumers was 2.17 [95% CI: 1.34-3.51], Ptrend<0.001) [115]. Further, a recent meta-analysis of dairy products and calcium intake and prostate cancer risk found that high intake of dairy products, milk, low-fat milk, cheese, and total, dietary, and dairy calcium may increase the risk of prostate cancer [18].

Intake of red and processed meat

Evidence in favor of an association of red and processed meat with prostate cancer has been inconsistent [116, 117]. Such inconsistency may be attributed to lack of standardised definitions for meat exposures. However, a large individual participant meta-analysis with standardised definitions for both outcome and exposure (Up to N^{cases} = 52,683, $N^{overall}$ = 842,149) found no overall association of red or processed meat with prostate cancer risk, and a modest positive association was observed for advanced tumours (T4, N1, M1, or fatal cases not initially diagnosed with localised disease) [19].

Intake of poultry

A World Cancer Research Fund (WCRF) report suggested that there may be a positive association of total poultry intake and prostate cancer risk [20], and a subsequent study found a significant positive association of intake of poultry with skin and prostate cancer risk (HR for the top vs bottom tertile was 2.26 [95% CI: 1.36, 3.76, P for trend = 0.003]) [118]. Further, a recent large individual participant meta-analysis (see section 2.4.3) found higher intake of poultry was associated with a modest significant reduced risk of advanced and fatal prostate cancer [19]. Although the biological mechanism for this association is unclear, a role for higher heterocyclic amine content in poultry has been suggested.

Selenium

A 2014 meta-analysis ($N^{overall}$ = 3,559) found little evidence that circulating concentrations of selenium were associated with prostate cancer risk [21]. However, a subsequent individual participant meta-analysis of 15 prospective studies (Up to N^{cases} = 4,527, $N^{controls}$ = 6,021) found nail concentrations of selenium were associated with a reduced risk of prostate cancer, and that this association may be specific to aggressive tumours [119].

Lycopene

A meta-analysis from 2015 (Up to N^{cases} = 11,239, $N^{controls}$ = 18,541) found no significant evidence that circulating lycopene concentrations were associated with prostate cancer risk overall. Notably, however, there was evidence that lycopene may be specifically associated with a reduced risk for advanced

prostate cancer and aggressive disease ($N^{advanced}$ = 1,654, $N^{aggressive}$ = 1,741). Lycopene may act as an antioxidant or may inhibit the cell cycle behaviours [120].

a subsequent individual participant meta-analysis of 15 prospective studies ($N^{advanced}$ = 1,654, $N^{aggressive}$ = 1,741) found nail concentrations of selenium were associated with a reduced risk of prostate cancer, and that this association may be specific to aggressive tumours [119].

2.4.4 Anthropometry

A meta-analysis ($N^{overall}$ = 79,387) in the most recent WCRF report shows a dose-response, statistically significant 4% increased risk of prostate cancer per 5 cm increase in height (RR 1.04, [95% CI 1.03-1.05]) [21]. Further, a subsequent prospective analysis in the EPIC cohort found a significant association of height with risk for death from prostate cancer (HR for the top vs bottom quintile was 1.43 [95% CI: 1.14-1.80, P for trend = 0.001]) [121]. However, in contrast to the majority of observational literature, a large Mendelian randomisation study (Up to N^{cases} = 20,848, $N^{controls}$ = 20,214) found no significant association of height with risk of prostate cancer overall, by tumour subgroup, or for death from prostate cancer [22]. The absence of an association may indicate there is no true association of genetically defined variation in adult height with prostate cancer or instead that environmental (early life) factors that co-vary with height explain the previously reported positive association of height with prostate cancer risk.

In a recent meta-analysis of available research by 2014 from the WCRF, there was a modest increased risk of advanced prostate cancer associated with BMI, waist circumference, and waist-hip ratio (RR per 5 kg/m^2 increase in BMI 1.08 [95% CI 1.04-1.12], RR per 10 cm in waist circumference 1.12 [95% CI 1.04-1.21], RR per 0.1 units in waist-hip ratio 1.15 [95% CI 1.03-1.28]) [21]. Further, a subsequent large prospective analysis in the EPIC cohort found an elevated risk of high grade prostate cancer and death from prostate cancer among men with higher body fatness (BMI, waist circumference, and hip circumference)(HR for death from prostate cancer for the top vs bottom quintile were 1.35 [95% CI 1.09-1.68], 1.55 [1.23-1.96], and 1.43 [1.14-1.79] for BMI, waist circumference, and hip circumference, respectively) [121]. However, two recent large Mendelian randomisation studies did not find a significant association of BMI and waist-hip ratio with prostate cancer risk, by

tumour subgroup, or for death from prostate cancer [22, 122], which may suggest that at least a proportion of the previously reported association of body fatness with prostate cancer risk may be due to covariance with other environmental, perhaps early life, factors. Indeed, one of the Mendelian randomisation studies reports a significant increased risk of aggressive prostate cancer associated with increased birth weight [122].

2.4.5 Other factors

Diabetes

Results from both a 2013 meta-analysis of 45 studies [123] and a subsequent large prospective analysis in the EPIC cohort [124] suggest a significantly reduced risk of prostate cancer among men with type 2 diabetes compared to men without type 2 diabetes (RR 0.86 [95% CI 0.80-0.92] & RR 0.74 [95% CI 0.63-0.86], respectively). It also appears that duration of diabetes is inversely associated with prostate cancer risk [123].

Men with severe type 2 diabetes also have lower circulating testosterone concentrations that may result from a detrimental effect of hyperglycemia on testosterone production, which may implicate the androgen pathway in these results [125]. However, both men with low circulating testosterone concentrations and men with diabetes have been found to have significantly reduced circulating PSA concentrations. As such, men diagnosed with diabetes may also be subject to an inverse detection bias due to a reduced likelihood of being referred for biopsy after a PSA test.

Sexually transmitted diseases

Two meta-analyses of sexually transmitted disease (STD) and prostate cancer risk [25, 26] have suggested that men who report having had an STD may be at a significantly greater risk of being diagnosed with prostate cancer compared to men who do not report having an STD (RR: 1.49 [95% CI 1.19-1.92]) [26]. However, there is apparent heterogeneity by specific STD, with gonorrhea demonstrating the strongest and most consistent association [26, 25]. Further, men whose partners report having an STD are at an increased risk of being diagnosed with prostate cancer compared to men whose partners do not report having an STD (RR: 2.06 [95% CI 1.02-4.19]) [25].

When interpreting these results, it is necessary to acknowledge possible sources of bias both by study design and due to the culturally sensitive nature of reporting an STD. Dennis et al. (2002) [25] observed heterogeneity by study design and no stratification by study design was reported in Gandini et al. (2014) [26], as such it is possible that results were affected by differences in case ascertainment. It is also possible that data were subject to reporter bias; it has been extensively reported that the accurate estimation of sexual history is sensitive to data collection methods, and that, in particular, questionnaires may lead to bias among male reporters [126].

Vasectomy

A meta-analysis of five prospective cohort studies [33] and a subsequent analysis in the Cancer Prevention Study II and a Candadian population-based matched cohort study found no significant elevated risk of prostate cancer associated with vasectomy [34, 24]. However, a large recent investigation in the Health Professionals Follow-Up study has reported a significant increase in risk of high grade (RR: 1.22 [95% CI 1.03-1.45]) and advanced stage prostate cancer (RR: 1.20 [95% CI 1.03-1.40]) associated with having a vasectomy [23]. Much of the previous research into the association of vasectomy and prostate cancer risk has been from studies with a small number of exposed cases (≤ 150) and little is understood about possible biological mechanisms. As such, the role of vasectomy as a risk factor remains unclear.

2.5 Predicting prostate cancer risk

As discussed in section 2.3, a common criticism of PSA testing is that it is highly non-specific for aggressive and lethal disease; one estimate suggests almost 800 men need to be screened and approximately 30 diagnosed to prevent a single prostate cancer specific death [76]. Although previous literature on risk stratification has reported on the utility of risk stratification for prostate cancer by ethnicity, family history, and age [84, 127, 82], it is unclear whether such strategies differentially identify men at high risk of aggressive disease compared to prostate cancer overall [128]. There is, however, some evidence that PSA may meaningfully stratify risk of aggressive and lethal disease [129, 72, 74]. For example, up to 50% of men with lethal prostate cancer by 75 yrs have PSA in the top decile at age 50 yrs [129].

2.5.1 The four Kallikrein Risk score

The Kallikrein risk score (KLK) is a four biomarker panel (total, intact, and free PSA, HK2, and age) developed to predict risk of high grade disease in the ProtecT trial [46]. There is some evidence that KLK improves upon the predictive accuracy of a PSA-based model for outcome of prostate biopsy in both screened and un-screened populations, and in men with a previous negative biopsy. Additionally, it may improve prediction of high grade, metastatic, and lethal prostate cancer, and may help with predicting prostate cancer in men with PSA in the *grey zone* (3-7 ng/ml) [73, 46, 42, 72, 130, 131, 132]. Further, KLK has shown utility for risk stratification for distant metastases. Compared to biopsying all men with PSA >3 ng/ml, a strategy that biopsied all men with PSA >2 ng/ml and KLK >10% in the Västerbotten Intervention Project would produce a reduction in biopsy rates of 45% and detect a comparable number of metastatic cases within 10 years (56 compared with 57 per 10,000 men) [72].

2.6 Classification of prostate cancer for present analyses

Prostate cancer is classified using the TNM (tumour, nodes, metastases) tumour grading system (TNM) [133] and the Gleason score [134], which were created in the mid twentieth century to describe the clinical stage and pathological grade of prostate cancer, respectively. As discussed above, diagnostic practices for prostate cancer have changed due to the prevalence of PSA testing, however, there have also been changes and improvements in prostate biopsy techniques. Although the systematic sextant biopsy protocol has been standard procedure since 1989 [135], there has been a recent trend towards an increased number of biopsy cores, which is associated with an increased number of cancers detected [136]. Further, there has been a trend towards the allocation of higher grade to cancers that would have previously been classified as low risk [133, 137, 138], and a commensurate update to scoring recommendations for prostate cancer risk by Gleason score [139].

Changes to the classification of prostate cancer over time can introduce systematic error into the estimation of an association between an exposure and case status for prospective studies. Although many prospective studies

aim to standardise the classification of prostate cancer and have collected information on PSA testing, this has yet to be done within EPIC due to the difficulties surrounding its international multi-centre design. Nevertheless, this issue has been partly addressed by the adoption of a stringent stage and grade classification, as detailed below.

In the EPIC study and the PRACTICAL study TNM was ascertained by reference to pathology reports and cancer registry data to define two stage categories: localised and advanced, as below.

- Localised (TNM: T1-T2 and N0/Nx and M0/Mx, or stage coded in the recruitment centre as localised)
- Advanced (TNM: T3-T4 and/or N1-N3 and/or M1, or stage coded in the recruitment centre as metastatic)

Information on Gleason score was used to define two stage categories: low-intermediate grade and high grade, as below.

- Low-intermediate grade (Gleason score less than 8, or grade coded as well differentiated, or moderately or poorly differentiated)
- High grade (Gleason score of 8+, or grade coded as undifferentiated)

All analyses that follow in this thesis will use the aforementioned definitions.

2.7 Conclusion

Prostate cancer is the second most highly incident cancer in men, globally. Although prostate cancer incidence rose between 1990 and 2012 for the majority of countries, it should be noted there was heterogeneity for temporal trends between developing and developed nations, which may be attributed to diagnostic practices, such as the frequency of PSA testing.

Germline genetic polymorphisms are one possible mechanism through which age (particularly early onset disease), family history, and ethnicity may be associated with prostate cancer aetiology. To date, however, it is unclear via which pathway many of the genetic risk variants relate to the incidence of prostate cancer. Of these variants, the strongest candidate SNP,

rs10993994, may act via the downstream association of the protein coded by the MSMB gene, MSP. As there is only one prospective analysis of the association of MSP with rs10993994 or with prostate cancer, more research is need to understand both rs10993994 and MSP with relation to prostate cancer risk.

Of the established risk factors for prostate cancer only IGF-I has the potential to be modifiable through its association with the intake of protein from dairy produce. As with the other implicated dietary factors, however, it is not yet clear whether the intake of protein from dairy sources is robustly associated with risk for prostate cancer. Although novel statistical methods such as Mendelian randomisation may aid in establishing greater confidence for dietary risk factors for prostate cancer, such methods require genetic instrumental variables that are not yet available.

Regarding other potential modifiable risk factors, much observational data suggests a positive association of body fatness with risk for prostate cancer. However, a number of recent large Mendelian randomisation studies that do not support the association, and so, as with other putative risk factors discussed above, its role in prostate cancer aetiology remains unclear. In contrast to many of the other putative risk factors discussed above, there are relatively few data from large prospective cohorts on the association of vasectomy and prostate cancer risk. Given its widespread use as a cheap and highly effective form of male sterilisation, further research is necessary.

It is notable that little of the research into risk factors for prostate cancer has been translated into advances in risk prediction. Although there is some support for the utility of the four Kallikrein score for the improvement of specificity of prostate cancer diagnosis, it has yet to be widely adopted in clinical practice. One potential reason for the slow adoption may be relatively sparse evidence for the association of its constituent proteins with prostate cancer risk; namely, there is little evidence in men without elevated PSA for the association of HK2 with prostate cancer. Given the potential benefits of the four Kallikrein score, it is important to understand better the association of its constituent proteins with prostate cancer risk.

The following thesis will aim to address gaps in our understanding of a number of these factors with relation to prostate cancer risk.

Figure 2.1: Estimated age-standardised prostate cancer incidence per 100,000 worldwide in 2012

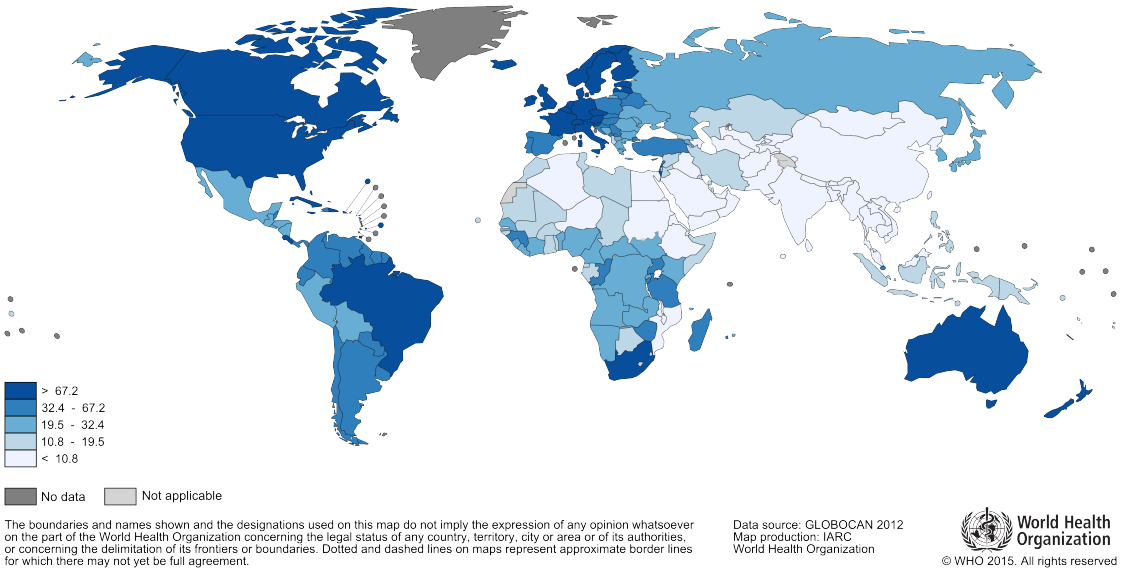
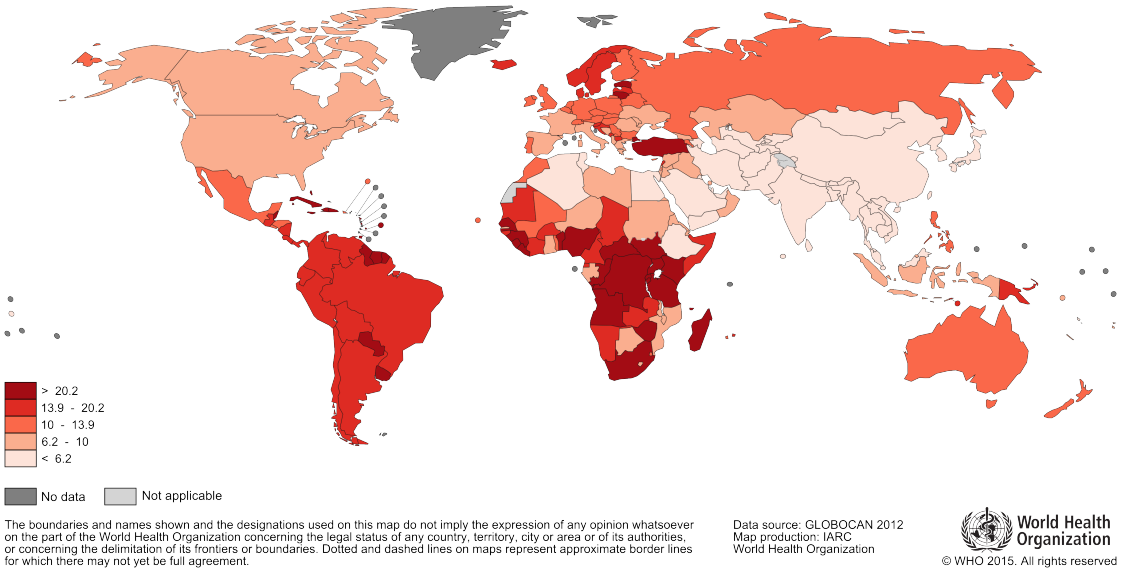


Figure 2.2: Estimated age-standardised prostate cancer mortality per 100,000 worldwide in 2012



Chapter 3

Studies Used in this Thesis

3.1 Summary

Data from two large prospective cohorts and one large consortium were used in the analyses for this thesis. What follows is a brief description of the design and methods for the European Prospective Investigation into Cancer and Nutrition [140], UK Biobank [141], and the PRACTICAL consortium [12].

3.2 The European Prospective Investigation into Cancer and Nutrition (EPIC)

3.2.1 Introduction

The EPIC cohort was established in 1992 [140] with the principal aim of investigating the prospective association of a variety of nutritional and lifestyle factors with various cancer endpoints in men and women from a large European population. However, with the increasing availability of affordable technologies, EPIC has also investigated genetic [142], molecular [143], and metabolomic risk factors [144], extracted from blood samples taken at baseline, in relation to a more diverse range of endpoints including, for example, cardiovascular disease and diabetes.

3.2.2 Design and Methods

EPIC is an ongoing multi-centre prospective cohort study. The prospective design has included the collection of lifestyle and anthropometric information as part of a baseline questionnaire, and interview data on diet and non-dietary variables, as well as blood samples for long-term storage from apparently healthy populations. Participants are followed up over time for the occurrence of cancer and other diseases, and for overall mortality. In addition, at regular intervals, follow-up questionnaires are used to collect updated information on exposures suspected to be related to disease risk.

The EPIC cohort consists of 519,978 participants (366,521 women and 153,457 men), from 23 centres located in 10 European countries - enrolment of subjects for all EPIC centres was between 1992 and 2000. EPIC began with 17 recruitment centres in seven countries (France, Germany, Greece, Italy, The Netherlands, Spain and the UK). Subsequently, centres in three

Scandinavian countries joined (Denmark, Norway, and Sweden), and one additional centre in Italy (Naples) as they were already engaged in broadly similar prospective research [140].

3.2.3 Recruitment

Data

Source populations were generally identified by geographic administrative boundaries, constituting a sample of convenience, and invited in person, by phone, or by mail to complete baseline EPIC questionnaires. At enrolment participants were mostly aged between 35 and 79 years. Only some of the centres have maintained records of the number of individuals invited to participate, however, estimates of response rates were between approximately 22% and 60% across centers [145, 146, 147, 148].

There were some exceptions to how participants were identified: the French cohort was recruited from members of teachers health insurance records; Italian and Spanish cohorts were, in part, recruited from members of a local blood donor association; Utrecht (The Netherlands) and Florence (Italy) centres included women invited for a local population-based breast cancer screening programme; and half of the Oxford (UK) cohort was recruited from participants that did not consume meat, including vegans, lacto-ovo vegetarians and fish eaters. Additionally, in France, Norway, Utrecht (The Netherlands) and Naples (Italy) only women were recruited [140].

Once invited, individuals who agreed to participate provided informed consent and were mailed a questionnaire [1] on diet, lifestyle-factors, socio-demographic characteristics, and medical history. In general, participants completed the questionnaire at home and were subsequently invited to a recruitment centre for examination. Differences in procedure were observed: although Greek participants were initially invited by mail, due to poor recruitment numbers active recruitment was initiated, and participants completed an interviewer-administered questionnaire; in Denmark and in Malmö (Sweden), participants completed dietary questionnaires at home and later provided information on lifestyle-factors, socio-demographic characteristics, and medical history at recruitment centres; in Umeå (Sweden) both diet, and lifestyle-factors, socio-demographic characteristics, and medical history

[1] See Appendix B

questionnaires were completed at a recruitment centre; in Norway, although participants were initially mailed a questionnaire unrelated to EPIC, they were subsequently mailed and completed EPIC questionnaires, and had blood samples mailed to the recruitment centre in Tromsø [140].

Examination at recruitment centres included the collection of questionnaires, venepuncture, measurement of blood pressure, and measurement of anthropometry.

Dietary assessment

Diet was assessed by one of three methods that had been developed and validated by a series of studies within the various source populations [149, 150].

1. In Germany [151], Greece, northern Italy, and The Netherlands self-administrated quantitative dietary questionnaires, containing up to 260 food items and that estimated individual average portions systematically, were used. In France, Ragusa (south Italy), and Spain a similar questionnaire that was instead structured by meals was used. Additionally, to address concerns of compliance, recruitment centres in Spain and the Ragusa centre (Italy) performed computer-assisted, face-to-face dietary interviews.

2. In Denmark [152], Norway, Naples (Italy), and Umeå (Sweden) a semi-quantitative food-frequency questionnaire (with the same standard portions assigned to all subjects) was used.

3. In Malmö [153] (Sweden) and the UK [154] combined dietary methods were used. Both the Cambridge and Oxford centres used a semi-quantitative food-frequency questionnaire and a seven-day record. In Malmö a method combining a short non-quantitative food-frequency questionnaire with 14-day record of hot meals consumed was used.

Non-dietary questionnaire data

Questionnaires were used to collect information on a large number of lifestyle and socio-demographic characteristics, and medical history, that were suspected to be related to nutritional status or cancer risk. A common set of questions and possible answers were agreed upon and translated for the

seven initial EPIC countries [140]. Questions included: current and past occupation that might have led to exposure to carcinogens; history of previous illness, disorders or surgical operations; lifetime history of tobacco smoking; lifetime history of consumption of alcoholic beverages; and physical activity (occupational, walking, cycling, gardening, housework, physical exercise, climbing stairs).

Independent questionnaires were developed for Danish and Swedish centres as they joined EPIC at a later stage. Nonetheless, questionnaires for these centres covered largely the same variables, and a comprehensive recoding scheme was developed to minimise the disparity in the measurement of questionnaire variables between EPIC recruitment centres [140].

Anthropometric measurements

Height, weight, and waist and hip circumference were measured for almost all participants using a similar protocol. Exceptions were: in Umeå (Sweden) only weight and height were measured; in Oxford (UK) weight, height, and waist and hip circumference were measured only for a subset of participants, however, self-reported weight and height were available for all participants; for Oxford (UK) self-reported waist and hip circumference was also available; and in Norway only self-reported weight and height were available. Additionally, in Denmark, Italy, Germany, Greece, Italy, Denmark, and Spain sitting height was measured [140].

Blood collection and storage

Blood samples were collected from 385,747 EPIC study participants and later separated into plasma, serum, white blood cells, and erythrocytes. There were minor differences in collection and storage procedures among the seven initial EPIC countries. However, there were greater differences in collection procedures between the seven original EPIC countries and the three Scandinavian countries that later joined EPIC.

To ensure standardisation among the initial seven EPIC countries, all materials (syringes, straws, etc.) were purchased centrally and then distributed to participating centres. All blood samples were aliquoted into 28 plastic straws that each contained 0.5 ml (12 plasma with sodium citrate, eight serum, four erythrocytes, and four buffy coat for DNA). The 24 straws were then split into two mirrored halves; one set of aliquots was stored locally and

the other set was transported to IARC to be stored in liquid nitrogen (at -196°C) at a central repository. A slight deviation from these procedures was observed of Oxford centre samples; the blood samples were delivered by mail from GP practices to a central laboratory.

Swedish and Danish blood samples were stored in tubes rather than plastic straws and so, for practical reasons, are stored in local repositories rather than at the central IARC repository. Swedish samples are kept in freezers at - 70°C, and Danish samples are stored in nitrogen vapour at - 150°C [140].

In total, the IARC repository contains approximately 3.8 million straws from 275,861 participants.

Data entry and storage

EPIC questionnaire data including centre-specific and EPIC-wide standardised variables and formats are stored centrally at IARC in the EPIC ORACLE database. Personal identifying information is stored locally at recruitment centres [140].

3.2.4 Ethics

All participants gave written and informed consent. Approval for the EPIC study was obtained from local ethical boards for participating institutions and from IARC.

3.2.5 Follow-up

Changes in lifestyle and health conditions

After being initially recruited, members of the EPIC cohort are contacted at regular intervals to gather information on a variety of variables that may be expected to change over time. These variables include, among others: smoking status, alcohol consumption, vasectomy status, and weight. Additionally, information on whether participants had suffered any major diseases is collected [140].

Cancer incidence and overall mortality

Follow-up for cancer diagnosis was obtained from national and regional cancer registries for Denmark, Italy, the Netherlands, Norway, Spain, Sweden, and the United Kingdom. In France and Greece follow-up was achieved by a

combination of health insurance records and cancer and pathology registries. For Germany active follow-up, including inquiries by mail or telephone to participants, municipal registries, regional health departments, physicians, and hospitals, was used. Information on death, including death from prostate cancer as the underlying cause, was obtained from mortality registries, active follow-up, and death certificates [140]. Details pertaining to the analysis-specific methods are elaborated in the methods section of each chapter.

3.3 UK Biobank

3.3.1 Introduction

The UK Biobank is a large prospective cohort study that was established, primarily, to investigate the lifestyle and genetic determinants of a range of health outcomes that occur in middle and later life [141].

3.3.2 Design and Methods

UK Biobank is an ongoing prospective cohort study of approximately 500,000 individuals (229,171 men and 273,461 women). Participants were aged between 40 and 69 years between 2006 and 2010, and located in proximity to one of 22 recruitment centres in England, Scotland, or Wales. The prospective design has included the collection of questionnaire data on, among others, diet, lifestyle, and anthropometric variables. Further, biological samples were collected from all participants at baseline, and have been used to conduct genotyping and a selection of biological assays for the entire cohort

Follow-up for common diet, lifestyle, and anthropometric variables is conducted at regular intervals via web-based questionnaires for large representative subsets of the cohort. In addition, ongoing enhanced phenotyping (for example, magnetic resonance imaging and accelerometer measurements for physical activity) is conducted on large subsets of the cohort. All participants are followed-up for health outcomes through linkage to national electronic health-related datasets [141].

3.3.3 Recruitment

Data

Postal invitations were sent to 9,238,453 individuals registered with the National Health Service (NHS) who were aged 40-69 years and also lived within approximately 25 miles of one of the 22 UK Biobank recruitment centres located throughout England, Wales and Scotland. Following invitation, 503,317 participants consented to join UK Biobank and went on to attend a recruitment centre between 2006 and 2010 - the overall response rate was 5.74% [155].

To optimise the accuracy and completeness of consent procedures, and data collection, computerised direct entry was used at recruitment centres. After participants completed consent procedures, the touch screen self-administered questionnaire [156] was used to collect the majority of baseline information (described in more detail below).

For information that was not easily collected by touch-screen (questions not answered by categorical or numerical responses, for example), a subsequent computer-assisted personal interview (CAPI) was used. To assist CAPI, a pre-visit aide memoire was provided to participants to reduce difficulty or inaccuracy in the recollection of certain information (medication, previous operations, family history of disease, or birth details, for example). During this portion of data collection, certain questions were asked only to participants that had given particular answers to previous touch-screen questions (if a participant indicates particular medical conditions, the interviewer was prompted to additional disease-specific questions, for example) [141].

Dietary assessment

UK Biobank used both a short, self-completed food frequency questionnaire (FFQ) at recruitment and a follow-up 24-hour dietary recall questionnaire (DRQ) to assess diet [157]. The FFQ was designed to rank participants at baseline according to commonly eaten food groups among a British population, and to collect information on the intake of some commonly consumed sources of nutrients [158]. The FFQ was also supplemented by a DRQ that was completed in large sub-cohorts on-line at regular intervals after recruitment - for a small number of participants, recruited towards the end of the recruitment period, the DRQ was also completed at recruitment centres.

Non-dietary questionnaire data

Questions that concern socioeconomic status were included on housing tenure, car ownership, household income, household structure, employment status and current occupation, ethnicity and country of birth, qualifications and school leaving age. Questions included were mostly sourced/adapted from general population surveys (such as the 2001 Census [159] and the Health Survey for England [160]).

Smoking behaviour questions were adapted from various longitudinal epidemiological studies and surveys, and covered duration and dose of smoking behaviour. Alcohol consumption was investigated with quantity-frequency type questions, and include beverage specificity to assist in accurate reporting. For both smoking and alcohol exposure, reasons for recent cessation were investigated to account for possible reverse causality [141].

Questions relating to birth weight, breastfeeding, maternal smoking, childhood body size, and residence at birth were included. In addition, a limited family history among first degree relatives, for a given participant, of common serious illnesses, such as cancer or cardiovascular disease, as well as questions about being a twin or other multiple order birth were included [141].

Medical history, reproductive history for women, general health questions, self-reported disability, as well as some limited phenotype information (related to skin and hair colour, chronic pain and chest pain, wheeze) were assessed. To avoid inaccuracy, this was done via CAPI and administered by trained interviewers [141].

Questions that concern environmental exposures covered a variety of topics, for example: current address, residence at birth, occupation at baseline and other workplace factors, passive smoke exposure, indoor air pollution, and mobile phone use were collected [161, 162]. Blood and urine sample collection were also used to quantify environmental exposures, such as cotinine for cigarette smoke.

Physical activity was assessed by a validated survey instrument [163], and was intended to rank physical activity (vigorous, moderate, and walking). Common sedentary activities were collected to provide a composite measure of physical activity [164, 165], and follow-up information for physical activity was collected for approximately 20,000 participants who attended a repeat assessment. Additionally, objectively measured physical activity was collected on a sub-sample of the population using accelerometers [166].

Information on psychological state and cognitive function was collected via a paired-associated learning questions to assess global cognition and reaction time tests for touch-screen administration [167, 168].

Anthropometric measurements

Blood pressure was measured using the Omron HEM-7015IT digital blood pressure monitor. Participants were measured twice with a short rest between measurements (about one minute). Blood pressure was subsequently remeasured in large sub-cohorts of UK Biobank (∼25,000) to allow for the correction of regression dilution. Weight and bio-impedance were measured using the Tanita BC-418 MA body composition analyser, which accurately measures weight to within 0.1 kg. Standing and sitting height (shoeless) were measured using a Seca 202 height measure. Hip circumference and waist circumference (at the level of the umbilicus) were measured using a Wessex non-stretchable sprung tape measure. Right and left hand grip strengths were measured once each using a Jamar J00105 hydraulic hand dynamometer. The Vitalograph Pneumotrac 6800 spirometer was used to collect information on Spirometry - up to three measurements of lung function were taken to improve accuracy. Calcaneal bone density in the left heel was measured using the Norland McCue Contact Ultrasound Bone Analyser (CUBA), which provides a measure of Broadband Ultrasound Attenuation (BUA) [141].

Blood collection and storage

A protocol for the collection of biological samples was created to ensure standardised collection across recruitment centres, and that the widest range of assays could be measured in future research. During the baseline assessment visit a 40-50 ml blood sample was collected from all participants using a vacutainer system. Blood was then separated into tubes as described in Table 3.1. With the exception of acid citrate dextrose (maintained at 18°C), all tubes were maintained at 4°C until shipped to central processing laboratories. Blood samples for each individual were aliquoted into approximately 24 1.4 ml samples for long-term frozen storage (see table). After recruitment, UK Biobank stores about 15 million 1.4 ml aliquot tubes. Samples are stored in two geographically separate locations to protect resources against unforeseen consequences (see Table 3.2). A "working" archive holds approximately

9 million samples at -80°C, while a "back up" archive holds approximately 6 million samples in liquid nitrogen (-196°C) [141].

Data entry and storage

UK Biobank holds data of many sorts on participants, and ranges from simple demographic information to more complex clinical data such as ECG results and eye scans. UK Biobank does not provide information such as NHS number, name, or postal address, as these could be used to identify individual participants. Instead UK Biobank provides data, on an application-specific basis, via reverse-anonymised form encoded using an irreversible algorithm. Specifically, UK Biobank retains an internal database that enables UK Biobank (and only UK Biobank) to reverse anonymise the data [169].

3.3.4 Ethics

All participants provided written informed consent. Approval for the UK Biobank was obtained from the North West Multi-centre Research Ethics Committee for local NHS Primary Care Trusts, the National Information Governance Board for Health and Social Care in England and Wales, and the Community Health Index Advisory Group in Scotland [141].

3.3.5 Follow-up

Changes in lifestyle and health conditions

Following initial assessment at recruitment, large subsets of participants (approximately 25,000) are contacted at regular intervals every 2-3 years. A large proportion of variables included in follow-up assessment are repeats of those previously included at baseline to allow for correction of regression dilution. However, repeat assessment was also used to collect additional information on a variety of exposures such as the DRQ (see section) and additional information on physical activity [141].

Disease incidence and mortality

Detailed access to past and future medical, and other health-related records was obtained for all participants by reference to NHS number in England and Wales and the Community Health number in Scotland. In practice, these identifiers were used to link to a variety of data sources and systems

to ascertain death, disease occurrence and other health-related information among participants during long-term follow-up, these include: death and cancer registries, and general practice and hospital activity records. Additional identifiers (name, date of birth, for example) were also obtained to enable linkage to other types of health-related information (occupational health records, for example). To reduce loss of follow-up, information including telephone number and e-mail address were collected [141].

3.4 PRACTICAL

3.4.1 Introduction

The Prostate Cancer Association Group to Investigate Cancer Associated Alterations in the Genome (PRACTICAL) consortium was established in September 2008 to gain insight into the genetic architecture and mechanisms of prostate cancer risk[170]. The Consortium currently consists of 123 different study groups, incorporating sites in the EU, Australia, China, Japan, India, Africa, Canada and USA. At present, it has access to samples from over 120,000 prostate cancer cases and 100,000 controls from two large genotyping projects: the iCOGS array [12] and the OncoArray [37].

3.4.2 Criteria for joining PRACTICAL and participating studies

To join PRACTICAL a given study must include at least 150 cases of prostate cancer and 150 suitably matched controls and have blood samples available for genotyping. However, for studies within emerging nations study size requirement may be relaxed to at least 50 cases and 50 matched controls. Cases may be selected based on age, family history, or tumour characteristics but must have invasive prostate cancer, and only one member per family should be included. All controls must be male and matched on ethnicity, and preferably, also matched by geographical region. Although it is not essential to include age-matched controls, where age matching is performed controls should be cancer free within five years of their index case. Note, the controls should not be from men unaffected by prostate cancer in the family [170].

A brief description of the studies within PRACTICAL that are included in analyses for this thesis can be found in Table 3.3. The description includes: the name of a given study, the number of cases and controls it contributes, the study design, the location of the study, and, where applicable, a reference to further study details. Due to the nature of the PRACTICAL consortium, some included studies do not have existing publications and so lack references to further information; in this case, further information can be found in the supplementary materials [171] for *Eeles et al. (2013)* [12]

3.4.3 Available phenotypic and epidemiological data

A comprehensive data dictionary has been created using questionnaires from studies in PRACTICAL. The data dictionary currently includes common clinical, pathological, survival, and epidemiological variables. Variables are included if they are available in at least half of participating studies.

Core variables are defined as: ID and study ID, case status, ethnicity, date of birth, age at interview/diagnosis, family history of prostate cancer and number of relatives affected by prostate cancer, date of last follow-up, months between diagnosis and last follow-up, vital status, cause of death, and date of death.

Pathological/clinical data were available from studies for: date of blood draw, Gleason's score, cancer stage, PSA at diagnosis, international prostate symptom score, method of detection for prostate cancer, prostate cancer reatment, TMPRSS2-ERG fusion status, BRCA mutation, and MMR mutation.

Epidemiological data were available from studies for: country of origin, educational attainment, marital status, occupational history, physical activity, smoking status, caffeine consumption, alcohol intake, acne, vasectomy status, baldness, family history of breast cancer, barium enema use, history of X-ray use, hypertension, heart disease, diabetes, prostatitis, benign prostatic hyperplasia, sexual activity, BMI, hand pattern, and medication use [172].

3.4.4 Genetic data

iCOGS

iCOGS is a custom Illumina iSelect genotyping array, designed as part of the Collaborative Oncological Gene-Environment Study (COGS) that was created to investigate genetic polymorphism associated with three hormone related cancers: breast, ovarian, and prostate. The array includes SNPs associated with: risk for these cancers in genome-wide association studies (GWAS); breast or ovarian cancer risk in carriers of BRCA1 or BRCA2; subtypes of disease (for example, aggressive prostate cancer); survival after diagnosis; related quantitative traits; and functional candidate variants, including rare variants in known cancer susceptibility loci [12].

OncoArray

The OncoArray is a custom Illumina SNP genotyping array that was designed to evaluate genetic variants for their association with the risk of breast, ovarian, prostate, colorectal and lung cancer, as part of the GAME-ON initiative. The array includes approximately 600k SNPs with a genome-wide backbone. In addition, it includes SNPs that are believed associated with: the 5 cancers; SNPs associated with ancestry; quantitative traits such as obesity, physical activity, non-steroidal anti-inflammatory drug use, hormone use, diet, smoking, and alcohol; pharmacogenetics; and fine-mapping of common cancer susceptibility loci [37].

3.5 Outcome of incident prostate cancer

Prostate cancer case status was defined as code C61 according to the 10^{th} revision of the International Classification of Diseases [173].

Table 3.1: UK Biobank blood collection and transport temperature protocol

Type of sample	Collection priority	Volume collected (ml)	Transport temperature (°C)
EDTA	1	9	4
EDTA (PST[a])	2	8	4
Clot activator (SST[a])	3	8	4
EDTA	4	9	4
Acid citrate dextrose	5	6	18
EDTA	6	4	4

[a] Plasma separation tube

[b] Serum separation tube

Table 3.2: UK Biobank blood collection and storage protocol

Vacutainer tube	Fractions	Number of aliquots	
		-80°C	Liquid N_2
EDTA x 2	Plasma	6	2
	Buffy coat	2	2
	Red cells	-	2
EDTA (PST)	Plasma	3	1
Clot activator (SST)	Serum	3	1
Acid citrate dextrose	DMSO blood	-	2

[a] Plasma separation tube

[b] Serum separation tube

Table 3.3: Studies contributing to PRACTICAL

Study Name	Cases/Controls	Design	Location
Aarhus Prostate Cancer Study	1,077/545	Hospital-based case-control, retrospective	Aarhus, Denmark [174]
The Agricultural Health Study	471/1,179	Nested case-control, prospective	Iowa and North Carolina, U.S. [175]
Alpha-Tocopherol Beta-Carotene	1,279/1,915	Nested case-control, prospective	Various regions, Finland [176]
French Prostate Case-Control Study	922/645	Hospital-based case-control, retrospective	Paris, France [13]
City Of Hope	257/259	Hospital-based case-control, retrospective	Los Angeles, U.S.
the Cohort of Swedish Men	2,293/1,122	Population-based cohort, prospective	Västmanland and Örebro, Sweden [177]
Copenhagen Prostate Cancer Study 1	536/258	Nested case-control, prospective	Copenhagen, Denmark [178]
Copenhagen Prostate Cancer Study 2	444/228	Nested case-control, prospective	Copenhagen, Denmark [179]
EPIC	635/693	Nested case-control, prospective	See 3.2 [140]
ERSPC[a]	71/65	Population-based randomised trial	Multi-centre, Europe [76]
ESTHER[b]	324/315	Nested case-control, prospective	Saarland, Germany [180]
Fred Hutchinson Prostate Cancer Studies	407/388	Case-control, retrospective	King County, U.S. [181]
Institut fuer Humangengetik Ulm	146/149	Hospital-based case-control, retrospective	Germany
Health Professionals Follow-up Study	1,168/1,044	Nested case-control, prospective	Massachusetts, U.S. [182]
IMPACT[c]	49/867	Case-control, prospective	U.K./U.S. [183]
Portuguese Oncology Institute of Porto	374/180	Hospital-based case-control, retrospective	Portugal [184]
Katholieke Universiteit Leuven	164/103	Hospital-based case-control, retrospective	Belgium [185]
Los Angeles Prostate Cancer Study	440/282	Case-control, retrospective	Los Angeles County, U.S.
Multi Case Control Study-Spain	520/397	Hospital-based case-control, retrospective	Spain [186]
Melbourne Collaborative Cohort Study	714/316	Nested case-control, prospective	Melbourne, Australia [187]
Multiethnic Cohort Study	598/642	Nested case-control, prospective	Hawaii and California, U.S. [27]
The Moffitt Group	403/203	Hospital-based case-control, retrospective	Tampa, Florida [188]
Prostate Cancer study, Sofia	192/89	Hospital-based case-control, retrospective	Sofia, Bulgaria [12]
Physicians Health Study	622/257	Nested case-control, prospective	Massachusetts, U.S. [189]
PLCO[d]	678/980	Population-based randomised trial	U.S. [78]
The Poland Group	484/317	Case-control, retrospective	Szczecin, Poland
Prostate cancer genetics in Galicia	129/100	Case-control	Galicia, Spain
Progression in cancer of the prostate	659/236	Population-based case-control, retrospective	Sweden [190]
PROFILE[e]	13/21	Hospital-based case-control, retrospective	U.K.

Prostate Cancer Group, Santiago	673/322	Hospital-based case-control, retrospective	Santiago, Spain [191]
Mechanisms of progression and Treatment	838/11	Case-control, retrospective	U.K. [192]
The QldMen and the Red Cross study	3,282/1,241	Hospital-based case-control, retrospective	Queensland, Australia [193]
Study of Epidemiology and Risk Factors	2,511/1,442	Hospital-based case-control, retrospective	U.K. [194]
San Francisco Prostate Cancer Study	279/205	Population-based case-control, retrospective	San Fransisco, U.S. [195]
Aggressive prostate disease	40/170	Hospital-based case-control, retrospective	Guernsey
Stockholm 2	3,019/1,481	Population-based case-control, retrospective	Stockholm, Sweden [196]
Genetic Predisposition to Prostate Cancer	2,421/1,183	Population-based case-control, retrospective	Finland [12]
Princess Margaret Biopsy Database	668/455	Hospital-based biopsy cohort, prospective	Toronto, Canada [197]
U.K. Genetic Prostate Cancer Study	11,972/6,932	Case-control, retrospective	U.K. [198]
Familial Prostate Cancer Study Germany	457/178	Population-based case-control, retrospective	Germany [199]
Total	**48,471/29,866**		

[a] Randomized Study of Screening for Prostate Cancer

[b] Epidemiological investigations of the chances of preventing, recognizing early and optimally treating chronic diseases in an elderly population

[c] Identification of Men with a genetic predisposition to ProstAte Cancer: Targeted screening in men at a higher genetic risk and controls

[d] Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial

[e] Germline genetic profiling: correlation with targeted prostate cancer screening and treatment The Pilot Profile Study

Chapter 4

Vasectomy and prostate cancer risk in the European Prospective Investigation into Cancer and Nutrition

4.1 Introduction

Vasectomy has been implicated as a risk factor for total and aggressive prostate cancer by some [23, 200, 201, 202] but not all cohort studies [33, 34, 24, 203, 204]. Notably, no study has successfully supported a biological rationale for the reported risk of vasectomy associated with prostate cancer.

4.1.1 Vasectomy

Vasectomy is a common method of birth control for men in which the vas deferens is cut, blocked, or sealed to prevent sperm from traveling from the testes to the seminal fluid [205]. Estimates suggest between 40 and 60 million men globally have opted for this procedure as a form of contraceptive [206]. The two most widely used surgical methods when performing a vasectomy are an incisional method and a no-scalpel technique. While the incisional technique requires one to two cuts (between 1 and 2 cm in length), the no-scalpel technique uses a custom, pointed forceps-type tool to puncture the scrotal skin [205]. Two randomised control trials, that compared these two techniques, found no difference in efficacy for preventing conception. However, the no-scalpel procedure was associated with fewer hematomas and post-surgery infections, less bleeding and pain, and faster return to normal sexual activity [207, 208].

4.1.2 Factors associated with having a vasectomy

A number of lifestyle factors have been associated with having a vasectomy from published cohort studies. For example, in a US study compared to men without a vasectomy, men with vasectomies have been reported as more likely to be married (92% vs. 87%) [200], to have remarried (15% vs. 10%) [202], and less likely to be college educated (38% vs. 46%) [202]. Further, compared to men without a vasectomy, vasectomised men have been reported to be more likely to drink alcohol (44% vs. 36%, drink >2 g/day) [200] and less likely to abstain from alcohol consumption (19% vs. 22%) [201]. Men with vasectomies were also more likely to have a history of prostate-specific antigen (PSA) testing than men without a vasectomy (67% vs. 59%) [23], and less likely to have never had a PSA test (12% vs. 16%) [34]. Additionally, men with a vasectomy have been reported to have a higher, age-adjusted,

prevalence of human papilloma virus (22% vs. 14%) than men without vasectomies [23].

4.1.3 Epidemiological studies of vasectomy and prostate cancer

A meta-analysis of 22 observational studies of vasectomy and prostate cancer found a relative risk of 1.37 (95% CI 1.15-1.62) [33] for men with a vasectomy compared to those without a vasectomy. However, 17 of these studies used a retrospective case-control design, which Dennis et al. (2001) partitions into population-based and hospital-based (HR: 1.14 [95% CI 0.93-1.39] HR: 1.92 [95% CI 1.37-2.67], respectively). The relative risk for the five prospective cohort studies was 1.22 (95% CI 0.90-1.64) [33]. Further, two subsequent, large prospective studies have also found no significant association of vasectomy with prostate cancer risk overall, by tumour subtype, or for fatal prostate cancer [34, 24].

It is possible that the differences in the association of vasectomy with prostate cancer by study design may be explained by a detection bias [206, 33, 209, 210]; men who receive a vasectomy may also be more likely to be individuals that highly monitor their health, and thus, be more likely to have a PSA test, which may lead to an increase in the diagnosis of prostate cancer that does not reflect a true association of vasectomy with prostate cancer. Indeed, as described in section 4.1.2, there is evidence that, compared to men without a vasectomy, vasectomised men are more likely to have a history of PSA testing (67% vs. 59%) [23, 200, 34]. However, many of the studies to date have not been able to fully account for differences in the use of PSA tests among participants.

The Health Professionals Follow-Up study has recently reported a significant increase in the risk of total prostate cancer and aggressive prostate cancer associated with having a vasectomy of 1.10 [95% CI 1.04-1.17], and 1.22 [95% CI 1.03-1.45], respectively, with 24 years of follow-up in the largest cohort to date with 1,524 cases with vasectomy [23]. Further, this study found a significant association of vasectomy with risk of fatal prostate cancer in a highly screened sub-cohort (1.56; [95% CI 1.03-2.36]).

4.1.4 Proposed mechanisms

There is little consensus for an underlying biological mechanism to explain an association of vasectomy with prostate cancer. Although Howards (1993) [211] suggested that vasectomy might disrupt the regulation of endogenous growth factors, the only evidence of increased cell proliferation post-vasectomy (PV) is in rat models [212]. Research on immunological changes PV has suggested that antisperm antibodies may be elevated in vasectomised compared to non-vasectomised men [213, 214, 215]. However, there is no evidence, as yet, that sperm sera-induced immunological response PV explain an association of vasectomy with prostate cancer. It is also possible that sex hormones differ by vasectomy status [216]; dihydrotestosterone has been found elevated in vasectomised men compared to non-vasectomised men [213], and testosterone was slightly elevated among men post-vasectomy compared to pre-vasectomy [217]. However, many subsequent studies have failed to find differences in sex hormone levels by vasectomy status [23, 217, 218]. In addition, there may be differential regulation of seminal proteins among men with a vasectomy compared to men without a vasectomy [219].

4.2 Aim of Study

The current chapter will investigate the association between vasectomy and prostate cancer risk in EPIC, with a focus on tumour stage and grade, and death from prostate cancer. It will also look at the cross-sectional association of vasectomy with PSA-testing, and circulating concentrations of seminal proteins, insulin-like growth factors, and steroid sex hormones. Finally, current available evidence for an association of vasectomy with prostate cancer risk from prospective studies will be summarised by inverse-variance weighted meta-analysis.

4.3 Methods

4.3.1 Study Population

These analyses use data from the EPIC cohort [140]. Although complete details on the cohort can be found in Chapter 3, see below for specific details that are pertinent to the analysis of vasectomy and prostate cancer risk.

This chapter includes 84,753 men from Denmark, Germany, Spain, and the United Kingdom who provided information on vasectomy status at recruitment; data on vasectomy were missing for 712 men for these countries. Information on vasectomy was not available for men in Italy, The Netherlands, or Sweden (n = 45,960). Additionally, in the Greek recruitment center, only four men had had a vasectomy and there were no exposed incident cases of prostate cancer; therefore, because all analyses were stratified by recruitment center, Greek participants were excluded (n = 10,814).

Analyses suggest this study was adequately powered to discover an association of vasectomy with prostate cancer risk; 8,948 exposed (5% exposed cases) and 50,645 unexposed participants would be necessary to discover a 10% increased relative risk [23] with 80% power (α=0.05) [220].

For 6,771 (97.3%) of 6,961 men in the EPIC-Oxford cohort who completed a follow-up questionnaire, 10 years after recruitment, history of PSA testing and age at PSA test, and additional information vasectomy status were collected.

Information on cancer diagnosis was obtained from national and regional registries for Denmark, Spain, and the United Kingdom. For Germany, active follow-up, including inquiries by mail or telephone to participants, municipal registries, regional health departments, physicians, and hospitals was used. Information on death, including death due to prostate cancer as the underlying cause, was obtained from death certificates; available evidence suggests information on death due to prostate cancer is accurate [221, 222, 223, 224]. For analyses of incidence, follow-up continued from date of recruitment to date of any primary cancer diagnosis, death, or last completed follow-up (Denmark, December 31, 2012; Germany, January 5, 2011; Spain, October 19, 2013; and the United Kingdom, December 31, 2012), whichever was first. For analyses of death due to prostate cancer, follow-up continued until death or date of last completed follow-up. During the follow-up period, 4,377 men developed prostate cancer (International Classification of Diseases 10th revision codes, C61 [173]).

Stage information was available for 2,749 cases (63.3%): 1,733 were localised (tumor-node-metastasis [TNM] staging of T1-T2 and N0/Nx and M0/Mx, or stage coded in the recruitment center as localised); 1,000 were identified as advanced prostate cancer (T3-T4 and/or N1-N3 and/or M1, or stage coded in the recruitment center as metastatic). Grade information was

available for 2,986 cases (68.2%): 2,438 were low-intermediate grade (Gleason score less than 8, or grade coded as well differentiated, or moderately or poorly differentiated) and 544 were identified as high-grade prostate cancer (Gleason score of 8+, or grade coded as undifferentiated). There was a small difference in the frequency of vasectomy between patients with prostate cancer who did and did not have tumor subtype information; thirteen percent of participants with information on tumor stage had had a vasectomy, compared with 17% of participants without tumor stage information; and 14% of participants with information on tumor grade had had a vasectomy, compared with 16% of participants without tumor grade information. By the end of follow-up, 15,285 men had died, of whom 632 had died of prostate cancer.

4.3.2 Laboratory Assays

Assay data were available for men in the EPIC cohort who had been selected as controls in a selection of published and unpublished, matched nested case-control study of prostate cancer. Broadly, each control had been selected at random from the cohort of men who were alive and free of cancer (excluding nonmelanoma skin cancer) at the time of diagnosis of their index case, using an incidence density sampling protocol (further details on matching methods can be found in Travis et al) [225]. Follow-up for index prostate cancer cases was conducted in three phases, occurring approximately in 2004, 2008, and 2010. For all assay measurements, laboratory personnel were blind to the case status of the samples and each sample was analyzed together with, at minimum, duplicate quality control samples.

Seminal Analytes

Immunoassay measurements for total PSA, free PSA [226] intact PSA, human kallikrein 2 (HK2) [227, 228] and microseminoprotein-β (MSP) [229, 230] were conducted in plasma samples from 1,469 men from phases 1, 2, and 3 on the AutoDELFIA 1235 automatic immunoassay system (PerkinElmer, Turku, Finland) at the Wallenberg Research Laboratories, Department of Translational Medicine, Lund University, Skåne University Hospital, Sweden. All intra- and interassay coefficients of variation were $< 9\%$.

Insulin-like Growth Factors

Serum samples of IGF-1 and IGFBP-3, for controls selected for phase 1 of the EPIC study, were assayed in singleton at IARC, Lyon (France) using the DSL-10-5600 ACTIVE ELISA from Diagnostic Systems Laboratories (DSL). Serum samples of IGF-1, for controls selected for phase 2 of the EPIC study, were assayed in duplicate at the Cancer Epidemiology Unit (CEU) laboratory in Oxford (UK) using the DSL-10-5600 ACTIVE ELISA from DSL. For IGF-1, assays also included a step to dissociate IGF-I from IGF-I-binding proteins before measurement. Serum samples of IGFBP-2 were assayed by radioimmunoassay, and IGFBP-1 by immunoradiometric assay, from Diagnostic System Laboratories (Webster, TX). All intra- and interassay coefficients of variation were $< 13\%$, with exception for the interbatch coefficient of variation for IGFBP-1, which was 24.2% [231, 232, 233]. While detailed information was unavailable for IFG-2, assay methods are broadly similar to those described here [234].

Endogenous Sex Steroid Hormones

Serum testosterone concentrations were measured by radioimmunoassays (Immunotech, Marseilles, France). Androstenedione concentrations were measured by a radioimmunoassay with a double antibody system for the separation of free and bound antigen (Diagnostic Systems Laboratories Inc., Webster, TX, USA). SHBG was measured by a solid phase sandwich IRMA (Cis-Bio International, Gif-sur-Yvette, France). All hormone assays were performed by the laboratory of the Hormones and Cancer Team at IARC. All intra- and interassay coefficients of variation were $< 12\%$ [235]. Details were unavailable for androstanediol glucuronide.

4.3.3 Statistical Analyses

Cox proportional hazards models were used to estimate the hazard ratios (HRs) and 95% confidence intervals (95% CIs) for prostate cancer incidence, and separately for death from prostate cancer using age as the underlying time variable. Entry age was defined as the participant's age at recruitment, and exit age was age at diagnosis of prostate cancer, death, loss to follow-up or censoring at the end of follow-up period for each center, whichever was first. The slope of the Schoenfeld residuals over time was used to verify the

proportionality of hazards. All models were stratified by age at enrolment (<50, 50-54, 55-59, 60-64, 65-69, ≥70 years) and EPIC recruitment center. Multivariable models were adjusted for factors suspected to be associated with prostate cancer or vasectomy, including education (less than university, university graduate), smoking status (never, former, current), body mass index (BMI) (<20, 20-24, 25-29, ≥30 kg/m2), alcohol intake (<8, 8-15, 16-39, ≥40 g/d ethanol), and physical activity (inactive, moderately inactive, moderately active, active). Missing values were assigned to separate categories for education (3.8%), smoking status (1.5%), BMI (0.6%), and physical activity (0.8%), and missing indicators were used in the statistical models. Additional analyses were conducted adjusting for marital status (single, married, divorced, widower), however, this information was only available for the UK and Germany.

Subgroup analyses were conducted according to age at vasectomy (<38 vs. ≥ 38 years), time since vasectomy (<25 vs. ≥ 25 years), by median BMI and alcohol consumption, physical activity, marital status, educational attainment, and smoking status. Additional analyses were conducted by tumour subtypes: stage (localised vs. advanced) and tumour grade (low-intermediate vs. high). Tests for heterogeneity in the association between vasectomy and prostate cancer were likelihood ratio tests for subgroup analyses and competing risk methods [236] for stage and grade analyses. Country-specific associations were estimated using the aforementioned Cox regression models and tests for heterogeneity were by likelihood ratio test.

For a subset of 6,771 men in the EPIC-Oxford subcohort, for whom data on PSA testing were available, we investigated the suggestion that any association between vasectomy and prostate cancer is influenced by differences in the use of PSA testing in men who have undergone vasectomy [33]. We used multivariable logistic regression to estimate the association of vasectomy with having a PSA test, adjusted for BMI and age at completion of questionnaire.

We also evaluated the cross-sectional association of vasectomy status with naturally logarithm-transformed analyte concentrations (MSP, PSA [total, free, intact, free-to-total], HK2, IGF-1, IGF-2, IGFBP-1, IGFBP-2, IGFBP-3, androstenedione, androstanediol glucuronide, testosterone, and SHBG) using analysis of variance to compare geometric means adjusted for age at recruitment, BMI, recruitment centre, and laboratory batch in a subset of

1,469 men [1] without prostate cancer.

Much of the evidence on prostate cancer risk and vasectomy, to date, has been from studies with a low number of incident cases (< 150) [200, 201, 203, 204, 202], and so risk estimates have been subject to substantial uncertainty. However, a number of recent, large cohort studies provide robust, but inconsistent, estimates of the association of vasectomy with prostate cancer. Where summary estimates are given combining our results with those from these large cohorts to date [23, 34, 24], study-specific results were combined using the method of inverse variance weighted least squares.

4.4 Results

Overall, 84,743 men were followed up for a median of 15.4 years (range, 0-20 years), of whom 4,377 developed prostate cancer. The mean age at recruitment was 53 years, which ranged from 50 years in Spain to 56 years in Denmark. The mean age at diagnosis of prostate cancer was 68 years, with a range of 65 years in Germany to 71 years in the United Kingdom. The proportion of men with self-reported vasectomy was 15% (n = 12,712), which ranged from 4.1% (n = 863) in Germany to 20.5% (n = 4,640) in the United Kingdom. For the 97.9% (n = 12,455) of men who had undergone vasectomy and who also provided age at vasectomy, median age at vasectomy was 38 years.

Compared to men without a vasectomy, men with a vasectomy were, on average, younger at recruitment (52 years vs. 54 years), had a lower education level (university graduate, 26% vs. 33%), and were more physically active (19% vs. 14%). Men with a vasectomy were more likely to be married (91% vs. 80%). Vasectomy status also varied significantly by smoking status and alcohol consumption, although the magnitude of the differences was small (Table 6.1). Additionally, an analysis in the EPIC-Oxford sub-cohort showed that men who had undergone a vasectomy were 54% more likely to have had a PSA test when compared with men without a vasectomy (OR: 1.54, [95% CI 1.35-1.76]). In men from the EPIC-Oxford subcohort updated information on vasectomy status were available and found that 5.1% of men without a vasectomy at baseline reported having had a vasectomy during the 10 years after recruitment.

[1] Numbers for available samples differ moderately by analyte.

4.4.1 Vasectomy and prostate cancer

Of the 4,377 men with prostate cancer for whom vasectomy status was available, 641 (14.6%) had a self-reported vasectomy at recruitment. Vasectomy was not significantly associated with prostate cancer risk after stratification by recruitment center and age at recruitment (HR: 1.05, [95% CI 0.96-1.15]). Additional adjustment for BMI, smoking status, marital status, educational attainment, alcohol consumption, physical activity, and protein from dairy sources did not alter results (HR: 1.05, [95% CI 0.96-1.15]; Table 6.2). No evidence of heterogeneity was found in the association between vasectomy and prostate cancer by the stage of disease (P = 0.6) or recruitment country (P = 0.09; Table 6.3). However, there was evidence of heterogeneity by tumor grade (P = 0.02; Table 6.2); vasectomy was associated with an increased risk of low- intermediate grade (HR: 1.14, [95% CI 1.01-1.29]) but not of high-grade prostate cancer (HR: 0.83, [95% CI 0.64-1.07]). Additionally, there was no significant association of vasectomy with death due to prostate cancer (HR: 0.88, [95% CI 0.68-1.12]; Table 6.2).

There was significant heterogeneity in the association by median age at vasectomy (38 years) (P = 0.04; Table 6.2); compared to men who had not had a vasectomy, men who had a vasectomy below the median age were at a significantly increased risk of prostate cancer (HR: 1.18, [95% CI 1.03-1.35]), whereas, there was no significant association with prostate cancer in men who had a vasectomy above the median age (HR: 0.99, [95% CI 0.89-1.09]). There was also significant heterogeneity for the association of vasectomy with prostate cancer by median-defined strata of alcohol consumption (P = 0.03); in men with below median alcohol consumption those who had had a vasectomy were at a significantly increased risk of prostate cancer (HR: 1.16, [95% CI 1.02-1.31]) compared to men without a vasectomy, while for men with above median alcohol consumption vasectomy was not associated with prostate cancer (HR: 0.96, [95% CI 0.84-1.08]). No heterogeneity was observed for subgroup analyses by BMI, physical activity, marital status educational attainment, or smoking status (not shown). There was no heterogeneity in the association with prostate cancer risk by time since vasectomy (P = 0.9; Table 6.2).

4.4.2 Circulating concentrations of analytes and vasectomy

When compared to men without a vasectomy, men with a vasectomy had significantly higher concentrations of MSP (multivariable-adjusted geometric mean, ng/ml: 14.2 [95% CI 12.9-15.6] vs. 12.8 [95% CI 12.4-13.1], P = 0.03) and IGFBP-2 (multivariable-adjusted geometric mean, ng/ml: 389.5 [95% CI, 347.8-436.1] vs. 338.8 [95% CI, 325.2-352.9], P = 0.03), and significantly lower levels of androstenedione (multivariable-adjusted geometric mean, nmol/L: 3.9 [95% CI, 3.6-4.4] vs. 4.6 [95% CI, 4.5-4.8], P = 0.01). No significant differences by vasectomy status were observed by PSA (total, free, intact, or free-to-total), HK2, IGF-I, IGF-II, IGFBP-1, IGFBP-3, androstanediol glucuronide, testosterone, or SHBG, all P >0.05 (see Table 6.4).

4.4.3 Pooled evidence from existing large cohort studies and current results

An inverse variance weighted meta-analysis of three large cohort studies for the association of vasectomy with prostate cancer risk and current results shows a significant association (RR: 1.04, [95% CI 1.01-1.08]). However, there was little evidence to suggest an association of vasectomy with risk for high grade (RR: 1.01, [95% CI 0.91-1.14]) or advanced stage disease (RR: 1.06, [95% CI 0.97-1.17]), or for death from prostate cancer (RR: 1.03, [95% CI, 0.96-1.10]; Figure 4.1).

4.5 Discussion

In this large prospective European study, vasectomy was not associated with risk of prostate cancer overall, with risk for high grade or advanced stage tumours, or for death from prostate cancer. However, there was some evidence that vasectomy may be associated with an elevated risk of low-intermediate grade disease, and that having had a vasectomy is associated with also having had PSA test. Additionally, although a meta-analysis of available evidence from large cohort studies found a significant association of vasectomy with prostate cancer risk, no significant association was observed for high grade or advanced stage disease, or death from prostate cancer.

Further, there were significantly higher concentrations of MSP and IGFBP-2, and significantly lower concentrations of androstanediol glucuronide in men with a vasectomy compared to men without a vasectomy. No differences were found for PSA analytes or HK2, or other insulin-like growth factor or endogenous sex steroid hormones by vasectomy status.

Three of the eight previous cohort studies on vasectomy have reported an increased risk of prostate cancer in men with vasectomies [23, 200, 201]. However, aside from the recent Health Professionals Follow-Up Study (HPFS) [23], the CPS-II cohort [34], and a population based matched cohort study of residents in Ontario [24] (PBCO), all studies have had a low number of incident cases (< 150), and so risk estimates have been subject to substantial uncertainty. The results from HPFS cohort suggested a modest, 10% elevated risk of overall prostate cancer, and an elevated risk for aggressive tumour subtypes, with a 22% increased risk of high grade and a 20% increased risk of advanced stage tumours for men with vasectomies compared to men without vasectomies. In contrast, the current study, CPS-II [34], and PBCO [24] found no significant association of vasectomy with prostate cancer overall, high-grade or advanced-stage disease, or death from prostate cancer. Furthermore, the weight of current evidence from a pooled estimate of current evidence does not support the previously hypothesized role of vasectomy as a risk factor for more aggressive tumors (Figure 4.1).

There is no established biological rationale for an association of vasectomy with prostate cancer risk [35]. During a vasectomy, the vas deferens is cut, blocked, or sealed to prevent the sperm from reaching the seminal fluid. Although previous studies have investigated a series of theoretical mechanisms that include immunological response [237], changes to cell proliferation [212], and endocrine function [23, 213], the biological significance of these pathways in humans is unclear.

4.5.1 Circulating concentrations of analytes and vasectomy

Seminal analytes

This study addressed a recent suggestion that there may be differential regulation of seminal analytes in men after vasectomy [23, 219]; among 1,469 men without prostate cancer, we found little evidence that vasectomy is associated

with different blood plasma concentrations of PSA variants or HK2. However, significantly higher blood plasma concentrations of MSP were found in men with vasectomy compared with those without. MSP is a protein abundant in the seminal fluid [39] that has been found at significantly higher concentrations in the seminal plasma in infertile men when compared with fertile men [238]. However, circulating concentrations have been previously inversely associated with prostate cancer [27] and so our observation of higher circulating concentrations of MSP in men who had undergone vasectomy does not provide evidence in favor of vasectomy as a risk factor of prostate cancer.

Insulin-like growth factors

There were significantly higher concentrations of IGFBP-2 among men with a vasectomy compared to men without a vasectomy. IGFBP-2 is produced in the liver and other tissues by prostate epithelial cells and is present in the seminal plasma [239, 240, 241, 242]. There is evidence that serum concentrations of IGFBP-2 are elevated in prostate cancer patients, and that concentrations are positively predicted by stage of tumour [240]. However, IGFBP-2 concentrations have also been shown positively associated with less aggressive disease and with reduced risk of progression [243, 244]. As such, it remains unclear if these data are evidence in favor of a cell proliferation hypothesis for vasectomy and prostate cancer.

Endogenous sex steroid hormones

Androstenedione is an endogenous androgen steroid hormone that may act as a precursor for testosterone production in prostate tissue [245]. A previous publication using EPIC data reported a lower risk of advanced prostate cancer associated with increased concentrations of androstenedione [246], and a recent review suggests that reduced androstenedione may be a precursor for androgen production in castrate-resistant prostate cancer cells [247, 245]. Although there is evidence that androstenedione is found at reduced concentrations in vasectomised men [248], the majority of prior research finds no evidence that androstenedione concentrations differ significantly by vasectomy status [23, 213, 217, 218]. As the difference we observed for the men in the EPIC cohort was modest, it is not clear that the differential regulation of androstenedione provides support for a sex hormone/androgen pathway hypothesis for vasectomy and prostate cancer.

4.5.2 Health monitoring behaviours and vasectomy

A significantly increased risk of low-intermediate grade prostate cancer in men who had had a vasectomy in the current study might be at least partly explained by differences in the use of PSA testing. Men who receive a vasectomy may be more likely to attend health-care services and have their PSA level tested, and thus also be more likely to be diagnosed with prostate cancer, especially low-grade disease. Evidence for this hypothesis comes both from our finding that, in the EPIC-Oxford subcohort, men with a vasectomy were more likely to have had a PSA test than men without a vasectomy, and from other previous cohort studies, which also found that men with a vasectomy were more likely to have a history of PSA testing than men without a vasectomy [23, 200, 34]. Nonetheless, it is notable that the HPFS [23] found a significant association of vasectomy with death from prostate cancer only in a highly screened sub-cohort (HR, 1.56; 95% CI 1.03 to 2.36). However, for this sub-cohort all other associations of vasectomy with prostate cancer overall and by tumour subtype were null, and so this finding provides only tenuous evidence for an association outwith health monitoring behaviours and PSA testing. There was also some evidence for heterogeneity in the association of vasectomy with prostate cancer risk by age at vasectomy and alcohol intake, but the implications of these subgroup analyses are unclear.

4.5.3 Limitations

The lack of information on PSA testing history has been previously raised as a limitation for investigations into vasectomy and prostate cancer risk. Evidence from the European Randomized Study of Screening for Prostate Cancer suggests that when PSA testing is offered to all men, it reduces prostate cancer mortality by approximately 28% at 13 years of follow-up [249]. If we assume that the use of PSA testing is 20% among men who have not had a vasectomy and 40% among men who have (data from EPIC-Oxford), it is possible that increased screening in the latter could result in a 5.6% reduced risk of death due to prostate cancer. This suggests that although it is possible that an adverse effect of vasectomy on the risk of potentially lethal prostate cancer is being partly masked by a beneficial effect of increased PSA testing, any such bias is likely small. Nevertheless, it remains a limitation of the current investigation that, because of limited information on PSA testing,

we were unable to more fully address the role of PSA testing in the proposed association of vasectomy with prostate cancer.

Due to the lack of updated data collection for vasectomy status, it is possible that our results were biased by misclassification of men who had undergone vasectomy as being nonvasectomized. However, for the EPIC-Oxford subcohort, updated data on vasectomy status were available and showed that 5.1% of men without a vasectomy at baseline reported having had a vasectomy during the 10 years after recruitment. Furthermore, a recent report suggested that a small misclassification of men who had undergone vasectomy as nonvasectomized would likely result in only a minimal underestimate of any association [34] of vasectomy with prostate cancer risk.

4.6 Conclusions

Ultimately, this investigation of 84,753 men in the EPIC cohort did not find a significant association between vasectomy and overall prostate cancer, high-grade or advanced-stage tumors, or death due to prostate cancer. The small increase in the risk of low-intermediate grade prostate cancer in men who had had a vasectomy may be due to differences in health-monitoring behaviors. Further, a pooled estimate of available evidence finds that although vasectomy may be associated with a small increased risk of being diagnosed with prostate cancer, there is no association with high-grade or advanced-stage tumors, or death due to prostate cancer.

Table 4.1: Characteristics of control participants by vasectomy status in EPIC

	No Vasectomy (N=72,041)	Vasectomy (N=12,712)
Age at recruitment, yrs	54.0 (48.0, 60.0)*	52.00 (47.0, 57.0)*
Weight at recruitment, kg	80.5 (73.3, 88.5)*	80.1 (73.4, 88.0)*
Height at recruitment, kg	175.0 (170.0, 179.8)*	175.0 (170.5, 180.0)*
BMI at recruitment, kg/m^2	26.4 (24.2, 28.9)*	26.2 (24.2, 28.4)*
Smoking status		
Never	23,535 (33%)	3,945 (32%)
Former	27,332 (38%)	4,796 (38%)
Current	20,497 (29%)	3,761 (30%)
Alcohol consumption, g/d		
<8	22,271 (31%)	3,952 (31%)
8-15	14,297 (20%)	2,712 (21%)
16-39	20,898 (29%)	3,572 (28%)
40+	14,574 (20%)	2,476 (19%)
Physical activity		
Inactive	14,538 (20%)	2,465 (20%)
Moderately inactive	19,599 (27%)	3,239 (26%)
Moderately active	27,220 (38%)	4,393 (35%)
Active	10,117 (14%)	2,419 (19%)
Marital status		
Single	4,090 (11%)	66 (1.2%)
Married	29,799 (80%)	4,816 (91%)
Divorced	2,788 (7.4%)	380 (7.1%)
Widower	779 (2.1%)	58 (1.1%)
Educational attainment		
Primary or less	22,541 (32%)	4,052 (34%)
Secondary	17,814 (26%)	3,773 (32%)
Technical	5,701 (8.2%)	1,004 (8.4%)
Degree	23,116 (33%)	3,129 (26%)

* Data given as medians (interquartile range)

Body mass index (BMI)

Numbers may not sum to total as a result of missing values

Table 4.2: Hazard ratio (HR) and 95% confidence intervals (CI) for vasectomy and prostate cancer in 84,753 men in EPIC

Variable	Cases	Minimally adjusted HR (95% CI)[1]	Multivariable adjusted, HR (95% CI)[2]	p-heterogenity
Total prostate cancer				
Without vasectomy	3,736	1.00 (reference)	1.00 (reference)	
With vasectomy	641	1.05 (0.96-1.15)	1.05 (0.96-1.15)	
Age at vasectomy				
Without vasectomy	2,532	1.00 (reference)	1.00 (reference)	
With vasectomy				
<38 years	246	1.17 (1.02-1.34)	1.18 (1.03-1.35)	
≥ 38 years	395	0.99 (0.89-1.10)	0.99 (0.89-1.09)	0.04[3]
Years since vasectomy				
Without vasectomy	3,736	1.00 (reference)	1.00 (reference)	
With vasectomy				
<25 years	258	1.07 (0.94-1.22)	1.07 (0.94-1.22)	
≥ 25 years	378	1.06 (0.95-1.18)	1.06 (0.95-1.19)	0.9[3]
Localised prostate cancer				
Without vasectomy	1,521	1.00 (reference)	1.00 (reference)	
With vasectomy	212	1.09 (0.93-1.27)	1.09 (0.93-1.27)	
Advanced prostate cancer				
Without vasectomy	850	1.00 (reference)	1.00 (reference)	
With vasectomy	150	1.01 (0.85-1.22)	1.01 (0.84-1.21)	0.7[4]
Low grade prostate cancer				
Without vasectomy	2,090	1.00 (reference)	1.00 (reference)	
With vasectomy	348	1.15 (1.02-1.29)	1.14 (1.01-1.29)	
High grade prostate cancer				
Without vasectomy	475	1.00 (reference)	1.00 (reference)	
With vasectomy	69	0.82 (0.63-1.07)	0.83 (0.64-1.07)	0.02[4]
Fatal prostate cancer				
Without vasectomy	555	1.00 (reference)	1.00 (reference)	
With vasectomy	77	0.87 (0.68-1.11)	0.88 (0.68-1.12)	

[1] From a Cox proportional hazards model stratified by recruitment centre and age at recruitment

[2] As model[1], with adjustment for BMI, smoking, marital status, education, alcohol consumption, physical activity, and protein from dairy sources

[3] Test for heterogeneity was by likelihood ratio test

[4] Test for heterogeneity was by competing risks method

Table 4.3: Hazard ratio (HR) and 95% confidence intervals (CI) for vasectomy and prostate cancer in 84,753 men by recruitment country in EPIC

Country	Cases	Minimally adjusted HR (95% CI)[1]	Multivariable adjusted, HR (95% CI)[2]	*p*-heterogeneity
Spain				
Without vasectomy	600	1.00 (reference)	1.00 (reference)	
With vasectomy	64	1.23 (0.93-1.62)	1.23 (0.93-1.63)	
Germany				
Without vasectomy	788	1.00 (reference)	1.00 (reference)	
With vasectomy	37	1.49 (1.07-2.09)	1.48 (1.06-2.08)	
UK				
Without vasectomy	812	1.00 (reference)	1.00 (reference)	
With vasectomy	200	0.98 (0.83-1.15)	0.96 (0.82-1.13)	
Denmark				
Without vasectomy	1,536	1.00 (reference)	1.00 (reference)	
With vasectomy	340	1.03 (0.91-1.16)	1.03 (0.92-1.17)	0.09[3]

[1] From a Cox proportional hazard models stratified by recruitment centre and age at recruitment

[2] As model[1], with adjustment for BMI, smoking status, marital status, educational attainment, alcohol consumption, physical activity, and protein from dairy sources

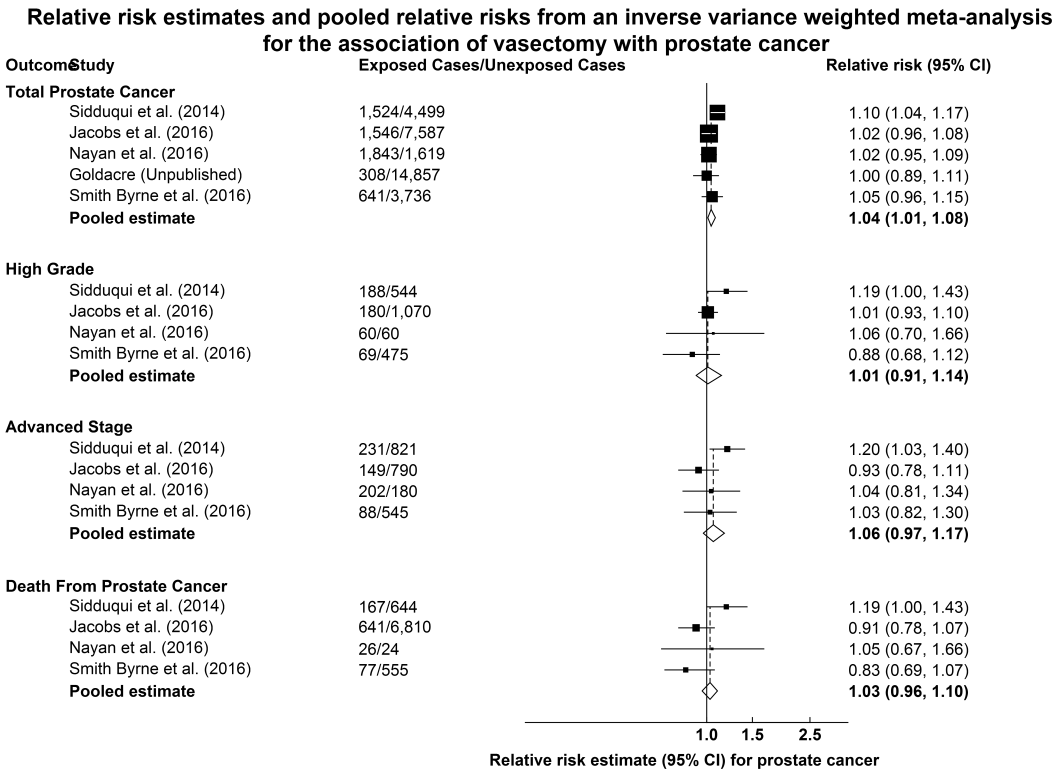
[3] Test for heterogeneity was by likelihood ratio test

Table 4.4: Adjusted geometric mean concentrations [1] of analytes in control men by vasectomy status

		Vasectomy Status		
		Yes	No	
Analytes	**Vasectomy/No Vasectomy, n**	**Mean (95% CI)**	**Mean (95% CI)**	***p*-value** [1]
IGF-1, nmol/l	111/949	20.9 (22.33-13.1)	20.2 (19.75-20.60)	0.2
IGFBP-1, ng/ml	68/482	5.9 (4.4-7.9)	5.3 (4.8-5.9)	0.6
IGF-2, ng/ml	68/482	849.8 (792.2-911.5)	868.6 (846.9-890.9)	0.6
IGFBP-2, ng/ml	68/482	389.5 (347.8-436.1)	338.8 (325.2-352.9)	0.03
IGFBP-3, ng/ml	38/430	128.1 (118.9-137.9)	126.5 (123.8-129.2)	0.8
Androstenedione, nmol/l	38/431	3.9 (3.6-4.4)	4.6 (4.5-4.8)	0.01
Androstanediol glucuronide, nmol/l	38/433	14.3 (11.8-17.2)	13.1 (12.5-13.9)	0.4
Testosterone, nmol/l	34/393	13.8 (12.0-15.8)	15.6 (14.9-16.2)	0.1
SHBG, nmol/l	38/407	38.3 (33.5-43.7)	42.6 (40.9-44.3)	0.1
MSP, ng/ml	142/1327	14.2 (12.9-15.6)	12.8 (12.4-13.1)	0.03
Total PSA, ng/ml	142/1328	0.86 (0.75-0.98)	0.87 (0.83-0.91)	0.9
Intact PSA, ng/ml	142/1328	0.12 (0.10-0.14)	0.13 (0.12-0.14)	0.3
Free PSA, ng/ml	142/1328	0.27 (0.24-0.30)	0.28 (0.27-0.29)	0.7
Free-to-total PSA, ng/ml	142/1325	31.3 (29.3-33.4)	31.7 (31.1-32.4)	0.7
HK2, ng/ml	142/1328	0.029 (0.025-0.33)	0.029 (0.027-0.029)	0.9

[1] *p*-values are calculated from analyses of variance adjusted for age, BMI, recruitment centre, and laboratory batch

Figure 4.1: Relative risk estimates and pooled relative risks from an inverse variance weighted meta-analysis for the association of vasectomy with prostate cancer



Chapter 5

Microseminoprotein-β levels and prostate cancer risk in the European Prospective Investigation into Cancer and Nutrition

5.1 Introduction

A previous prospective study in the Multi-Ethnic Cohort (MEC) [27] found a significant protective assocation of microseminoprotein-β (MSP) with prostate cancer. However, due to limited follow-up and an unclear association of MSP with prostate-specific antigen (PSA) the relationship of MSP with prostate cancer aetiology remains unclear. This chapter with investigate the association of MSP with subsequent risk of prostate cancer, and but will not address the discriminative properties of MSP.

5.1.1 rs10993994 and the MSMB gene

rs10993994 is a single nucleotide polymorphism located 57 base-pairs upstream in the 5' promoter region of the MSMB gene on chromosome 10, which encodes MSP [39]. A genome-wide association study (GWAS) found a significant increased risk of prostate cancer associated with rs10993994 [13, 27]. In particular, in a large replication study, T homozygotes had a 57% elevated risk of prostate cancer compared to C homozygotes - for heterozygotes, a 21% increased risk was observed [36]. Further, fine-mapping studies of the MSMB region confirm the association with prostate cancer risk is attributable primarily to rs10993994 and not alternative surrounding nucleotide polymorphisms [250]. Notably, the association of rs10993994 with prostate cancer has been observed for men of both European and African decent [251], which may serve as evidence of an rs10993994 role in cross-ethnic prostate cancer aetiology.

Functional studies to date suggest that rs10993994 is a regulator of MSMB expression. Promoter activity in the MSMB gene was reduced by 87% in cells homozygous for the prostate cancer risk T allele compared with those homozygous for the C allele in LNCa cell lines. Further, when cell lines were treated with synthetic androgen R1881 a dose-dependent increase in promoter activity was observed only for carriers of the C allele [252]. These results are commensurate with findings from tumour cell lines that show significantly higher expression in the MSMB gene for CC and CT genotypes than for TT carriers [253, 252].

5.1.2 Microseminoprotein-β

MSP is a cysteine rich non-glycosylated protein, also commonly referred to as prostate inhibitory protein or prostate secretory protein of 94 amino acids, and is one of the three most abundantly secreted proteins by the prostate epithelium into the seminal fluid [39]. However, there is also evidence that it is present at lower concentrations in the secretions from the respiratory tract; immunohistochemical staining suggests MSP is secreted in the epithelium and sub-mucosal glands of the bronchi and trachea. Additionally, MSP has been found at low concentrations in the stomach and the uterine cervix [254, 255].

To date, the precise biological function of MSP remains unclear. Several studies suggest it may be involved in tumour suppression [256, 257, 258, 259, 260, 261] or pathogen defense [262]. However, MSP is also found on the surface of spermatozoa and may inhibit spontaneous acrosome reactions. As such, it has also been associated with semen quality and reproduction [263, 238].

5.1.3 Factors associated with Microseminoprotein-β

The strongest known predictor of MSP concentrations is rs10993994 genotype; up to 38% and 23% of the variance in blood and semen concentrations of MSP in healthy males can be explained by rs10993994 genotype, respectively, with the lowest MSP concentrations observed in T homozygotes [250].

In blood plasma, MEC found significant heterogeneity in MSP concentrations by ethnicity; concentrations were highest in Caucasian men (21 ng/ml) and lowest in Native Hawaiian men (15 ng/ml) [27]. Further, MSP is found at higher concentrations in older men and men with lower body mass index (BMI) [27].

When measured in the tracheobronchial tree in epithelial secretory cells, MSP expression is found at up to 2.5 fold higher levels in smokers compared to non-smokers [264]. However, there is no evidence to date that smoking status predicts MSP concentration in blood plasma or serum.

5.1.4 Microseminoprotein-β and prostate cancer

In MEC [27] there was a 2% increased risk of prostate cancer per unit decrease in MSP (ng/ml). Further, circulating MSP concentrations have been shown

to be significantly lower in men with aggressive disease [265, 266] and there is tentative evidence that suggests circulating concentrations of MSP are lower in men with Gleason scores of 7 or higher when compared to men with low grade disease (Gleason scores < 7) [265]. However, tissue MSP expression has also been found to be lower in localized prostate tumors than in advanced tumors, and there is some evidence that men with lower MSP expression may have longer progression-free survival times than men with high MSP expression profiles [267, 268, 269]. Additionally, there is evidence at a mechanistic level that MSP may act to inhibit prostate cancer tumour growth by inducing apoptosis [270].

5.1.5 Aim of Study

This chapter will evaluate whether MSP concentration is associated with subsequent risk of prostate cancer overall, and if this association varies by tumour characteristics for men in the European Prospective Investigation into Cancer and Nutrition (EPIC). It will then investigate the association of MSMB rs10993994 with circulating concentrations of MSP in EPIC and proceed to use this genetic variant as an instrument for MSP to test its causal role through Mendelian randomization (MR) analyses by combining EPIC-derived estimates with published data from the PRACTICAL consortium [12].

5.2 Methods

5.2.1 Study Population

These analyses use data from the following countries within the EPIC cohort: Germany, Greece, Italy, The Netherlands, Spain, and the United Kingdom[140]. For complete details on the cohort disease, please see Chapter 3.

5.2.2 Follow-up for cancer incidence and vital status

Cancer incidence was identified through record linkage to regional or national registries in most countries. For Germany and Greece, combined health insurance records, regional health departments, municipality registries, hospital or physician-based cancer and pathology records, or mail or phone call-based follow-up were used. Follow-up procedures continued to date of prostate cancer diagnosis, death, or last follow-up completed (from 31 December 2007 to 14 June 2010 according to recruitment center).

Cases were defined as men who were diagnosed with prostate cancer (International Classification of Diseases 10th revision code C61 [173]) after the date of blood collection and before the end of the study period, as determined by the latest date of follow-up in each study center. An incidence density sampling protocol was used to select control participants at random from the cohort of men who were alive and free of cancer (except non-melanoma skin cancer) at the time of diagnosis of the index case and who were matched on study center, length of follow-up, age at blood collection ($\pm$ 6 months), time of blood collection ($\pm$ 1 hour), and duration of fasting at blood collection ($<$ 3, 3-6, $>$ 6 hours). For the current analyses participants were 1,871 cases with 1,871 matched controls.

Analysis suggested this study would be sufficiently powered to discover an association of MSP with prostate cancer risk as was previously observed in MEC [27]; 500 cases and 500 controls would be needed to discover a 30% reduced relative risk of prostate cancer for the top fourth compared to the bottom fourth of MSP concentrations with 80% power (α=0.05) [220].

Information on tumor stage at diagnosis was available for 1,263 cases (67.5%): 886 were localized (tumor-node-metastasis [TNM] staging score of T1-T2 and N0/Nx and M0/Mx, or stage coded in the recruitment centre as

localized); 377 were identified as advanced prostate cancer (T3-T4 and/or N1-N3 and/or M1, or stage coded in the recruitment centre as metastatic). Information on tumor grade at diagnosis was available for 1,554 cases (85.1%): 1,357 were low-intermediate grade (Gleason score < 8, or grade coded as well, moderately, or poorly differentiated) and 197 were identified as high-grade prostate cancer (Gleason score ≥ 8, or grade coded as undifferentiated).

5.2.3 Assessment of MSP and PSA

Immunoassay measurements for total PSA [226] and MSP[271] were conducted on the AutoDelfia®1235 automatic immunoassay system at the Wallenberg Research Laboratories, Department of Translational Medicine, Lund University, Skåne University Hospital, Malmö, Sweden, and with all measurements conducted blinded to case status. We measured total PSA using the dual-label DELFIA Prostatus®total PSA-Assay (Perkin-Elmer, Turku, Finland) [226] calibrated against the WHO 96/670 (PSA-WHO) standard. Production and purification of the polyclonal rabbit anti-MSMB antibody, protocols for biotinylation and Europium labeling of the anti-MSP antibody, and performance of the MSP-immunoassay were performed as previously reported[27, 271].

In a pilot study, the concordance between measurements of MSP and PSA analyte concentration in serum and citrated plasma samples was assessed ($N = 25$ for serum and plasma, respectively), and a high concordance was observed ($r = 0.98$ and 0.99, for MSP and total PSA, respectively). Additionally, the temporal reproducibility of analyte concentrations was assessed between samples drawn at five year intervals from 49 and 40 individuals for MSP and PSA, respectively. There were no significant differences between serum concentrations of MSP ($p = 0.4$) or PSA ($p = 0.8$) drawn at five year intervals, and the intra-class correlation coefficient was 0.11 (95% CI: 0.00-0.38) for MSP and 0.23 (95% CI 0.00-0.53) for PSA. Low ICC may be expected as pilot was conducted in controls with concentrations very close to lower limits of detection where assay is likely moderately less accurate. Given the wider availability of plasma samples for the EPIC cohort, all assays for the nested case-control study were performed using plasma.

Quality control samples were inserted into each assay batch and analysed in duplicate; samples reflected low, medium, and high values according to the calibration chart (for PSA range was 0.05-250 ng/ml and 0.2-90 ng/ml

for MSP). The average inter-assay coefficients of variation (CVs) for MSP and PSA were 8% and 14% for phase 1, 5% and 9% for phase 2, and 15% and 7% for phase 3, respectively. An additional 286 "blinded" quality control samples were inserted. Quality control samples were pooled sodium citrate plasma from at least two males. All intra- and inter-assay CVs were $< 9\%$. The detectable ranges for MSP were 0.2 to 90 ng/ml and 0.1 to 250 ng/ml for PSA.

5.2.4 Statistical Analyses

Plasma concentrations below the lower limits of detection for MSP and PSA were set to half of the lowest value of detection (PSA, N = 7) while concentrations above the upper levels of detection were set to the highest detectable value for that particular analyte (MSP, N = 82; PSA, N = 65). Pearsons chi-squared tests for differences and paired t-tests for categorical and continuous variables, respectively, were conducted between matched case-control sets for anthropometric and lifestyle characteristics. Analysis of variance was used to assess differences in analyte concentrations in controls by strata of selected characteristics, and by country and study phase (prostate cancer follow-up was conducted in three waves, occurring approximately in 2004, 2008, and 2010). Additionally, analysis of variance was used to test for differences in mean analyte concentrations by days in post for samples from the EPIC-Oxford sub-cohort. To conform to parametric model assumptions, log transformations were applied to MSP and PSA concentrations and results are presented as geometric means adjusted for age at blood collection, body mass index (BMI), recruitment centre, and laboratory batch. The relationship between transformed analyte concentrations was examined using partial correlation adjusted for exact age, BMI, recruitment centre, and laboratory batch.

Conditional logistic regression models were used to examine the association of MSP concentration with risk of prostate cancer, conditioned on the matching factors (listed above), and adjusted initially for BMI and exact age, and also for fourth of PSA concentration in a further adjusted model (these adjustment factors were chosen after additional adjustment was shown not to materially alter the results, see Table 6.1). Conditional logistic regression models were repeated in subgroups defined according to study phase, time between blood collection and diagnosis ($\leq$ 7, 7-10, $\geq$10 years), age at

blood collection (≤60, >60 years), age at diagnosis (≤65, >65 years), and smoking status (never, previous, current). In addition, we conducted analyses by prostate tumor stage (localized, advanced), and histological grade (low-intermediate or high grade). Due to the strong positive predictive value of PSA, an additional analysis stratified by median PSA concentration was unconditional and adjusted for exact age, BMI, recruitment centre, fourth of PSA concentration, and matching factors. For all regression models linear trend was tested by entering a pseudo-continuous variable equal to the medians of the fourths of MSP concentration. For subgroup analyses, likelihood ratio tests were used to test for heterogeneity of the association of MSP concentration with risk of prostate cancer.

Genotype data were available for rs10993994 for a subset of 1,068 EPIC cases and 1,186 EPIC controls from the iCogs [12], OncoArray [37], and BPC3 [38] genotyping projects. Logistic regression models were used to investigate the association of rs10993994 genotype with prostate cancer risk.

I investigated the potential causal role of MSP in prostate cancer risk using MR analyses. To do this, I used a published summary estimate of the association of rs10993994 with prostate cancer risk from the large genotyping project, iCogs, in the international consortium, PRACTICAL [36], and using an additional set of EPIC prostate cancer cases and controls from the OncoArray chip [37] and BPC3 genotyping [38]. Summary estimates for the association of rs10993994 with MSP were calculated using these EPIC data [37, 38]. I used the MR-Base platform to do a phenome-wide association scan for rs10993994 with over 850 traits [1] to check for pleiotropy [61]. Two-sample MR estimates for the association of MSP with prostate cancer risk were then calculated separately using summary estimates for each of PRACTICAL (iCogs) [36] and EPIC-derived rs10993994-prostate cancer risk estimates with the EPIC-derived rs10993994-MSP estimate, which were then combined using the inverse-variance weighted method. To address possible confounding by PSA concentration, we conducted sensitivity analyses by using the summary association of rs10993994 with residuals from a linear regression of log total PSA on MSP, also calculated within EPIC.

All statistical tests are two-sided and were conducted using STATA software version 14 (College Station, TX: StataCorp LP).

[1] Phenotypes include risk factors for various diseases, diseases, metabolites, and measures of immune function, for example.

5.3 Results

Data from 1,871 cases and 1,871 matched controls from three study phases (contributing 246, 648, 977 cases, respectively) were included in analyses. The median age at blood collection was 58 years (range, 39 to 79 years), and, for cases, the median time between blood collection and diagnosis was 8.3 years. No significant differences were observed in selected baseline characteristics between cases and controls(see Table 5.1).

Mean MSP concentration (ng/ml) at blood collection did not differ significantly between cases and controls (adjusted geometric means were 12.8; [95% CI 12.5-13.2], and 12.9; [95% CI 12.6-13.2], respectively, $p = 0.7$). In contrast, mean PSA concentration (ng/ml) measured at blood collection was about three-fold higher in cases than in controls (adjusted geometric mean = 2.4; [95% CI 2.3-2.5] and 0.8; [95% CI 0.8-0.9] respectively, p <0.0001 5.2). No significant differences were observed for MSP or PSA concentrations by days in the post for samples from the EPIC-Oxford sub-cohort ($p = 0.6$ & $p = 0.7$, respectively). Additional case characteristics can be found in Table 5.3.

MSP concentration in controls was higher in men who were older at blood collection, not married, had a normal/low BMI, or low intake of alcohol, and had a highest educational attainment of secondary school when compared to both primary school and degree level or higher ($p < 0.05$ for all). Compared to never smokers, men who smoked greater than the median number of cigarettes per day (N = 15) had 30% higher MSP concentrations with evidence of a significant linear trend (p-trend < 0.0001; Table 5.4). Additionally, there were significant differences in MSP concentrations by recruitment country and by study phase (see Table 5.5). PSA concentration was positively associated with age at blood collection and educational attainment, was higher in men with blood collected between midnight and 10 am, and lower in men with higher BMI and diabetes (see Table 5.3). Further, there were significant differences in PSA concentrations by recruitment country and by study phase (see Table 5.5).

MSP and PSA concentrations were weakly, but significantly, positively correlated in both cases and controls (partial correlation $r = 0.3$ and $r = 0.2$, respectively, $p < 0.0001$). Moreover, we observed modest differences in this correlation by country with the lowest correlation for control men in

Spain ($r = 0.04$) and highest correlation in the UK ($r = 0.31$) (see Table 5.6).

MSP concentration was not associated with risk of prostate cancer after adjustment for age at blood collection and BMI (OR for highest versus lowest fourth = 0.98; [95% CI 0.82-1.19], p-trend across the medians of the fourths = 0.9). However, given the *a priori* expectation of an association of PSA with MSP [27], we subsequently adjusted for PSA concentration by fourths; after adjustment for PSA, MSP concentration was significantly associated with prostate cancer risk (OR for highest versus lowest fourth = 0.65; [95%CI 0.51-0.84], p-trend = 0.001; Table 5.7 & Figure 5.1). There was weak heterogeneity of the association of MSP with prostate cancer by recruitment country (p-heterogeneity = 0.02; Table 5.8) and some evidence of heterogeneity by age at diagnosis (p-heterogeneity = 0.03; Table 5.7). Given the strong association of MSP concentration with smoking status, we conducted exploratory analyses of the MSP-risk association by smoking status (never, previous, current) there was no significant heterogeneity in the association by smoking status (p-heterogeneity = 0.6; Table **??**).

The association of MSP concentration with prostate cancer did not differ by tumour stage or grade, time to diagnosis, or age at blood collection (all p-heterogeneity > 0.05; Table 5.7). Additional analyses were conducted to evaluate the sensitivity of the estimated associations to the conservative delimitation of grade as low-intermediate compared to high at a cut-off of Gleason score 8. However, when high grade was defined instead as Gleason score of 7 or higher the results were not materially altered and no significant heterogeneity was observed (Table 5.7).

In a subset of 1,068 cases and 1,186 controls with available MSMB rs10993994 genotype data no significant deviation from Hardy-Weinberg Equilibrium was observed (D = -0.005). There was a 6.09 ng/ml; [95% CI 5.56-6.61] per allele decrease in MSP concentration - highest concentrations observed for CC homozygotes and rs10993994 explained 42% of the variability of MSP. Further, in controls only, there was a 0.22 ng/ml; [95% CI 0.09-0.35] per allele increase in PSA concentrations, with the highest concentrations observed for TT homozygotes (Table 5.9). rs10993994 genotype was significantly associated with prostate cancer risk (OR in TT versus CC = 1.37; [95% CI 1.08-1.75], p-trend = 0.006, Table 5.10).

No significant association of rs10993994 genotype was observed with potential confounders beyond that of PSA concentrations in controls in EPIC (Table 5.11 & Table 5.12), or after correction for multiple testing in MR-Base [61](Table A.1). An inverse-variance weighted MR provided evidence that a one unit increase in circulating plasma MSP concentrations (ng/ml) is associated with a 4% reduction in risk of prostate cancer (OR 0.96; [95% CI 0.95-0.97]), which compared with a 2% reduced risk per unit increase in MSP in the observational study (OR 0.98; [95% CI 0.97-0.99]) (Table 5.13 and Figure 5.2). Results were not materially altered by adjustment for total PSA (Table 5.13).

5.4 Discussion

In this large prospective study, we found evidence of a lower risk for prostate cancer in men with higher circulating concentrations of the prostate protein MSP, but only after adjustment for circulating PSA concentrations. MSP is a protein in the immunoglobulin binding factor family that is primarily secreted by epithelial cells into the seminal plasma, and which may have a role in tumor suppression [257] [23] and pathogen defense [262]. These findings are in partial agreement with findings from the only other published prospective investigation in the MEC study [27], and with results from a retrospective case-control study [265]; both found a significant inverse association between circulating MSP concentration and prostate cancer risk without adjustment for PSA. All other previous studies of MSP and prostate cancer risk have been cross-sectional, measuring urinary or blood concentration collected from cases after diagnosis, in which the potential for reverse causality is larger [266, 272, 267, 268, 269, 256, 257, 258, 273, 274]. Furthermore, this is the first study to use an MR approach, and identifies a potentially causal relationship of MSP with prostate cancer.

The strong positive association of circulating PSA concentration with prostate cancer risk [72], and the moderate positive correlation between MSP and PSA [27] may produce a negative confounding effect that masks a true underlying association of MSP with prostate cancer risk. This hypothesis is consistent with findings from the MEC [27] study, which reported an increase in the strength of association of MSP levels with prostate cancer risk after adjustment for PSA concentration.

In accordance with findings from MEC [27], we found no evidence that the association of MSP with risk differed by prostate cancer stage or grade. However, there are relatively small numbers of cases in subgroups defined by tumour characteristics and the analyses by stage and grade of tumour have limited power to evaluate heterogeneity between and associations by tumour subtype.

In addition to the prostate epithelium, MSP is secreted at lower levels by epithelial cells in the tracheobronchial tree [254, 275]. Smoking has been associated with as much as a 2.5 fold increase in expression of MSP in the airway epithelium when compared to non-smokers [264]. As such, variation in circulating MSP concentrations may be due to secretory/goblet cell hyperplasia and not prostatic health. Therefore, some of the variation in circulating MSP concentrations may be due to smoking-induced secretory cell hyperplasia in the respiratory tract. To our knowledge, we are the only study to report higher levels of MSP (approximately 30%) among current smokers compared to men with no history of smoking. However, we found no evidence of heterogeneity in the MSP-risk association by smoking status.

Short follow-up time (3.8 years) and thus reverse causality was previously suggested in MEC as a possible explanation for the association of MSP with prostate cancer [27]. However, the present study confirms this previously observed association with more than double the follow-up time (8.3 years), and MEC observed no heterogeneity for an MSP association with prostate cancer by history of PSA testing [27]. Further, while it is not possible, using observational analyses, to exclude reverse causality due to pre-existing subclinical prostate cancer, this is possible using MR methods.

The strength of the current MR result stems primarily from the use of rs10993994 as an instrumental variable; rs10993994 lies in the promotor region of the MSMB region, the locus that encodes synthesis of MSP, and, as confirmed in the current study, rs10993994 is strongly associated with both circulating MSP concentrations and prostate cancer risk [27, 36]. Although there is some inconsistent evidence that rs10993994 may be associated with levels of NCOA4 mRNA (located within 16 kb of rs10993994) in benign prostate tissue [276, 277], a recent review of MSP function [278] and our phenome-wide scan of more than 850 traits using MR-Base [61] suggest the rs10993994 genetic association is specific to MSP. Further, rs10993994 is a cis-acting variant in the promoter region of the gene (MSMB) that codes for

MSP. This is perhaps the most robust scenario of MR, as there is a relatively high degree of certainty that the genetic instrument in the causal variant for the biomarker of interest [279, 280]. Nonetheless, as we observe a modest association of rs10993994 with circulating plasma concentrations of PSA in controls, it remains possible that PSA may confound these results, for example, through its use in diagnosing prostate cancer. However, this is unlikely for two reasons. First, given that the association of rs10993994 with PSA is present only in controls (i.e. after stratifying on disease status) and that MR results were materially robust to adjustment for PSA concentration, the association of rs10993994 with PSA may also arise from collider bias: i.e. the association of rs10993994 with PSA is induced by stratifying on prostate cancer status. Such collider bias [281] should not invalidate the results of the naïve MR analysis where PSA is not conditioned upon. Indeed, given no substantial differences in the MR estimates are found after conditioning on PSA concentration, collider bias is not a likely influence on these findings. Second, for the biological role of PSA to confound these findings PSA would have to be causal to prostate cancer development; although PSA is strongly associated with prostate cancer [72, 44], there is as yet no strong evidence for an aetiological role of PSA in prostate cancer biology.

5.5 Conclusion

Using both observational data from a prospective nested case-control study and MR, the current study supports a protective role of MSP in the development of prostate cancer. Experimental studies are needed to elucidate the mechanisms through which MSP may influence prostate cancer development. If the predictive effect of MSP is shown to be true from randomized clinical trials, therapies that raise MSP levels may provide novel opportunities for the treatment and prevention of prostate cancer.

Figure 5.1: Multi-variable adjusted odds ratios (95% CI) for prostate cancer by fourth of plasma microseminoprotein-β (MSP) concentration

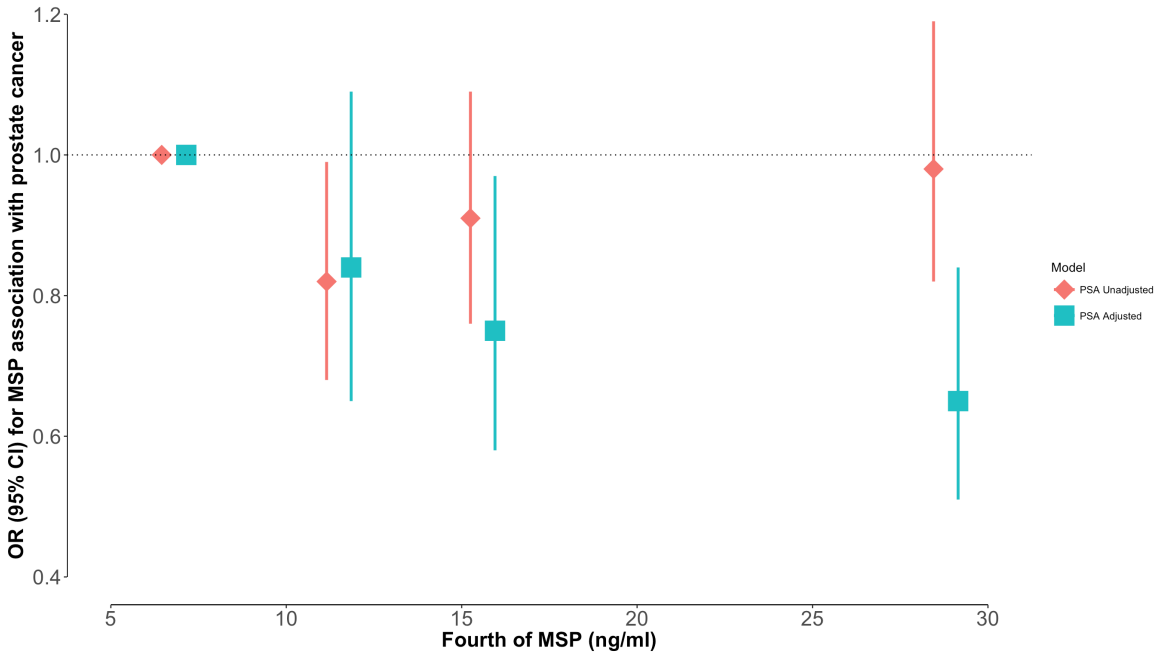


Figure 5.2: Microseminoprotein-β (MSP) association (Per unit (ng/ml)) with prostate cancer from observational data in EPIC additionally adjusted for prostate-specific antigen (PSA), and from MR that combines PRACTICAL and EPIC estimates

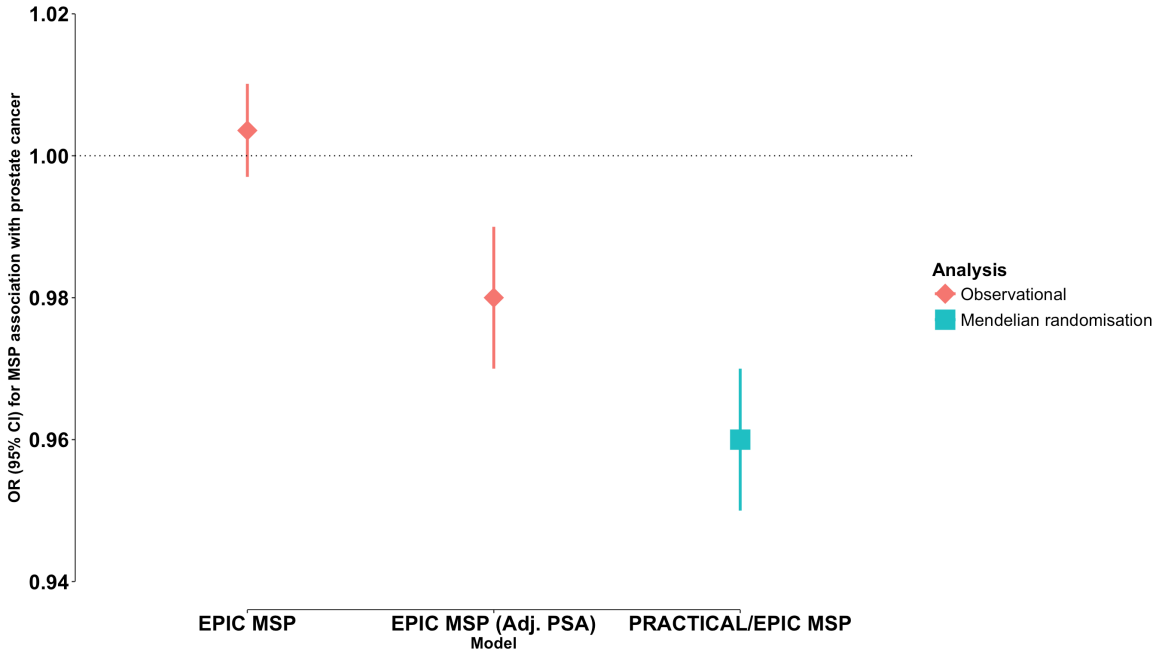


Table 5.1: Sensitivity analysis of odds ratio (95% CI) for prostate cancer associated with fourth of microseminoprotein-β (MSP) (lowest compared to highest fourth)

Model	Odds Ratio[c] (95% CI)	p-trend[d]
Basic Model[a]	0.65 (0.51-0.84)	0.003
Adjusted Models		
Basic + smoking status	0.65 (0.50-0.85)	0.004
Basic + alcohol consumption	0.65 (0.51-0.84)	0.003
Basic + marital status	0.66 (0.51-0.85)	0.003
Basic + total physical activity	0.65 (0.51-0.84)	0.003
Basic + educational attainment	0.66 (0.51-0.85)	0.003
Fully Adjusted model[b]	0.66 (0.51-0.86)	0.005

[a] Minimally adjusted model: total PSA, age at blood collection, and BMI

[b] Basic model + smoking status, alcohol consumption, marital status, total physical activity, and educational attainment

[c] Odds ratio for comparison between highest and lowest quartile of MSP

[d] Test for trend obtained by pseudo continuous variable of median concentration within each fourth of microsemonoprotein-β concentration

Table 5.2: Characteristics of control participants and men who developed prostate cancer in EPIC

Characteristic	Controls (n = 1,871)	Cases (n = 1,871)	p[a]
Age at Blood Collection, years (SD)	58.3 (6.9)	58.3 (6.9)	0.5
Weight, kg (SD)	80.2 (11.6)	80.2 (11.5)	0.5
Height, cm (SD)	172.9 (7.2)	172.5 (7.1)	0.9
BMI[b], kg/m^2	26.9 (3.5)	27.1 (3.5)	0.2
Smoking Status, n (%)			
Never	578 (31.5)	621 (34.5)	
Previous	826 (45.1)	792 (43.9)	
Current	431 (23.4)	389 (21.6)	0.1
Alcohol, n (%)			
<8	657 (34.8)	682 (36.4)	
8-15	368 (19.9)	363 (19.4)	
16-39	542 (28.9)	491 (26.2)	
>40	304 (16.3)	335 (17.9)	0.3
Physical Activity, n (%)			
Inactive	277 (15.1)	268 (14.9)	
Moderately Inactive	533 (29.1)	525 (29.4)	
Active	1,025 (55.9)	996 (55.7)	0.9
Marital Status, n (%)			
Married/Cohabitating	1,377 (89.7)	1,333 (88.6)	
Not Married/ Not Cohabitating	160 (10.4)	172 (11.4)	0.3
Educational attainment, n (%)			
Primary/None	687 (38.3)	668 (38.3)	
Secondary	633 (35.4)	596 (34.1)	
Degree	471 (26.3)	482 (27.6)	0.6
Geometric mean analyte concentration			
MSP, ng/ml (95% CI)	12.8 (12.5-13.2)	12.9 (12.6-13.2)	0.7
MSP adj. PSA, ng/ml (95% CI)	12.9 (12.6-13.3)	12.8 (12.5-13.1)	0.5
PSA, ng/ml (95% CI)	0.8 (0.8-0.9)	2.4 (2.3-2.5)	<0.0001

[a] Significance calculated from analysis of variance and chi-square for continuous and categorical variables, respectively

[b] Body mass index (BMI), microsemonoprotein-β (MSP), prostate-specific antigen (PSA)

Table 5.3: Characteristics of men who developed prostate cancer in EPIC

Characteristic	Cases (n = 1,871)
Time to diagnosis, n (%)	
≤ 2 years	81 (4.4)
2 to 4 years	111 (5.9)
4 to 6 years	244 (13.1)
6 to 8 years	375 (20.2)
≥ 8 years	1,049 (56.4)
Year of diagnosis, median (range)	2004 (1994-2009)
Age at diagnosis, years (SD)	66.9 (6.9)
TNM-Code[a]	
Tumour	
T1	178 (19.5)
T2	532 (58.3)
T3	183 (20.1)
T4	19 (2.1)
Nodes	
N0	613 (92.3)
N1	46 (6.9)
N2	4 (0.1)
N3	1 (0.01)
Metastases	
M0	522 (94.7)
M1	29 (5.3)
EPIC Stage Information[a]	
Localized	787 (79.2)
Metastatic	207 (20.8)
Tumour Grade	
Gleason Grade[a]	
≤ 6	452 (55.7)
7	218 (28.6)
≥ 8	120 (15.7)
EPIC Grade Information[a]	
Well Differentiated	139 (16.6)
Moderately Differentiated	503 (60.1)
Poorly Differentiated	191 (22.8)
Undifferentiated	4 (0.1)
PSA[b] (ng/ml) at diagnosis	
<3	20 (3.6)
≥ 3 and <10	335 (59.8)
≥ 10 and <50	187 (33.4)
≥ 50	18 (3.2)

[a] TNM-code and EPIC stage, and Gleasons grade and EPIC grade are not mutually exclusive

[b] Prostate-specific antigen (PSA)

Table 5.4: Adjusted geometric means[a] of microseminoprotein-β (MSP) and prostate-specific antigen (PSA) concentration (ng/ml) in controls by selected characteristics in EPIC

Factor and Subset	N	Mean (95% CI)	p-diff/trend[b]	N	Mean (95% CI)	p-diff/trend[b]
Age at blood collection, years						
<50 years	195	11.7 (10.8-12.6)		195	0.6 (0.5-0.6)	
50 years to 55 years	339	12.3 (11.6-13.1)		339	0.7 (0.6-0.7)	
55 years to 59 years	503	12.2 (11.7-12.9)		503	0.8 (0.7-0.9)	
60 years to 64 years	549	12.7 (12.1-13.3)		549	0.9 (0.9-1.0)	
65 years to 69 years	169	15.4 (14.2-16.8)		169	1.3 (1.1-1.4)	
>70 years	116	16.8 (15.2-18.6)	<0.0001/<0.0001	116	1.6 (1.3-1.8)	<0.0001/<0.0001
Time of blood collection, hours						
0000-0959	305	13.0 (12.1-13.9)		306	0.9 (0.8-1.0)	
1000-1259	267	13.6 (12.6-14.6)		267	0.8 (0.7-0.9)	
1300-2359	1,176	12.4 (12.0-12.9)	0.1	1,176	0.8 (0.8-0.9)	0.03
Marital status						
Married/Cohabitating	1,377	12.9 (12.5-13.2)		1,377	0.8 (0.8-0.9)	
Not married/ Not Cohabitating	160	14.3 (13.1-15.6)	0.01	160	0.9 (0.8-1.0)	0.9
Educational attainment						
Primary/none	687	12.3 (11.8-12.8)		687	0.8 (0.7-0.8)	
Secondary	633	13.2 (12.7-13.8)		633	0.9 (0.8-0.9)	
Degree	471	12.8 (12.2-13.5)	0.02	471	0.9 (0.8-0.9)	0.02
BMI, kg/m2						
16-24	542	13.7 (13.1-14.3)		542	0.9 (0.8-0.9)	
25-29	1,003	12.8 (12.3-13.2)		1,003	0.8 (0.8-0.9)	
>30	317	11.7 (10.9-12.4)	0.0007/0.001	317	0.7 (0.7-0.8)	0.02/<0.001
Smoking status						
Never	578	11.9 (11.4-12.4)		578	0.9 (0.8-0.9)	
Previous	826	12.1 (11.7-12.6)		826	0.8 (0.8-0.9)	
Current (<15 cigarettes)	181	15.8 (14.6-17.1)		181	0.9 (0.8-0.9)	
Current ($\geq$ 15 cigarettes)	154	17.0 (15.6-18.6)	<0.0001/<0.0001	154	0.8 (0.7-0.9)	0.9
Alcohol consumption, g/d						
<8	657	13.6 (13.0-14.2)		657	0.9 (0.8-0.9)	
8 to 15	368	13.0 (12.3-13.8)		368	0.9 (0.8-0.9)	
16 to 39	542	12.3 (11.8-12.9)		542	0.8 (0.8-0.9)	
>40	304	11.9 (11.2-12.7)	0.002/<0.0001	304	0.8 (0.7-0.9)	0.3
Diabetic						
No	1,760	12.9 (12.5-13.2)		1,760	0.9 (0.8-0.9)	
Yes	97	12.1 (10.8-13.5)	0.5	97	0.7 (0.6-0.8)	0.02

[a] All means adjusted for age at blood collection, body mass index (BMI), recruitment centre, and laboratory batch

[b] Test for difference by analysis of variance; test for trend by entering a pseudo-continuous variable for given factor in linear regression

Table 5.5: Adjusted geometric mean[a] plasma microseminoprotein-β (MSP) and prostate-specific antigen (PSA) concentration (ng/ml) among controls by country and study phase in EPIC

MSP (ng/ml)	Germany	Greece	Italy	The Netherlands	Spain	UK	Total	p^{d}
Phase 1								
Number	100	6	40	11	32	57	246	
MSP[a]	9.8 (8.6-11.1)	6.8 (4.1-11.3)	11.0 (9.1-13.4)	13.6 (9.3-20.0)	10.1 (8.1-12.7)	13.9 (11.5-16.8)	10.9 (10.1-11.8)	0.02
PSA[a]	0.7 (0.6-0.9)	0.7 (0.4-1.3)	0.9 (0.7-1.1)	0.9 (0.5-1.5)	0.9 (0.7-1.3)	0.9 (0.7-1.1)	0.8 (0.7-0.9)	0.6
Follow-up time[b]	37.8 (33.4-42.1)	28.2 (10.3-45.9)	40.2 (33.3-47.1)	-	57.9 (50.3-65.7)	43.8 (37.4-50.3)	42.1 (38.8-45.3)	
Phase 2								
Number	206	31	84	23	108	196	648	
MSP[a]	14.1 (13.0-15.2)	16.4 (13.7-19.6)	12.1 (10.8-13.5)	16.9 (13.8-20.8)	13.7 (12.4-15.0)	17.1 (15.8-18.5)	14.8 (14.2-15.4)	0.001
PSA[a]	0.9 (0.8-1.0)	0.7 (0.5-0.9)	0.9 (0.7-1.0)	0.8 (0.6-1.1)	0.9 (0.8-1.1)	1.1 (0.9-1.2)	0.9 (0.8-1.0)	0.04
Follow-up time[b]	75.5 (72.5-78.6)	64.2 (56.4-71.9)	85.8 (81.1-90.5)	84.2 (75.1-93.2)	109.4 (105.2-113.6)	87.5 (84.4-90.6)	85.8 (83.9-87.7)	
Phase 3								
Number	383	38	148	74	132	202	977	
MSP[a]	10.6 (9.7-11.7)	9.4 (7.8-11.2)	14.4 (12.7- 16.3)	13.5 (11.9-15.3)	13.8 (12.1-15.8)	12.9 (11.8-14.1)	12.1 (11.7-12.6)	0.002
PSA[a]	0.8 (0.7-0.9)	0.7 (0.6-0.9)	0.7 (0.6- 0.8)	0.9 (0.7-1.0)	0.7 (0.6-0.9)	0.9 (0.8-1.0)	0.8 (0.7-0.9)	0.04
Follow-up time[b]	112.7 (110.3-115.1)	95.3 (87.7-102.8)	119.5 (115.6-123.4)	129.4 (123.9-134	130.2 (126.1-134.3)	124.1 (120.8-127.4)	119.0 (117.5-120.6)	
p^{c}							<0.0001	
Total								
Number	689	75	272	108	272	455	1,871	
MSP[a]	11.7 (11.3-12.3)	11.6 (10.2-13.2)	12.8 (11.9-13.7)	14.5 (13.0-16.1)	12.8 (11.9-13.7)	14.5 (13.7-15.3)	12.7 (12.4-13.1)	0.0001
PSA[a]	0.8 (0.7-0.9)	0.7 (0.5-0.8)	0.8 (0.7-0.9)	0.9 (0.7-1.0)	0.9 (0.8-1.0)	1.0 (0.9-1.1)	0.8 (0.8-0.9)	0.0001

[a] Adjustment was made for age at blood collection, body mass index, and laboratory batch, (95% CI)

[b] Approximate minimum time controls known to be free of prostate cancer, defined from corresponding matched case values of time to diagnosis

[c] For significant differences in adjusted mean concentration of MSP by study phase by analysis of variance

[d] Significance of difference across country for mean concentration of MSP by analysis of variance

Table 5.6: Partial Pearson's correlation[a] between microseminoprotein-β and prostate-specific antigen by recruitment country among controls in EPIC

	Germany	Greece	Italy	The Netherlands	Spain	The United Kingdom
Overall	0.17*	0.06	0.02	0.23*	0.04	0.31*
Excluding						
First year	0.17*	0.06	0.05	0.24*	0.04	0.31*
First 5 years	0.13*	0.08	0.06	0.19	0.05	0.29*
First 10 years	0.19*	0.15	-0.01	0.31*	-0.01	0.22*

[a] Adjusted for age at recruitment, recruitment centre, body mass index, and laboratory batch

* Significant at $p < 0.05$

Table 5.7: Multi-variable adjusted odds ratios (95% CI) for prostate cancer by fourth of plasma microseminoprotein-β (MSP) concentration, subdivided by selected factors in EPIC

		Fourth of Microseminoprotein-β concentration, ng/ml					
		1	2	3	4	p-trend[c]	p-heterogeneity[d]
Overall	Cases/controls, n	508/468	402/464	458/468	501/469		
	Mean MSP, ng/ml (range)	7 (1-9)	12 (9-13)	16 (13-18)	23 (18-90)		
	Adjusted RR (95% CI)[a]	1 (reference)	0.82 (0.68-0.99)	0.91 (0.76-1.09)	0.98 (0.82-1.19)	0.9	
	Adjusted RR (95% CI)[b]	1 (reference)	0.84 (0.65-1.09)	0.75 (0.58-0.97)	0.65 (0.51-0.84)	0.001	
Study Phase							
Phase 1	Cases/controls, n	72/88	53/67	61/47	60/44		
	Adjusted RR (95% CI)[b]	1 (reference)	0.62 (0.22-1.75)	0.96 (0.38-2.39)	0.52 (0.19-1.36)	0.2	
Phase 2	Cases/controls, n	130/111	121/135	162/173	233/227		
	Adjusted RR (95% CI)[b]	1 (reference)	0.77 (0.45-1.32)	0.59 (0.37-0.98)	0.59 (0.38-0.94)	0.03	
Phase 3	Cases/controls, n	306/269	228/262	235/248	208/198		
	Adjusted RR (95% CI)[b]	1 (reference)	0.85 (0.61-1.17)	0.77 (0.56-1.07)	0.65 (0.46-0.91)	0.01	0.9
Time to diagnosis							
<7 years	Cases/controls, n	146/157	124/147	163/143	176/162		
	Adjusted RR (95% CI)[b]	1 (reference)	0.85 (0.47-1.56)	0.79 (0.44-1.41)	0.47 (0.27-0.83)	0.004	
7 years to 10 years	Cases/controls, n	188/146	149/170	153/185	202/191		
	Adjusted RR (95% CI)[b]	1 (reference)	0.68 (0.43-1.07)	0.55 (0.36-0.84)	0.65 (0.43-0.96)	0.1	
>10 years	Cases/controls, n	174/165	129/147	142/140	123/116		
	Adjusted RR (95% CI)[b]	1 (reference)	0.93 (0.62-1.39)	0.89 (0.61-1.32)	0.76 (0.49-1.18)	0.2	0.4
Age at blood collection							
<60 years	Cases/controls, n	314/279	242/287	260/270	214/194		
	Adjusted RR (95% CI)[b]	1 (reference)	0.79 (0.56-1.13)	0.65 (0.46-0.91)	0.68 (0.46-0.99)	0.03	
$\geqslant$ 60 years	Cases/controls, n	189/184	155/171	193/194	281/269		
	Adjusted RR (95% CI)[b]	1 (reference)	0.89 (0.58-1.35)	0.89 (0.59-1.32)	0.63 (0.44-0.90)	0.007	0.5
Age at diagnosis							
<65 years	Cases/controls, n	218/194	177/193	183/191	146/146		
	Adjusted RR (95% CI)[b]	1 (reference)	0.77 (0.47-1.25)	0.45 (0.27-0.74)	0.45 (0.26-0.78)	0.001	
$\geqslant$ 65 years	Cases/controls, n	290/274	225/271	275/277	355/323		
	Adjusted RR (95% CI)[b]	1 (reference)	0.84 (0.61-1.16)	0.89 (0.67-1.21)	0.74 (0.55-0.99)	0.05	0.03

Stage							
Localised	Cases/controls, n	243/229	194/221	214/209	235/225		
	Adjusted RR (95% CI)b	1 (reference)	0.86 (0.57-1.28)	0.77 (0.52-1.15)	0.64 (0.44-0.92)	0.02	
Advanced	Cases/controls, n	110/95	87/81	91/109	89/92		
	Adjusted RR (95% CI)b	1 (reference)	0.79 (0.44-1.43)	0.45 (0.25-0.79)	0.45 (0.24-0.82)	0.002	0.2
Grade, ≥ 8							
Low-intermediate	Cases/controls, n	384/349	281/338	346/354	345/315		
	Adjusted RR (95% CI)b	1 (reference)	0.83 (0.59-1.15)	0.76 (0.56-1.02)	0.63 (0.46-0.86)	0.004	
High	Cases/controls, n	53/49	44/48	36/38	63/62		
	Adjusted RR (95% CI)b	1 (reference)	0.86 (0.42-1.76)	0.68 (0.32-1.46)	0.73 (0.39-1.42)	0.3	0.7
Grade, ≥ 7							
Low-intermediate	Cases/controls, n	263/249	190/234	239/246	235/196		
	Adjusted RR (95% CI)b	1 (reference)	0.94 (0.59-1.51)	0.99 (0.65-1.50)	0.84 (0.54-1.29)	0.2	
High	Cases/controls, n	173/146	133/149	139/144	172/178		
	Adjusted RR (95% CI)b	1 (reference)	0.81 (0.39-1.18)	0.72 (0.47-1.13)	0.55 (0.36-0.83)	0.004	0.2
Smoking status							
Never	Cases/controls, n	195/163	141/152	145/151	139/115		
	Adjusted RR (95% CI)[b]	1 (reference)	0.88 (0.59-1.31)	0.63 (0.43-0.91)	0.62 (0.42-0.92)	0.005	
Previous	Cases/controls, n	227/231	189/222	186/200	190/178		
	Adjusted RR (95% CI)[b]	1 (reference)	0.84 (0.61-1.16)	0.74 (0.54-1.03)	0.64 (0.46-0.89)	0.007	
Current	Cases/controls, n	71/72	61/80	110/110	147/168		
	Adjusted RR (95% CI)[b]	1 (reference)	0.98 (0.56-1.71)	0.99 (0.61-1.62)	0.84 (0.52-1.35)	0.4	0.6
By median PSA							
Below median PSA	Cases/controls, n	48/259	44/288	40/214	28/170		
	Adjusted RR (95% CI)[b]	1 (reference)	0.84 (0.53-1.33)	0.96 (0.59-1.55)	0.91 (0.54-1.55)	0.8	
Above median PSA	Cases/controls, n	460/210	359/176	418/254	473/299		
	Adjusted RR (95% CI)[b]	1 (reference)	0.92 (0.71-1.19)	0.71 (0.56-0.90)	0.55 (0.43-0.69)	<0.0001	0.02

[a] Estimates are from logistic regression conditioned on the matching variables (see methods) with adjustment for age, and body mass index

[b] Additional to model [a], adjustment was made for total PSA

[c] A categorical variable was replaced with a continuous variable equal to median concentration by fourth of plasma MSP

[d] Test for heterogeneity in the trends by likelihood ratio test

Table 5.8: Multi-variable adjusted odds ratios (95% CI) for prostate cancer by fourth of plasma microseminoprotein-β concentration by recruitment country in EPIC

		Fourth of microseminoprotein-β concentration, ng/ml					
		1	2	3	4	p-trend[b]	p-heterogeneity[c]
Country							
Greece	Cases/controls, n	18/22	14/19	23/20	20/14		
	Adjusted RR (95% CI)[a]	1 (reference)	1.02 (0.21-4.95)	0.72 (0.18-2.90)	1.12 (0.27-4.66)	0.9	
The Netherlands	Cases/controls, n	29/23	23/26	32/28	24/31		
	Adjusted RR (95% CI)[a]	1 (reference)	0.75 (0.25-2.25)	1.09 (0.39-3.03)	0.27 (0.08-0.94)	0.07	
Spain	Cases/controls, n	66/77	64/62	70/75	71/57		
	Adjusted RR (95% CI)[a]	1 (reference)	1.17 (0.64-2.16)	1.19 (0.66-2.19)	1.58 (0.86-2.92)	0.2	
Italy	Cases/controls, n	81/74	66/76	61/70	64/52		
	Adjusted RR (95% CI)[a]	1 (reference)	1.21 (0.56-2.58)	0.75 (0.38-1.49)	1.05 (0.51-2.19)	0.8	
The United Kingdom	Cases/controls, n	82/72	89/93	109/115	175/175		
	Adjusted RR (95% CI)[a]	1 (reference)	0.91 (0.49-1.69)	0.68 (0.38-1.22)	0.49 (0.29-0.87)	0.004	
Germany	Cases/controls, n	232/200	146/187	163/160	146/140		
	Adjusted RR (95% CI)[a]	1 (reference)	0.60 (0.38-0.93)	0.51 (0.32-0.81)	0.46 (0.29-0.72)	<0.0001	0.02

[a] Model adjusted for age at recruitment, body mass index, and total PSA

[b] A categorical variable was replaced with a continuous variable equal to median concentration by fourth of plasma MSP

[c] Test for heterogeneity in the trends using likelihood ratio test

Table 5.9: Adjusted geometric means[a] of microseminoprotein-β (MSP) and prostate-specific antigen (PSA) concentration (ng/ml) in controls by rs10993994 genotype in EPIC

	rs10993994 genotype			
	C,C	C,T	T,T	p-diff[b]
Case/Control, n	298/398	545/571	215/208	
MSP (ng/ml)[a]				
Control	18.1 (17.4-18.9)	12.9 (12.5-13.4)	5.9 (5.6-6.3)	<0.0001
Case	19.2 (18.3-20.1)	13.2 (12.7-13.7)	6.3 (5.9-6.7)	<0.0001
Total PSA (ng/ml)[a]				
Control	0.7 (0.7-0.8)	0.8 (0.7-0.9)	0.9 (0.9-1.1)	0.0004
Case	2.3 (2.1-2.5)	2.4 (2.3-2.6)	2.3 (2.1-2.6)	0.6

[a] Adjusted for age, body mass index, recruitment centre, and laboratory batch
[b] Difference calculated from analysis of variance

Table 5.10: Odds ratios (95% CI) for prostate cancer by rs10993994 in EPIC

rs10993994	Case/Control	OR (95% CI)[a]	p-trend [b]
C,C	301/400	1(reference)	
C,T	552/578	1.27 (1.05-1.53)	
T,T	215/208	1.37 (1.08-1.75)	0.006

[a] Results from a univariate logistic regression

[b] A T allele count was entered as a continuous variable in a logistic regression

Table 5.11: Per allele (rs10993994) difference in total PSA by smoking status at recruitment separately for prostate cancer cases and controls

	Smoking status		
	Never	Previous	Current
Controls			
N	376	507	285
Per allele change in total PSA (95% CI)	0.05 (-0.15-0.25)	0.40 (0.15-0.65)	0.13 (-0.05-0.32)
p-value*	0.62	0.002	0.16
Cases			
N	375	441	240
Per allele change in total PSA (95% CI)	0.30 (-1.54-2.14)	0.13 (-1.87-2.12)	-1.22 (-2.48-0.03)
p-value*	0.8	0.9	0.06

* *p*-value from a linear regression of rs10993994 genotype on total PSA concentration

Table 5.12: Various characteristics of participants by rs10993994 genotype in EPIC

Characteristic	CC	CT	TT	p-value*
N	701	1,130	423	
Age at Blood Collection, years**	57.94 (57.46 - 58.42)	58 (57.62 - 58.38)	57.91 (57.3 - 58.52)	0.9
Weight, kg **	80.36 (79.5 - 81.22)	80.14 (79.49 - 80.79)	80.08 (78.98 - 81.18)	0.9
Height, cm **	172.34 (171.82 - 172.86)	172.28 (171.88 - 172.68)	172.38 (171.67 - 173.09)	0.9
BMI, kg/m2**	27.12 (26.86 - 27.38)	27.12 (26.92 - 27.32)	27.05 (26.73 - 27.37)	0.9
Smoking Status, n (%)				
Never	226 (32.24)	394 (34.87)	131 (30.97)	
Previous	314 (44.79)	452 (40)	182 (43.03)	
Current	148 (21.11)	271 (23.98)	106 (25.06)	0.2
Alcohol, n (%)				
<8	224 (31.95)	365 (32.3)	138 (32.62)	
8-15	140 (19.97)	220 (19.47)	76 (17.97)	
16-39	212 (30.24)	330 (29.2)	133 (31.44)	
>40	125 (17.83)	214 (18.94)	76 (17.97)	0.9
Physical Activity, n (%)				
Inactive	94 (13.41)	186 (16.46)	73 (17.26)	
Moderately Inactive	220 (31.38)	318 (28.14)	113 (26.71)	
Active	379 (54.07)	618 (54.69)	235 (55.56)	0.2
Marital Status, n (%)				
Married/Cohabitating	502 (71.61)	824 (72.92)	297 (70.21)	
Not Married/Cohabitating	63 (8.99)	100 (8.85)	39 (9.22)	0.9
Educational attainment, n (%)				
Primary/none	274 (39.09)	450 (39.82)	186 (43.97)	
Secondary	224 (31.95)	346 (30.62)	121 (28.61)	
Degree	191 (27.25)	309 (27.35)	108 (25.53)	0.5

* p-values are from analysis of variance models and chi square,

** Geometric means are presented with 95% confidence intervals

Numbers may not add to total due to missing values

Table 5.13: Per unit MSP (ng/ml) odds ratio (OR) for prostate cancer for IV estimates and Mendelian randomisation results using inverse-variance with and without adjustment for circulating concentrations of total PSA (ng/ml) in EPIC

Study	**Per unit OR**	**Adj. SE**
MSP		
PRACTICAL for incident cancer	0.96 (0.95-0.98)	0.006
EPIC (Exc. PRACTICAL) for incident cancer	0.97 (0.95-0.99)	0.009
All Pooled	**0.96 (0.95-0.97)**	
MSP (adj. PSA)		
PRACTICAL for incident cancer[36]	0.97 (0.96-0.98)	0.006
EPIC (Exc. PRACTICAL) for incident cancer	0.98 (0.95-0.99)	0.008
All Pooled	**0.97 (0.96-0.98)**	

Chapter 6

Human Kallikrein 2 levels and prostate cancer risk in the European Prospective Investigation into Cancer and Nutrition

6.1 Introduction

To date, research into HK2 has been primarily concerned with assessing the ability of HK2 to improve the discrimination for prostate cancer when included in a PSA and age based prediction model [47, 48, 49, 50, 51, 52, 282, 283, 284, 53]. There has been, however, no investigation, in men without elevated PSA, of whether the prospective association of HK2 with prostate cancer risk is independent of highly collinear circulating PSA concentrations.

6.1.1 Human Kallikrein 2

Human kallikrein 2 (HK2) is a serine protease coded for by a kallikrein gene, KLK2. Although HK2 is expressed primarily by the prostate epithelium and is present in the prostatic fluid, it is also expressed at lower levels in amniotic fluid, breast milk, and saliva [285, 286]. HK2 is frequently co-expressed with prostate-specific antigen (PSA) [285], and shares 80% sequence homology with PSA [287]. Despite these similarities, HK2 is present at much lower concentrations than PSA (approximately 1% of PSA) in prostate tissue, semen, and plasma and serum samples [285, 43], and levels of hK2 mRNA transcript expression are half those of total PSA [288]. Further, there is some evidence from *in vitro* studies that HK2 and PSA differ in their enzymatic activity, and that HK2 is able to function as a protease to activate itself [289, 290, 291].

The function of HK2 is not yet clear. There is some evidence that HK2 has a role in seminal clot liquefaction after ejaculation [292], and that HK2 may activate pro-PSA, *in vitro* [293, 294]. However, there is also evidence that the transcription of the KLK gene family is regulated by steroid hormones that include androgens and estrogens [295]. Further, the only known protein substrate of HK2 is the ARA70, which is an androgen receptor coregulator, and this suggests that HK2 may be involved in regulation of androgen receptor-related tissue growth [292].

6.1.2 Factors associated with Human Kallikrein 2

A PubMed search using the key terms *"Human Kallikrein 2" OR "HK2" OR "KLK2"* returned 2,253 articles. None of these articles contained any cross-sectional analysis on the association of HK2 with common epidemiological factors such as age, height, or weight.

6.1.3 Human Kallikrein 2 and prostate cancer

Tissue expression of HK2

Initial evidence for an association of HK2 with prostate cancer came from expression in tissue studies, and focused on the differences in expression patterns of HK2 compared to total PSA [296, 291, 297]. Darson et al. found that HK2 was most intensely expressed in prostate adenocarcinoma, while PSA was most intensely expressed in benign prostate tissue [296]. Subsequently, Darson et al. (1999) found a monotonic increase in the percentage of cells expressing HK2 by tumour sub-type: 46.4% for benign cells, 83.2% for primary cancer, and 87.0% for lymph node metastases, which led to the conclusion that HK2 levels may be most strongly associated with metastatic prostate cancer. However, note that this study also reported a monotonic increase for cells expressing total PSA: 63.2% for benign cells, 75.2% for primary cancer, and 85.9% for lymph node metastases [297].

Case-only studies of HK2

A large number of case-only studies that compared the circulating concentrations of HK2 in patients with differing tumour profiles, radical prostatectomy, and also benign prostate hyperplasia (BPH) emerged following these expression studies. Most [47, 48, 49, 50, 51, 52, 282, 283, 284], but not all [298] of these studies supported statistically significantly higher concentrations of HK2 in more severe tumour subtypes when compared to less aggressive subtypes, or to BPH. Further, when compared to total PSA, circulating HK2 concentrations were found to positively correlate more strongly than PSA with the volume of high grade tumours (Pearson's r, 0.56 vs. 0.36 for Gleason's primary grade ≥ 4, for total PSA and HK2, respectively) [299].

Although, on balance, case-only studies have found that circulating HK2 concentration provides greater discrimination [1] for more severe tumour subtypes when compared to total PSA (PSA), there has been substantial variation in findings between studies (percentage difference for area under the

[1] Owing to the aim of the majority of available research into HK2 being disease prediction, few studies on HK2 and prostate cancer report relative risks, and instead report a measure of discrimination, area under the receiver operating curve (AUC). AUC is a rank order statistic that estimates the probability that, if presented with a case and a control, a model will rank the risk of disease as higher for a case compared to a control.

curve (AUC) HK2 vs. PSA range -4 to 17) [47, 48, 49, 50, 51, 52, 282, 283, 284, 53].

Only three of these case-only studies have considered an association of HK2 indepedent from PSA: after adjustment for PSA, one study found no statistically significant association of HK2 with advanced prostate cancer compared to localised prostate cancer [53]; a study of biochemical recurrence (BCR) following radical prostatectomy reported no association of HK2 with BCR after adjustment for PSA, unless analyses were restricted to men with PSA 10 ng/ml prior to RP [284]; and a study that built a predictive model for minimal prostate cancer compared to moderate or high-risk prostate cancer used a backward stepwise procedure that did not find a significant improvement of a model with PSA when HK2 was included [53].

Prospective case-control studies of HK2 and prostate cancer risk

With the exception of two studies from the Malmö Preventative Project, all case-control studies that have reported on the association of HK2 with prostate cancer risk have been in men with an elevated PSA (2.5 ng/ml). The majority of such studies have been of men in screening projects, such as the European Randomized Study of Screening for Prostate Cancer (ERSPC) [42, 43, 44, 45, 46], or the Göteborg Screening Programme (GSP) [40, 41]. Studies from both ERSPC and GSP have generally found higher concentrations of HK2 in men who developed prostate cancer compared to those without prostate cancer. However, less consistency is observed for AUC statistics, which have been reported for three different methods of evaluation, described in turn. GSP did not observe greater discrimination in a univariate model for prostate cancer cases from negative biopsy controls (HK2: 0.67 vs. PSA: 0.70) [40, 41]. Studies using ERSPC data have reported AUC in two ways: as AUC from a model with PSA and age at recruitment, with and without HK2, and have found an increase in AUC ranging from 1% to 5% when HK2 was included for prostate cancer overall[46, 45, 44], and either a 1% increase [46] or no increase [44] for high grade disease; or as a change in AUC when HK2 was excluded from a larger model of kallikrein proteins (HK2, free, intact, and total PSA, and age), which suggests a 1% reduction in AUC for prostate cancer overall, and between a 2% and 6% reduction for high grade disease when HK2 was excluded from analyses [42, 43].

Two further studies, from the Malmö Preventative Project [300] and a multi-centre trial of a four kallikrein prediction score [301], used the same method to evaluate the contribution of HK2 to AUC, and found no reduction in discrimination for prostate cancer overall when HK2 was excluded, and between a 2% and 3% reduction for high grade disease when HK2 was excluded. Subsequently, a study in the Stockholm 3 cohort found a 1% lower AUC for high grade from a univariate HK2 model compared to a univariate PSA model [302].

Lastly, two prospective studies in the Malmö Preventative Project found moderately consistent univariate associations of HK2 with incident prostate cancer risk (RR:1.75, 95% CI 1.45-2.11 & RR:1.25, 95% CI 1.13-1.38) [28, 29]. However, to my knowledge, only one study has investigated the association of HK2 with prostate cancer adjusted for PSA; the Princess Margaret Hospital cohort found a significant increased risk for prostate cancer with HK2 (univariate, lowest vs highest fourth RR: 5.83, 95% CI, 2.81-12.1, and PSA adjusted, lowest vs highest fourth RR: 6.72 95% CI, 2.90-15.6). However, it is worth noting that controls had elevated total PSA with a mean of 10 ng/ml [265]; given mean total PSA among controls in prospective studies has previously been reported as ranging from 0.6 to 1.0 ng/ml [72, 74], these concentrations may raise concern over the controls included in this study.

As such, although there is substantial evidence that expression and circulating concentration of HK2 may be higher in cases than in controls, and that HK2 concentration may be higher in more severe prostate cancer subtypes, it is not clear whether this is due to the co-localisation of HK2 with PSA or due to an independent association of HK2 with prostate cancer risk (either as a risk factor or as a marker of disease). Indeed, there is minimal evidence that a model of prostate cancer risk calculated using PSA and HK2 combined compared to PSA alone improves discrimination for both prostate cancer overall (1% to 4%) and for high grade disease (1% to 6%).

6.1.4 Aim of Study

This chapter will evaluate whether HK2 concentration is associated, independent of total PSA concentrations, with subsequent risk of prostate cancer overall, and whether this association varies by tumour characteristics in the largest prospective study to date using data from the European Prospective Investigation into Cancer and Nutrition (EPIC).

6.2 Methods

6.2.1 Study Population

These analyses are based on data from the countries within the EPIC cohort that had samples stored at the central IARC Biobank: Germany, Greece, Italy, The Netherlands, Spain, and the United Kingdom [140]; or in the Malmö centre for Swedish men. For complete details on the EPIC cohort, please see Chapter 3.

6.2.2 Follow-up for cancer incidence and vital status

Cancer incidence was identified through record linkage to regional or national registries in most countries. For Germany and Greece, combined health insurance records, regional health departments, municipality registries, hospital or physician-based cancer and pathology records, or mail or phone call-based follow-up were used. Follow-up procedures continued to date of prostate cancer diagnosis, death, or last follow-up completed (from 31 December 2007 to 14 June 2010 according to recruitment center).

Cases were defined as men who were diagnosed with prostate cancer as the first incident malignancy (International Classification of Diseases 10th revision code C61 [173]) after the date of blood collection and before the end of the study period, as determined by the latest date of follow-up in each study center. For Germany, Greece, Italy, The Netherlands, Spain, and the United Kingdom an incidence density sampling protocol was used to select control participants at random from the cohort of men who were alive and free of cancer (except non-melanoma skin cancer) at the time of diagnosis of the index case, and who were matched on study centre, length of follow-up, age at blood collection ($\pm$ 6 months), time of blood collection ($\pm$ 1 hour), and duration of fasting at blood collection ($\leq$ 3, 3-6, $\geq$ 6 hours). For Sweden (Malmö), controls were matched on length of follow-up and age at blood collection ($\pm$ 6 months) - time of blood collection and fasting status were unavailable. For the current analyses participants were 2,867 cases with 2,867 matched controls.

No reliable prior estimates exist for the expected association of HK2 with prostate cancer risk after adjustment for PSA. However, compared to a previously reported association in the the Princess Margret cohort (lowest vs highest fourth RR: 6.72 95% CI, 2.90-15.6) [265], this study was adequately

powered to detect even a modest association of HK2 with prostate cancer risk after adjustment for PSA; 392 cases and 392 controls would be needed to discover a 50% increased relative risk of prostate cancer for the top fourth compared to the bottom fourth of MSP concentrations with 80% power (α=0.05) [220].

Stage information at diagnosis was available for 1,532 cases (67.5%): 1,054 were localised (tumor-node-metastasis [TNM] staging score of T1-T2 and N0/Nx and M0/Mx, or stage coded in the recruitment centre as localised); 478 were identified as advanced prostate cancer (T3-T4 and/or N1-N3 and/or M1, or stage coded in the recruitment centre as metastatic). Tumor grade information at diagnosis was available for 1,799 cases (85.1%): 1,574 were low-intermediate grade (Gleason score < 8, or grade coded as well, moderately, or poorly differentiated) and 225 were identified as high grade prostate cancer (Gleason score ≥ 8, or grade coded as undifferentiated).

6.2.3 Assessment of Human Kallikrein 2 and Prostate-specific Antigen

Immunoassay measurements for total PSA and HK2 [226] were conducted on the AutoDelfia®1235 automatic immunoassay system at the Wallenberg Research Laboratories, Department of Translational Medicine, Lund University, Skåne University Hospital, Malmö, Sweden, and with all measurements conducted blinded to case status. We measured total PSA using the dual-label DELFIA Prostatus®total PSA-Assay (Perkin-Elmer, Turku, Finland)[226] calibrated against the WHO 96/670 (PSA-WHO) standard.

In a pilot study, the concordance between measurements of HK2 and PSA analyte concentration in serum and citrated plasma samples was assessed (N = 25 for serum and plasma, respectively), and a high concordance was observed (r = 0.98 and 0.99, for HK2 and total PSA, respectively). Additionally, the temporal reproducibility of analyte concentrations was assessed between samples drawn at five year intervals from 49 and 40 individuals for HK2 and PSA, respectively. There were no statistically significant differences between plasma concentrations of HK2 (p = 0.09) or PSA (p = 0.8) drawn at five year intervals in controls using a paired t-test, and the intra-class correlation coefficient (ICC) was 0.08 (95% CI: 0.00-0.41) for HK2 and 0.23 (95% CI 0.00-0.53) for PSA. Low ICC may be expected as pilot was conducted in controls with concentrations very close to lower limits of detection where

assays are likely moderately less accurate. Given the wider availability of plasma samples for the EPIC cohort, all assays for the nested case-control study were performed using plasma.

Quality control samples were inserted into each assay batch and analysed in duplicate; samples reflected low, medium, and high values according to the calibration chart. The average inter-assay coefficients of variation (CVs) for HK2 and PSA were 11% and 14% for phase 1, 7% and 9% for phase 2, and 9% and 7% for phase 3, respectively. An additional 286 "blinded" quality control samples were inserted. Quality control samples were pooled sodium citrate plasma from at least two males. All intra- and inter-assay CVs were $< 11\%$. The detectable ranges for HK2 were 0.002 to 2.27 ng/ml and 0.1 to 250 ng/ml for PSA.

6.2.4 Statistical Analyses

Plasma concentrations below the lower limits of detection for HK2 and PSA were set to half of the lowest value of detection (PSA, N = 9) while concentrations above the upper levels of detection were set to the highest detectable value for that particular analyte (HK2, N = 83; PSA, N = 75). Pearsons chi-squared tests for differences and paired t-tests for categorical and continuous variables, respectively, were conducted between matched case-control sets for anthropometric and lifestyle characteristics. Analysis of variance was used to assess differences in analyte concentrations in controls by strata of selected characteristics, and by country and, (for all countries except Sweden where data were collected in one set) by study phase (for Germany, Greece, Italy, The Netherlands, Spain, and the United Kingdom prostate cancer follow-up was conducted in three waves, occurring approximately in 2004, 2008, and 2010). Additionally, analysis of variance was used to test for differences in mean analyte concentrations by days in post for samples from the EPIC-Oxford sub-cohort. To conform to parametric model assumptions, log transformations were applied to HK2 and PSA concentrations and results are presented as geometric means adjusted for age at blood collection, BMI, and recruitment centre. The relationship between log transformed analyte concentrations was examined using partial correlation adjusted for age, BMI, and recruitment centre.

Conditional logistic regression models were used to examine the association of HK2 concentration with risk of prostate cancer, conditioned on

the matching factors (listed above). A sensitivity analysis was conducted to assess the impact of additional adjustment for: smoking status, alcohol consumption, marital status, physical activity, and educational attainment, on top of a base model with exact age and BMI. Due to apparent sensitivity of estimates to adjustment, all subsequent models were fully adjusted for all aforementioned variables (see Table 6.1).

Given the high collinearity of PSA and HK2 [285], the previously reported strong positive association of PSA with prostate cancer [72], and the use of PSA testing in clinical practice [92], there was an *a priori* concern of residual confounding for a HK2 association with prostate cancer following adjustment for circulating PSA concentrations. Thus, further sensitivity analysis and model checking was conducted to ensure no such residual confounding was due to the manner by which PSA was entered into regression models. Firstly, PSA was entered as a categorical variable with fourths delimited in controls - a common practice for biomarker studies, in general [246, 27]. Secondly, PSA was entered as a restricted cubic spline with knots at the tertiles - the most common manner that PSA, specifically, is modeled in regression analyses due to its extreme non-linear relationship with prostate cancer risk and its extreme positive skew [42, 43, 44, 45, 46]. Best model fit was assessed by reference to McFadden's R^2, and Akaike and Bayesian [303] Information Criteria (AIC and BIC, respectively). Further, differences in model fit were evaluated by comparing mean absolute error between models with PSA (categorical) and PSA (spline) models at deciles of PSA and HK2.

A further concern when modeling two highly collinear variables is that outlier values may have undue influence, and, that although ordinary least squares (OLS) estimates may be unbiased, their variances will likely be inflated. Ridge regression is a commonly used method to estimate model coefficients for highly collinear variables, and achieves this by adding a regularisation parameter, λ, to the OLS loss function that penalises extreme parameter values. Although this implies that the ridge estimator may no longer be *unbiased*, the variance of the ridge estimate can be substantially lower such that the mean square error is also less. Consequently, the ridge regression may be less likely to falsely reject a small, true association when adjusting for highly collinear covariates [304]. As such, ridge regression was additionally used to investigate the association of HK2 with prostate cancer risk overall, with adjustment variables as previously specified (with exception

that input variables were standardised [305] as the ridge regression is sensitive to difference in scale for variables [304]) with and without adjustment for PSA (spline).

Conditional logistic regression models then were repeated in subgroups defined according to time between blood collection and diagnosis ($\leq$ 7, 7-10, $\geq$10 years), and age at blood collection ($\leq$60, >60 years). Due to the strong positive association of HK2 with age at diagnosis, an analysis by fourth of HK2 concentrations stratified by categories defined according to age at diagnosis was not possible; instead the association of HK2 (linear continuous) is reported by strata of age at diagnosis ($\leq$65, >65 years). In addition, we conducted analyses by prostate tumour stage (localised, advanced), histological grade (low-intermediate or high grade). For all conditional logistic regression models linear trend was tested by entering a pseudo-continuous variable equal to the medians of the fourths of HK2 concentration. For subgroup and subtype analyses, likelihood ratio tests were used to test for heterogeneity of the association of HK2 concentration with risk of prostate cancer.

Much of the previous research into the association of HK2 with prostate cancer research has related to the improved discrimination that it may provide for prostate cancer overall or for high grade prostate cancer, when added to a PSA-based predictive model; predictive models have typically been PSA + age, however, models have also included other PSA subforms. As such, logistic regression was additionally used to estimate the probability of prostate cancer overall and for high grade tumours using a model with PSA (including aforementioned adjustment factors) and separately with PSA and HK2 (including aforementioned adjustment factors). AUC were then calculated for each model to assess the additional discrimination contributed by the inclusion of HK2. All AUC comparisons were by DeLong method [306].

All statistical tests are two-sided and were conducted using STATA software version 14 (College Station, TX: StataCorp LP) or R software version 3.32.

6.3 Results

Data from 2,867 prostate cancer cases and 2,867 matched controls were included in analyses. The median age at blood collection was 58 years (range, 39 to 79 years), and, for cases, the median time between blood collection and diagnosis was 8.4 years. No significant differences were observed in selected baseline characteristics between cases and controls (see Table 6.2).

Mean HK2 concentration (ng/ml) at blood collection was significantly higher among cases compared to controls (adjusted geometric mean = 0.047; [95% CI 0.045-0.048] and 0.028; [95% CI 0.027-0.029] respectively, $p < 0.0001$) before adjustment for PSA; after adjustment for PSA, HK2 did not differ significantly between cases and controls (adjusted geometric means were 0.036; [95% CI 0.035-0.037], and 0.037; [95% CI 0.036-0.037], respectively, $p = 0.4$). Mean PSA concentration (ng/ml) measured at blood collection was about three-fold higher in cases than in controls (adjusted geometric mean = 2.5; [95% CI 2.4-2.6] and 0.9; [95% CI $p < 0.0001$, see Table 6.2). No significant differences were observed for HK2 or PSA concentrations by days in the post for samples from the EPIC-Oxford sub-cohort ($p = 0.7$ & $p = 0.2$, respectively). Additional case characteristics can be found in Table 6.3.

HK2 concentration in controls was higher in men who were older at blood collection, married, had a normal/low BMI or low intake of alcohol, never smoked, and had a highest educational attainment of degree level when compared to both primary school, technical, and degree level ($p < 0.05$ for all) (see Table 6.4). PSA concentration was positively associated with age at blood collection and educational attainment, and was lower in men with greater BMI and diabetes (see Table 6.4). Further, there were significant differences in HK2 and PSA concentrations by recruitment country and by study phase (see Table 6.5).

HK2 and PSA concentrations were strongly significantly, positively correlated in both cases and controls (Partial correlation $r = 0.8$ and $r = 0.8$, respectively, $p < 0.0001$). Moreover, there were modest differences in this correlation by country with the lowest correlation for control men in Spain ($r = 0.54$) and highest correlation in Greece ($r = 0.80$) (see Table 6.6).

HK2 concentration was strongly associated with risk of prostate cancer after adjustment for age at blood collection, BMI, smoking status, alcohol

consumption, marital status, total physical activity, and educational attainment (OR for highest versus lowest fourth = 7.09, [95% CI 5.82-8.65], p-trend across the medians of the fourths = 0.001). However, given the *a priori* expectation of an association of PSA with HK2 [285, 43], and PSA with prostate cancer [72], we additionally adjusted for PSA.

Initial adjustment for PSA was completed by entering PSA as a categorical variable with fourths delimited by the PSA distribution among controls, and led to a 76% attenuation in the effect estimate (OR for highest versus lowest fourth = 1.70; [95% CI 1.32-2.19], p-trend across the medians of the fourths = 0.001). Compared to entering PSA in fourths, a model with PSA entered instead as a cubic spline greatly improved model fit ($BIC^{fourths}$ - BIC^{spline} = 194, where evidence is considered very strong with BIC difference > 10 [303]), while also attenuating HK2 effect estimates by 82% (see Table 6.8). Further investigation demonstrated that the improved fit for a model with PSA entered as a cubic spline was due mostly to the reduced error (up to 44%) in model fit for men with high concentrations of HK2 and PSA (see Table 6.7). After adjustment for PSA (spline), HK2 concentration was not significantly associated with prostate cancer risk (OR for highest versus lowest fourth = 1.29; [95% CI 0.98-1.67], p-trend = 0.1) (see Table 6.9. & Figure 6.1). I further conducted a ridge regression, with all variables standardized [305] to assess whether a small, true, association of HK2 with prostate cancer risk was falsely rejected due to inflated variances in OLS caused by high collinearity between HK2 and PSA. A fully adjusted model, excluding PSA, found a strong association of HK2 with prostate cancer risk (RR for unit (ng/ml) increase in HK2: 4.33; [95% CI 3.09-6.08]). After additional adjustment for PSA (spline), no evidence for an association of HK2 with prostate cancer remained (RR for unit (ng/ml) increase in HK2: 1.24; [95% CI 0.96-1.54]).

The association of HK2 concentration with prostate cancer did not differ by tumour stage or grade (all p-heterogeneity > 0.05; Table 6.9). Further, no heterogeneity in the association of HK2 with prostate cancer risk was observed by time to diagnosis or age at blood collection (all $p > 0.05$, see Table 6.9). Further, the association of HK2 did not differ by age at diagnosis (RR for < 65: 1.88; [95% CI 0.003-1136.38] & ≥ 65: 2.40; [95% CI 0.33-17.32] $p = 0.9$), and there was no heterogeneity of the association of HK2

with prostate cancer risk by recruitment country (p-heterogeneity = 0.5; Table 6.10).

Additionally, I investigated the ability of HK2 to improve upon the discrimination of a model with PSA (spline) and age (plus other adjustment factors, see Table 6.1) for prostate cancer overall and for high grade tumours. No difference in AUC was found between a PSA-based model and a PSA + HK2 model for prostate cancer overall (AUC, 0.816 vs. 0.816, see Figure 6.2) or for high grade tumours (AUC, 0.752 vs. 0.752, see Figure 6.3).

6.4 Discussion

In this large prospective study, there was little or no evidence of a higher risk for prostate cancer in men with higher circulating concentrations of the prostate protein HK2 after adjustment for circulating PSA concentrations. Further, there was no evidence that the additional inclusion of HK2 in a model with PSA, among other common epidemiological factors, improves model discrimination for prostate cancer overall, or for high grade tumours.

Only one study has presented prospective evidence for an association of HK2 after adjustment for circulating PSA concentrations [265]. However, control men had notably high average PSA concentrations (10 ng/ml), and so it is unclear whether these analyses represent an investigation of HK2 and prostate cancer risk among men comparable to an average male population where PSA among cancer-free men is typically below 1 ng/ml [72].

Other prior research into the association of HK2 with prostate cancer risk has focused primarily on prediction, and so results have been presented as AUC rather than relative risks. These studies have suggested a small improvement for a model that includes HK2 together with PSA compared to a PSA-only model for prostate cancer overall (1% to 4%), and a more modest improvement for high grade prostate cancer (1% to 6%) [42, 43, 44, 45, 46, 40, 41]. In contrast, the current study found no improvement of discrimination for overall or for high grade prostate cancer when HK2 was included in a model that contained PSA, age, BMI, and other epidemiological factors (see Table 6.1).

It is possible that the matched design of the current study affected the accurate derivation of the AUC statistic; when matching variables, such as age, are positively associated with prediction variables, the matching process may

artificially reduce variance for predictor variables, make an ordinal ranking more difficult, and lead to an underestimate AUC statistics [307, 308]. However, if this were true for current results, a downward bias for AUC would be expected. Estimated discrimination from these data for a PSA model are similar in magnitude to previous estimates for unmatched designs, and so it does not appear that the matched design is likely to have influenced current AUC estimates. Further, it is not clear that a matched design would lead to an underestimate for the discrimination of a HK2 and PSA model, while not underestimating discrimination for a PSA-only model. As such, it does not appears that the matched design for this study biased the AUC present here.

Even in the presence of evidence in favor of a functional important role of HK2 in the biology of prostate cancer distinct from PSA [289, 290, 291], the strong positive correlation between HK2 and PSA [27], and their extreme positive skew, may have hindered the accurate estimation of the association of HK2 with prostate cancer risk. When multicollinearity occurs, outlier values may have undue influence, and although the OLS estimate may be unbiased, their variances will likely be inflated. Thus, it is possible that a small true, independent, association of HK2 and prostate cancer risk may have been incorrectly rejected [304]. However, HK2 estimates, or their variances, derived using ridge regression, a commonly used technique to more accurately estimate regression coefficients in the presence of extreme collinearity, did not differ materially from those calculated by OLS. As such, I do not believe that the non-significant association of HK2 with prostate cancer from EPIC has resulted from error induced by high collinearity between PSA and HK2.

While initial tissue expression work and case series analyses suggested that HK2 may be more strongly associated with high grade and advanced prostate cancer when compared to low grade or localised tumours, or BPH [296, 291, 297], no previous prospective study has considered the heterogeneity of association of HK2 with prostate cancer by tumour characteristics. In the current study, there was no evidence of heterogeneity in the association of HK2 with prostate cancer risk by tumour stage or grade. However, there are relatively small numbers of cases in subgroups defined by tumour characteristics and the analyses by stage and grade of tumour have limited power to evaluate heterogeneity between and associations by tumour subtype. Given the strong correlation between PSA and HK2 [285, 43], the use of PSA in

clinical practice [92], and the putative reduction in mortality associated with PSA screening [249, 77], follow-up time may be an important factor in the analysis of the HK2 and prostate cancer risk. In a cohort where many cases may be those men diagnosed with prostate cancer after a referral for biopsy from an elevated PSA value, PSA, and thus HK2, at recruitment will likely be strongly associated with prostate cancer. More specifically, we would expect PSA and HK2 to be more strongly associated with cases diagnosed soon after recruitment. However, there is little evidence from the current research for heterogeneity of the association of HK2 by follow-up time or that HK2 may be strongly associated with low grade or localised prostate cancer. As such, it is not clear that the use of PSA in clinical practice has confounded the association of HK2 with prostate cancer risk in the current study.

To the best of my knowledge, this is the first large scale prospective study to investigate the association of circulating HK2 concentrations with common epidemiological factors: age, marital status, educational attainment, BMI, smoking status, and alcohol consumption. Many of these associations (age, alcohol consumption, smoking status, and BMI) may not be surprising given the positive association of HK2 and PSA, and the association of these factors with PSA, which has been attributed to decreased testosterone levels [309, 310] and hemodilution [311]. However, the association of HK2 with marital status is less easily interpreted. Much of the research on HK2 to date has been done in men with elevated PSA, using negative-biopsy controls, and without adjustment for common epidemiological factors; it is possible that previously reported positive associations of HK2 with prostate cancer were influenced by unaccounted for variation in other factors that may influence HK2 among men with abnormally functioning prostates, as indicated by elevated PSA.

6.5 Conclusion

In this large prospective European nested case-control study there was little evidence for an association of HK2 with prostate cancer risk overall, or with risk for advanced stage or high grade tumours independent of circulating PSA concentrations. In contrast to previous results from case-control studies, there was no evidence to suggest the addition of HK2 to a predictive model of prostate cancer risk using PSA leads to improved discrimination.

Figure 6.1: Multi-variable adjusted odds ratios (95% CI) for prostate cancer by fourth of plasma human kallikrein 2 (HK2) concentration

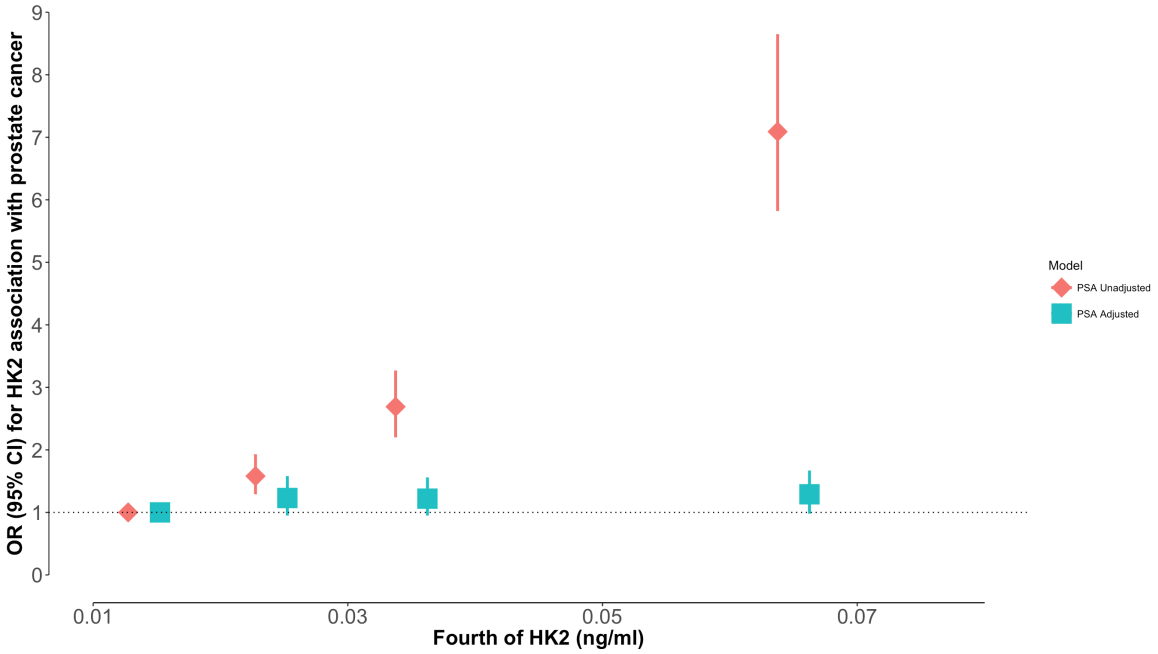


Figure 6.2: Receiver operator curves and area under the curve statistics for a prostate-specific antigen (PSA) + Age model and a PSA + human kallikrein 2 (HK2) + Age model for prostate cancer overall

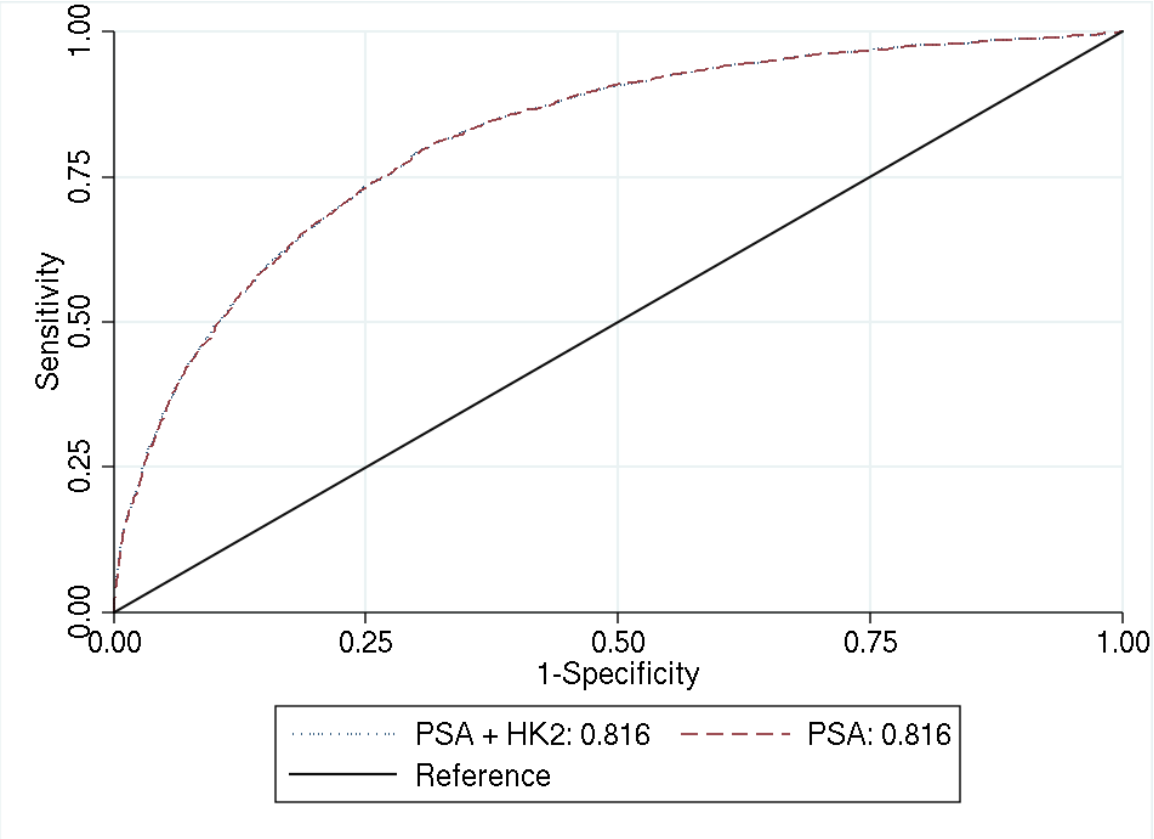


Figure 6.3: Receiver operator curves and area under the curve statistics for a prostate-specific antigen (PSA) + Age model and a PSA + human kallikrein 2 (HK2) + Age model for high grade prostate cancer

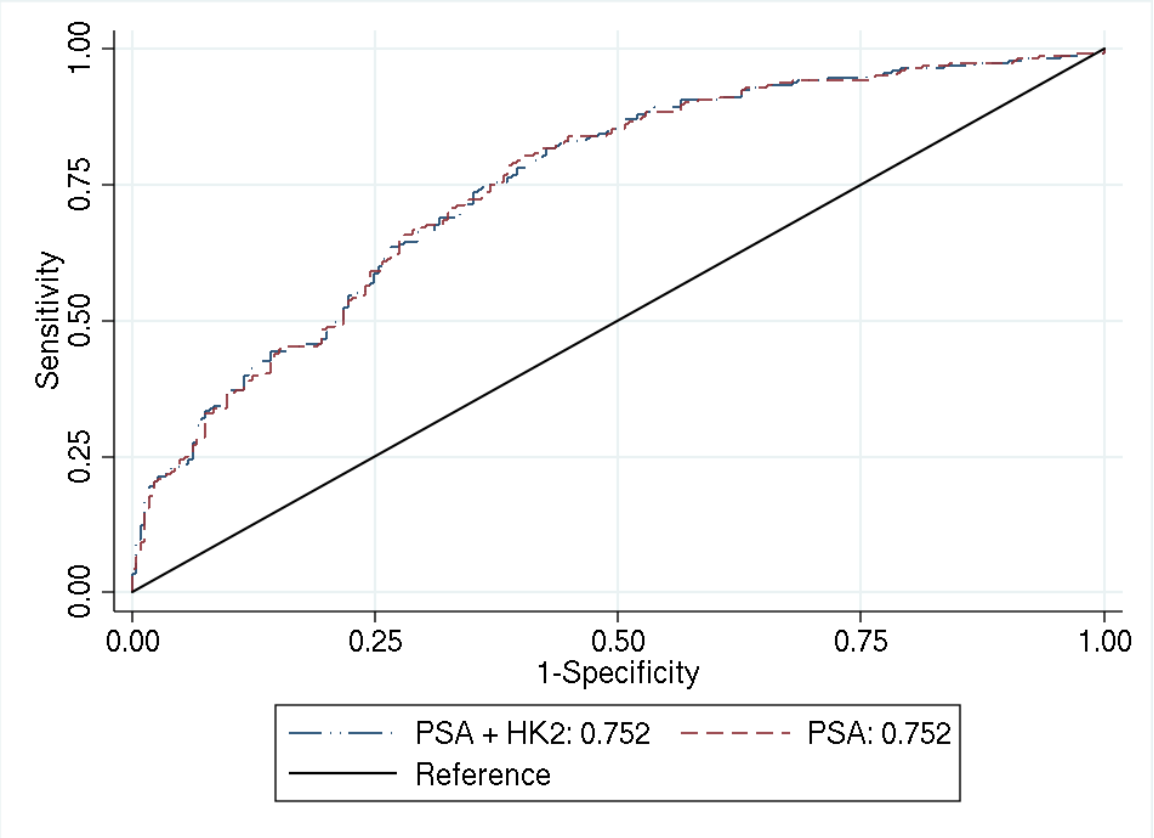


Table 6.1: Sensitivity analysis of odds ratio (95% CI) for prostate cancer associated with fourth of human kallikrein 2 (HK2) (lowest compared to highest fourth)

Model	Odds Ratio(95% CI) c	p-trend[d]
Basic Model[a]	1.29 (1.00-1.67)	0.08
Adjusted Models		
Basic + smoking status	1.29 (1.00-1.67)	0.09
Basic + alcohol consumption	1.30 (1.01-1.69)	0.07
Basic + marital status	1.29 (1.00-1.67)	0.08
Basic + total physical activity	1.29 (1.00-1.67)	0.08
Basic + educational attainment	1.21 (0.88-1.66)	0.2
Fully Adjusted model[b]	1.22 (0.89-1.68)	0.3

[a] Minimally adjusted model: total PSA, age at blood collection, and BMI

[b] Basic model + smoking status, alcohol consumption, marital status, total physical activity, and educational attainment

[c] Odds ratio for comparison between highest and lowest quartile of MSP

[d] Test for trend obtained by pseudo continuous variable of median concentration within each fourth of human kallikrein 2 concentration

Table 6.2: Characteristics of control participants and men who developed prostate cancer in EPIC

Characteristic	Controls (n = 2,867)	Cases (n = 2,867)	p[a]
Age at Blood Collection, years (SD)	58.9 (6.8)	58.9 (6.8)	0.4
Weight, kg (SD)	80.2 (11.5)	80.4 (11.6)	0.4
Height, cm (SD)	174.1 (7.2)	172.0 (7.2)	0.8
BMI[b], kg/m^2	26.5 (3.4)	26.5 (3.5)	0.3
Smoking Status, n (%)			
Never	936 (32.9)	956 (33.7)	
Previous	1,243 (43.8)	1,264 (44.6)	
Current	660 (23.3)	617 (21.8)	0.4
Alcohol, n (%)			
<8	1,031 (35.9)	1,064 (37.1)	
8-15	613 (21.4)	597 (20.8)	
16-39	874 (30.5)	807 (28.2)	
>40	349 (12.2)	399 (13.9)	0.1
Physical Activity, n (%)			
Inactive	489 (17.3)	473 (16.8)	
Moderately Inactive	977 (34.5)	1,012 (35.8)	
Active	1,365 (48.2)	1,339 (47.4.7)	0.6
Marital Status, n (%)			
Married/Cohabitating	1,435 (59.6)	1,390 (58.4)	
Not Married/Not Cohabitating	971 (40.4)	991 (41.6)	0.4
Educational attainment, n (%)			
Primary/None	1,140 (40.8)	1,120 (40.3)	
Technical	636 (22.8)	603 (21.7)	
Secondary	299 (10.7)	305 (10.9)	
Degree	718 (25.7)	752 (27.1)	0.6
Geometric mean analyte concentration			
HK2, ng/ml (95% CI)	0.028 (0.027-0.029)	0.047 (0.045-0.048)	<0.0001
HK2 adj. PSA, ng/ml (95% CI)	0.037 (0.036-0.037)	0.036 (0.035-0.037)	0.4
PSA, ng/ml (95% CI)	0.9 (0.8-0.9)	2.5 (2.4-2.6)	<0.0001

[a] Significance calculated from analysis of variance and chi-square for continuous and categorical variables, respectively

[b] Body mass index (BMI), human kallikrein 2 (HK2), prostate-specific antigen (PSA)

Table 6.3: Characteristics of men who developed prostate cancer in EPIC

Characteristic	Cases (n = 2,867)
Time to diagnosis, n (%)	
≤ 2 years	141 (4.5)
2 to 4 years	190 (6.7)
4 to 6 years	352 (12.3)
6 to 8 years	535 (18.7)
≥ 8 years	1,639 (57.4)
Year of diagnosis, median (range)	2004 (1991-2013)
Age at diagnosis, years (SD)	67.9 (6.8)
TNM-Code[a]	
Tumour	
T1	311 (24.9)
T2	648 (51.8)
T3	262 (20.9)
T4	28 (2.2)
Nodes	
N0	727 (92.8)
N1	51 (6.5)
N2	4 (0.1)
N3	1 (0.01)
Metastases	
M0	749 (92.2)
M1	63 (7.7)
EPIC Stage Information[a]	
Localised	765 (79.0)
Metastatic	203 (20.9)
Tumour Grade	
Gleason Grade[a]	
≤ 6	703 (50.2)
7	484 (34.5)
≥ 8	214 (15.3)
EPIC Grade Information[a]	
Well Differentiated	132 (16.2)
Moderately Differentiated	496 (60.8)
Poorly Differentiated	184 (22.6)
Undifferentiated	4 (0.5)
PSA[b] (ng/ml) at diagnosis	
<3	33 (3.6)
≥ 3 and <10	473 (51.8)
≥ 10 and <50	325 (35.5)
≥ 50	82 (8.9)

[a] TNM-code and EPIC stage, and Gleasons grade and EPIC grade are not mutually exclusive

[b] Prostate-specific antigen (PSA)

Table 6.4: Adjusted geometric means[a] of human kallikrein 2 (HK2) and prostate-specific antigen (PSA) concentration (ng/ml) in controls by selected characteristics in EPIC

Factor and Subset	N	Mean HK2 (95% CI)	p-diff/trend[b]	N	Mean PSA (95% CI)	p-diff/trend[b]
Age at blood collection, years						
<50 years	260	0.023 (0.021-0.025)		260	0.6 (0.5-0.7)	
50 years to 55 years	503	0.024 (0.023-0.025)		503	0.7 (0.6-0.7)	
55 years to 59 years	730	0.026 (0.025-0.028)		730	0.8 (0.8-0.9)	
60 years to 64 years	837	0.029 (0.027-0.030)		837	0.9 (0.9-1.0)	
65 years to 69 years	315	0.035 (0.032-0.038)		315	1.3 (1.2-1.4)	
>70 years	222	0.045 (0.041-0.049)	<0.0001/<0.0001	222	1.5 (1.5-1.8)	<0.0001/<0.0001
Marital status						
Married/Cohabitating	1,435	0.029 (0.028-0.031)		1,435	0.9 (0.9-0.9)	
Not married/ Not Cohabitating	971	0.027 (0.026-0.028)	0.007	971	0.9 (0.9-1.0)	0.7
Educational attainment						
Primary/None	1,140	0.026 (0.025-0.027)		1,140	0.8 (0.8-0.9)	
Technical	636	0.030 (0.029-0.032)		636	0.9 (0.9-1.0)	
Secondary	299	0.028 (0.026-0.031)		299	0.9 (0.8-1.0)	
Degree	718	0.029 (0.027-0.030)	0.0004	718	0.9 (0.9-1.0)	0.004
BMI, kg/m2						
16-24	1,007	0.029 (0.028-0.031)		1,007	0.9 (0.9-1.0)	
25-29	1,459	0.028 (0.027-0.029)		1,459	0.9 (0.9-1.0)	
>30	401	0.026 (0.024-0.028)	0.01/0.01	401	0.8 (0.7-0.9)	0.002/0.6
Smoking status						
Never	936	0.030 (0.029-0.031)		936	0.9 (0.9-1.0)	
Previous	1,243	0.028 (0.027-0.029)		1,243	0.9 (0.9-0.9)	
Current	660	0.027 (0.025-0.028)	<0.0001/<0.0001	660	0.9 (0.9-1.0)	0.9
Alcohol consumption, g/d						
<8	1,031	0.031 (0.029-0.032)		1.031	0.9 (0.9-1.0)	
8 to 15	613	0.029 (0.027-0.030)		613	0.9 (0.8-1.0)	
16 to 39	874	0.026 (0.025-0.028)		874	0.9 (0.8-0.9)	
>40	349	0.025 (0.023-0.027)	0.002/<0.0001	349	0.9 (0.8-0.9)	0.1/0.1
Diabetic						
No	2,606	0.028 (0.028-0.029)		2,606	0.9 (0.9-0.9)	
Yes	115	0.025 (0.022-0.028)	0.05	115	0.7 (0.6-0.8)	0.002

[a] All means adjusted for age at blood collection, body mass index (BMI), recruitment centre, and laboratory batch

[b] Test for difference by analysis of variance; test for trend by entering a pseudo-continuous variable for given factor in linear regression

Table 6.5: Adjusted geometric mean[a] plasma human kallikrein 2 (HK2) and prostate-specific antigen (PSA) concentration (ng/ml) in controls by country and study phase in EPIC

HK2 (ng/ml)	Germany	Greece	Italy	The Netherlands	Spain	Sweden	UK	Total	p^d
Phase 1									
Number	100	6	40	11	32	-	57	246	
HK2[a]	0.020 (0.018-0.023)	0.025 (0.014-0.044)	0.029 (0.024-0.037)	0.032 (0.021-0.049)	0.021 (0.016-0.027)	-	0.029 (0.024-0.036)	0.025 (0.022-0.027)	0.07
PSA[a]	0.7 (0.6-0.9)	0.7 (0.4-1.3)	0.9 (0.7-1.1)	0.9 (0.5-1.5)	0.9 (0.7-1.3)	-	0.9 (0.7-1.1)	0.8 (0.7-0.9)	0.6
Follow-up[b]	37.8 (33.4-42.1)	28.2 (10.3-45.9)	40.2 (33.3-47.1)	-	57.9 (50.3-65.7)	-	43.8 (37.4-50.3)	42.1 (38.8-45.3)	
Phase 2									
Number	206	31	84	23	108	-	196	648	
HK2[a]	0.026 (0.022-0.031)	0.027 (0.020-0.035)	0.024 (0.021-0.029)	0.031 (0.023-0.043)	0.031 (0.026-0.036)	-	0.037 (0.033-0.041)	0.031 (0.029-0.034)	0.0003
PSA[a]	0.9 (0.8-1.0)	0.7 (0.5-0.9)	0.9 (0.7-1.0)	0.8 (0.6-1.1)	0.9 (0.8-1.1)	-	1.1 (0.9-1.2)	0.9 (0.8-1.0)	0.04
Follow-up[b]	75.5 (72.5-78.6)	64.2 (56.4-71.9)	85.8 (81.1-90.5)	84.2 (75.1-93.2)	109.4 (105.2-113.6)	-	87.5 (84.4-90.6)	85.8 (83.9-87.7)	
Phase 3									
Number	383	38	148	74	132	-	202	977	
HK2[a]	0.026 (0.024-0.028)	0.028 (0.022-0.035)	0.025 (0.022-0.029)	0.025 (0.021-0.030)	0.023 (0.019-0.027)	-	0.037 (0.032-0.042)	0.026 (0.025-0.028)	<0.0001
PSA[a]	0.8 (0.7-0.9)	0.7 (0.6-0.9)	0.7 (0.6- 0.8)	0.9 (0.7-1.0)	0.7 (0.6-0.9)	-	0.9 (0.8-1.0)	0.8 (0.7-0.9)	0.04
Follow-up[b]	112.7 (110.3-115.1)	95.3 (87.7-102.8)	119.5 (115.6-123.4)	129.4 (123.9-134	130.2 (126.1-134.3)	-	124.1 (120.8-127.4)	119.0 (117.5-120.6)	
p^c								<0.0001	
Total									
Number	689	75	272	108	272	1,042	455	1,871	
HK2[a]	0.025 (0.023-0.026)	0.028 (0.024-0.032)	0.026 (0.024-0.028)	0.028 (0.024-0.031)	0.026 (0.024-0.028)	0.028 (0.027-0.029)	0.036 (0.034-0.039)	12.7 (12.4-13.1)	0.0001
PSA[a]	0.8 (0.7-0.9)	0.7 (0.5-0.8)	0.8 (0.7-0.9)	0.9 (0.7-1.0)	0.9 (0.8-1.0)	0.9 (0.9-1.0)	1.0 (0.9-1.1)	0.8 (0.8-0.9)	0.0001

[a] Adjustment was made for age at blood collection, body mass index, and laboratory batch, (95% CI)

[b] Approximate minimum time controls known to be free of prostate cancer, defined from corresponding matched case values of time to diagnosis

[c] For significant differences in adjusted mean concentration of MSP by study phase by analysis of variance

[d] Significance of difference across country for mean concentration of MSP by analysis of variance

Table 6.6: Partial correlation[a] between human kallikrein 2 and prostate-specific antigen by recruitment country in controls in EPIC

	Germany	Greece	Italy	The Netherlands	Spain	Sweden	UK
Overall	0.62*	0.80*	0.66*	0.64*	0.54*	0.71*	0.67*
Excluding							
First year	0.61*	0.77*	0.65*	0.65*	0.54*	0.70*	0.68*
First 5 years	0.57*	0.76*	0.61*	0.63*	0.51*	0.63*	0.64*
First 10 years	0.52*	0.69*	0.48*	0.71*	0.53*	0.61	0.56*

[a] Adjusted for age at recruitment, recruitment centre, and laboratory batch

* Significant at $p < 0.05$

Table 6.7: Mean absolute error (MAE) by tenth of HK2/PSA for a HK2 model of prostate cancer risk with PSA modeled in fourths variable compared to as a spline with knots at tertiles.

Decile	1	2	3	4	5	6	7	8	9	10
HK2	0.002-0.014	0.015-0.020	0.021-0.026	0.027-0.030	0.031-0.035	0.036-0.043	0.044-0.050	0.051-0.064	0.065-0.090	0.091-2.27
MAE for Model A[a]	0.24	0.23	0.27	0.28	0.29	0.28	0.29	0.30	0.30	0.28
MAE for Model B[b]	0.24	0.23	0.26	0.26	0.27	0.27	0.25	0.26	0.24	0.18
Percentage reduction in MAE	-	-	3.7	7.1	6.9	3.6	13.8	13.3	20.0	35.7
PSA	0.025-0.43	0.44-0.62	0.63-0.83	0.84-1.09	1.10-1.42	1.43-1.83	1.84-2.46	2.47-3.44	3.45-5.67	5.68-250
MAE for Model A[a]	0.15	0.20	0.27	0.32	0.34	0.35	0.31	0.30	0.27	0.27
MAE for Model B[b]	0.15	0.20	0.27	0.30	0.31	0.31	0.28	0.26	0.21	0.15
Percentage reduction in MAE	-	-	-	6.3	8.8	11.4	9.7	13.3	22.2	44.4

[a] Model is PSA as quartile with adjustment for confounders

[b] Model is PSA as splines with knots at the teriles with adjustment for confounders

Table 6.8: Statistics for model fit used to compare a model that fit PSA in fourth variable compared to one that fit PSA as splines with knots at the tertiles

Model	A[a]	B[b]	Difference
McFadden's R^2	0.38	0.43	0.05
BIC	2493	2293	194
AIC	2453	2265	188

[a] Model is PSA as quartile with adjustment for confounders

[b] Model is PSA as splines with knots at the teriles with adjustment for confounders

Table 6.9: Multi-variable adjusted odds ratios (95% CI) for prostate cancer by fourth of plasma human kallikrein 2 (HK2) concentration, subdivided by tumour subtype and subgroup in EPIC

		Fourth of Human Kallikrein 2 concentration, ng/ml					
		1	2	3	4	p-trend[d]	p-het[e]
Overall	Cases/controls, n	233/660	413/734	688/760	1,533/713		
	Mean MSP, ng/ml (range)	0.014 (0.002-0.018)	0.024 (0.019-0.028)	0.035 (0.029-0.043)	0.065 (0.044-2.67)		
	Adjusted RR (95% CI)[a]	1 (reference)	1.58 (1.29-1.93)	2.69 (2.20-3.27)	7.09 (5.82-8.65)	0.001	
	Adjusted RR (95% CI)[b]	1 (reference)	1.15 (0.88-1.48)	1.16 (0.89-1.49)	1.70 (1.32-2.19)	0.001	
	Adjusted RR (95% CI)[c]	1 (reference)	1.23 (0.95-1.58)	1.22 (0.95-1.56)	1.29 (0.98-1.67)	0.1	
Time to diagnosis							
<7 years	Cases/controls, n	37/204	79/230	155/245	675/267		
	Adjusted RR (95% CI)[c]	1 (reference)	1.36 (0.62-2.94)	0.98 (0.48-2.02)	1.33 (0.65-2.77)	0.3	
7 years to 10 years	Cases/controls, n	76/206	137/232	250/251	463/237		
	Adjusted RR (95% CI)[c]	1 (reference)	1.12 (0.72-1.75)	1.14 (0.73-1.78)	1.15 (0.73-1.81)	0.8	
>10 years	Cases/controls, n	120/250	197/272	283/264	395/209		
	Adjusted RR (95% CI)[c]	1 (reference)	1.22 (0.88-1.70)	1.31 (0.93-1.84)	1.26 (0.88-1.83)	0.4	0.9
Age at blood collection							
<60 years	Cases/controls, n	106/306	170/260	257/272	478/173		
	Adjusted RR (95% CI)[c]	1 (reference)	1.15 (0.77-1.72)	1.05 (0.69-1.58)	1.52 (0.98-2.34)	0.04	
$\geqslant$ 60 years	Cases/controls, n	126/352	240/468	427/481	1,043/535		
	Adjusted RR (95% CI)[c]	1 (reference)	1.22 (0.88-1.71)	1.27 (0.92-1.77)	1.19 (0.85-1.67)	0.7	0.5
Stage							
Localised	Cases/controls, n	83/272	136/238	241/262	594/281		
	Adjusted RR (95% CI)[c]	1 (reference)	1.29 (0.81-2.09)	1.31 (0.83-2.08)	1.21 (0.76-1.94)	0.2	
Advanced	Cases/controls, n	36/113	55/115	97/127	290/123		
	Adjusted RR (95% CI)[b]	1 (reference)	0.99 (0.49-2.01)	1.08 (0.53-2.19)	1.22 (0.58-2.54)	0.5	0.9
Grade							
Low-intermediate	Cases/controls, n	129/415	200/389	370/412	875/358		
	Adjusted RR (95% CI)[c]	1 (reference)	1.13 (0.78-1.64)	1.06 (0.74-1.52)	1.34 (0.93-1.94)	0.09	
High	Cases/controls, n	19/46	35/39	47/61	124/79		
	Adjusted RR (95% CI)[c]	1 (reference)	1.79 (0.68-4.73)	1.63 (0.64-4.16)	0.99 (0.38-2.63)	0.5	0.9

[a] Estimates are from logistic regression conditioned on the matching variables (see methods) with adjustment for age, and body mass index

[b] Additional to model [a], adjustment was made for total PSA as fourths

[c] Additional to model [a], adjustment was made for total PSA as splines at the tertiles

[d] A categorical variable was replaced with a continuous variable equal to median concentration by fourth of plasma HK2

[e] Test for heterogeneity in the trends by likelihood ratio test

Table 6.10: Multi-variable adjusted odds ratios (95% CI) for prostate cancer by fourth of plasma human kallikrein 2 concentration by recruitment country

		Fourth of human kallikrein 2 concentration, ng/ml					
		1	2	3	4	p-trend[b]	p-heterogeneity[c]
Country							
Greece	Cases/controls, n	5/17	9/19	16/18	43/19		
	Adjusted RR (95% CI)[a]	1 (reference)	0.48 (0.04-6.35)	3.58 (0.14-94.46)	0.67 (0.04-11.56)	0.9	
The Netherlands	Cases/controls, n	7/30	17/23	23/35	58/17		
	Adjusted RR (95% CI)[a]	1 (reference)	1.41 (0.26-7.80)	0.74 (0.11-4.95)	4.61 (0.67-31.76)	0.1	
Spain	Cases/controls, n	40/77	46/69	55/68	126/53		
	Adjusted RR (95% CI)[a]	1 (reference)	1.14 (0.54-2.42)	0.80 (0.39-1.67)	1.17 (0.57-2.39)	0.7	
Italy	Cases/controls, n	26/75	41/70	73/81	129/43		
	Adjusted RR (95% CI)[a]	1 (reference)	0.99 (0.39-2.49)	1.09 (0.43-2.77)	1.27 (0.46-3.58)	0.7	
The United Kingdom	Cases/controls, n	21/55	31/67	77/203	313/203		
	Adjusted RR (95% CI)[a]	1 (reference)	1.83 (0.71-4.68)	0.87 (0.38-1.99)	0.84 (0.38-1.85)	0.3	
Germany	Cases/controls, n	65/211	101/185	168/127	335/127		
	Adjusted RR (95% CI)[a]	1 (reference)	0.93 (0.54-1.58)	1.26 (0.74-2.16)	1.25 (0.71-2.21)	0.3	
Sweden	Cases/controls, n	69/195	168/301	276/295	529/251		
	Adjusted RR (95% CI)[a]	1 (reference)	1.45 (0.94-2.23)	1.55 (0.99-2.39)	1.46 (0.92-2.33)	0.4	0.5

[a] Model adjusted for age at recruitment, body mass index, and total PSA (spline)

[b] A categorical variable was replaced with a continuous variable equal to median concentration by fourth of plasma HK2

[c] Test for heterogeneity in the trends by country using likelihood ratio test

Chapter 7

Genetic polymorphism (rs4988235) in the lactase gene, prostate cancer risk in the PRACTICAL consortium and the intake of dairy products in the UK Biobank

7.1 Introduction

The intake of protein from dairy sources has been associated with an increased risk of prostate cancer [18, 98, 312, 313], which may act via insulin-like growth factor I [231] or calcium intake [18]. A germline polymorphism, rs4988235, strongly associated with both the intake of dairy produce [55] and the ability to digest latctose sugar [314] may be a risk factor for prostate cancer. rs4988235 may provide an opportunity to explore a Mendelian randomization approach to investigate the association between the intake of dairy produce consumption and prostate cancer risk [279].

7.1.1 Lactase persistence, lactose intolerance, and rs4988235

The inability to digest the lactose sugar in many dairy products, commonly referred to as lactose intolerance, depends on lactase enzyme (lactase-phlorizin hydrolase) activity in the intestinal wall. The symptoms of lactose intolerance, such as diarrhea, abdominal pain, or flatulence [315], result from bacterial fermentation of undigested lactose in the colon and are believed to result in the avoidance of dairy products among affected individuals. The predominant genetic polymorphism previously associated with lactase enzyme activity in adult European populations is a C/T single nucleotide polymorphism (SNP) on chromosome 2 (rs4988235), upstream of the lactase coding sequence in the MCM6 gene [54]. In adults, the ancestral C allele causes the down-regulation of lactase enzyme activity (lactase non-persistence) and the reduced tolerance for lactose-rich food [316]. In contrast, the T allele conveys a persistent ability to digest and absorb lactase throughout life [314]. rs4988235 is believed to act by increasing promoter [317, 318, 319] and enhancer activity [320] in the LCT gene.

Notably, however, there is evidence that the lactase enzyme is still produced by homozygous C individuals, albeit at approximately one quarter of levels in homozygous T individuals [321]. Further, rs4988235 has been shown to have a low positive predictive value (20%) for clinical lactose intolerance as defined by the gold standard, hydrogen-methane breath test [322], and symptoms of lactose intolerance have been previously reported as equally prevalent among homozygous C and heterozygous CT individuals [323]. As such, although there may be strong functional evidence to suggest that rs4988235

causes differences in the production of the lactase enzyme [321, 324, 320], it may not fully define the ability to tolerate lactose products.

7.1.2 The intake of dairy produce with rs4988235

The lactase persistence T allele exists at varying frequencies in different populations [325]; in general, among European populations, the T allele is found at highest frequency among northwestern populations, such as the Swedish ($> 80\%$), and at lower frequency among southeastern populations, such as the Greek ($< 10\%$). The differences in T allele frequency are believed to have resulted from recent selective sweeps on genetic variation, probably due to the selective advantage conferred by dairy consumption in the context of the spread of dairy farming [55, 326]. Indeed, there is some evidence of this from ecological research, which has found that, on average, countries with a high frequency of the lactase persistence T allele tend to consume a greater quantity of dairy produce (kg per year per capita) [327].

Due to evidence of positive selection for the lactase persistence allele, research into rs4988235 has focused primarily on its association with the intake of dairy produce. A significantly higher percentage of non-consumers of dairy produce are found among homozygous C individuals compared to homozygous T (18% vs. 12%) [55]. Differences in the consumption of dairy produce by rs4988235 genotype appear to be due principally to differences in the consumption of dairy milk and milk-based beverages. As much as a 50% increase in the consumption of milk and milk beverages has been observed in men homozygous for the T allele compared to men homozygous for the C allele in European populations [55, 56, 57, 58, 59]. However, no such difference was observed for the consumption other dairy produce such as yogurt, cheese, or ice cream by rs4988235 [55, 56]. Given the low lactose content in many dairy products, this may not be a surprising result. Aside from some select processed cheeses, such as cheese slices that may contain 5-7g per serving, the average lactose content for cheese is negligible at 0.1g per serving. Yogurt and ice cream may contain slightly higher amounts of lactose and have ranges between 3-5g per serving. However, dairy milk has, by far, the highest lactose content of dairy products at 6-13g per serving, and may be as high as 52g per serving for some skimmed dried or fortified milk products [328]. There is also some evidence that the lactase persistence T

allele, in particular, is associated with a higher intake of calcium from milk (244 vs 207, mg/day) [329].

7.1.3 Other factors associated with rs4988235

A recent systematic review suggests that rs4988235 is also associated with with a small increase in BMI and a greater likelihood of being over-weight or obese; compared to homozygous C or heterozygote individuals, homozygous T individuals were, on average, 0.17 (0.07 to 0.27) BMI units heavier, and were 9% (2% to 17%) more likely to be over-weight or obese [57]. Additionally, there was evidence to suggest that, compared to homozygous C individuals, homozygous T individuals may be associated with: a 8% (6% to 10%) increased risk of ischemic heart disease; have 0.056 (0.048 to 0.064) mmol/l lower LDL cholesterol; and 0.028 (0.028 to 0.036) mmol/l lower HDL cholesterol [330]. Additionally, compared to homozygous C individuals, homozygous T individuals have been found to be more likely to be educated to secondary school or higher [55].

7.1.4 rs4899235 and prostate cancer

A recent meta-analysis of all published data (4,783 cases and 3,188 controls) on the association of rs4988235 with prostate cancer found no significant elevated risk of prostate cancer (OR: 1.12, 95% CI: 0.96-1.32) [55]. However, it remains possible that previous studies were under-powered to detect the modest expected association of rs4988235 with prostate cancer; one estimate suggested 30,000 cases and 30,000 controls would be necessary to detect a 2% increased relative risk with 80% power [55].

7.1.5 Aim of Study

This chapter has two primary questions:

1. Is the lactase persistence SNP, rs4988235, associated with subsequent risk of prostate cancer overall, and does any association vary by tumour characteristics among men in the PRACTICAL consortium.

2. How does the intake of dairy produce vary by rs4988235 in the UK Biobank, the largest cross-sectional study to date, and how do these findings aid the interpretation of any association of rs4988235 with prostate cancer risk.

7.2 Methods

7.2.1 Study Population

These analyses use data from the PRACTICAL consortium and the UK Biobank. For complete details on the studies that constitute the PRACTICAL consortium and the UK Biobank cohort, please see Chapter 3.

7.2.2 Follow-up for cancer incidence and vital status in the PRACTICAL consortium

Cases were defined as men who were diagnosed with prostate cancer (International Classification of Diseases 10th revision code C61 [173]). For the current analyses participants were 48,471 cases with 29,866 controls. Information for advanced stage disease (T3-T4 and/or N1-N3 and/or M1) was available for 4,927 men. Information for high-grade disease (Gleason score $\geq$8) was available for 5,940 men. In addition, 3,876 men were known to have died of prostate cancer as an underlying cause.

7.2.3 Genotyping and Imputation

PRACTICAL

Genotyping for study samples in PRACTICAL was done using a custom Illumina array as part of the OncoArray project, which is a collaborative project on the risk of breast, ovarian, prostate, colorectal and lung cancer [37]. The array, of approximately 600,000 markers, consisted of a GWAS backbone (approximately 260,660 markers in the Illumina HumanCore) and markers selected from the disease consortia representing the main cancer sites involved in the project.

All analysis centres genotyped a common set of HapMap samples so that strand alignment and integrity of imputation could be compared. Principal components were derived based on a subset of 2,318 markers identified as informative for ancestry. The first two principal components were then used to estimate the proportion of ancestry relative to the European, and Asian HapMap populations. The cut-offs used to group samples by ancestry were: European: > 0.8 European ancestry and Asian: > 0.4 Asian ancestry.

Within and between study duplicates, and close relatives were identified, and excluded, by calculating a concordance matrix for all individuals; samples

with concordance > 0.86 were flagged as duplicates and concordance between 0.74 and 0.86 were flagged as relatives. Samples passed genotyping quality control steps if more than 95% of SNPs had valid calls. After additional manual review of the cluster plots for SNPs failing to achieve 95% call rates, a total of 494,763 SNPs were retained. Overall, 97% of samples had call rates of 95% or higher. To adjust for potential (intra-continental) population stratification in main analyses, principal components analysis was performed using data from 33,661 uncorrelated SNPs (which included the 2,318 SNPs for continental ancestry) with a major allele frequency of at least 0.05 and maximum correlation of 0.1 in OncoArray.

Directly genotyped rs4988235 was not available as it was was excluded at quality control stages for OncoArray data. All current analyses used rs4988235 allele dosage from two-stage imputation based on the 1000 genomes release 3 reference panel, and calculated using SHAPEIT ($http://www/shapeit.fr/$) and IMPUTE.V2 ($https://mathgen.stats.ok.ac.uk/impute/impute\_v2.html$). SHAPEIT was used to derive phased genotypes using the default parameters. IMPUTE.V2 was then used to perform imputation using 5 megabase non-overlapping intervals for the whole genome - variants whose minor allele frequency was less than 0.001 were excluded from imputation. A high quality of imputation for rs4988235 was observed, $r^2 = 0.84$.

UK Biobank

Genotyping was performed by the UK Biobank, and genotyping, quality control and imputation procedures are described in detail elsewhere [331]. In brief, DNA was extracted from buffy coat samples collected from all participants. Participant DNA was genotyped on two arrays: UK BiLEVE and UKB Axiom, which had more than 95% common content, and approximately 800,000 SNPs. Genotype was called using the Affymetrix Power Tools software in 33 batches of roughly 4,700 samples. Samples with high missingness or heterozygosity (480), short runs of homozygosity (8), related individuals (1,856), and sex mismatches (191), were removed. Genotypes for 152,736 samples passed sample quality control (approximately 99.9% of total samples). SNPs were excluded if they did not pass quality control filters across all 33 genotyping batches, or had missingness greater than 90%. Batch effects were identified through frequency and HardyWeinberg equilibrium (HWE) tests ($P < 10^{-6}$). In total, 806,466 SNPs passed quality control in at least

one batch (approximately 99% of the array content). Population structure was modelled using principal component analysis in a subset of high quality samples with low missingness (< 1.5%) and high frequency SNPs (> 2.5%, 100,000 SNPs) of European descent.

7.2.4 Calculation of the intake of dairy milk in the UK Biobank

New variables were created for the weight (in grams) from the 24-hour dietary assessments for dairy milk, cheese, yogurt, and ice cream. This was done by using the steering file of the Oxford WebQ which contains the serving size of each food item listed in the 24-hour dietary assessment, in grams. To estimate daily intake, the serving size in grams was multiplied by the frequency reported in the 24-hour dietary assessment. The top frequency category was open ended and differed by food group, these were coded so that 3+ = 3, 4+=4, 5+=5, 6+ = 6. Less than one was coded as 0.5 (Table 8.1).

Information on never consumers of milk, milk type consumed, cheese intake, and spread used were collected from the baseline touchscreen questionnaire, each of which is now described in turn. Never consumers of milk was a binary variable where never consumers were coded as 1 and ever consumers were coded as 0. Milk type consumed was used to create two binary variables: the first variable was coded as 1 if a man elected to consume dairy milk and 0 if a man elected to consume soy milk, or other (non-dairy) types of milk, or never/rarely consuming milk; the second variables was coded as 1 if a man elected to consume dairy milk and 0 if a man elected to consume soy milk, or never/rarely consuming milk. Cheese intake was a binary variable where never consumers of cheese were coded as 1 and ever consumers of cheese were coded as 0. Spread used was a binary variable where never/rarely uses spread were coded as 1 and ever uses butter spread were coded as 0.

7.2.5 Statistical Analyses

PRACTICAL

Individuals not of genetically European ancestry, and studies with less than 10 cases and 10 controls were excluded. Where data were available, for a selection of epidemiological factors, means and 95% confidence intervals (CI)

were calculated for continuous variables, and N and percentages presented for categorical variables by hard-called rs4988235 genotype from available imputed allele dosage. Imputed allele dosage is a probabilistic score that ranges from 0 to 2, where 0, 1, and 2 indicate complete certainty in the homozygous (CC), heterozygous (CT), and homozygous (TT) genotypes, respectively. For the purpose of cross-tabulation with common epidemiological factors, however, we have hard-called genotypes by collapsing across genotype uncertainty to create a rs4988235 genotype variable (CC: 0-0.4, CT: 0.8-1.2, TT: 1.6-2). Hard-called rs4988235 genotype were also used to calculate T allele frequency overall, and range by study, for cases and controls by prostate cancer endpoint:

$$Freq_T = \frac{2N_{TT} + N_{CT}}{2N_{Total}} \tag{7.1}$$

Poisson regression, adjusted for principal components, was used to estimate the relative risk (95%CI) of having family members with a history of prostate cancer associated with continuous rs4988235 allele dosage within each study; study-level estimates were then pooled by the inverse-variance weighted method.

Unconditional logistic regression was used to estimate the association of rs4988235 allele dosage with prostate cancer risk overall, for high grade and advanced stage tumours, and for death from prostate cancer. Additional analyses were conducted in prospective studies only to assess the impact of potential differences, on risk estimates, in the intake of dairy products between cases and controls recruited at different time points for retrospective case-control studies.

All analyses were study-specific, and adjusted for population stratification using the first seven principal components. Study-level estimates were then pooled by the inverse-variance weighted method to derive an overall measure of the association of rs4988235 with prostate cancer within the PRACTICAL consortium.

UK Biobank

Analyses were only in men of European ancestry, and, due to low quality imputation of rs4988235 in the UK Biobank cohort, only directly genotyped data were used.

The percentage of: non-consumers for dairy milk, ice cream, yogurt, and cheese from the 24-hour recall questionnaires; and non-consumers of milk and consumers of soy milk, non-consumers of cheese, and butter-based spreads, from the baseline touchscreen questionnaire were derived for each rs4988235 genotype by taking adjusted proportions from a logistic regression holding other parameters at the mean. Estimates for the difference in percentages between CC and TT genotypes were derived by bootstrapped difference between marginal estimates. Confidence intervals were calculated by the Efron bias-corrected and accelerated method to correct for possible bias and skewness in the bootstrapped samples [332].

The intake of dairy produce in the UK Biobank from the 24-hour recall questionnaire are pseudo-continuous variables; this is due largely to the format of data and the frequency of consumption, and affects less frequently consumed dairy produce, such as yogurt and ice cream, most. A consequence of the nature of these dietary intake variables is that they have a low number of unique values and extreme positive skews. These qualities result in non-normally distributed, non-constant variance when deriving mean estimated intake of dairy produce by rs4988235 genotype due to the extreme violation of homoscedasticity. This results in unreliable estimates for standard errors, and in the case of the more extreme violations, such as for ice cream or yogurt, may also result in unreliable estimates for mean intake [333]. A general solution may be to bootstrap a common least squares estimate to derive more reliable standard errors. However, boostrap resampling can be sensitive to long tailed error distributions, in this case due to extreme outliers, and may still produce biased standard errors.

As such, current analyses of the intake of dairy produce by rs4988235 genotype will be by bootstrapped robust regression, which uses an iteratively reweighted least squares algorithm to weight observations as an inverse function of their residual error defined by Huber's M-estimator [334]. Estimates for the difference in percentages between CC and TT genotypes were derived by bootstrapped difference between marginal estimates. Confidence intervals were calculated by the Efron bias-corrected and accelerated method to correct for possible bias and skewness in the bootstrapped samples [332].

Analyses for the intake of dairy produce from 24-hour recall questionnaires were conducted both for intake of dairy produce overall, and for intake of dairy produce excluding non-consumers. Note, however, that I was unable

to compute confidence intervals for the intake of ice cream or yogurt overall; due to the large proportion of zeros (> 50%), the Huber's M-estimator cannot converge on a non-zero estimate. Instead arithmetic means are presented alone and no difference is estimated.

All aforementioned cross-sectional analyses were conducted using two models: an initial model stratified by region within UK Biobank; and a model that stratified by region but additionally adjusted for the first 10 principal components conducted (as described above). Further, it is common methodological practice in genome-wide association studies (GWAS) to perform an inverse-variance rank normal transformation on continuous data to allow data to conform to model assumptions. As I present results from a GWAS for dairy milk in the following chapter, results for dairy milk are also presented after inverse-variance rank normal transformation. All bootstrapping was for 1000 repetitions.

All statistical tests are two-sided and were conducted using STATA software version 14 (College Station, TX: StataCorp LP).

7.3 Results

PRACTICAL

Among controls, there was little evidence of differences by rs4988235 genotype for age at recruitment, weight, height, or body mass index. Compared to homozygous C individuals, homozygous T individuals were slightly less likely to be current smokers (23.8% vs 27.6%), and slightly more likely to have a family history of prostate cancer (15.9% vs 13.0%) (Table 7.2). However, Poisson regression, adjusted for principal components, found no association of rs4988235 allele dosage with number of family members with prostate cancer (RR: 0.98; [95% CI 0.97-1.01). As with previous studies, we observed a strong south-east to north-west cline for T allele frequency within studies from Europe, ranging from 8% in Greece to 78% in Sweden (Figure 7.1). High T allele frequencies were observed for non-European studies: USA (64%), Canada (56%), and Australia (68%). We also observed a modest difference in the T allele frequency between controls (0.67) and prostate cancer cases overall (0.69), high grade (0.69) and advanced tumours (0.70), and for death from prostate cancer (0.70) (see Table 7.3).

A pooled estimate from all included studies found no significant association of rs4988235 allele dosage with prostate cancer overall (OR: 1.01; [95% CI 0.99-1.04]), high grade or advanced stage tumours (OR: 0.99; [95% CI 0.95-1.04] and OR: 0.99; [95% CI 0.93-1.04], respectively), or death from prostate cancer (OR: 0.96; [95% CI 0.90-1.03]). After adjustment for principal components, results were materially unchanged for all endpoints: prostate cancer overall (OR: 1.01; [95% CI 0.98-1.04], high grade and advanced stage tumours (OR: 1.00; [95% CI 0.95-1.05] and OR: 0.98; [95% CI 0.92-1.04], respectively), and death from prostate cancer (OR: 0.96; [95% CI 0.90-1.03]. Additional analyses that included only data from prospective studies, adjusting for principal components, also found no significant association of rs4988235 allele dosage with prostate cancer overall (OR: 1.04; [95% CI 0.99-1.08], high grade or advanced stage tumours (OR: 1.02; [95% CI 0.92-1.14] and OR: 0.95; [95% CI 0.82-1.10], respectively), or death from prostate cancer (OR: 0.97; 95% CI 0.85-1.12]).

UK Biobank

No difference was observed in percentage of non-consumers of dairy milk, ice cream, yogurt, or cheese between CC and TT homozygotes after stratification by region (difference: 1.1% [95% CI: -0.4-1.9], 0.08% [95% CI: -1.0-2.3], 2.0% [95% CI: -1.0-4.5], and 2.7% [95% CI: -0.2-5.5], respectively). Results were materially unaltered after further adjustment for principal components (difference: 0.1% [95% CI: -1.0-1.0], 1.0% [95% CI: -1.0-2.7], 1.7% [95% CI:-1.0-3.7], and 0.02% [95% CI:-0.6-1.4], respectively) (Table 7.4 & Table 7.5).

A modest difference in the intake of dairy milk was observed between TT and CC genotypes (TT: 217.7 g/d, [95% CI 214.4-220.6] vs. CC: 192.2 g/d, [95% CI 186.2-198.3], difference: 25.5 g/d [95% CI 18.8-32.2]) for a model stratified by region. After additional adjustment for principal components an approximately 20% smaller difference was observed in the intake of dairy milk by rs4988235 genotype (TT: 217.0 g/d, [95% CI 213.9-220.1] vs. CC: 198.6 g/d, [95% CI 191.4-205.7], difference: 19.9 g/d [95% CI 12.8-26.9]). When analyses were restricted to only current consumers, the magnitude of the difference in the intake of dairy milk by rs4988235 was not altered for either the model stratified by region or stratified by region and adjusted for principal components. However, the removal of non-consumers did increase estimated mean dairy milk intake by approximately 15 g/d. Results were

materially similar for the inverse variance rank normal transformed variable for dairy milk. No differences were observed by rs4988235 genotype for ice cream, yogurt, or cheese intake for stratified or adjusted models with or without the inclusion of non-consumers (Table 7.4 & Table 7.5).

For a model stratified by region, a modest difference in the percentage of men consuming soy milk or other (non-dairy) types of milk, or never/rarely consuming milks in general was observed among homozygous C men compared to homozygous T (TT: 6.3%, [95% CI 6.0-6.7] vs. CC: 8.3%, [95% CI 7.3-9.3], difference: 2.0% [95% CI 0.09-2.4]). Results for this model were not materially altered when analyses were restricted to exclude consumers of other (non-dairy) types of milk (TT: 5.3%, [95% CI 5.1-5.6] vs. CC: 7.1%, [95% CI 6.1-7.9], difference: 1.8% [95% CI 0.08-2.6]). After additional adjustment for principal components these results were attenuated by approximately 40%; the percentage of men consuming soy milk or other (non-dairy) types of milk, or never/rarely consuming milks in general remained higher among C homozygotes compared to T homozygotes (TT: 6.4%, [95% CI 6.1-6.8] vs. CC: 7.6%, [95% CI 6.6-8.6], difference: 1.2% [95% CI 0.04-2.4]), and similar results were observed when models were restricted to exclude consumers of other (non-dairy) types of milk (TT: 5.5%, [95% CI 5.2-5.8] vs. CC: 6.4%, [95% CI 5.5-7.3], difference: 0.9% [95% CI 0.03-2.1]) (Table **??**).

After stratification by region, men homozygous for the lactase intolerance allele (C) were more likely to be never consumers of dairy milk than men homozygous for the T allele (TT: 2.0%, [95% CI 1.8-2.2] vs. CC: 2.6%, [95% CI 2.0-3.2], difference: 0.6% [95% CI 0.03-0.8]); however, after further adjustment for principal components the confidence in this difference was moderately attenuated (TT: 2.0%, [95% CI 1.8-2.2] vs. CC: 2.6%, [95% CI 2.0-3.2], difference: 0.6% [95% CI -0.01-0.7]).

Further, after stratification by region, C homozygotes were approximately 4% more likely to be never consumers of butter spreads than T homozygotes (TT: 18.1%, [95% CI 17.4-18.9] vs. CC: 22.1%, [95% CI 19.9-24.3], difference: 4.0% [95% CI 1.0-5.6]); results were unchanged after additional adjustment for principal components (TT: 18.1%, [95% CI 17.4-18.9] vs. CC: 22.1%, [95% CI 19.9-24.3], difference: 4.0% [95% CI 1.0-5.5]). No differences were observed by rs4988235 for the intake of cheese from the baseline touchscreen questionnaire (Table 7.6).

7.4 Discussion

In the PRACTICAL prostate cancer genetics consortium rs4988235 T allele dosage was not associated with prostate cancer risk overall, high grade or advanced stage tumours, or death from prostate cancer. When analyses were restricted to data from prospective cohort studies the results were materially unchanged. Subsequent analyses, in the UK Biobank, suggested that rs4988235 may only be associated with a modest, 25-19 g/d, difference in the intake of dairy milk, and may not be associated with the intake of other dairy produce, such as ice cream, yogurt, or cheese. The null association with prostate cancer risk may thus be due to a possibly weak association of rs4988235 with the intake of dairy. However, the null association with ice cream is in contrast to a previous study of rs4988235 with dairy intake, which found significantly higher ice crease consumption among individuals homozygous for the lactase persistence T allele [335].

A summary of existing evidence on the association of rs4988235 with prostate cancer risk from a meta-analysis of the two previous studies to date did not support rs4988235 as a risk factor for prostate cancer (OR: 1.12; [95% CI 0.96-1.32])[55]. However, Travis et al. (2013) [55] noted that analyses may have been under-powered. Previous research suggested that a 35 g/day increased intake of protein from dairy may result in an approximately 32% increased risk for prostate cancer. Further, the difference in the intake of protein from dairy between homozygous C and and homozygous T individuals has been estimated as approximately 2.2 g/day [55]. As such, an association of rs4988235 that occurs via the intake of dairy products may be expected to be 2% for T compared to C homozygotes. To detect such a relative risk, with 80% power, Travis et al. (2013)[55] suggest that 30,000 cases and 30,000 may be needed. As such, the present null association of rs4988235 with prostate cancer risk, with 48,471 cases and 29,866 controls, may be seen as a robust investigation of the association for rs4988235 with prostate cancer risk overall. However, it is possible that there were differences in both the frequency of rs4988235 T allele frequency and the intake of dairy milk by cases status for case-control studies in PRACTICAL that recruited controls separately from cases. Given there is evidence that the intake of dairy products has differed over time [60], it is possible that such

differences in the intake of dairy products by case status led to an attenuation of the association of rs4998235 with prostate cancer risk as a marker for dairy intake. However, we do not believe this to have been the case as when analyses were restricted to only prospective studies conclusions were unchanged.

Previous studies [55, 56] did not investigate the association of rs4988235 with prostate cancer by tumour subtype or for death from prostate cancer. The current study finds little evidence to suggest rs4988235 is associated with an increased risk of high grade or advanced stage prostate cancer, or death from prostate cancer. Given we believe that any association of rs4988235 with prostate cancer acts via the intake of dairy produce, these results are not in opposition with a recent non-significant meta-analysis of dairy produce with both advanced prostate cancer (RR: 0.92; [95% CI 0.79-1.08]) and death from prostate cancer (RR: 1.11; [95% CI 0.97-1.27]) [18].

Current results may also stem from differences in the association of rs4988235 with various dairy produce, or from difficulties associated with the stable measurement of dairy produce from questionnaires. Firstly, the association of rs4988235 with dairy products appears to be largely driven by the association of rs4988235 with dairy milk [55] rather than with dairy protein or other dairy products, which likely results from the relatively low lactose content of other dairy products [328]; if any potential association of dairy products with prostate cancer is, in fact, due to differences in the intake of dairy protein, we may not expect rs4988235 to be associated with prostate cancer risk. Secondly, while all studies find that rs4988235 is positively associated with the intake of dairy milk, estimates have been shown to differ by as much as two-fold (g/d), which may indicate measurement error; although a previous study in the EPIC cohort [55] found rs4988235 accounts for approximately a 33 g/d difference in the intake of dairy milk between C and T homozygotes and a recent meta-analysis suggests the difference may be as low as 24 g/day [330], a study of genetically European Brazilians finds milk intake by rs4988235 genotype may be as large as 67 g/day[57]. Results from this study, which suggest rs4988235 is associated with a 19-25 g/d difference in milk consumption are in agreement with results from the previous and largest study by Bergholdt et al (2015) [336] that estimates the difference in milk intake as a modest 24 g/d.

Further, this study finds only a small difference in the percentage of never consumers of milk by rs4988235 genotype, which was not significant after adjustment for principal components, and only weak evidence that men homozygous for the lactase intolerance allele consume a greater percentage of soy milk. As such, it is possible that null results for rs4988235 and prostate cancer risk derive from the potentially small association with the intake of dairy produce, primarily driven by milk consumption.

Moreover, it does not appear that rs4988235 alone defines current ability to tolerate lactose produce. There is evidence that: C homozygotes have low-level lactase enzyme activity [321, 337]; that they consume low levels of a variety of dairy products [55]; that adverse symptoms of lactose intolerance may be similar for homozygous C and heterozygous CT individuals [323]; and that rs4988235 genotype has poor positive predictive value for lactase persistence when tested against a hydrogen breath test [322]. Additionally, the current analyses find a 20% reduction in estimated dairy milk intake by rs4988235 genotype after adjustment for principal components. Stratifying by region likely captures some of the differences that may exist between men recruitment at different locations in Britain in the consumption of dairy produce, and indeed, some of the variation in the rs4988235 T allele frequency. However, stratifying by region alone is not an appropriate way to take full account of population stratification, and to ensure that the association of rs4988235 with dairy products in estimated in an otherwise genetically homogenous population [338]. As such, the observed 20% reduction in the estimated intake of dairy milk after adjustment for principal components likely reflects differences in the ancestry of men within region; one further explanation, however, is that these differences in ancestry may account for a proportion of the intake of dairy milk, which suggests there may be additional germline variation that determines the intake of dairy milk that is, as yet, undiscovered.

Therefore, should a true association exist between dairy products and prostate cancer risk, as rs4988235 appears to predict a modest difference only in the intake of dairy milk, it may not explain a large enough proportion of the behavioral variance of intake of dairy products as a whole, or for protein from dairy specifically, to have a detectable association with prostate cancer risk, even in this large collaborative consortium. This may be of particular concern given many previous prostate cancer risk estimates are for a comparison of

men consuming little or no dairy produce to those who consume often and a lot. In this case we may not expect to observe an association of rs4988235 with prostate cancer risk as it appears to be a proxy for a moderately low consumption compared to a moderately elevated consumption.

One additional limitation of the current study may be the diversity of studies within the PRACTICAL consortium. Even among men in populations with high T allele frequency, the consumption of dairy produce may vary by as much as three-fold [327, 330]. As such, cultural and social differences in the consumption of dairy products across the numerous studies from differing countries may have led to unexplained variation in the consumption of dairy even among T homozygotes. This may have masked any potential association of rs4988235 with prostate cancer risk. It remains a limitation of the current study that dietary information was unavailable in PRACTICAL, and that further investigation of this limitation was not possible.

7.5 Conclusion

Ultimately, this large investigation with 48,471 cases and 29,866 controls finds no association of the lactase variant, rs4988235, with prostate cancer. However, previous research, and current analyses in the UK Biobank, suggest both that rs4988235 may account for only a small difference in the intake of dairy milk, and that any estimate is likely subject to substantial measurement error. As such, while these findings may be robust evidence against an association of rs4988235 with prostate cancer, we do not believe they provide evidence against an association of dairy produce and prostate cancer risk.

Future research should explore the potential for alternative genetic variation that predicts the intake of dairy milk and dairy protein from a genome-wide association study of dairy intake. Results from such an analysis may help calculate a polygenic score to be used as an instrument for a future potential MR study, in particular, a score that more strongly predicts the full distribution of dairy produce intake.

Table 7.1: Dietary variables from the Oxford WebQ 24-hour dietary assessment included in each touchscreen food group

Food group	**Dietary variables included in each food group and weightings perserving in grams**
Dairy Milk	Glasses of milk (259g), hot chocolate (260g), milk with black tea (35g), milk with rooibos tea (35g),milk with green tea (35g), milk with herbal tea (35g), milk with other tea (35g), milk with instant coffee (25g), milk with filter coffee (25g), milk with espresso coffee (25g), milk with other coffee (25g), cappuccino (190g), latte (190g), porridge (79g), milk added to cereal (100g), cream sauce (30g), white sauce (54g), pudding (175g), and low calorie hot chocolate (130g)
Cheese	Low fat hard cheese (40g), hard cheese (40g), soft cheese (40g), blue cheese (35g), low fat cheese spread (15g), cheese spread (15g), cottage cheese (60), feta (40g), mozzarella (40g), goat cheese (40g), other cheese (40g), and the cheese disaggregated from cheese sauce using the recipe provided in McCance and Widdowson's The Composition of Food (8) (10g), cheese disaggregated from pizza using standard topping sizes (40g), and cheese disaggregated from cheesecake using standard recipes (55g)
Yogurt	Yogurt (125g), yogurt drink (250g), and yogurt smoothie (125g)
Ice Cream	Ice cream (120g)

Table 7.2: Characteristics of controls by rs4988235 genotype*, and PSA, and free to total PSA ratio at diagnosis for cases within PRACTICAL.

	rs4988235 genotype		
	CC	CT	TT
Controls			
Age at recruitment**, yrs	60.1 (59.8-60.4)	59.9 (59.8-60.2)	59.9 (59.8-60.1)
Weight**, kg	83.4 (82.7-83.9)	84.3 (83.8-84.7)	84.6 (84.3-84.9)
Height**, cm	174.4 (173.9-174.9)	175.7 (175.4-176.0)	176.3 (176.0-176.6)
BMI, kg/m^2**	27.1 (26.9-27.3)	27.6 (26.6-28.5)	26.9 (26.8-27.1)
Smoking Status, n (%)			
Never	579 (34.1)	1,593 (34.7)	2,328 (36.3)
Previous	652 (38.4)	1,729 (37.7)	2,562 (39.9)
Current	469 (27.6)	1,268 (27.6)	1,527 (23.8)
Family History, n (%)			
No	1,640 (87.0)	3,928 (86.9)	4,979 (84.1)
Yes	245 (13.0)	594 (13.1)	941 (15.9)
Cases			
PSA*** at diagnosis, ng/ml**	8.7 (8.4 - 9.1)	9.1 (8.9 - 9.3)	9.0 (8.9 - 9.2)
Free to total PSA ratio**	10.9 (9.8 - 12.3)	10.8 (10.1 - 11.7)	10.8 (10.1 - 11.7)

* T allele frequency requires the categorisation of a dosage allele score that ranges from 0 to 2 into genotypes at cut points of 0-0.4, 0.8-1.2, 1.6-2 for CC, CT & TT.

** Values are arithmetic means and 95% confidence intervals

*** PSA, prostate-specific antigen

Table 7.3: Median T allele frequency [range across studies] for rs4988235 in cases and controls, and case/control allele frequency difference within PRACTICAL

Outcome	**Case / Control***	**Case [Range]**	**Control [Range]**	**Difference**
Total Prostate Cancer	36,698 / 23,751	0.69 [0.30 - 0.86]	0.67 [0.21 - 0.81]	0.02
High Grade	4,976 / 21,846	0.69 [0.32 - 0.83]	0.68 [0.21 - 0.81]	0.01
Advanced Stage	4,087 / 19,252	0.70 [0.29 - 0.82]	0.67 [0.21 - 0.81]	0.03
Death from Prostate Cancer	2,649 / 17,995	0.70 [0.25 - 0.80]	0.68 [0.36 - 0.78]	0.02

* T allele frequency requires the categorisation of a dosage allele score that ranges from 0 to 2 into genotypes at cut points of 0-0.4, 0.8-1.2, 1.6-2 for CC, CT & TT.

Figure 7.1: T allele frequency for rs4988235 for available European studies within the PRACTICAL consortium

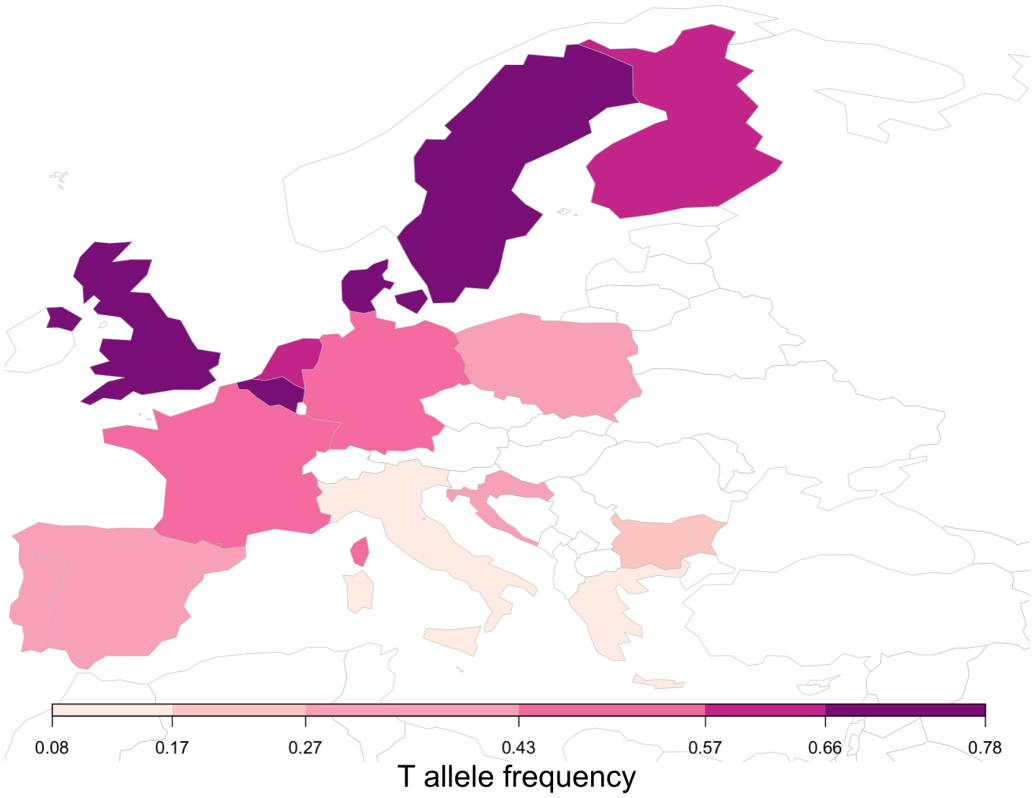


Figure 7.2: Per-allele (T allele) relative risk of prostate cancer overall, high grade and advanced stage, and death from prostate cancer within the PRACTICAL consortium

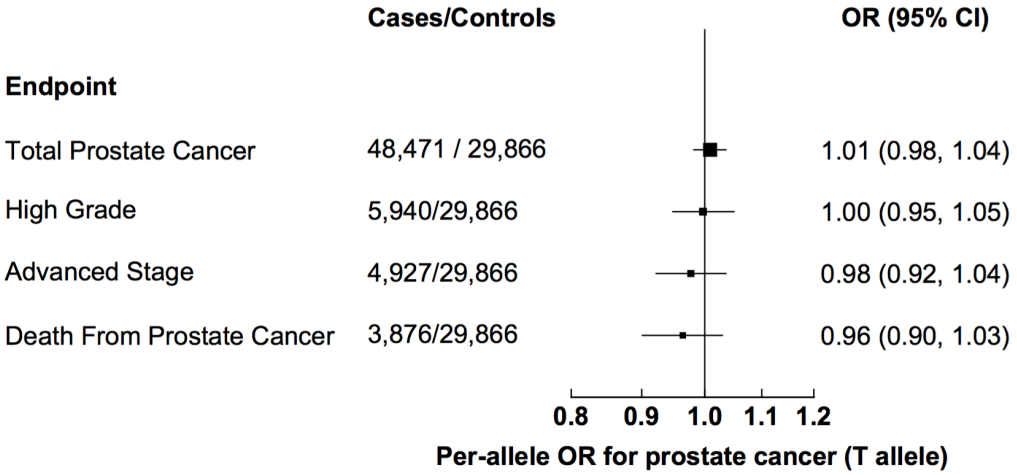


Figure 7.3: Per-allele (T allele) relative risk of prostate cancer overall from prospective studies within the PRACTICAL consortium

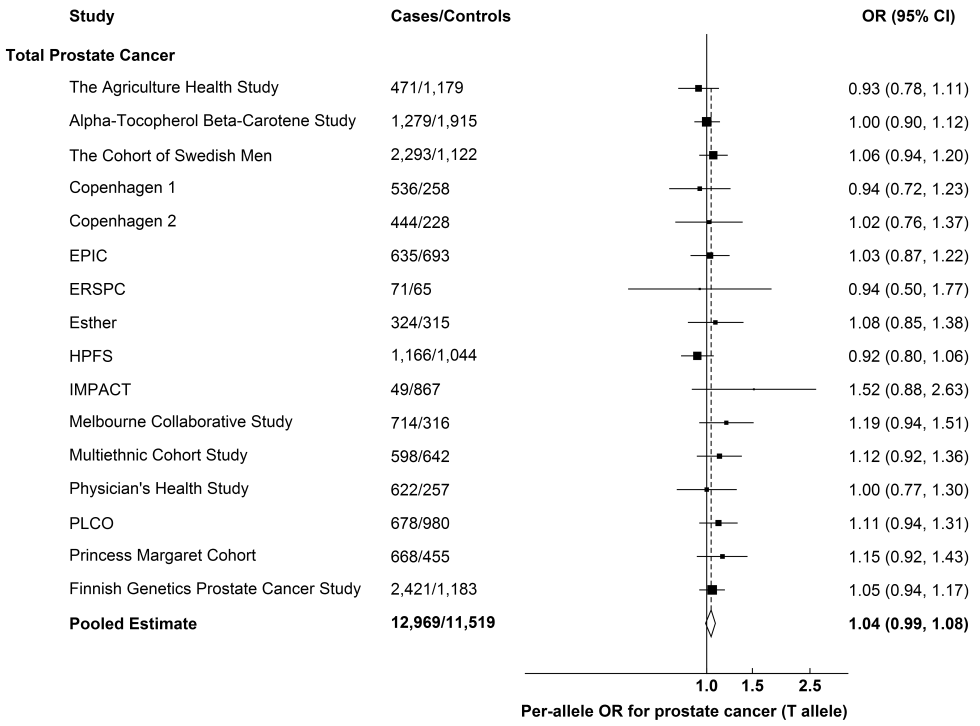


Figure 7.4: Per-allele (T allele) relative risk of high grade prostate cancer within the PRACTICAL consortium

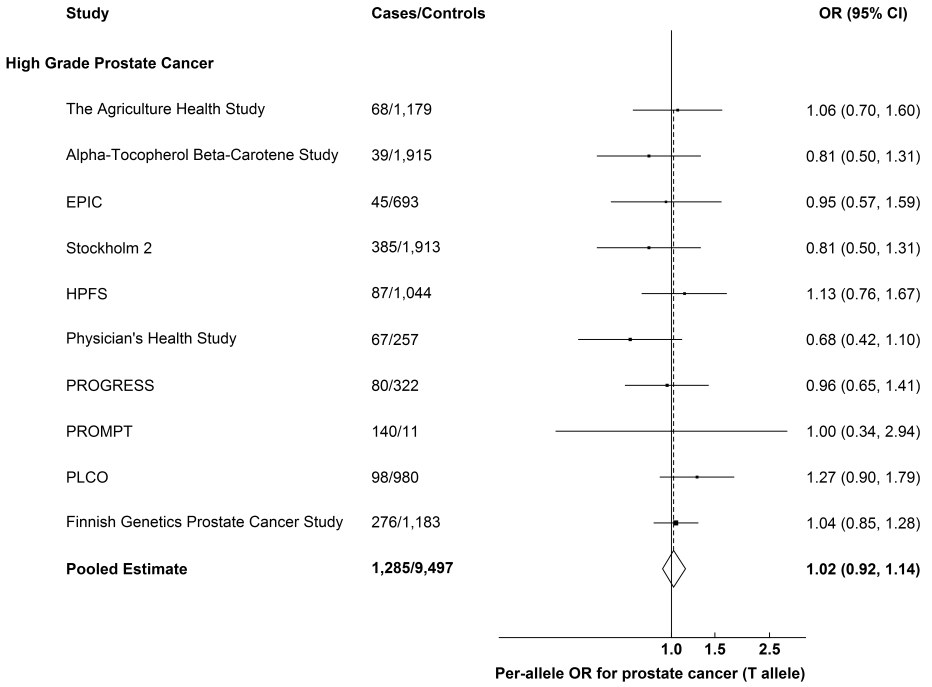


Figure 7.5: Per-allele (T allele) relative risk of advanced stage prostate cancer within the PRACTICAL consortium

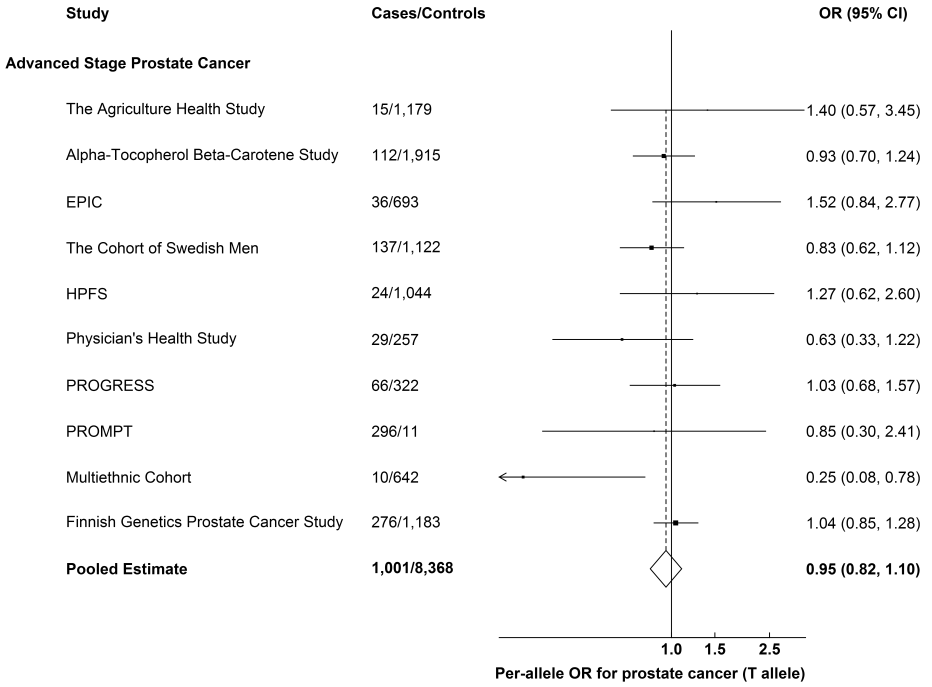


Figure 7.6: Per-allele (T allele) relative risk of death from prostate cancer within the PRACTICAL consortium

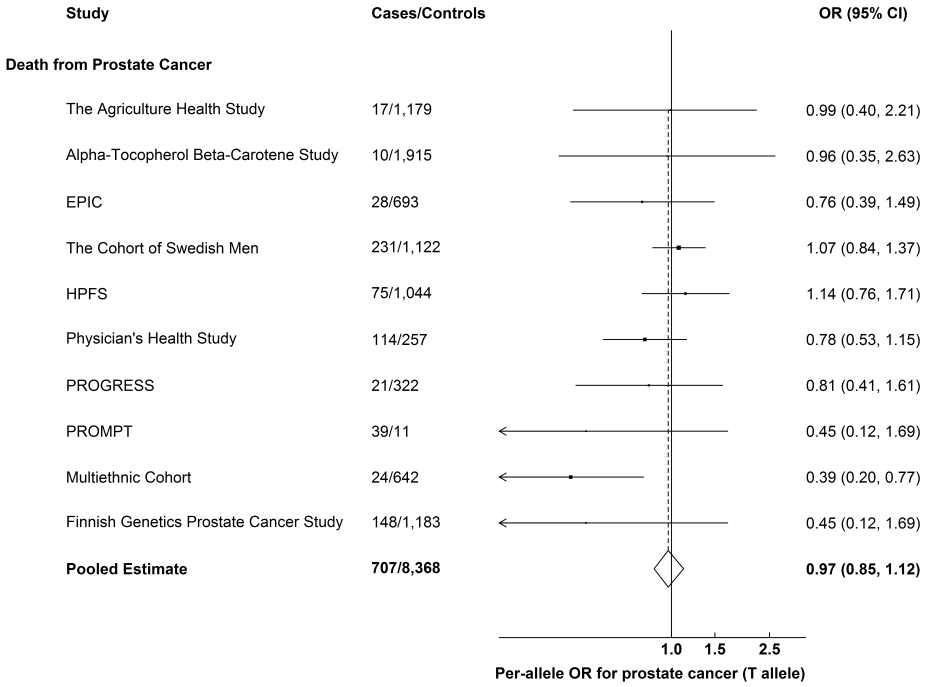


Table 7.4: Mean (95% CI) intake of dairy products by rs4988235 genotype in the UK Biobank

	rs4988235 genotype			
	TT	TC	CC	TT-CC difference
N	9,718	6,574	1,411	
Stratified for region				
Dairy Milk				
% non-consumers,	7.2 (6.7-7.7)	7.0 (6.4-7.6)	8.3 (6.9-9.7)	1.1 (-0.4-1.9)
Intake, g/d, by robust regression	217.7 (214.4-220.6)	215.6 (212.1-219.2)	192.2 (186.2-198.3)	25.5 (18.8-32.2)
Intake in consumers, g/d, by robust regression	233.3 (230.6-236.0)	230.7 (227.0-234.4)	208.9 (202.2-215.6)	24.4 (17.1-31.8)
Dairy Milk (Inverse-variance Rank Normal Transformation)				
Intake, g/d, by robust regression	0.03 (0.01-0.05)	0.02 (-0.01-0.04)	-0.17 (-0.22–0.11)	0.20 (0.14-0.25)
Intake in consumers, g/d, by robust regression	0.04 (0.02-0.06)	0.02 (-0.01-0.05)	-0.17 (-0.23–0.11)	0.21 (0.15-0.27)
Ice Cream				
% non-consumers, %	78.9 (78.2-79.8)	79.0 (78.0-80.0)	78.1 (75.9-80.3)	0.08 (-1.0-2.3)
Intake, g/d, by robust regression	14.4	13.8	13.4	-
Intake in consumers, g/d, by robust regression	60.9 (58.8-63.1)	61.1 (58.9-63.3)	56.3 (51.7-60.9)	4.6 (-0.1-9.3)
Yogurt				
% non-consumers, %	57.1 (56.1-58.1)	57.6 (56.3-58.8)	55.1 (52.5-57.8)	2.0 (-1.0-4.5)
Intake, g/d, by robust regression	45.5	45.8	46.4	-
Intake in consumers, g/d, by robust regression	92.1 (90.3-94.9)	93.7 (91.7-95.7)	88.6 (84.1-93.1)	3.5 (-0.5-7.5)
Cheese				
% non-consumers, %	35.7 (34.8-36.7)	35.6 (34.5-36.8)	33.0 (30.5-35.5)	2.7 (0.2-5.5)
Intake, g/d, by robust regression	18.2 (17.6-18.9)	18.2 (17.5-18.9)	19.0 (17.7-20.3)	0.8 (-0.4-1.9)
Intake in consumers, g/d, by robust regression	29.5 (29.1-29.9)	29.4 (28.9-30.0)	29.3 (27.9-30.5)	0.2 (-0.6-1.1)

Table 7.5: Mean (95% CI) intake of dairy products by rs4988235 genotype in the UK Biobank

	rs4988235 genotype			
	TT	TC	CC	TT-CC difference
N	9,718	6,574	1,411	
Region stratified and adjusted for principal components				
Dairy Milk				
% non-consumers, %	7.3 (6.8-7.9)	7.0 (6.4-7.6)	7.4 (5.9-8.8)	0.1 (-1.0-1.0)
Intake, g/d, by robust regression	217.0 (213.9-220.1)	215.1 (211.6-218.7)	198.6 (191.4-205.7)	19.9 (12.8-26.9)
Intake in consumers, g/d, by robust regression	232.8 (229.5-236.2)	230.1 (226.8-233.5)	213.3 (205.4-221.1)	19.5 (12.5-28.2)
Dairy Milk (Inverse-variance Rank Normal Transformation)				
Intake, g/d, by robust regression	0.03 (0.01-0.05)	0.01 (-0.01-0.04)	-0.12 (-0.017–0.06)	0.15 (0.09-0.21)
Intake in consumers, g/d, by robust regression	0.04 (0.02-0.06)	0.02 (-0.01-0.04)	-0.13 (-0.19–0.07)	0.17 (0.11-0.24)
Ice Cream				
% non-consumers, %	78.9 (78.2-79.8)	79.1 (78.1-80.1)	77.9 (75.5-80.2)	1.0 (-1.0-2.7)
Intake, g/d, by robust regression	14.4	13.8	13.4	-
Intake in consumers, g/d, by robust regression	60.8 (58.9-62.7)	61.2 (58.8-63.5)	57.5 (52.3-62.6)	3.8 (-0.7-8.2)
Yogurt				
% non-consumers, %	56.8 (55.8-57.8)	57.7 (56.6-58.9)	56.1 (53.3-58.8)	1.7 (-1.0-3.7)
Intake, g/d, by robust regression	45.5	45.8	46.4	-
Intake in consumers, g/d, by robust regression	92.7 (90.1-94.4)	94.5 (92.5-96.6)	89.7 (85.5-93.9)	3.0 (-0.9-6.6)
Cheese				
% non-consumers, %	35.4 (34.4-36.3)	35.9 (34.8-37.1)	34.1 (31.4-36.7)	0.02 (-0.6-1.4)
Intake, g/d, by robust regression	18.2 (17.6-18.8)	17.9 (17.2-18.6)	18.2 (17.0-19.5)	0.01 (-1.3-1.5)
Intake in consumers, g/d, by robust regression	29.7 (29.2-30.2)	29.7 (29.0-30.3)	28.9 (27.7-30.3)	0.8 (-0.5-1.6)

Table 7.6: Mean percentage (95% CI) of consumers of dairy-related dietary produce by rs4988235 genotype in the UK Biobank

	rs4988235 genotype			
	TT	TC	CC	TT-CC difference
Region stratified				
Dairy milk				
N	23,057	15,315	3,142	
%	2.0 (1.8-2.2)	2.0 (1.8-2.3)	2.6 (2.0-3.2)	0.6 (0.03-0.8)
Soy milk or never/rarely consuming milks [including other (non-dairy) types of milk]				
N	23,094	15,345	3,148	
%	6.3 (6.0-6.7)	6.6 (6.2-6.9)	8.3 (7.3-9.3)	2.0 (0.09-2.4)
Soy milk or never/rarely consuming milks [excluding other (non-dairy) types of milk]				
N	22,853	15,170	3,106	
%	5.3 (5.1-5.6)	5.5 (5.1-5.8)	7.1 (6.1-7.9)	1.8 (0.08-2.6)
Cheese				
N	22,613	15,014	3,061	
%	2.9 (2.7-3.2)	2.7 (2.4-2.9)	2.6 (1.9-3.1)	0.3 (-0.002-1.0)
Spread				
N	9,803	6,590	1,363	
%	18.1 (17.4-18.9)	19.4 (18.4-20.3)	22.1 (19.9-24.3)	4.0 (1.0-5.6)
Region stratified and adjusted for principal components				
Dairy milk				
N	23,057	15,315	3,142	
%	2.0 (1.8-2.2)	2.0 (1.8-2.3)	2.6 (2.0-3.2)	0.6 (-0.01-0.7)
Soy milk or never/rarely consuming milks [including other (non-dairy) types of milk]				
N	23,094	15,345	3,148	
%	6.4 (6.1-6.8)	6.5 (6.2-6.9)	7.6 (6.6-8.6)	1.2 (0.04-2.4)
Soy milk or never/rarely consuming milks [excluding other (non-dairy) types of milk]				
N	22,853	15,170	3,106	
%	5.5 (5.2-5.8)	5.5 (5.1-5.8)	6.4 (5.5-7.3)	0.9 (0.03-2.1)
Cheese				
N	22,613	15,014	3,061	
%	2.9 (2.7-3.2)	2.7 (2.4-2.9)	2.6 (2.0-3.1)	0.3 (-0.01-0.8)
Spread				
N	9,803	6,590	1,363	
%	18.1 (17.4-18.9)	19.4 (18.4-20.3)	22.1 (19.9-24.3)	4.0 (1.1-5.5)

Chapter 8

Genome-wide association study for the intake of dairy milk in the UK Biobank

8.1 Introduction

This chapter will investigate the association of germline genetic variation with the intake of dairy milk in a European population, with the aim of identifying variants that may contribute to a genetic instrument for use in a Mendelian randomisation study of dairy intake with subsequently developing disease, such as prostate cancer. Such novel methods may aid the investigation of dairy intake with prostate cancer risk given the acknowledged difficulty with measurement error and within person variability when recording dietary intake.

8.1.1 Dairy produce and the intake of dairy milk

Dairy produce intake makes up a moderate proportion of the diet of the British population. The Family Food Survey suggests the British population consumed approximately two litres of milk-based dairy products per person per week in 2015, which is estimated as 11.3% of daily energy intake. Further, it appears that dairy product consumption is a reasonably stable behaviour for this British population. Whole milk intake was only 15 ml greater in 2015 compared to 2012 per person per week. Over the same time-period the consumption of skimmed and semi-skimmed milk products fell by approximately 8%, and the consumption of cheese fluctuated with no clear trend [60].

8.1.2 Factors associated with dairy milk

Genetic

Dairy intake has principally been associated with a SNP on chromosome 2 in the MCM6 gene, rs4988235, which lies upstream of the gene responsible for the production of the lactase enzyme. In adults, carriers of the ancestral C allele have down-regulated lactase enzyme activity, and a reduced tolerance for lactose-rich foods [316], which contrasts the ability, of individuals homozygous for the T allele to digest and absorb lactose throughout life [314]. rs4988235 is believed to act by increasing promoter [317, 318, 319] and enhancer activity [320] in the lactase gene.

A study in EPIC found that a significantly higher percentage of non-consumers of dairy produce are C homozygotes compared to T homozygotes

for rs4988235 (18% vs. 12%) [55]. However, research to date indicates that the differences in the total intake of dairy produce by rs4988235 are largely due to the differences in the consumption of dairy milk; individuals homozygous for the lactase persistence allele (T) consume, on average, between 19 g/d and 67 g/d more dairy milk when compared to individuals homozygous for the allele that does not confer lactase persistence (C) [55, 330, 57].

There is some limited evidence that lactase persistence can also be driven by an alternative polymorphism [339, 340, 341]. However, this evidence comes primarily from non-European populations. Given the analyses in this chapter will focus exclusively on individuals of European ancestry, such genetic variation will not be further discussed.

Non-genetic

Socioeconomic status (SES: education, income, and/or occupation) is likely associated with the intake of dairy. Although there is evidence that various dairy products may be associated differently with SES - whole milk may be more highly consumed in low SES individuals [342, 343, 344] - on balance a recent meta-analysis suggested that high SES individuals were more likely to consume a greater amount of dairy produce overall [345]. Additionally, there is some evidence that high consumers of dairy milk are more likely to be never smokers [346].

8.1.3 The intake of dairy milk in relation to disease risk

Prostate cancer

The intake of dairy milk has previously been associated with prostate cancer risk; whole milk was strongly positively associated with risk of death from prostate cancer in the Physicians' Health Study (HR for 1 serving/d vs non-consumers was 2.17; [95% CI 1.34-3.51], p-trend<0.001) [115]. Further, a recent meta-analysis of dairy products and prostate cancer risk found that high intake of whole and low-fat milk may be associated with an increased risk of prostate cancer in prospective studies (RR per 200 g/d was 1.03; [95% CI 1.00-1.06] and 1.06; [95% CI 1.01-1.11], respectively). However, there is also evidence that the association of milk with prostate cancer risk is non-linear and plateaus at 300-400 g/d, and so it is possible that a continuous linear

estimate is an underestimate of the association of the intake of dairy with prostate cancer risk [18]. The intake of dairy may affect the risk of prostate cancer through its association with increased concentrations of growth factors (such as IGF-I) [231] or through the increased intake of calcium from dairy produce [18].

Other diseases

Dairy milk may have a protective association with overweight and obesity [347], diabetes [348, 349], systolic blood pressure [350], and risk of hypertension [351]. However, subsequent Mendelian randomisation analyses have failed to support these associations [336, 59]. There is also evidence that the intake of dairy milk may be positively associated with bone mass [352], insulin resistance [353], and insulin-like growth factors [354, 355]. Additionally, recent meta-analyses from prospective cohort studies suggest that the intake of dairy milk may have a modest protective association with cardiovascular disease [356] and colorectal cancer [357].

8.1.4 Aim of Study

This chapter will use data from the interim release of UK Biobank genetics data to conduct GWAS for dairy milk intake. Conditional analyses will be conducted for putative variants that associate with the intake of dairy milk to assess their independence from any other SNP associations. SNPs that associate independently with the intake of dairy milk will then be analysed in an independent sample from the UK Biobank to determine their replicability.

8.2 Methods

8.2.1 Study Population

These analyses are in men with available data on the intake of dairy milk and imputed genotype data within the UK Biobank cohort. For the current analysis, individuals of non-white ethnicity were excluded to avoid confounding effects. For full details on the cohort see Chapter 3.

8.2.2 Genotyping and quality control

Of the ∼500,000 individuals in the UK Biobank cohort, ∼153,000 (∼73,000 men) had genetic data available from an interim genetic data release at the time of writing. The discovery GWAS presented in this thesis was conducted on men with available dietary data who were also included in the interim data set (∼22,000 men). As of 27/07/2017 concerns were raised for data from the full imputation for genetic data in UK Biobank by the UK Biobank Access Team in a email sent to researchers; while there was confidence for variants imputed using the Haplotype Reference Consortium (HRC) panel, the imputation was not found to be robust for ∼40 million sites imputed using the UK10K + 1000 Genomes panel. As such, SNPs significantly associated with the intake of dairy milk at $p < 10^{-5}$, and that were imputed in the full UK Biobank genetic data set using the HCR panel, were subsequently tested for replication in men with available dietary and genetic data but that had not been included in the interim genetic data set (∼50,000 men).

Genotyping was performed by the UK Biobank, and genotyping, quality control and imputation procedures are described in detail elsewhere [331]. In brief, DNA was extracted from buffy coat samples collected from all participants. Participant DNA was genotyped on two arrays: UK BiLEVE and UKB Axiom, which had more than 95% common content, and approximately 800,000 SNPs. Genotype was called using the Affymetrix Power Tools software in 33 batches of roughly 4,700 samples each. Samples from both the interim and full data sets with high missingness or heterozygosity, short runs of homozygosity, related individuals (third cousin or closer), and sex mismatches were removed. Further, SNPs were excluded if they did not pass quality control filters across all 33 genotyping batches, or had missingness greater than 90%. Batch effects were identified through frequency and HardyWeinberg equilibrium (HWE) tests ($p < 10^{-12}$). Population structure was modelled using principal component analysis in a subset of high quality samples with low missingness (< 1.5%) and high frequency SNPs (> 2.5%, 100,000 SNPs) of European descent. For the interim data set, imputation of autosomal SNPs was performed using a merged reference panel of the Phase 3 1000 Genome Project and the UK10K using IMPUTE3 [358]. Data was prephased using SHAPEIT3 [359]. In total, ∼73,000,000 SNPs, short indels and large structural variants were imputed. As previously described, imputation for the full genetic data was performed using two reference panels,

the HRC and the UK10K + 1000 Genomes panels, which would amount to ∼80,000,000 SNPs. However, due to imputation concerns for the UK10K + 1000 Genomes panel, ∼40,000,000 SNPs were available from the HRC panel after imputation. Post-imputation quality control was performed as previously outlined and an information score cutoff of 0.1 was applied. For GWAS, we applied a further set of inclusion/exclusion criteria: genetically Caucasian men (inclusion, variable number = 22006), UK Biobank recommended relatedness (exclusion, variable number = 22011) and genomic exclusions (variable number = 22010), HWE $< 10^{-6}$ (exclusion), SNP missing call rate < 0.05 (exclusion), SNP missing rate < 0.05 (exclusion), and minor allele frequency (MAF) ≥ 0.01 (exclusion). After quality control procedures, 22,041 men with dietary and genetic data from the interim data set were used in a discovery GWAS. Further, 50,701 men with dietary and genetic data from the full UK Biobank genetic data set (imputation only from HRC panel) were available for a replication study of the SNPs significant at $p < 10^{-5}$ from the discovery GWAS.

8.2.3 Calculation of the intake of dairy milk

I created new variables for the weight (in grams), from each of the 24-hour dietary recall questionnaires, of the intake of dairy milk. This was done by using the steering file of the Oxford WebQ, which contains the serving size of each food item listed in the 24-hour dietary assessment in grams. To estimate daily intake, the serving size in grams was multiplied by the frequency reported in the 24-hour dietary assessment. The top frequency category was open ended and differed by food group; these were coded so that 3+=3, 4+=4, 5+=5, 6+=6. Less than one was coded as 0.5. Details of the individual dietary variables from the 24-hour dietary assessment that formed each food group are given in Table 8.1. Estimated intake of dairy milk from each questionnaire available for men was then used to generate an average intake of dairy milk using all available dietary data for each man.

8.2.4 Statistical Analyses

Arithmetic mean intake of dairy milk among men within the GWAS and replication data sets was calculated. However, due to the pseudo-continuous nature of dietary data from the 24-hour recall questionnaires, confidence

intervals were calculated by the Efron bias-corrected and accelerated method to correct for possible bias and skewness in the bootstrapped samples [332]. The association of age with the intake of dairy milk was calculated using linear regression.

For all genetic analyses, an inverse rank normal transform was applied to the residuals of a linear regression of age and age squared on the intake of dairy milk to control for possible effect modification by age, and to conform to the assumptions of parametric models used. Genetic association analysis was performed in SNPTEST [360] with the 'expected method' [1] using an additive genetic model adjusted for the first 10 principal components of ancestry, and stratified by genotyping array.

$p < 10^{-8}$ was used to identify robust genetic associations with dairy milk intake. However, due to the suggestion that the genome-wide significant threshold may be too stringent, a lower threshold was also used to discover tentative associations that likely require further investigation ($p < 10^{-5}$) [361]. The genomic inflation factor (λ =median(χ^2)/0.456) and quantile-quantile (Q-Q) plots were used to compare the genome-wide distribution of the test statistic with the expected null distribution, which aids in the identification of unknown familial relationships, a poorly calibrated test statistic, a systemic technical bias, or population stratification. All test statistics were adjusted for genomic inflation ($\chi^2_{adjusted} = \chi^2/\lambda$). In order to distinguish independent GWAS signals, SNPs associated at $p < 10^{-5}$ were subjected to conditional SNP analyses performed using SNPTEST for all variants within a 200kb region around the lead SNP for that locus. Specifically, each SNP was retested, using the model as previously described, for an association with the intake of dairy milk with the lead locus SNP as a covariate to assess the independence of the association of a given SNP with dairy milk intake.

Where possible[2], SNP associations ($p < 10^{-5}$) were included in a replication study using men with available data from the full UK Biobank genetic data and available information on the intake of dairy milk, but who were not included in the previous discovery GWAS. Replication was conducted using SNPTEST using the model as described for the discovery GWAS above. Due to the minor differences expected in the imputation between the interim and

[1]The expected method uses the allele dosage as expected genotype counts to test for an association with a quantitative phenotype. This method provides a good approximation that incorporates genotype uncertainty when estimating modest expected associations.

[2]Replication was only possible for SNPs that were imputed using the HRC panel

full genetic data set, and as a check for the integrity of code used in these analyses, SNPs associated at $p < 10^{-5}$ were also re-analyzed in the full genetic data in men who had previously been included in the discovery GWAS using the interim data.

8.3 Results

The mean intake of dairy milk in the 22,041 UKB individuals who contributed to the discovery GWAS analysis was 232.9 g/d, [95% CI 230-235.1]. This was very similar to the mean intake of dairy milk in men from the replication sample of 55,701 (231.4 g/d, [95% CI 230.3-232.5]). The mean age of participants was 56.9, (95% CI 56.8-57.0) and 56.7, (95% CI 56.6-56.8) for the discovery GWAS and replication samples, respectively. Age at recruitment accounted for 16% of the variation in the intake of dairy milk. Further, the association of age at recruitment with the intake of dairy milk was 0.88, (95% CI 0.63-1.15) and 0.80, (95% CI 0.72-1.01) g/day change in dairy milk intake per year older at recruitment for the discovery GWAS and replication samples, respectively.

No deviation from the null was observed for genome-wide association statistics (λ=1.002), which provided little evidence in favor of population stratification. The Q-Q plot for GWAS analyses showed fewer SNPs significant $p < 10^{-5}$ than might be expected under the null, which may be moderate evidence that the GWAS was underpowered to discover any potential association of germline variation with dairy milk intake (Figure 8.1).

No SNPs were genome-wide significant ($p<5 \times 10^{-8}$)(Figure 8.2). There were, however, 23 loci with SNPs significant at $p < 10^{-5}$, four of which had SNPs borderline genome-wide significant at $p < 10^{-7}$ (Table A.2). There is evidence that 5×10^{-8} may be an overly conservative threshold [361] and the GWAS Q-Q plot suggested that this study may be underpowered. As such, additional analyses were run for all SNPs at these 23 loci, conditioned on the lead SNP at that locus, which suggest that the lead SNP for each locus was the independent SNP driving the association with dairy milk intake.

Due to their borderline significance, the 4 SNPs significant at $p < 10^{-7}$ were of particular interest and were investigated in more detail: rs1316538 (6:130573882, MAF: 0.42) that may be involved in the non-coding Y_RNA

region on chromosome 6, rs62469670 (7:130573882, MAF: 0.06) in an intergenic region on chromosome 7, rs150080038 (2:19919107, MAF: 0.02) that may be involved in nonsense-mediated mRNA decay in the WDR35 region on chromosome 2, and rs4647869 in a promoter flanking region on chromosome 17 (17:76515685, MAF: 0.21). Further, the MR-Base platform provided evidence for phenotypes previously associated with these four SNPs [61]. Only one SNP was associated with phenotypes linked to dairy produce; rs150080038 has been associated with metabolites that include lactate, creatine, and mono-unsaturated fatty acids. rs1316538, rs62469670, and rs4647869 have previously been associated with a variety of risk factors for disease, metabolites, and immune function, such as sleep duration, chronotype, HDL cholesterol, heptanoate, and CD4 immune cells. However, it is not clear that these relate to the intake of dairy milk. Regional association plots for the four SNPs significant at $p < 10^{-7}$ are shown in Figures 8.3, 8.4, 8.5, and 8.6.

Of the 23 SNPs found to associate with dairy milk at $p < 10^{-5}$, 15 were available in the HRC panel used to impute data for the independent replication sample. These 15 SNPs were subsequently regressed against the intake of dairy milk using the same model specification as the discovery GWAS. None of the 15 SNPs met the conventional ($P < 0.05$) significance threshold for an association with the intake of dairy milk. It is unusual that none of the preliminary associations for these 15 SNPs with the intake of dairy milk replicated; to ensure the results were not due to errors in coding or due to the minor differences in imputation methods between the interim data set and the full UK Biobank genetic data, these 15 SNPs were regressed against dairy milk intake in the full Biobank genetic data but for men who were present in the discovery GWAS set. Associations for each of the 15 SNPs were replicated in men from the discovery GWAS using the new data, which suggests that replication results are robust (Table 8.2).

8.4 Discussion

In this study no SNPs were identified as associated with the intake of dairy milk below the genome-wide significance threshold. Of the 23 SNPs that were associated with dairy milk intake at the more tentative $p < 10^{-5}$ threshold in a discovery sample of ∼22,000 men, 15 were also available in a replication

sample of ∼50,000 men. However, none of these putative associations were successfully replicated at even the conventional significance threshold ($P < 0.05$). As such, these analyses have not identified any novel germline variation that is robustly associated with the intake of dairy milk.

The genetic variation most robustly associated with the intake of dairy milk is the lactase persistence SNP, rs4988235 [55, 326]. rs4988235 is also a strong candidate for recent positive selection [362]. This has largely been attributed to the benefit conveyed by the cultural practice of dairy farming in northern Europe that, in conjunction with selective pressure from the environment, led to a change in the frequency of the T allele that facilitates the production of the lactase enzyme, and thus the tolerance of lactose sugars in food, into adulthood [55, 326]. This scenario has been described as gene-culture co-evolution [362]. However, the positive selection on rs4988235 allele frequency may also be due to the benefits of conferred from increased calcium intake from dairy produce for populations in northern climates with low vitamin D exposure from sunlight [363]. It is possible that additional germline polymorphisms associated with the intake of dairy milk have been under similar recent positive selection; estimates suggest as much as 10% of known polymorphisms may have been subject to recent selective pressure [364].

Positive selection has varied effects on both the linkage disequilibrium (LD) structure of a chromosome [365] and Hardy-Weinberg Equilibrium for a given SNP [366] - both may impact the likelihood that variants are included in GWAS analyses. Multiple imputation for genetic data requires a comprehensive measure of the LD structure for a given population. Imputation methods resolve this by calculating imputed genotypes by comparing the expected allele frequencies to an appropriate reference population. However, SNPs under recent positive selection will likely have distinct LD structure as some function of the environment [365] - for example, the difference in the frequency of rs4988235 T allele along the north-south cline due to the differences in dairy farming practices and the strong environmental pressure to consume dairy produce [55]. If the reference panel used for imputation does not adequately account for the LD structure surrounding a SNP that has been under recent positive selection, it is possible that it would not be well imputed. This may have contributed to the poor imputation of rs4988235 in UK Biobank described in chapter 7, and may also have led to other variants

under positive selection being inaccurately imputed or excluded altogether from the data set. Further, where SNPs of interest are robustly imputed it is not likely that SNPs under recent positive selection would be in HWE, and so will be excluded as part of quality control procedures in preparation for a GWAS.

As such, it is possible that the discovery GWAS would not have been able to investigate potential SNPs of interest if they associate with the intake of dairy milk due to recent positive selection and gene-culture co-evolution. Indeed, it is notable that rs4988235 would not have been discovered using this GWAS as it was removed from the analysis due to strongly violated HWE. Given the strong existing evidence that germline polymorphism associated with dairy milk may be candidates for positive selection these issues should be considered in future research. Further, while it is not advisable to use poorly imputed variants for GWAS analyses, it may be advisable to also consider directly genotyped data or to relax HWE assumptions for regions of the genome where there is evidence to suggest gene-culture co-evolution may have occurred with respect to dairy milk intake [362].

The effects of measurement error of dietary exposures are well described in the observational epidemiological literature [32]. However, while GWAS are greatly concerned with the measurement error for genetic variants, there is typically less attention given to the potential role for phenotypic measurement error in these studies [3]. Nonetheless, measurement error of a phenotypic trait reduces the power to detect genetic associations [367]. There are serveral potential ways in which measurement error may have affected the statistical power for these analyses. First, men may have misreported the amount of dairy milk that was consumed during the previous day or the recorded daily consumption may not have been a typical example of dairy milk intake, both of which would likely lead to an attenuation of any potential association of dairy milk intake with germline polymorphism. Second, the measure of dairy milk intake used in this study is the mean intake from the total number of repeat 24-hour recall questionnaires available for a given individual (a maximum of five). There is likely less measurement error for the estimated dairy milk intake from men with a greater number of repeat questionnaires. Due to the vast number of SNPs in the discovery GWAS,

[3] I have only been able to find two research articles that have investigated the role of phenotype measurement error in GWAS [367, 368].

it is not possible to be certain that each allele for each SNPs considered in the discovery GWAS was equally likely to have men with each number of repeat questionnaires. Thus, it is possible that some SNPs' alleles have differentially elevated measurement error, which would lead to potential heteroskedasticity, and make it more difficult to detect the mean difference in dairy milk intake by SNP allele. Third, differences in the potential measurement error described above for dairy milk intake between the discovery and replication samples for GWAS may have contributed to the lack of replication for putative variants significant at $p < 10^{-5}$; a simulation study found that the correlation between allele effect estimates from GWAS can be as low as $r = 0.5$ where measurement or measurement error structure differs between analysis samples, by SNPs analysed [368]. However, the expected concordance between allele effect estimates should approximate the square of the correlation between dairy milk intake in the discovery and replication samples [368], and any effect of such measurement error should, on average, be modest [367]. Given the remarkable similarity of the intake of dairy milk in both samples, it appears unlikely that measurement error alone accounts for the lack of replication for SNPs in this study.

The discovery sample used to generate the GWAS estimates in this study was smaller than initially intended because, due to unforeseen circumstances, the full genetic data for UK Biobank were unavailable; men with available dairy and genetic data from the interim UK genetic data release were used instead. The Q-Q plot for the current GWAS suggested that fewer results at $p < 10^{-5}$ were observed than would be expected under the null hypothesis. As such, although the discovery GWAS constitutes the second largest investigation of germline polymorphisms and dairy milk intake [4], it is possible that these analyses were under-powered to discover germline polymorphisms that determine the intake of dairy milk.

Despite the null results for germline polymorphism and dairy milk intake, this investigation has several strengths. The UK Biobank cohort is large and drawn from an ancestrally similar population [369] from a relatively restricted age range (40-69). The benefit of these characteristics may be demonstrated by the lack of deviation from the null for genome-wide association statistics that indicated little evidence of population stratification. Further, all

[4] A large Danish cohort (∼90,000) analysed the association of rs4988235 with dairy milk intake [336]

individuals resided in the UK at recruitment. Given there are socio-cultural factors that contribute to the amount and frequency of the consumption of dairy milk, a sample of men exposed to similar norms is advantageous to reduce the influence of extraneous cultural differences on analyses. Moreover, all dietary and genetic data were subjected to the same, but separate, quality control procedures, which likely contributed positively to the ability of this study to detect any association of dairy milk intake with genetic variants.

Given the continued importance of genetic instruments for potential MR analyses, future research should re-run the analyses for the association of dairy milk intake once the full UK Biobank genetic data set for all men with dietary data (∼70,000). However, if analyses of dairy milk intake in future also included available data for women in UK Biobank, the discovery sample could be as large as ∼140,000 individuals with at least one 24 hour-recall questionnaire. Such a large discovery sample would substantially improve the power to detect the association of germline polymorphism with the intake of dairy milk. Crucially, however, any associations that may result from such a large discovery GWAS should be also replicated in an independent sample, such as EPIC-Interact [370] or the Fenland study [371]. In particular, the Fenland study has very rich dietary data from food diaries, and so would also likely provide greater detail into how any germline polymorphism may associate with dairy milk subdivided by dairy milk characteristics, such as the fat content of milk.

8.5 Conclusion

No SNPs were genome-wide significant in a discovery GWAS for the intake of dairy milk in the UK Biobank. Putative associations for SNPs significant at the lower tentative genome-wide significance threshold ($p < 10^{-5}$) were not replicated when investigated in an independent sample of men from the UK Biobank. It is possible that imputation or quality control procedures, such as HWE, may have hindered the discovery of SNP-dairy milk associations that resulted from gene-culture co-evolution, and should be investigated in the future. It is also possible that measurement error or insufficient sample size contributed to these null results. Future studies should re-run a discovery GWAS for the intake of dairy milk once the UK Biobank full genetic data set for UK Biobank becomes available.

Table 8.1: Dietary variables from the Oxford WebQ 24-hour dietary assessment included in each touchscreen food group

Food group	**Dietary variables included in each food group and weightings in grams**
Dairy Milk	Glasses of milk (259g), hot chocolate (260g), milk with black tea (35g), milk with rooibos tea (35g), milk with green tea (35g), milk with herbal tea (35g), milk with other tea (35g), milk with instant coffee (25g), milk with filter coffee (25g), milk with espresso coffee (25g), milk with other coffee (25g), cappuccino (190g), latte (190g), porridge (79g), milk added to cereal (100g), cream sauce (30g), white sauce (54g), pudding (175g), and low calorie hot chocolate (130g)

Figure 8.1: Quantile-quantile of dairy milk intake GWAS in UKB. N=22,041 individuals. λ=1.002 for GWAS analyses

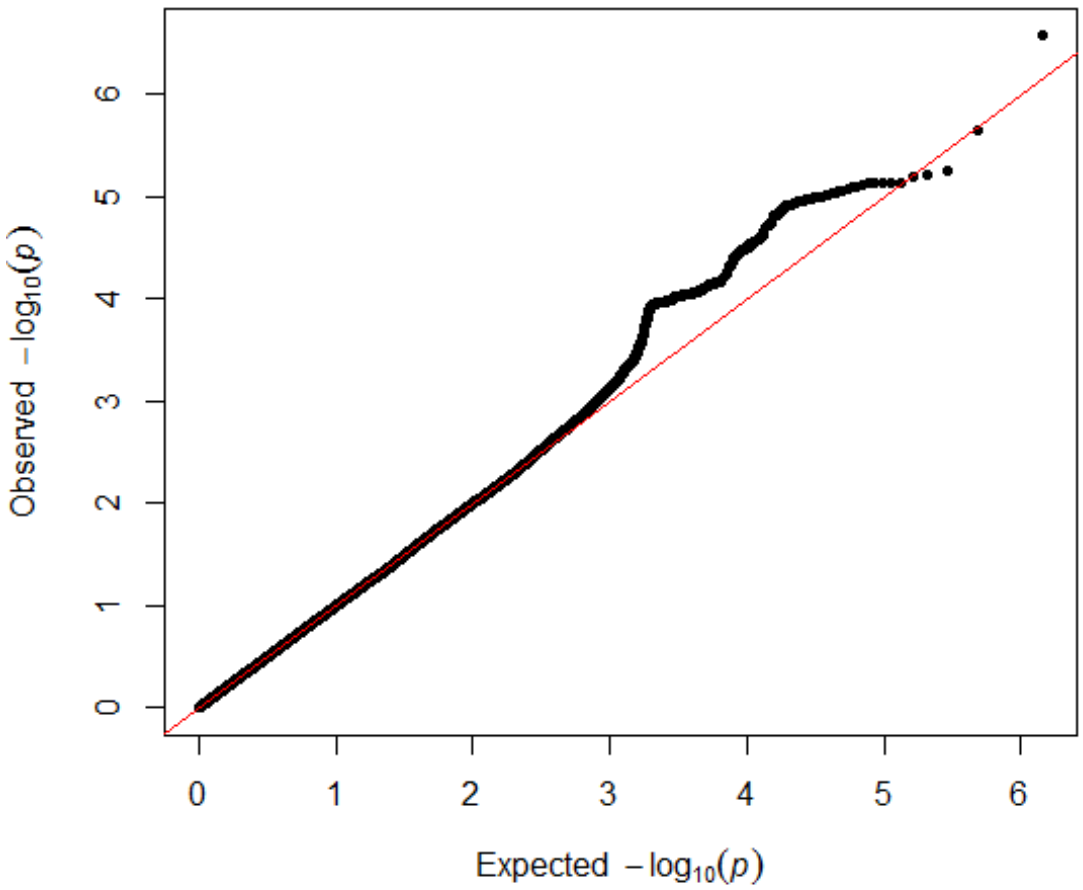


Table 8.2: SNPs associated with the intake of dairy milk and results of replication in an independent sample in the UK Biobank

Chromosome	SNP	Allele A	Allele B	Beta[a]	P[a]	In Full dataset	P[b]	Beta[b]	P[c]	Beta[c]
2	rs150080038	G	A	-0.237595	2.67E-07	Yes	3.77E-07	-0.22	0.364865	-0.03
2	rs34128006	A	AT	-0.0472842	2.20E-06	No				
2	rs1987446	T	C	-0.0580089	5.52E-06	Yes	4.67E-07	-0.06	0.0616328	0.02
3	rs9820500	C	T	-0.064445	3.36E-06	Yes	2.67E-06	-0.05	0.878865	-0.001
3	rs73040193	A	G	0.242878	6.99E-06	No				
3	rs114342505	C	T	0.164091	5.77E-06	Yes	2.46E-06	0.14	0.745773	-0.01
3	rs72558061	T	G	0.155833	4.69E-06	Yes	2.06E-06	0.12	0.851772	-0.004
6	rs1316538	C	G	-0.0507008	1.12E-07	Yes	1.42E-06	-0.04	0.276492	-0.01
7	rs62469670	T	G	0.0715656	2.46E-07	Yes	2.28E-07	0.06	0.980112	-0.0002
9	rs7029703	C	T	-0.0425765	6.84E-06	Yes	3.72E-06	-0.04	0.275312	-0.01
12	rs12367733	T	C	-0.143609	2.59E-06	Yes	3.02E-06	-0.14	0.361209	0.02
12	rs11180909	C	T	-0.151124	1.22E-06	Yes	1.68E-06	-0.14	0.613087	0.01
14	rs146106168	C	T	0.195123	9.40E-06	Yes	3.50E-06	0.17	0.305519	-0.03
16	rs72778535	G	A	0.0804969	1.18E-06	Yes	2.51E-06	0.07	0.292872	0.01
17	rs4647869	C	A	0.058601	3.50E-07	Yes	2.47E-07	0.04	0.903174	0.001
17	rs4792352	T	C	0.0434281	4.21E-06	Yes	3.26E-06	0.04	0.105628	-0.01
20	rs77410568	A	G	0.204402	7.05E-06	No				
21	rs71317650	G	A	-0.233565	2.32E-06	No				
21	rs148462362	TC	T	0.110927	6.71E-06	No				
21	rs77409573	A	G	0.155487	4.39E-06	No				
21	rs62230080	C	T	-0.121612	2.05E-06	No				
22	rs114463484	A	G	-0.101943	6.08E-06	Yes	5.30E-06	-0.09	0.832605	-0.003

[a] Results from the initial discovery GWAS sample using interim genetic data

[b] Results from the analysis of men who were in the initial discovery GWAS sample using full UK Biobank genetic data

[c] Results from the analysis of men who were not in the initial discovery GWAS sample using full UK Biobank genetic data

Figure 8.2: Manhattan plot of genome-wide association study (GWAS) of dairy milk intake in UK Biobank (UKB; N=22,041). Blue line for suggestive significance ($P < 10^{-5}$)

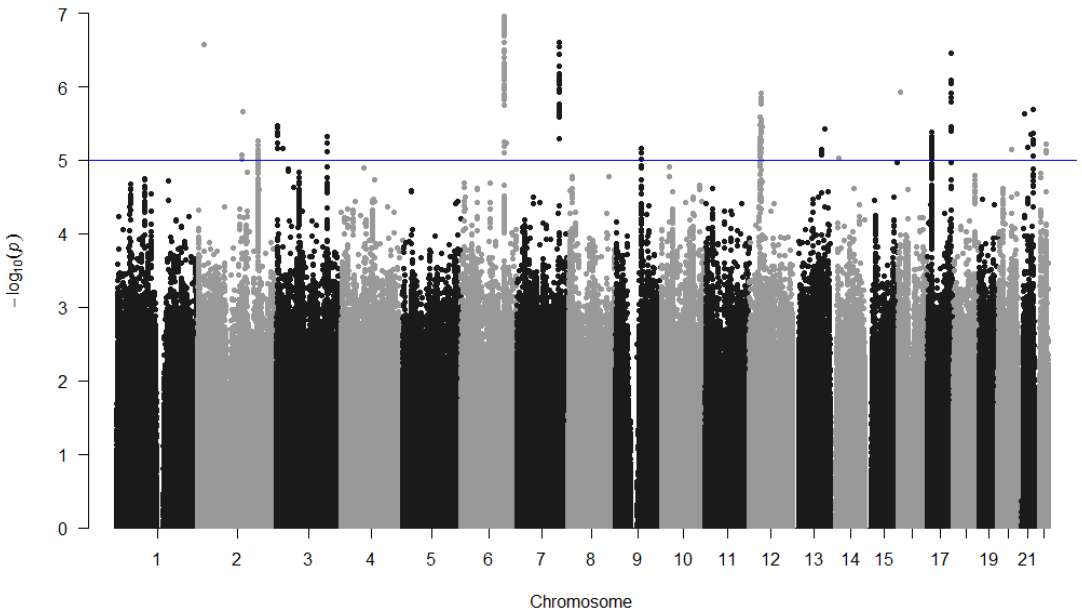


Figure 8.3: Locus plot for rs1316538 on chromosome 6 with gene annotation and recombination rates

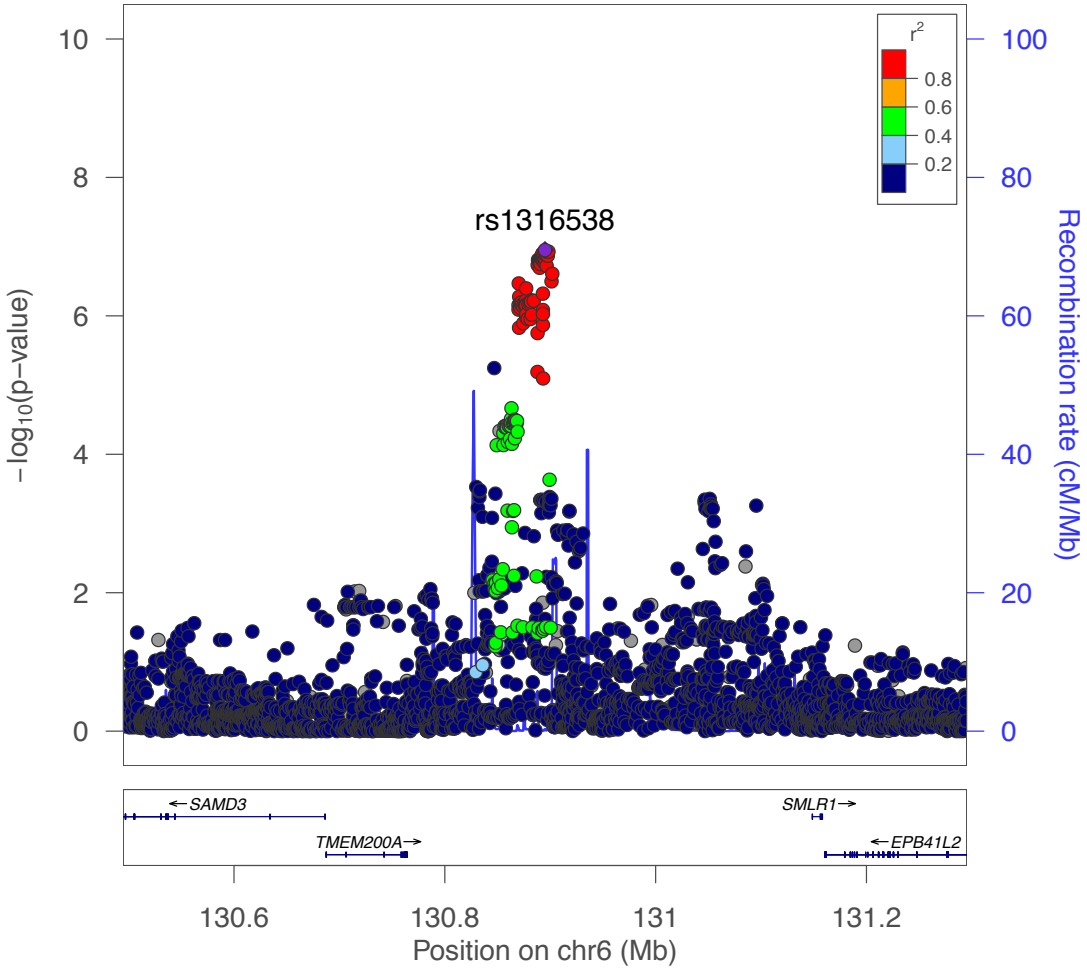


Figure 8.4: Locus plot for rs62469670 on chromosome 7 with gene annotation and recombination rates

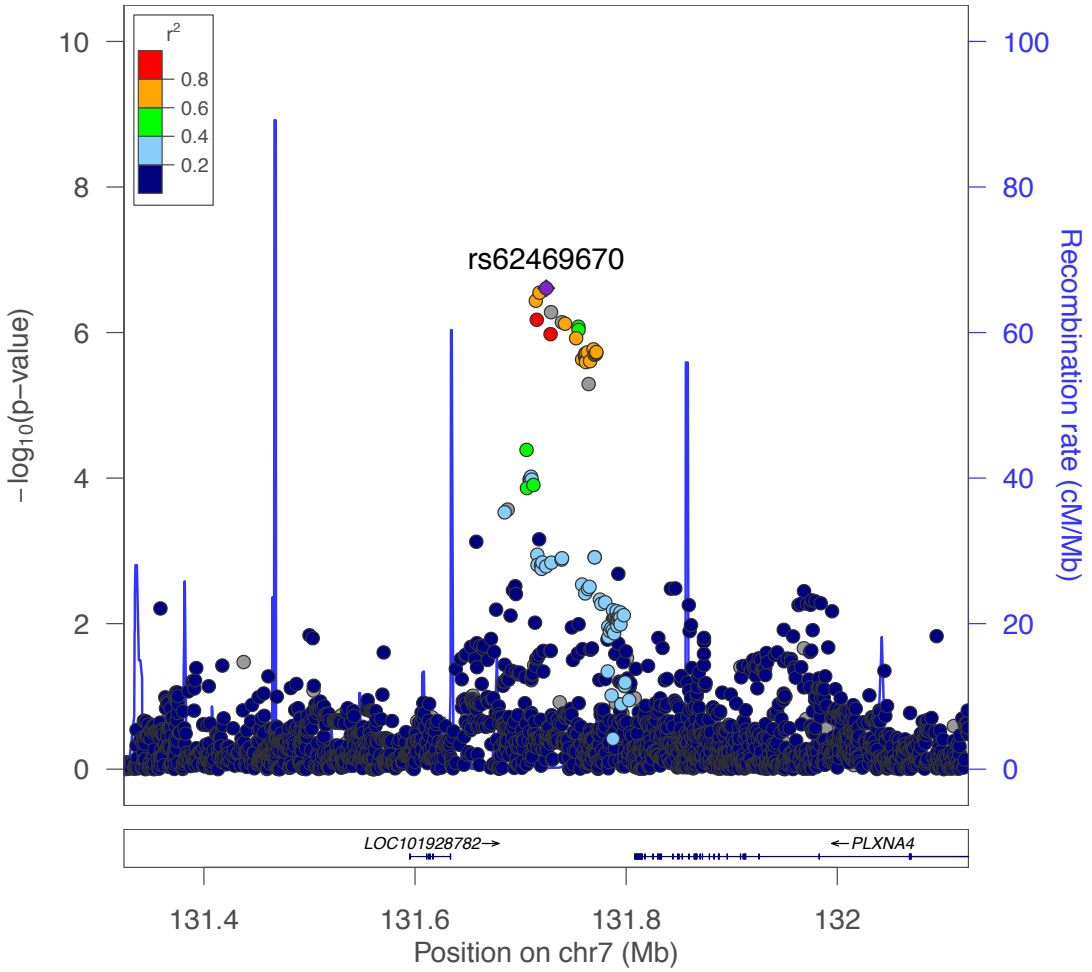


Figure 8.5: Locus plot for rs150080038 on chromosome 2 with gene annotation and recombination rates

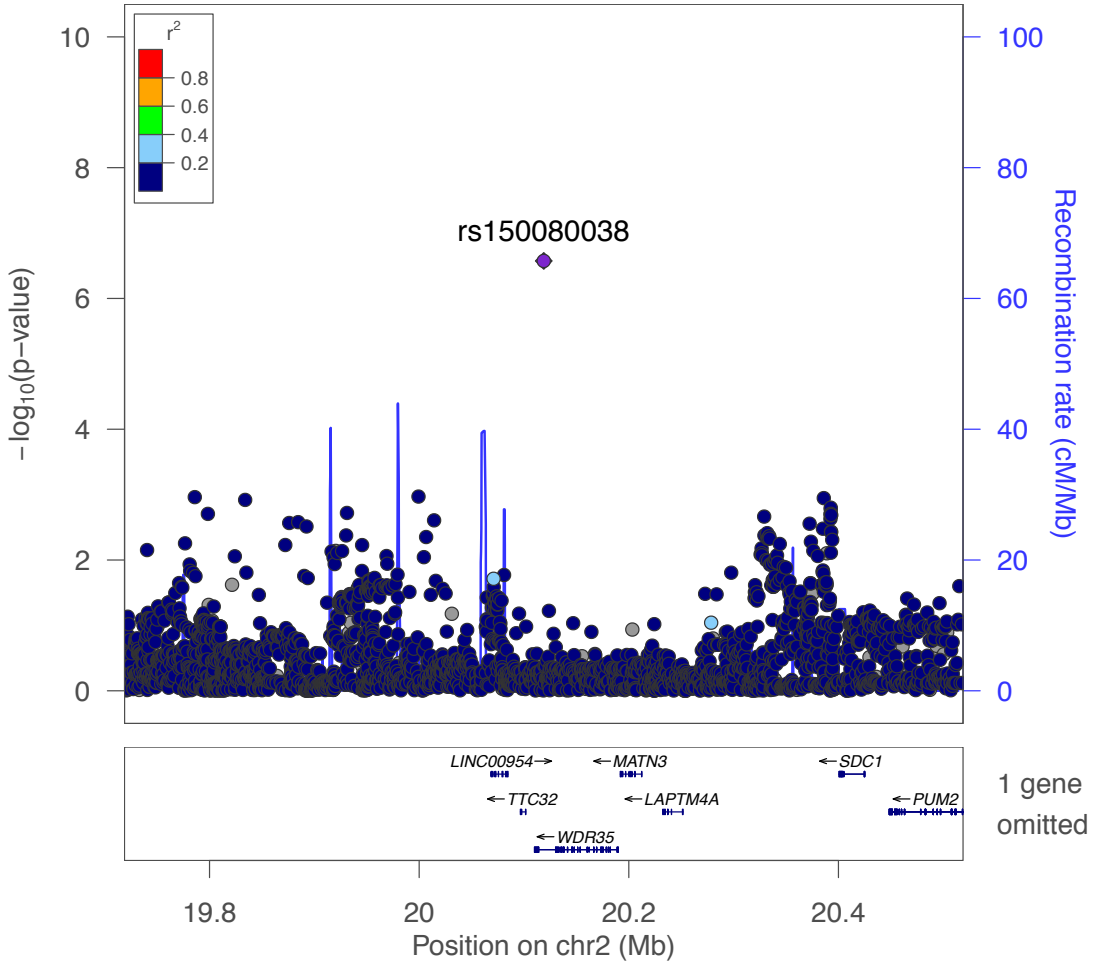
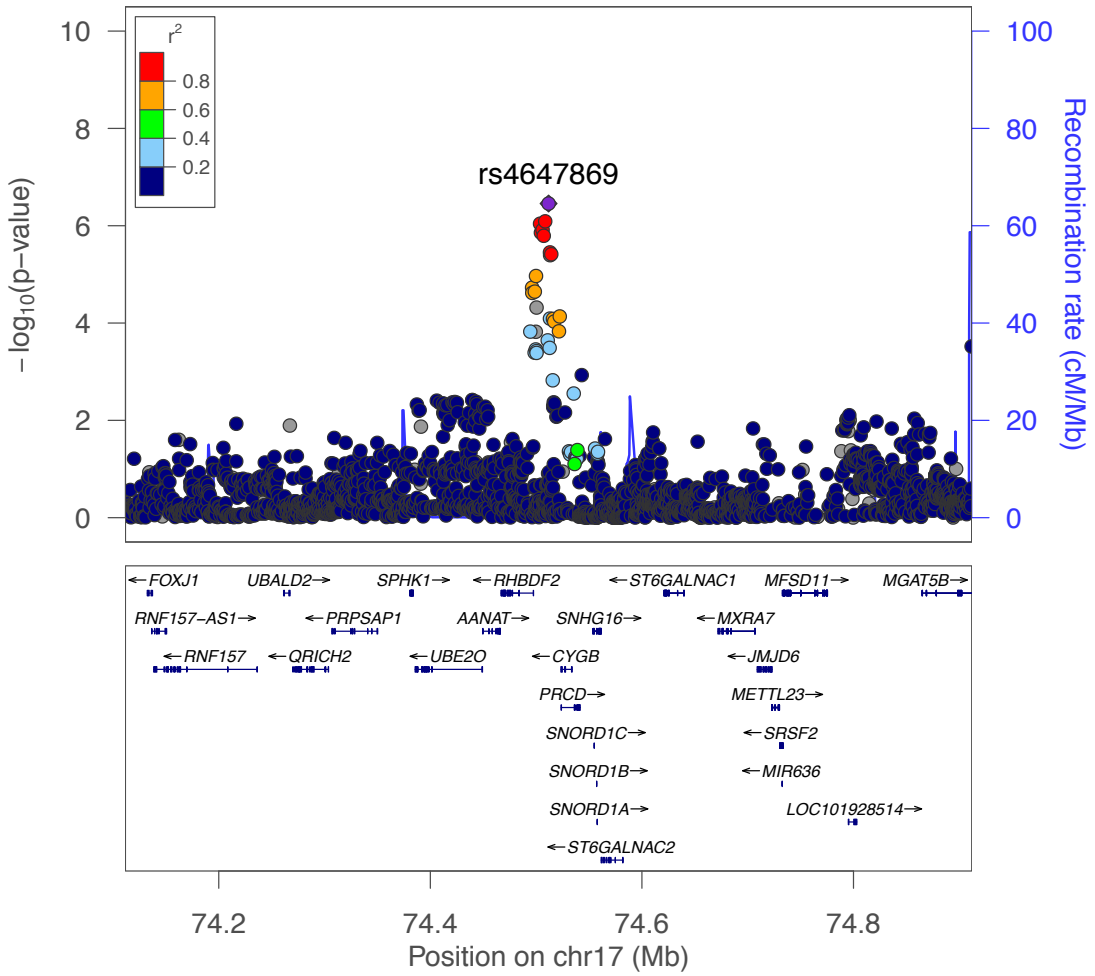


Figure 8.6: Locus plot for rs4647869 on chromosome 17 with gene annotation and recombination rates



Chapter 9

Discussion

9.1 Overview

The aetiology of prostate cancer is poorly understood. While there is some evidence in favour of numerous putative risk factors - vasectomy, microseminoprotein-β (MSP), human kallikrein 2 (HK2), and the lactase persistence SNP (rs4988235), for example - this evidence has often been generated using studies of modest size, retrospective design, or prospective studies with short follow-up time. The characteristics of such research have led to: inconsistent findings that may be due to health seeking behaviours or case mix between studies due to diffences in PSA testing practices in the case of vasectomy; uncertainty about whether findings for MSP were due to reverse causality; a lack of clear evidence that any association of HK2 with prostate cancer was independent of circulating PSA concentrations; and investigations that were likely underpowered to discover an association of rs4988235 with prostate cancer. The primary value of understanding the association of rs4988235 with prostate cancer is as a marker of dairy intake; as such, it is also pertinent to note that there is a poor understanding of how germline polymorphism, including rs4988235, determines the intake of dairy produce. Large, prospective studies are necessary to better estimate the independent associations for these putative risk factors with prostate cancer risk, and to characterise the association of genetic variation with dairy produce to facilitate novel statistical methods in prostate cancer epidemiology, such as Mendelian randomisation analyses (MR).

9.2 Aim of thesis

In this thesis, I aimed to further our understanding of the aetiology of prostate cancer. To that end, I investigated the association of lifestyle factors (vasectomy status), biochemical factors (circulating MSP and HK2 concentrations), and genetic variants (rs10993994 and rs4988235, as markers of MSP and dairy intake, respectively) with prostate cancer risk, with a focus on high grade and advanced stage prostate cancer, and death from prostate cancer. Given the reported association of dairy products with prostate cancer risk [18], the thesis was also concerned with how germline polymorphism determines the intake of dairy produce, primarily to explore potential genetic instruments for future MR analyses [279]. This included a cross-sectional

analysis of rs4988235 with the intake of dairy produce, and a genome-wide association study for the intake of dairy milk.

9.3 Main findings from this thesis

- In chapter 4 I investigated the association of vasectomy with prostate cancer risk in EPIC. Vasectomy was not associated with risk of prostate cancer overall, with risk for high grade or advanced stage tumours, or for death from prostate cancer. There was, however, some evidence that vasectomy may be associated with an increased risk for low-intermediate grade tumours, and that having a vasectomy was associated with also having a PSA test - together these results suggest that any previously reported association of vasectomy with prostate cancer risk may have been, at least partly, due to health seeking behaviours. Moreover, a meta-analysis combining these results with all available evidence from large prospective cohorts only found a significantly elevated risk for prostate cancer overall, and did not support an association of vasectomy with high grade or advanced stage disease, or death from prostate cancer. Additionally, although there were significantly higher concentrations of MSP and IGFBP-2, and lower concentrations of androstanediol glucuronide in men with a vasectomy compared to men without a vasectomy, these differences were modest and do not strongly support a biological difference in the biomarker profile of men by vasectomy status.

- The association of prediagnostic circulating concentrations of MSP with prostate cancer risk was examined in chapter 5. In a large European nested case-control study there was strong evidence that men with higher concentrations of MSP had a lower risk of prostate cancer. These results were subsequently confirmed by MR analyses that also found a significantly lower risk of prostate cancer in men with higher circulating concentrations of MSP. These analyses represent the largest prospective study of MSP and prostate cancer to date, and the only evidence of MSP as a potentially causal risk factor for prostate cancer from a MR design. Although there was no evidence that the association of MSP with risk differed by prostate cancer stage or grade, there were small

numbers of cases in subgroups defined by tumour subtype and so analyses may have had limited power to detect heterogeneity. Further, this study is the first to report that circulating concentrations of MSP are significantly higher (approximately 30%) among current smokers when compared to men without a history of smoking. However, there was no evidence of heterogeneity in the MSP-risk association by smoking status.

- In chapter 6 I estimated the association of prediagnostic circulating concentrations of HK2 with prostate cancer risk in the EPIC cohort. In this large nested case-control study there was no strong evidence for an association of HK2 with overall prostate cancer risk after adjustment for total PSA concentrations. Further, there was little evidence for an association of HK2 with risk of high grade or advanced stage tumours after adjustment for total PSA. Although the extreme collinearity between HK2 and total PSA may have hindered the accurate estimation of the association of HK2 with prostate cancer risk, results from a ridge regression did not produce materially different estimates of the association of HK2 with prostate cancer risk. Additionally, there was no evidence that the inclusion of HK2 to a risk prediction model of circulating PSA and age improved model discrimination for prostate cancer overall, or for high grade tumours. Further, this is the first cross-sectional analysis of HK2, which shows that HK2 differs significantly by age, marital status, educational attainment, body mass index, smoking status, and alcohol consumption.

- The association of the lactase persistence SNP, rs4988235, as a marker of dairy intake, with prostate cancer risk in the PRACTICAL consortium was investigated in chapter 7. As a secondary aim, this chapter estimated difference in the intake of dairy produce by rs4988235 genotype in the UK Biobank. There was no evidence of an association of rs4988235 with prostate cancer risk overall, high grade or advanced stage tumours, or death from prostate cancer. Further, when analyses were restricted to only data from prospective cohort studies the results were materially unchanged. Analyses of dairy intake suggested that rs4988235 may be associated only with a modest, between 25 and 19 g/d, difference in the intake of dairy milk, and may not be associated

with the intake of other dairy produce, such as ice cream, yogurt, or cheese. If the previously observed association for dairy produce with prostate cancer risk is not driven primarily by the intake of dairy milk [18], but instead by the intake of protein from dairy [98], an association of rs4988235 with prostate cancer risk may not have been expected. It is also possible that alternative genetic variation drives the intake of dairy milk; a 20% reduction in estimated difference in dairy milk intake by rs4988235 genotype after adjustment for principal components may indicate population stratification but it may also signal that alternative germline polymorphisms explain a proportion of the variation in the intake of dairy milk.

- Chapter 9 was a GWAS for the intake of dairy milk in the UK Biobank. Due to data processing concerns in the full UK Biobank genetic data set a discovery GWAS was conducted in the initial interim genetic data release, from which the putative SNPs associated with dairy milk intake were further investigated in an independent replication sample, also in UK Biobank. No SNPs were genome-wide significant in a discovery GWAS in the UK Biobank. Putative associations for SNPs significant at the lower tentative genome-wide significance threshold ($p < 10^{-5}$) did not replicate when investigated in an independent sample of men from the UK Biobank. It is possible that imputation or quality control procedures, such as HWE, may have hindered the discovery of SNP-dairy milk associations that resulted from gene-culture co-evolution, and should be investigated in the future. It is also possible that measurement error or insufficient sample size contributed to these null results. Future studies should re-run a discovery GWAS for the intake of dairy milk once full genetic data set for UK Biobank becomes available.

9.4 Findings in context

The findings presented in this thesis (Chapters 4 to 8) provide novel insight into the relationships of lifestyle, biochemical, and genetic factors with prostate cancer risk, and further our understanding of how germline polymorphism determines the variance in the intake of dairy products.

Vasectomy is a commonly used form of male sterilization that has been performed globally in 40 million to 60 million men [206]. Much of the previous

research into the association of vasectomy with prostate cancer risk has been conducted in small studies, which were often retrospective in design. The results presented in this thesis join only four previous studies [1] that have investigated the association of vasectomy with prostate cancer risk prospectively with more than 150 exposed cases [23, 34, 24]. Further, a meta-analysis combining results from this thesis and the four previous studies did not find evidence for vasectomy as a risk factor for high grade, advanced stage, or death from prostate cancer. The small, borderline significant association for vasectomy and prostate cancer overall from this meta-analysis is likely driven by health-seeking behaviours and does not indicate a role of vasectomy in prostate cancer aetiology; men with a vasectomy are more likely to have had a PSA test compared to men without a vasectomy [23, 200, 34]. Moreover, a recent meta-analysis that also considered case-control and retrospective cohort results found that the association of vasectomy with prostate cancer overall was more likely to be null with increasingly robust study design [372]. Given these findings I do not believe that more research is necessary for the association of vasectomy with prostate cancer as it does not appear to be a likely risk factor.

MSP is one of the three most abundantly secreted proteins by the prostate epithelium into the seminal fluid [39], and has been implicated in tumour suppression [256, 257, 258, 259, 260, 261] and pathogen defense [262]. Prior research into the association of MSP with prostate cancer risk has largely been in cross-sectional or tissue studies that measured urinary or blood concentrations collected from cases after diagnosis. The only previous prospective analysis of MSP with prostate cancer risk found a significant inverse association of MSP with prostate cancer risk but it was unable to exclude the possible influence of sub-clinical prostate cancer on their results due to short average follow-up time (3.8 years) [27]. In contrast, analyses in this thesis support a strong inverse association of MSP with prostate cancer risk with more than double the follow-up time. Moreover, the analysis in this thesis leveraged genetic data available in the EPIC cohort to conduct an MR study that show a significant protective association that is unlikely to be due to reverse causality or other common epidemiological biases. As such, the research in this thesis highlights a potentially causal inverse association of MSP

[1] One unpublished study from the UK record linkage study

with prostate cancer, which may eventually prove useful for the prevention of prostate cancer.

Existing research into HK2 in relation to prostate cancer risk has been aimed primarily at assessing the ability of HK2 to improve the discrimination of prostate cancer among men with elevated PSA when compared to using PSA alone or a combination of other kallikrein biomarkers [47, 48, 49, 50, 51, 52, 282, 283, 284, 53]. To date, there has not been an investigation of the prospective association of HK2 with prostate cancer risk in men without elevated PSA that considered the association of HK2 independent of the highly collinear circulating PSA concentrations. The results presented in this thesis constitute the largest prospective study to date, and do not support an association of HK2 with prostate cancer risk independent of its association with circulating concentrations of PSA. Prior research had suggested a modest improvement (1%- 6%) in the discrimination of prostate cancer overall and for high grade tumors after the inclusion of HK2 in a PSA and age-based risk model. In contrast, this thesis finds no evidence that HK2 improves the area under the curve when included in such a prediction model for either prostate cancer overall or for high grade tumours.

The PRACTICAL consortium is the largest prostate cancer genetics data set in the world, and combines the data from many studies to provide a reliable estimate of the association between genetic polymorphism and prostate cancer risk. The largest previous analysis of rs4988235, as a marker for dairy intake [330, 18], was a meta-analysis [55] of all available data (N^{cases} = 4,783, $N^{controls}$ = 3,188), but was unable to look into a rs4988235 association by tumour subtypes due to small sample size in these subgroups. Data for the current analysis in the PRACTICAL consortium are almost an order of magnitude larger (N^{cases} = 48,471, $N^{controls}$ = 29,866), and, in contrast, have been sufficient to consider the association of rs4988235 with high grade and advanced stage tumours, and death from prostate cancer. While the current analyses can thus be considered the first adequately powered investigation [55] of rs4988235 and prostate cancer risk, these results are not necessarily a refutation of the association of dairy with prostate cancer risk; if the association of dairy products with prostate cancer is driven not by dairy milk intake but instead by protein intake from dairy products [98] we may not expect rs4988235 to be clearly associated with prostate cancer. A more in depth understanding of how rs4988235 associates with the intake of various

dairy products is required to better interpret the findings from this thesis in the PRACTICAL consortium.

The UK Biobank is one of the world's largest prospective cohorts with available data for genetic and dietary variables. Prior research into the association of the lactase persistence SNP, rs4988235, has largely been from small studies [55, 330] - with the exception of one large Danish study [336] - none of which adjusted for principal components of genetic variation (PC) [2]. As such, it has not yet been possible to assess the potential confounding by population stratification for the association of rs4988235 with the intake of dairy products research. Further, much of the research, to date, has focused on the intake of milk, with only infrequent estimation [55] of the differences for the intake of other dairy products by rs4988235 genotype; current analyses are the largest investigation of the intake of yogurt, ice cream, and cheese by rs4988235. Results for the intake of dairy milk by rs4988235 genotype presented in this thesis before adjustment for PC (CC vs TT: 25.5 g/d [95% CI 18.8-32.2]) were remarkably similar to a recent meta-analysis that found an approximate 25 g/d difference [330]. Following adjustment for PC, however, estimates of the difference in dairy milk intake by rs4988235 were reduced by approximately 20%, which may indicate the influence of population stratification. Such confounding may also have affected previous estimates of the association of rs4988235 with dairy milk intake. One additional explanation for the attenuated estimated difference for the intake of dairy milk by rs4988235 may be the existence of additional germline variation that may determine dairy milk intake, which should be investigated further. No difference in the intake of other dairy produce was observed by rs4988235 genotype, which likely reflects the relatively low lactose content of these food items.

There has been no previous investigation of how genome-wide germline variation may determine the intake of dairy milk. The majority of research focused on single candidate SNPs such as the lactase persistence SNP, rs4988235, and has been discussed above. While results from a discovery GWAS in a sample of $\sim$ 22,000 men in UK Biobank found 23 putative SNPs associated at a lower significance threshold ($p < 10^{-5}$), no SNPs were significant at the traditional genome-wide significance threshold ($p < 5 \text{ x } 10^{-8}$).

[2]It does not appear that principal components were available in previous studies of rs4988235 and the intake of dairy milk

Further, no putative significant SNPs from the discovery GWAS were successfully replicated in a large independent sample in UK Biobank ($\sim$ 50,000 men), and so this thesis did not discover any novel variation associated with the intake of dairy milk. As such, the strongest candidate for germline variation that determines the intake of dairy milk remains rs4988235. rs4988235 is believed to be associated with dairy milk intake due differences in the production of the lactase enzyme into adulthood that resulted from gene-culture co-coevolution [362]. It is likely that SNPs subjected to recent positive selection may not be included in GWAS due to how recent positive selection may result in allele frequencies that violate common QC procedures, such as Hardy-Weinberg equilibrium. As such, it is possible that if we expect novel germline variation that determines the intake of dairy milk to have been subjected to similar evolutionary pressures as rs4988235, SNPs associated with dairy milk intake may have been filtered out before analyses, and thus would not have been discovered in these analyses. Future research should consider carefully the mechanisms that may have led to germline polymorphism being associated with exposures to ensure the methodology used does not hinder the discovery of potential important genetic associations.

9.5 Methodological Considerations

Several methodological considerations have arisen in the process of preparing this thesis, and are discussed below.

9.5.1 Sample size and statistical power

The EPIC [140] and UK Biobank [141] cohorts, and the PRACTICAL consortium were sufficiently large to detect a modest association for risk factors with prostate cancer in the current thesis. In particular, the size of the PRACTICAL consortium has allowed, not only for the robust investigation of the association of rs4988235 with prostate cancer risk overall, for which previous studies have likely been under-powered [55], but also the investigation of rs4988235 with prostate cancer tumour subtypes, and fatal prostate cancer, which may be of greater clinical relevance. However, it remains possible that for analyses by tumour subtype there were insufficient numbers to adequately test for heterogeneity by stage or grade of disease. In particular,

this may have been the case for analyses by histological grade for the nested case-control of MSP, the smaller of the two nested case-control studies.

9.5.2 General considerations for measurement and confounding

Information on covariates used in this thesis was taken from the recruitment questionnaires for the EPIC and UK Biobank questionnaires. It is likely that exposures for some variables (such as body weight, alcohol consumption, or smoking status) changed over the duration of follow-up, which may have led to some measurement error. However, with the exception of analyses of HK2 and prostate cancer risk, adjustment for variables measured at recruitment made no material difference to estimates. As such, any potential error introduced is unlikely to have had a large effect on the relative risks reported in this thesis.

There was no information collected for socioeconomic status in the EPIC cohort - with the possible exception that educational attainment may capture a proportion of the variance attributable to socioeconomic status - which limited the ability to stratify analyses by level of social deprivation. Further, family history of prostate cancer was not collected. Social deprivation and family history of prostate cancer are known to be associated with a delayed presentation for prostate cancer diagnosis and differences in diagnostic practices and the degree of invasiveness for treatment received [373, 70]. As such, for research questions in this thesis where health seeking behaviours were especially pertinent (vasectomy and MSP), the lack of information at recruitment on socioeconomic status and family history of prostate cancer in the EPIC cohort likely hindered the ability to investigate potential confounding effects, and should be considered a limitation.

Participants in the EPIC [140] cohort were predominantly of European origin. Further, analyses in the PRACTICAL consortium and UK Biobank explicitly excluded men of non-European ancestry. Thus, it is possible that results from this thesis do not generalise to populations of differing ancestry.

9.5.3 Measurement of vasectomy

The ascertainment of vasectomy status in prospective cohort studies has been a source of criticism for past research [34, 23]. It is possible that men without

a vasectomy at recruitment go on to receive a vasectomy during follow-up, and so are misclassified as non-vasectomised due to the lack of complete follow-up for vasectomy status. However, the prevalence of vasectomy in the general population is relatively low in EPIC (∼15%) and men who receive a vasectomy typically have one before reaching middle age [34]. As the mean age at recruitment in the EPIC cohort was 54 years it appears unlikely that any misclassification of vasectomy status during follow-up would be substantial, which is supported by analyses in the EPIC-Oxford sub-cohort. For the EPIC-Oxford sub-cohort updated data were available on vasectomy status collected 10 years after recruitment. These data showed that only 5.1% of men without a vasectomy at baseline reported having had a vasectomy during the 10 years after recruitment. Moreover, a recent study estimated that a 5% misclassification of vasectomy status (e.g. if the true prevalence of vasectomy was 20% compared to 15%), the relative risk would be underestimated by only approximately 1% [34]. Thus, we believe that any misclassification would have resulted in a minimal underestimate of any association of vasectomy with prostate cancer risk.

9.5.4 Measurement of biochemical factors

In the EPIC cohort standardised blood collection and transport protocols were implemented in most centres (as described in chapter 3). However, the protocol was modified for the Oxford centre, and although protocols were similar, the Danish and Swedish centres also differed as they were added at a later date to the EPIC cohort [140]. For the Oxford blood samples, blood samples were delivered by post to the processing laboratory; for markers that may degrade under delayed processing procedures this protocol may have led to differences in the measured concentrations of biomarkers. However, no difference was observed in the average concentration of biomarkers used in this thesis by number of days in post in a sensitivity analysis, and so we do not believe that slight differences in blood collection and sample transport protocols have impacted results from this thesis.

All assays from the EPIC cohort in this thesis were conducted by trained laboratory persons who were blinded to the case status of samples. Quality control samples were included in all batches and intra- and inter-batch coefficients of variation were calculated. Further, for analyses in EPIC that

included MSP, HK2, and PSA, the availability of blood samples has facilitated a quality control pilot study that addressed a potential concern over the fact that some of the previous studies to investigate MSP [265, 266] and HK2 [47, 48, 49, 50] have used serum samples. We considered the concordance between MSP, HK2, and PSA measurements in serum and citrated plasma samples and found a high degree of concordance. Additionally, we found no significant difference in the mean concentration of these biomarkers between plasma samples drawn at five year intervals. Together these results allowed for confidence that analyses using assays from these two different media would be comparable, and would be able to make a robust contribution to the existing body of literature. Further, the procedures ensure a high degree of confidence can be had in the precision or inter-assay repeatability of analyte measurements in this thesis.

9.5.5 Measurement of genetic factors

All genotyping was standardized for data from UK Biobank [331] and the Oncoarray chip [37] as part of the PRACTICAL consortium. This is not true for genetic data available within the EPIC cohort; data were drawn from subgroups of the nested case-control in the EPIC cohort that had genotype data available from the iCogs [36], Oncoarray [37], and BPC3 genotyping arrays [38]. It is possible, though unlikely, that there were slight discrepancies in the called genotype frequencies for rs10993994 which would likely have attenuated any estimates derived. However, estimates derived for the association of rs10993994 with both MSP and prostate cancer were remarkably similar to previously reported results [13, 27], and so we believe the rs10993994 measurement to have been robust.

9.5.6 Measurement of diet

Diet is a complex exposure to measure with precision and reliability, and the potential for errors in measurement is well described elsewhere [32]. Analyses that concerned the dietary intake of dairy products in this thesis relied on available data from 24-hour recall questionnaires within the UK Biobank (dairy product intake was not investigated in detail for the baseline food frequency questionnaire). As such, the strength of these measurements is predicated on the assumption that dairy product intake on a given day may

approximate general consumption for men in UK Biobank. While this is, on average, not an unreasonable assumption for men in UK Biobank, it is also a likely source of modest measurement error for men with dairy consumption that varied over time. Note, this may have been a greater source of measurement error for dairy products that are consumed with a lower regularity such as yogurt or ice cream, and would attenuate any estimated associations in this thesis.

Moreover, the measure of dairy intake used in this study is the mean intake from the total number of repeat 24-hour recall questionnaires available for a given individual (a maximum of five). The availability of multiple questionnaires was a major strength for analyses in this thesis due to the reduction in measurement error for men with mean intake calculated from a greater number of 24-hour recall questionnaires. Once more, this was particularly beneficial for dairy products that are eaten with a relatively low frequency in the British population (ice cream and yogurt). However, for the analyses of germline variation, it was not possible to be certain that all SNP alleles were equally likely to have men with a mean dairy intake calculated from an equal number of questionnaires, and so there likely existed some heteroskedasticity for genetic analyses in this thesis. The effect of such differences in the measurement error for dairy intake by SNP allele would be to make it less likely to observe a significant mean difference in intake by SNP allele, which may have contributed to null results observed for GWAS analyses [367, 368]. However, for analyses of the lactase persisitence SNP, rs4988235, and dairy intake homoscedasticity was thoroughly considered in regression methods. Further, derived estimates were very similar to the most recent large investigation of dairy milk intake and rs4988235 [336] and so we do not believe it was a concern for the results from chapter 7.

9.6 Confounding

The interpretation of epidemiological studies are frequently subject to limitations due to confounding. A confounder is a variable that is associated with both an exposure and outcome of interest that results in a spurious association between exposure and outcome. For analyses of prostate cancer risk, the investigation of confounding is typically limited by the small number of confirmed risk factors for prostate cancer (age, family history of

prostate cancer, ethnicity, select germline genetic polymorphisms, and circulating concentrations of insulin-like growth factor one (IGF-I)). All analyses for prostate cancer risk were adjusted for age and all analyses were restricted to ethnically Caucasian men, and so these factors were not a likely concern for confounding for findings from this thesis. It is a limitation of this thesis that information on family history of prostate cancer, germline polymorphism, and IGF-I was not available for all men in the EPIC cohort, and so we were therefore not able to investigate possible confounding by these factors.

The precise biological functions of MSP, PSA, and HK2 remain poorly characterised, and it is likely that these biochemical factors may be regulated by numerous factors. Analyses in this thesis included a wide range of potential confounders due to the scope and size of the study populations, and the detailed covariate information described in recruitment questionnaires. Indeed such information was used to conduct sensitivity analysis to ascertain the potential effect of the inclusion of factors found to be associated with biochemical variables on any association for biochemical factors and prostate cancer risk. Nonetheless, it is not possible to fully exclude possible confounding by all factors or the influence of residual confounding by factors included as covariates in the observational analyses.

Analyses for the association of germline polymorphism and disease risk are typically not easily confounded. However, population stratification, which can be interpreted as confounding by ancestry, may severely affect the estimated relative risks for genetic factors and disease risk. Nonetheless, current analyses for the lactase SNP, rs4988235, and prostate cancer risk have been fully adjusted for the principal components of ancestry, and so we believe the role for confounding in these estimates is minimal. Similarly, the availability of detailed information on participant ancestry from principal components in UK Biobank has allowed for greater confidence that estimated intake of dairy products by rs4988235 were not due to residual confounding by population stratification. This was a potential weakness that was not addressed in previous studies of dairy intake and rs4988235.

History of PSA testing

Although history of PSA testing was available for an EPIC-Oxford sub-cohort follow-up questionnaire, for the majority of men in prostate cancer risk analyses within this thesis no information was available on the history of PSA

testing or PSA test results at diagnosis. Published estimates suggests 36% of men aged $\geq$ 50 years in Britain between 2006-2010 [92] have had at least one PSA test. Past research into vasectomy, MSP, and HK2, and prostate cancer risk has been subject to criticism pertaining to the potential influence of health seeking behaviours or PSA testing on the estimated relative risks. As such, it may be considered a limitation for some of the research in this thesis that we were not better able to investigate any heterogeneity and confounding of the associations for vasectomy, MSP, and HK2, by PSA testing history or PSA test results.

Nonetheless, an increased likelihood of having a PSA test was observed for vasectomised men compared to non-vasectomised men in an EPIC-Oxford sub-cohort follow-up questionnaire. Given the purported reduction in prostate cancer mortality associated with PSA testing, any effect of increased frequency of PSA testing in vasectomised men would likely be negative confounding; in chapter 4 we show that any potential adverse effect of vasectomy on the risk prostate cancer partly masked by a beneficial effect of increased PSA testing is likely small. Further, the only previous prospective study of MSP and prostate cancer risk found no heterogeneity by history of PSA testing [27]. There has, however, been little investigation into how the association of HK2 may vary by PSA testing history. Thus, while we believe that there has been little influence of PSA testing frequency or test results for the investigation of vasectomy and MSP, certainty for confounding of results for HK2 is not possible.

9.6.1 Reverse causality

A major consideration for any prospective epidemiological analyses is the role of follow-up time, and the potential that estimated associations for disease risk may have been due to reverse causality. Reverse causality occurs when the probability of the outcome is causally related to the exposure being studied. However, the EPIC cohort has up to an average of 15 years follow-up [140], which has allowed for the detailed prospective investigation of how associations with prostate cancer risk for vasectomy, MSP, and HK2 may vary by time between baseline (recruitment or blood collection) and diagnosis - no significant heterogeneity by time between baseline and diagnosis was observed for any prostate cancer risk analyses.

Further, nested case-control studies (MSP and HK2) in the EPIC cohort used an incidence density sampling protocol to select control participants from the same study centre as an index case. Samples were additionally matched on length of follow-up and age at blood collection ($\pm$ 6 months). Additionally, for all centres except Malmö, case-control pairs were also matched on time of blood collection ($\pm$ 1 hour), and duration of fasting at blood collection ($\leq$ 3, 3-6, $\geq$ 6 hours). This protocol has helped avoid the introduction of systematic differences in blood samples that may have otherwise led to bias by differences that may have existed between EPIC study centres for the investigation of MSP and HK2 and prostate cancer risk. However, it has also reduced the potential for matched controls to have latent prostate cancer at the time of blood collection, and thus reduced the potential influence of reverse causality for MSP and HK2 associations with prostate cancer risk.

With regard to the analysis of MSP and prostate cancer risk, the use of MR methods has precluded the possibility that any reported association was due to reverse causality. The online MR-base tool allowed for the phenome-wide investigation of rs10993994 with over 850 phenotypes [61], and confirmed previous assertions [278] that rs10993994 is likely the most robust case for a genetic instrument [279]. As such, the addition of significant MR analyses for MSP and prostate cancer risk to strong significant observational results suggest that any MSP association with prostate cancer is not likely due to common epidemiological biases, such as reverse causality.

9.6.2 Classification of prostate cancer

Prostate cancer classification and registry procedures were not standardised for all EPIC centres. Although population-based cancer and mortality registries were the source of data for cancer incidence, tumour characteristics and vital status for most centres in EPIC, this was not the case for German and Greek centres. For these centres a combination of methods are used that included health insurance records, local cancer and pathology registries, and active follow-up - in all cases self reported cancer incidence was verified via medical records. While it is clear that the classification of prostate cancer is likely individually robust for each EPIC centre, the lack of standardisation may have led to prostate cancer classification being affected differentially by the changes, over time, due to differences in diagnostic practices. This is

a particularly salient limitation given both the TNM stage system and the Gleason scoring method were initially devised approximately half a century ago, and so have been subject to improvements over time [133, 134].

TNM stage was initially informed only by physical and histological examination, using digital rectal examination, for example, or radiography. However, TNM score has subsequently used additional information from novel diagnostic techniques such as ultrasound or bone imagining [133]. Although recent additions to TNM scoring have led to improvement in the quality of stage information, it is notable that pathology specimens may return higher TNM scores than those based on clinical examination alone. In EPIC, information was not available on whether stage was from clinical or pathological examination, and so there is uncertainty surrounding potential variation in the precise classification of stage for cases within centres over time and between EPIC centres.

There have similarly been advances in the histological grading of prostate cancer [134]; the quality and amount of tissue collected have changed over time as the state of current best practice has evolved [136]. Changes include: the use of ultrasound-guided needle biopsy; the number of cores collected at biopsy; and the change in the practice of collecting tissue at diagnosis compared to collection via radical prostatectomy at surgery. It is also notable that the use of PSA testing in clinical practice has led to changes in the Gleason grade that men present with at diagnosis, such that many men present with low grade tumours [84, 81, 374]. These changes, on average, have resulted in an upward trend in the allocation of higher grades to cancers that may have previously been classified as having a lower Gleason grade [137]. Given the long available follow-up and the lack of standardisation across the EPIC cohort, it is possible that measured histological grade differs by year of diagnosis or recruitment centre.

The size and structure of the PRACTICAL consortium [170] is such that there may be similar concerns over the classification of tumour subtype. The PRACTICAL consortium consists of 123 different study groups, incorporating sites in the EU, Australia, Canada and USA, with cases that may have had a diagnosis anytime in the last 30 years. As such, there may also exist some differences in the classification of tumour characteristics for cases in the PRACTICAL consortium, which should be considered a limitation in the analysis of rs4988235 with prostate cancer risk by tumour subtype.

9.7 Recommendations for future research

Future research into the association of MSP and HK2 and prostate cancer risk should pool existing individual level data from available prospective cohorts to improve the statistical power to investigate associations among subgroups defined by tumour characteristics. To the author's best knowledge there are currently two other prospective cohorts with available data for MSP [27, 131]. However, only one of these studies, in the Multi-ethnic cohort (MEC) [27], has measurements in men without elevated PSA at recruitment. Given the covariance between PSA and MSP, it is likely that any pooling study would benefit from a sample of men without a restricted range of PSA. Further, MEC has black, Japanese, Latino, Native Hawaiian, and white populations, and so a pooled individual level analyses will likely produce risk estimates that may be more generalisable, on average, for men globally. Two prospective studies exist in men without elevated PSA for HK2 [28, 29] that could potentially be used in an individual level pooled analysis. Although both existing studies, and current analyses of HK2, are in men of European ancestry and so would likely not improve generalisability of results, such a pooled analysis would help improve statistical power to detect any small but real association of HK2 with prostate cancer risk that may be independent of circulating PSA concentrations for tumour subtypes.

Given the significant, potentially causal, association of MSP with prostate cancer risk from the MR analyses in this thesis, it is of principal interest to generate an independent replication. Such a replication may be achieved by using an estimate of the association of MSP with rs10993994 that exists within MEC [27], and an estimate of the association of rs10993994 with prostate cancer that will be available from the OncoArray project once the main publication has been accepted [37]. While rs10993994 has perhaps the strongest case for a robust MR instrument [279], a better understanding of how germline polymorphism determines circulating concentrations of MSP may allow for the use of methods that are more robust to assumptions of regular MR methods, such as pleiotropy, and give more confidence the association of MSP with prostate cancer risk. As such, future research should focus on completing a GWAS for MSP with a view to discovering additional germline polymorphism that determines circulating concentrations of MSP.

There is currently a poor understanding of the functional role of MSP in the body; the best hypotheses at present suggest MSP may be involved in tumour suppression [256, 257, 258, 259, 260, 261] or pathogen defense [262]. If the association of MSP with prostate cancer risk is supported by a further replication MR, it will be particularly important to better understand the possible mechanisms by which this putative protective association acts. Further, it will be necessary to understand what the potential effects of modifying the circulating concentrations of MSP in men would be for subsequent risk of prostate cancer. This could be achieved through a future large-scale randomised control trial if a method of modifying MSP levels in men becomes available.

General recommendations that follow from the research in this thesis into biochemical markers are to ascertain: the temporal reproducibility of these analytes in a larger sample, which may help future studies to correct for measurement error due to short-term biological variability or longer-term within-person variability due to changes in smoking behaviours, weight, or age; how circulating concentrations of these analytes covary with and whether these measurements are meaningful surrogates for tissue concentrations in the prostate; and whether prospective tissue concentrations of these analytes are associated with prostate cancer risk.

A prior investigation into the association of rs4988235 and prostate cancer risk suggested that approximately 30,000 cases and 30,000 controls would be necessary to detect a 2% per-allele increase in risk with 80% power [55]. The analyses in this thesis were conducted with 48,471 cases with 29,866. As such, we do not believe further research is necessary to better understand the association of rs4988235 with prostate cancer risk.

With regard to the role of germline polymorphism as a determinant for the intake of dairy produce, it appears that the lactase persistence SNP, rs4988235, is only strongly associated with the intake of dairy milk. However, rs4988235 does not appear to associate with prostate cancer risk. Given analyses in this thesis were not successful in identifying novel genetic variants associated with dairy milk and as dairy milk remains a putative risk factor for prostate cancer [18], future research should aim to discover further genetic variants associated with both dairy milk intake prostate cancer risk to facilitate the use of MR methods for dairy and prostate cancer risk analyses.

9.8 Conclusion

Studies contained in this thesis provide novel evidence on a number of putative risk factors for prostate cancer. The original analyses and the meta-analysis presented in this thesis for the association of vasectomy with prostate cancer risk provide convincing evidence that vasectomy should not be considered a risk factor for prostate cancer. The observational and MR results for the association of MSP with prostate cancer suggest that MSP is a potentially causal protective factor. Future research should focus on replicating these results and better understanding the biological function of MSP. In contrast, the research in this thesis does not support an association of HK2 with prostate cancer risk that is independent of circulating concentrations of PSA, or support past findings that, when added to a PSA and age-based model, HK2 improves the discrimination for prostate cancer overall and high grade tumours. Further, results from the present thesis do not find an association of the lactase persistence SNP, rs4988235, with prostate cancer risk; given we believe these analyses were adequately powered, more research into this genetic variant may not be necessary. Nonetheless, given the null results for the GWAS of dairy milk intake, future research should re-run a GWAS in UK Biobank once the full data becomes available.

Appendix A

Ancilliary Tables

Table A.1: Phenome-wide association scan for rs10993994 from MR-Base

Trait	Beta	SE	N	P
HOMA-B	0.01	0	36466	3.00E-03
Leucine	0	0	7351	4.00E-03
CD34 on Stem	0.06	0.02	460	4.00E-03
N-acetylornithine	0.01	0	7146	4.00E-03
Gamma-glutamylleucine	0	0	7354	6.00E-03
X-12556	0.01	0	7049	9.00E-03
Urea	0.01	0	7348	1.10E-02
X-14450–phenylalanylleucine	0.02	0.01	2520	1.10E-02
Total cholesterol in medium LDL	-0.03	0.01	21559	1.10E-02
Free cholesterol	-0.03	0.01	13496	1.30E-02
Phospholipids in medium LDL	-0.03	0.01	21558	1.40E-02
Taurocholate	0.02	0.01	3783	1.50E-02
B Mem:%IgE	0.05	0.02	460	1.50E-02
2-hydroxypalmitate	0	0	7348	1.60E-02
CD337 on NK early	-0.05	0.02	420	1.60E-02
Simple reaction time	-0.07	0.03	2378	1.70E-02
X-12717	0.03	0.01	1612	1.70E-02
Total cholesterol in large LDL	-0.03	0.01	21552	1.70E-02
X-04500	0.03	0.01	3393	1.70E-02
X-14086	-0.01	0.01	2651	1.80E-02
Myocardial infarction	0.03	0.01	171875	1.80E-02
Free cholesterol in large LDL	-0.03	0.01	21555	1.80E-02
CD123 on mDC	0.05	0.02	455	2.00E-02
Fasting insulin	0.01	0	38238	2.00E-02
Total cholesterol in LDL	-0.02	0.01	21559	2.00E-02
HOMA-IR	0.01	0	37037	2.10E-02
ADSGEGDFXAEGGGVR*	-0.01	0	5425	2.20E-02

Phospholipids in large LDL	-0.02	0.01	21550	2.30E-02
Iron	0.02	0.01	23986	2.40E-02
Valine	0	0	7359	2.70E-02
DSGEGDFXAEGGGVR*	-0.02	0.01	5229	2.70E-02
X-12544	-0.01	0.01	5581	2.80E-02
Allantoin	0.01	0	5344	3.10E-02
X-12056	0.02	0.01	3890	3.10E-02
Cigarettes smoked per day	0.19	0.09	68028	3.10E-02
Stachydrine	0.03	0.01	2779	3.20E-02
Citrate	0	0	7364	3.40E-02
Bilirubin (E,E)*	-0.01	0.01	7303	3.40E-02
B Mem:%IgG	-0.04	0.02	460	3.40E-02
Apolipoprotein B	-0.02	0.01	20689	3.40E-02
Proline	0	0	7367	3.50E-02
X-12094	0.01	0	7044	3.50E-02
X-13435	0.01	0	6543	3.60E-02
Cholesterol esters in large VLDL	-0.02	0.01	19273	3.90E-02
X-01911	0.01	0.01	6000	4.00E-02
Albumin	-0.02	0.01	18960	4.10E-02
Total cholesterol in IDL	-0.02	0.01	19273	4.20E-02
3-methylhistidine	0.02	0.01	5502	4.40E-02
Total lipids in large LDL	-0.02	0.01	19273	4.50E-02
X-12093	-0.01	0.01	2759	4.50E-02
X-12092	0.01	0.01	7089	4.70E-02
X-13671	-0.01	0	6883	4.90E-02
X-12040	0.03	0.01	528	4.90E-02
X-11537	-0.01	0.01	2684	5.00E-02
Alpha-hydroxyisovalerate	0.01	0	7222	5.00E-02
Aspartylphenylalanine	0.01	0.01	3838	5.20E-02
Pantothenate	0.01	0	7160	5.40E-02

Concentration of large LDL particles	-0.02	0.01	19273	5.40E-02
Xanthine	-0.01	0	6515	5.50E-02
X-11786–methylcysteine	0.01	0.01	2773	5.50E-02
Serum total cholesterol	-0.02	0.01	21491	5.60E-02
CD8:%Senescent	-0.04	0.02	460	5.60E-02
Kynurenine	0	0	7368	5.60E-02
Cholesterol esters in medium LDL	-0.02	0.01	19273	5.60E-02
Free cholesterol in IDL	-0.02	0.01	21559	5.90E-02
Major depressive disorder	-0.05	0.03	18759	5.90E-02
Total lipids in medium LDL	-0.02	0.01	19273	5.90E-02
CD32 on 11c+123+DC	-0.04	0.02	460	6.30E-02
Total lipids in IDL	-0.02	0.01	19273	6.70E-02
Concentration of medium LDL particles	-0.02	0.01	19273	6.80E-02
Autism	0.05	0.03	29415	7.00E-02
IgE+B:%27+20-38+	-0.04	0.02	455	7.00E-02
Mean cell volume	-0.05	0.03	30457	7.10E-02
CD4:%Act(38+)	0.04	0.02	459	7.10E-02
X-12216	-0.01	0	5080	7.10E-02
Phospholipids in IDL	-0.02	0.01	21559	7.30E-02
Free cholesterol to esterified cholesterol ratio	-0.02	0.01	13496	7.30E-02
CD8mem:%"TFH" (2)	0.04	0.02	447	7.40E-02
Caproate (6:0)	0	0	7363	7.40E-02
Mono:%11c-16-274+	-0.04	0.02	458	7.40E-02
Transferrin	0.02	0.01	23986	7.70E-02
2-hydroxystearate	0	0	7315	7.80E-02
Total cholesterol in small LDL	-0.02	0.01	21556	7.80E-02
X-11327	0	0	7229	8.00E-02
Hexadecanedioate	0.01	0	6447	8.00E-02
CD24 on IgA+ B	-0.04	0.02	456	8.10E-02
Insulin disposition index	0.04	0.02	5318	8.20E-02

Concentration of IDL particles	-0.02	0.01	19273	8.20E-02
Malate	0	0	6946	8.30E-02
X-11793–oxidized bilirubin*	-0.01	0	7166	8.50E-02
Rheumatoid arthritis	-0.04	0.03	22515	8.60E-02
CD8:%Naive	0.04	0.02	455	8.60E-02
Forearm bone mineral density	-0.03	0.02	10805	8.90E-02
Lymph:%DP T	-0.04	0.02	453	9.00E-02
CD4:%Eff(127-PD1-)	-0.04	0.02	459	9.20E-02
X-12253	-0.01	0.01	5694	9.70E-02
Leucylleucine	-0.01	0	3248	1.00E-01
X-11787	0	0	7362	1.03E-01
Myo-inositol	0	0	7354	1.05E-01
Mature B:%Memory	-0.03	0.02	458	1.07E-01
X-03094	0	0	7356	1.10E-01
Biliverdin	-0.01	0	6252	1.11E-01
X-06267	0	0	6678	1.13E-01
Alpha-tocopherol	0.01	0	7276	1.14E-01
Free cholesterol in small VLDL	-0.02	0.01	21559	1.15E-01
1-arachidonoylglycerophosphoinositol*	0	0	7351	1.18E-01
X-12116	-0.01	0.01	2859	1.18E-01
X-12095–N1-methyl-3-pyridone-4-carboxamide	0	0	7269	1.20E-01
Gamma-glutamylvaline	0	0	7307	1.21E-01
X-11444	0.01	0	7313	1.23E-01
Packed cell volume	-0.03	0.02	31255	1.23E-01
Histidine	0	0	7355	1.23E-01
Gamma-glutamylglutamine	0	0	7226	1.23E-01
X-12443	0.01	0.01	4999	1.27E-01
Cholesterol esters in medium VLDL	-0.02	0.01	19273	1.27E-01
3-indoxyl sulfate	-0.01	0	7339	1.28E-01
Dodecanedioate	0.01	0	6066	1.30E-01

Total cholesterol in medium VLDL	-0.02	0.01	21551	1.37E-01
Gamma-glutamylphenylalanine	0	0	7306	1.38E-01
X-11470	0.01	0	6857	1.38E-01
Total cholesterol in small VLDL	-0.02	0.01	17896	1.39E-01
Extreme height	0.04	0.03	15661	1.40E-01
Total lipids in small LDL	-0.02	0.01	19273	1.41E-01
Citrulline	0	0	7325	1.46E-01
X-12013	0.01	0.01	1863	1.46E-01
X-06351	0.01	0	4460	1.47E-01
Indolelactate	0	0	6939	1.51E-01
Ursodeoxycholate	-0.01	0.01	5162	1.51E-01
Phenyllactate (PLA)	0	0	5702	1.53E-01
Bilirubin (Z,Z)	-0.01	0.01	6393	1.54E-01
1-oleoylglycerophosphocholine	0	0	7364	1.54E-01
NKeff:%Act(335+)	0.03	0.02	460	1.55E-01
Succinylcarnitine	0	0	6552	1.56E-01
X-11540	-0.01	0.01	2656	1.56E-01
Sphingomyelins	-0.02	0.01	13476	1.56E-01
X-12456	0.01	0	4533	1.57E-01
NKeff:%2+337-R7	0.03	0.02	457	1.57E-01
DNT:%CD127+TEF	0.03	0.02	393	1.61E-01
Phospholipids in very small VLDL	-0.02	0.01	19273	1.61E-01
Lymph:%CD4	0.03	0.02	459	1.62E-01
CD4:%Act(DR+)	-0.03	0.02	455	1.62E-01
Corrected insulin response	0.03	0.02	5318	1.62E-01
2-methylbutyroylcarnitine	0	0	7007	1.64E-01
Mean cell haemoglobin	-0.01	0.01	28955	1.66E-01
Tyrosine	0	0	7358	1.67E-01
CD8mem:%"Th2"	-0.03	0.02	445	1.68E-01
CD8mem:%"Th17"	0.03	0.02	444	1.68E-01

Pyruvate	0	0	7237	1.71E-01
Inosine	-0.02	0.01	2639	1.72E-01
Docosahexaenoate (DHA; 22:6n3)	0	0	7369	1.76E-01
CD8mem:%PD1+R6+	0.03	0.02	444	1.77E-01
Schizophrenia	-0.01	0.01	82315	1.78E-01
X-14588	0	0	7330	1.79E-01
Transferrin Saturation	0.01	0.01	23986	1.80E-01
18:2, linoleic acid (LA)	-0.02	0.01	13527	1.80E-01
X-12627	-0.01	0	6987	1.81E-01
X-02973	0	0	7314	1.83E-01
2-tetradecenoyl carnitine	-0.01	0	6595	1.83E-01
Alanine	0	0	7344	1.84E-01
CD8:%TE	-0.03	0.02	460	1.85E-01
Phospholipids in small VLDL	-0.01	0.01	21551	1.86E-01
Concentration of small LDL particles	-0.01	0.01	19273	1.86E-01
Omega-6 fatty acids	-0.02	0.01	12139	1.88E-01
Internalizing problems	0.03	0.02	4596	1.89E-01
2-choice reaction time	-0.04	0.03	2602	1.90E-01
Insulin at 30 minutes	-0.03	0.02	5318	1.90E-01
CXCR3 on CD4n	-0.03	0.02	448	1.92E-01
CD8mem:%"TFH" (3)	0.03	0.02	446	1.93E-01
X-12441–12-hydroxyeicosatetraenoate (12-HETE)	0.01	0.01	2725	1.93E-01
3-hydroxybutyrate (BHBA)	0.01	0.01	7371	1.93E-01
Ascorbate (Vitamin C)	-0.01	0.01	2063	1.94E-01
Choline	0	0	7310	1.95E-01
X-04494	0	0	4689	1.95E-01
Gamma-glutamylthreonine*	0	0	3890	1.95E-01
NKT:%4+R5-	0.03	0.02	458	1.97E-01
CD8:%TM (1)	0.03	0.02	459	1.98E-01
Tryptophan betaine	0.01	0.01	7014	1.98E-01

Bipolar disorder	-0.03	0.02	16731	1.99E-01
Haemoglobin concentration	-0.01	0.01	33429	1.99E-01
X-12645	0.01	0	4908	1.99E-01
X-12007	0.01	0.01	1273	2.02E-01
Neuroblastoma	0.06	0.04	4881	2.03E-01
CD4Nv:%TFH (3)	0.03	0.02	448	2.07E-01
Cortisol	0	0	7346	2.07E-01
X-11521	0	0	6753	2.08E-01
Phenylalanine	0	0	7354	2.09E-01
X-13215	0	0	5885	2.10E-01
Type 2 diabetes	0.02	0.02	95272	2.10E-01
Total lipids in small VLDL	-0.01	0.01	19273	2.12E-01
CD123 on pDC	0.03	0.02	460	2.12E-01
CD8:%TM-like	0.03	0.02	454	2.13E-01
Parkinson's disease	0.05	0.04	5691	2.14E-01
Pseudouridine	0	0	7336	2.18E-01
Total fatty acids	-0.02	0.01	13505	2.18E-01
Serum cystatin C (eGFRcys)	0	0	33140	2.20E-01
Mean diameter for LDL particles	-0.01	0.01	19273	2.21E-01
22:6, docosahexaenoic acid	-0.02	0.01	13499	2.21E-01
X-12729	0.01	0.01	1649	2.23E-01
X-12734	0.01	0.01	5270	2.24E-01
CD8:%TM (2)	0.03	0.02	458	2.25E-01
CD8:%RTE	0.03	0.02	460	2.27E-01
Triglycerides in small VLDL	-0.01	0.01	21558	2.29E-01
CD4:%Treg(73+)	-0.03	0.02	457	2.29E-01
1-palmitoylglycerol (1-monopalmitin)	0	0	6999	2.30E-01
MDC:%X-Presenting	-0.02	0.02	459	2.30E-01
Stearoylcarnitine	0	0	6768	2.30E-01
Isovalerylcarnitine	0	0	7344	2.32E-01

Hexanoylcarnitine	0	0	7340	2.32E-01
X-14977–vanillin	-0.01	0	1782	2.33E-01
Bilirubin (E,Z or Z,E)*	-0.01	0.01	4961	2.33E-01
2-hydroxybutyrate (AHB)	0	0	7366	2.33E-01
X-12465	-0.01	0	5534	2.34E-01
CD27 on IgG+B	-0.02	0.02	459	2.35E-01
Oligoclonal band status	-0.1	0.08	3026	2.37E-01
HbA1C	0	0	46368	2.39E-01
X-11483	0.01	0.01	4274	2.39E-01
Mothers age at death	-0.01	0.01	75244	2.40E-01
Omega-9 and saturated fatty acids	-0.02	0.01	13506	2.40E-01
CD8:%Act(25+)	-0.02	0.02	455	2.42E-01
X-12990–docosapentaenoic acid (n6-DPA)	0.01	0.01	2545	2.43E-01
Total lipids in medium VLDL	-0.01	0.01	19273	2.43E-01
CD8mem:%R6+PD1+161-	0.02	0.02	447	2.43E-01
Concentration of very small VLDL particles	-0.01	0.01	19273	2.44E-01
X-10395	0	0	7340	2.45E-01
1-oleoylglycerol (1-monoolein)	-0.01	0.01	5436	2.45E-01
Cortisone	0	0	7136	2.46E-01
Erythrose	0	0	6830	2.46E-01
1-arachidonoylglycerophosphocholine*	0	0	7063	2.46E-01
HDL cholesterol	0.01	0.01	94311	2.46E-01
CD4:%Treg(39-73+)	-0.02	0.02	457	2.48E-01
Subjective well being	0	0	298420	2.50E-01
Palmitoyl sphingomyelin	0	0	7366	2.53E-01
Concentration of small VLDL particles	-0.01	0.01	19273	2.54E-01
T:%NKT	-0.02	0.02	452	2.56E-01
CD123 on 11c+123+DC	0.02	0.02	460	2.56E-01
Total lipids in very small VLDL	-0.01	0.01	19273	2.59E-01
Asparagine	0	0	7316	2.61E-01

2-oleoylglycerophosphocholine*	0	0	7100	2.61E-01
X-11805	0.01	0.01	2777	2.61E-01
Infant head circumference	0.02	0.01	10716	2.65E-01
3-(4-hydroxyphenyl)lactate	0	0	7346	2.65E-01
X-06246	0	0	6929	2.66E-01
Phosphatidylcholine and other cholines	-0.01	0.01	13542	2.66E-01
Total phosphoglycerides	-0.01	0.01	13519	2.67E-01
Glutamate	0	0	7357	2.68E-01
2-hydroxyglutarate	0	0	5932	2.68E-01
X-14541	0.01	0.01	1908	2.69E-01
4-choice reaction time	-0.03	0.03	2829	2.70E-01
X-13549	0	0	6920	2.71E-01
IgG+B:%20+24+27+	0.02	0.02	457	2.72E-01
Lactate	0	0	7365	2.73E-01
MDC (2):%CD1c+	-0.02	0.02	460	2.73E-01
MDC (2):%CD1c-	0.02	0.02	460	2.76E-01
X-12644	0	0	7347	2.77E-01
Octadecanedioate	0	0	6876	2.78E-01
Triglycerides in small HDL	-0.01	0.01	21558	2.82E-01
Ornithine	0	0	7307	2.82E-01
HWESASXX*	-0.01	0.01	7255	2.83E-01
CD4:%Naive	0.02	0.02	460	2.85E-01
Octanoylcarnitine	0	0	7345	2.85E-01
X-10506	0	0	7268	2.85E-01
NKeff:%337+	0.02	0.02	457	2.87E-01
Pipecolate	0	0	7345	2.88E-01
X-12786	0	0	6144	2.88E-01
Gamma-glutamyltyrosine	0	0	7037	2.90E-01
Glycoproteins	-0.01	0.01	16507	2.90E-01
X-13477	0	0	5694	2.90E-01

X-14658	0.01	0.01	3346	2.94E-01
EarlyB:%27+38-	-0.02	0.02	455	2.95E-01
3-carboxy-4-methyl-5-propyl-2-furanpropanoate (CMPF)	0.01	0.01	7363	2.96E-01
Cholate	0.01	0.01	5599	2.96E-01
Arachidonate (20:4n6)	0	0	7367	2.98E-01
Propionylcarnitine	0	0	7364	2.99E-01
5alpha-androstan-3beta,17beta-diol disulfate	0.01	0.01	6934	3.02E-01
CD4:%Act(PD1+)	-0.02	0.02	457	3.02E-01
X-10510	0	0	7347	3.03E-01
X-14189–leucylalanine	-0.01	0.01	2709	3.05E-01
Ischaemic stroke	-0.25	0.33	546	3.05E-01
X-12217	0	0	6109	3.06E-01
X-12749	0	0	6755	3.08E-01
X-11442	0	0	6712	3.09E-01
Insulin sensitivity index	0.02	0.02	5318	3.09E-01
Indolepropionate	0	0	7358	3.10E-01
IgG+B:%27+	0.02	0.02	459	3.11E-01
Ferritin	-0.01	0.01	23986	3.13E-01
CD4:%TM (1)	-0.02	0.02	458	3.15E-01
NKeff:%Act(314+)	0.02	0.02	454	3.16E-01
X-11261	0	0	7322	3.18E-01
Neuroticism	-0.01	0.01	160958	3.18E-01
1-stearoylglycerophosphoinositol	0	0	7251	3.19E-01
CD4:%Treg(39+73-)	-0.02	0.02	457	3.20E-01
LDL cholesterol	0.01	0.01	89888	3.21E-01
X-14056	0	0	6901	3.21E-01
CD4:%Treg(39+)	-0.02	0.02	459	3.22E-01
Homostachydrine*	0.01	0.01	1618	3.22E-01
CD8:%Act(25+38+)	-0.02	0.02	455	3.26E-01
MDCL%274+	-0.02	0.02	460	3.27E-01

Glycoprotein acetyls	-0.01	0.01	19270	3.28E-01
X-11334	0	0	5211	3.29E-01
Body mass index	0	0	236164	3.30E-01
Cis-4-decenoyl carnitine	0	0	7219	3.30E-01
Pyroglutamine*	0	0	7354	3.32E-01
X-14745	0	0	5292	3.34E-01
X-11799	-0.01	0.01	4817	3.35E-01
Amyotrophic lateral sclerosis	0	0	36052	3.35E-01
Pyridoxate	0	0	7263	3.36E-01
Acetate	-0.01	0.01	22521	3.36E-01
Nonadecanoate (19:0)	0	0	7335	3.38E-01
Glutamine	0	0	7372	3.38E-01
Glycerate	0	0	7333	3.42E-01
1-stearoylglycerophosphocholine	0	0	7368	3.45E-01
CD4:%TM (2)	-0.02	0.02	461	3.46E-01
Intracranial volume	1900.75	2018.1	11373	3.46E-01
ADpSGEGDFXAEGGGVR*	-0.01	0.01	3845	3.48E-01
Mean platelet volume	0	0	16284	3.48E-01
CD8mem:%"Th1*"	0.02	0.02	442	3.48E-01
CD8:%Act(25+38+RO+)	-0.02	0.02	453	3.48E-01
CD4:%SCM	-0.02	0.02	458	3.49E-01
Top 1 % survival	0	0	75244	3.50E-01
EarlyB:%27+	-0.02	0.02	456	3.50E-01
Neo-extraversion	0.08	0.08	17375	3.51E-01
Gamma-glutamylisoleucine*	0	0	5207	3.53E-01
Gamma-glutamylmethionine*	0	0	2212	3.53E-01
NKearly:%337+335+2-	0.02	0.02	454	3.54E-01
Guanosine	0.01	0.01	2315	3.54E-01
X-13496	0	0	7217	3.57E-01
Fructose	0	0	7333	3.60E-01

Plasma cortisol	0.01	0.01	12589	3.61E-01
CD4:%Treg(39+73+)	-0.02	0.02	455	3.63E-01
Pyroglutamylglycine	0.01	0.01	1558	3.64E-01
Hydroxyisovaleroyl carnitine	0	0	5306	3.64E-01
Total cholesterol in HDL	-0.01	0.01	21555	3.66E-01
10-nonadecenoate (19:1n9)	0	0	7345	3.67E-01
X-12776	0	0	7317	3.68E-01
X-12850	0	0.01	5856	3.68E-01
Homocitrulline	0	0	3950	3.69E-01
CD4nv:%preTh17	0.02	0.02	448	3.69E-01
X-09706	0	0	7256	3.72E-01
Apolipoprotein A-I	-0.01	0.01	20686	3.72E-01
Heart rate	0.05	0.05	87453	3.73E-01
Isoleucine	0	0	7352	3.77E-01
X-14304–leucylalanine	-0.01	0.01	2400	3.78E-01
Other polyunsaturated fatty acids than 18:2	-0.01	0.01	13549	3.79E-01
Docosapentaenoate (n3 DPA; 22:5n3)	0	0	7373	3.81E-01
1-palmitoylglycerophosphoinositol*	0	0	5979	3.81E-01
Heptanoate (7:0)	0	0	7355	3.83E-01
Oleoylcarnitine	0	0	7263	3.84E-01
Total lipids in small HDL	-0.01	0.01	19273	3.84E-01
Pro-hydroxy-pro	0	0	7344	3.89E-01
1-stearoylglycerophosphoethanolamine	0	0	6929	3.90E-01
X-04357	0	0	7064	3.92E-01
X-11550	0	0	7329	3.95E-01
X-05907	0	0	7286	3.95E-01
X-06226	0	0	7182	3.96E-01
Paget's disease	-0.05	0.06	3440	3.97E-01
X-12244–N-acetylcarnosine	0	0	6279	4.00E-01
Triglycerides in very small VLDL	-0.01	0.01	19273	4.00E-01

Triglycerides in IDL	-0.01	0.01	19273	4.01E-01
X-11905	0	0	4409	4.02E-01
X-08988	0	0	7329	4.08E-01
Weight	0.01	0.01	58315	4.10E-01
Carnitine	0	0	7349	4.11E-01
X-09108	0	0	6517	4.13E-01
X-14208–phenylalanylserine	-0.01	0.01	2432	4.14E-01
Indoleacetate	0	0	7180	4.17E-01
Phospholipids in medium VLDL	-0.01	0.01	21240	4.17E-01
X-11820	0	0	7263	4.20E-01
Triglycerides in chylomicrons and largest VLDL particles	0.01	0.01	21540	4.23E-01
Lymph:%CD8	-0.02	0.02	458	4.25E-01
Tryptophan	0	0	7355	4.25E-01
Mono-unsaturated fatty acids	-0.01	0.01	13535	4.28E-01
Pallidum volume	-1.58	2	13142	4.28E-01
4-androsten-3beta,17beta-diol disulfate 2*	0	0	7333	4.29E-01
X-10429	0	0	6388	4.30E-01
X-11876	0	0	4665	4.31E-01
Caprylate (8:0)	0	0	7355	4.32E-01
B Mem:%IgA	0.02	0.02	460	4.32E-01
X-12704	0.01	0.01	1766	4.33E-01
Cholesterol esters in very large HDL	0.01	0.01	19273	4.34E-01
Stearidonate (18:4n3)	0	0	7330	4.36E-01
1-palmitoylglycerophosphoethanolamine	0	0	7317	4.38E-01
X-12428	0.01	0.01	1428	4.39E-01
Waist circumference	0	0	153943	4.40E-01
MDC:%CD1c-	0.02	0.02	460	4.41E-01
X-11452	0	0.01	6150	4.43E-01
IgA nephropathy	0.06	0.08	5983	4.46E-01
X-12188	-0.01	0.01	824	4.47E-01

CD8mem:%"TFH" (1)	0.02	0.02	448	4.48E-01
Mannose	0	0	7345	4.49E-01
Triglycerides in very large HDL	-0.01	0.01	21536	4.54E-01
Cholesterol	0	0	7365	4.55E-01
X-11423–O-sulfo-L-tyrosine	0	0	7318	4.55E-01
Valerate	0	0	4004	4.56E-01
X-02249	0	0	7369	4.57E-01
Ratio of bisallylic groups to double bonds	0.01	0.01	13524	4.58E-01
Difference in height between childhood and adulthood	-0.01	0.02	5748	4.59E-01
Hip circumference	0	0	143204	4.60E-01
N2,N2-dimethylguanosine	0	0	4900	4.63E-01
7-methylguanine	0	0	5804	4.64E-01
X-12719	0.01	0.01	1492	4.66E-01
C-glycosyltryptophan*	0	0	7338	4.67E-01
Pancreatic cancer	-0.04	0.05	3835	4.70E-01
T:%Vd1	-0.02	0.02	456	4.70E-01
MDC:%CD1c+	-0.02	0.02	460	4.71E-01
NKeff:%314-R7-	-0.02	0.02	459	4.72E-01
X-12771	0	0	4442	4.72E-01
Erythronate*	0	0	7307	4.72E-01
Platelet count	0.34	0.47	66867	4.73E-01
1-linoleoylglycerophosphoethanolamine*	0	0	7369	4.73E-01
Deoxycholate	0	0.01	4934	4.74E-01
Free cholesterol in large HDL	-0.01	0.01	21559	4.76E-01
Taurochenodeoxycholate	0	0.01	5376	4.76E-01
Lysine	0	0	7363	4.77E-01
X-04495	0	0	7055	4.78E-01
Free cholesterol in medium VLDL	-0.01	0.01	21240	4.78E-01
Obesity class 1	0.01	0.01	98692	4.80E-01
15-methylpalmitate (isobar with 2-methylpalmitate)	0	0	6942	4.82E-01

Tetradecanedioate	0	0	5629	4.82E-01
Threitol	0	0	6960	4.88E-01
CD8:%TEM	-0.01	0.02	453	4.91E-01
X-11546	-0.01	0.01	1988	4.93E-01
X-11847	-0.01	0.01	5716	4.93E-01
Glycylvaline	0.01	0.01	2079	4.95E-01
X-14626	0	0	6464	4.98E-01
Caudate volume	3.48	5.16	13171	5.00E-01
Coronary heart disease	-0.01	0.02	30415	5.00E-01
CD8mem:%"pre-Th17" (2)	0.01	0.02	446	5.02E-01
3-(3-hydroxyphenyl)propionate	0.01	0.01	1098	5.02E-01
X-11374	0	0.01	2575	5.07E-01
X-11491	0	0.01	6200	5.10E-01
Serum total triglycerides	-0.01	0.01	21545	5.10E-01
X-12844	0	0	7322	5.13E-01
Phenylalanylphenylalanine	0	0	4754	5.13E-01
X-11445–5-alpha-pregnan-3beta,20alpha-disulfate	0.01	0.01	2537	5.13E-01
G speed factor	-0.02	0.03	2430	5.16E-01
X-03088	0	0	7034	5.16E-01
Concentration of small HDL particles	-0.01	0.01	19273	5.16E-01
Hypoxanthine	0	0	6941	5.17E-01
CD56 on 2+NK eff	0.01	0.02	459	5.17E-01
CD4:%Act(PD1+127+39-73-)	-0.01	0.02	457	5.20E-01
Squamous cell lung cancer	0.02	0.03	18313	5.23E-01
2-aminobutyrate	0	0	7365	5.24E-01
X-14625	0	0	7092	5.25E-01
X-12407	0.01	0.01	2660	5.29E-01
CD4mem:R6+	0.01	0.02	442	5.29E-01
Pentadecanoate (15:0)	0	0	7063	5.29E-01
Threonate	0	0	7336	5.29E-01

Eczema	-0.01	0.02	40531	5.30E-01
CD8mem:%"Th1"	0.01	0.02	448	5.31E-01
Methionine	0	0	7347	5.32E-01
Adrenate (22:4n6)	0	0	7330	5.32E-01
Incremental insulin at 30 minutes	-0.02	0.02	5318	5.35E-01
Cysteine	0	0	7257	5.36E-01
Lung cancer	0.01	0.02	27209	5.37E-01
Neo-neuroticism	0.06	0.1	17375	5.40E-01
Microalbuminuria	-0.01	0.02	54115	5.40E-01
Urinary albumin-to-creatinine ratio	0.02	0.03	5825	5.40E-01
AUCins/AUCglu	-0.02	0.02	5318	5.43E-01
Concentration of large HDL particles	-0.01	0.01	19273	5.49E-01
Arabinose	0	0	5502	5.50E-01
N-acetylalanine	0	0	7271	5.52E-01
Glycodeoxycholate	-0.01	0.01	1457	5.53E-01
Creatine	0	0	7373	5.54E-01
Cysteine-glutathione disulfide	0	0.01	1982	5.56E-01
CD39 on CD4 T	-0.01	0.02	460	5.56E-01
P-cresol sulfate	0	0.01	7311	5.57E-01
Omega-3 fatty acids	-0.01	0.01	13544	5.60E-01
1-arachidonoylglycerophosphoethanolamine*	0	0	7350	5.60E-01
DPT:%Exhausted	0.01	0.02	459	5.61E-01
Beta-hydroxyisovalerate	0	0	6864	5.64E-01
Stearate (18:0)	0	0	7355	5.66E-01
Palmitoylcarnitine	0	0	7259	5.72E-01
Phospholipids in chylomicrons and largest VLDL particles	0.01	0.01	21542	5.72E-01
CD4mem:%PD1-R6+	0.01	0.02	446	5.73E-01
5-oxoproline	0	0	7354	5.74E-01
Glucose	0	0	7325	5.76E-01
Total lipids in large VLDL	-0.01	0.01	18960	5.76E-01

Total cholesterol in large HDL	-0.01	0.01	21558	5.77E-01
X-11497	0	0	7063	5.78E-01
X-13553	0	0	3007	5.78E-01
Total lipids in large HDL	-0.01	0.01	19273	5.80E-01
X-08402	0	0	7284	5.81E-01
X-11792	-0.01	0.01	2391	5.81E-01
CD8:%DR+73+39-25-	0.01	0.02	455	5.81E-01
4-hydroxyhippurate	0	0.01	1291	5.83E-01
GdT:%CM (2)	0.01	0.02	453	5.84E-01
Eicosapentaenoate (EPA; 20:5n3)	0	0	7367	5.84E-01
Nucleus accumbens volume	0.63	1.16	13112	5.87E-01
CD8:%SCM	0.01	0.02	459	5.88E-01
CD27 on CD8 T	-0.01	0.02	460	5.90E-01
Obesity class 3	0.02	0.04	44810	5.90E-01
NKeff:%314-158a+	-0.01	0.02	461	5.91E-01
Laurate (12:0)	0	0	7346	5.93E-01
Ever vs never smoked	0.01	0.01	74035	5.95E-01
Fasting glucose	0	0	46186	5.97E-01
X-11422–xanthine	0	0	6042	5.97E-01
Lathosterol	0	0	5107	5.98E-01
Lymph:%NK	-0.01	0.02	460	5.99E-01
Fathers age at death	0	0	75244	6.00E-01
X-13548	0	0	5768	6.02E-01
X-12329	-0.01	0.01	831	6.04E-01
X-13619	0	0	7345	6.04E-01
Concentration of medium VLDL particles	-0.01	0.01	19273	6.04E-01
Total cholesterol in very large HDL	0.01	0.01	21540	6.05E-01
X-12855	0	0	4800	6.05E-01
X-13431–nonanoylcarnitine*	0	0	6234	6.06E-01
X-12728	-0.01	0.02	535	6.06E-01

CD27 on CD4 T	-0.01	0.02	460	6.07E-01
Phosphate	0	0	7341	6.08E-01
Phenylacetylglutamine	0	0.01	7364	6.12E-01
X-12442–5,8-tetradecadienoate	0	0	7334	6.15E-01
CD8:%39+	0.01	0.02	455	6.15E-01
X-11438	0	0	6658	6.17E-01
X-11850	0	0.01	4625	6.18E-01
Parents' age at death	0	0.01	75244	6.20E-01
X-12830	0	0.01	3239	6.22E-01
X-12231	0	0.01	4807	6.24E-01
X-12206	0	0	2757	6.24E-01
Myristoleate (14:1n5)	0	0	7355	6.26E-01
X-12236	0	0.01	1860	6.27E-01
Age of smoking initiation	0	0	47961	6.29E-01
Palmitate (16:0)	0	0	7352	6.29E-01
8-choice reaction time	0.02	0.04	1382	6.30E-01
Sleep duration	0	0	128266	6.30E-01
Glycine —	0	0	7356	6.32E-01
X-12740	0	0.01	3491	6.32E-01
Phospholipids in large HDL	-0.01	0.01	19273	6.32E-01
X-11204	0	0	7350	6.33E-01
Percent emphysema	-0.01	0.02	7667	6.34E-01
X-11247	0	0.01	6970	6.34E-01
1-myristoylglycerophosphocholine	0	0	7364	6.35E-01
X-11485	0	0.01	4311	6.35E-01
Ergothioneine	0	0.01	1805	6.35E-01
1-eicosadienoylglycerophosphocholine*	0	0	6506	6.38E-01
X-11552	0	0.01	2148	6.39E-01
Triglycerides in medium VLDL	0	0.01	21241	6.40E-01
IgE+B:%27+	0.01	0.02	459	6.40E-01

X-14205–alpha-glutamyltyrosine	0	0.01	1780	6.44E-01
NKearly:%337+	-0.01	0.02	458	6.47E-01
3-dehydrocarnitine*	0	0	7361	6.47E-01
X-12100–hydroxytryptophan*	0	0	7066	6.48E-01
Creatinine	0	0	7361	6.48E-01
CD4mem:%Th17	0.01	0.02	441	6.49E-01
X-11538	0	0	7355	6.51E-01
X-12816	0	0.01	4296	6.51E-01
CD8:%CM	0.01	0.02	458	6.51E-01
X-13429	0	0.01	5953	6.51E-01
CD3 on CD8 T	-0.01	0.02	458	6.52E-01
Heme*	0	0	5955	6.54E-01
X-05426	0	0.01	5984	6.55E-01
X-10810	0	0	6795	6.55E-01
Acetylphosphate	0	0	7345	6.56E-01
CD4mem:%cFTH (2)	0.01	0.02	447	6.56E-01
Gamma-glutamylglutamate	0	0.01	922	6.57E-01
Butyrylcarnitine	0	0	7349	6.59E-01
Acetylcarnitine	0	0	7356	6.59E-01
CD3 on 127+DN T	0.01	0.02	397	6.60E-01
CD4:%Exhausted	-0.01	0.02	456	6.61E-01
Selenium	0.01	0.03	2874	6.62E-01
X-12212	0	0.01	4323	6.62E-01
10-undecenoate (11:1n1)	0	0	7358	6.62E-01
Isovalerate	0	0	6735	6.64E-01
Multiple sclerosis	-0.03	0.07	1861	6.64E-01
Ulcerative colitis	0.01	0.02	27432	6.67E-01
X-11469	0	0.01	7330	6.68E-01
Triglycerides in very large VLDL	0	0.01	21548	6.68E-01
2-linoleoylglycerophosphocholine*	0	0	6538	6.71E-01

CD4nv:%Th1	0.01	0.02	447	6.71E-01
Trans-4-hydroxyproline	0	0	7356	6.72E-01
2-palmitoylglycerophosphocholine*	0	0	7261	6.73E-01
3-hydroxybutyrate	0	0.01	18712	6.73E-01
X-12712	-0.01	0.01	248	6.77E-01
Mannitol	0	0.01	5606	6.79E-01
MDC:%64-274-	0.01	0.02	460	6.80E-01
Years of schooling	0	0	106736	6.81E-01
Free cholesterol in medium HDL	0	0.01	21559	6.84E-01
Phospholipids in large VLDL	0	0.01	21239	6.84E-01
CD27 on IgA+B	-0.01	0.02	459	6.85E-01
2-stearoylglycerophosphocholine*	0	0	7291	6.86E-01
2hr glucose	-0.01	0.02	15234	6.86E-01
CD4mem:%Th22	0.01	0.02	442	6.86E-01
Serine	0	0	7349	6.87E-01
Glycochenodeoxycholate	0	0.01	6687	6.89E-01
CD4nv:%Th2	-0.01	0.02	448	6.93E-01
X-12696	0	0	6994	6.94E-01
CD4nv:%TFH (1)	0.01	0.02	445	6.94E-01
Glycerol	0	0	7352	6.95E-01
College completion	0	0.01	126559	6.96E-01
CD8mem:%CD31+	0.01	0.02	459	6.98E-01
Dihomo-linolenate (20:3n3 or n6)	0	0	7358	6.98E-01
NKeff:%Kir+ (4)	0.01	0.02	456	7.02E-01
Leptin	0	0.01	30192	7.02E-01
X-06350	0	0	4844	7.03E-01
X-04499–3,4-dihydroxybutyrate	0	0	6549	7.04E-01
Anorexia nervosa	-0.01	0.03	17767	7.05E-01
Age at menarche	0	0.01	182416	7.10E-01
CD4mem:%preTh17 (3)	-0.01	0.02	451	7.10E-01

Myristate (14:0)	0	0	7362	7.10E-01
Pelargonate (9:0)	0	0	7356	7.11E-01
X-11299	0	0.01	6904	7.14E-01
1,6-anhydroglucose	0	0.01	3422	7.14E-01
X-04498	0	0	6733	7.15E-01
MDC:%32+	-0.01	0.02	460	7.15E-01
X-11440	0	0.01	7250	7.16E-01
C-reactive protein	-0.01	0.01	10112	7.19E-01
Chronotype	0	0	128266	7.20E-01
Extreme body mass index	0.01	0.03	15591	7.20E-01
Thalamus volume	-2.34	6.57	13193	7.22E-01
X-12851	0	0.01	4515	7.23E-01
CD4mem:%Th1* (3)	0.01	0.02	453	7.23E-01
Triglycerides in large VLDL	0	0.01	21239	7.24E-01
AUCins	0.01	0.02	5318	7.25E-01
X-12510–2-aminooctanoic acid	0	0	7132	7.25E-01
1-palmitoylglycerophosphocholine	0	0	7354	7.25E-01
Decanoylcarnitine	0	0	7320	7.26E-01
X-12798	0	0	7119	7.26E-01
Linolenate [alpha or gamma; (18:3n3 or 6)]	0	0	7338	7.26E-01
Isobutyrylcarnitine	0	0	7365	7.27E-01
Chiro-inositol	0	0.01	2560	7.27E-01
3-methyl-2-oxobutyrate	0	0	7203	7.28E-01
1-palmitoleoylglycerophosphocholine*	0	0	7364	7.28E-01
Cholesterol esters in large HDL	0	0.01	19273	7.30E-01
1-docosahexaenoylglycerophosphocholine*	0	0	7350	7.33E-01
Glutaroyl carnitine	0	0	7256	7.33E-01
X-12435	-0.01	0.02	228	7.34E-01
3-phenylpropionate (hydrocinnamate)	0	0	5829	7.35E-01
Digit symbol	0.01	0.03	2956	7.36E-01

Glycerophosphorylcholine (GPC)	0	0	6782	7.42E-01
Total lipids in chylomicrons and largest VLDL particles	0	0.01	18960	7.45E-01
Free cholesterol in very large HDL	0	0.01	21542	7.47E-01
Dihomo-linoleate (20:2n6)	0	0	7353	7.48E-01
Extreme waist-to-hip ratio	0.01	0.03	9749	7.50E-01
Average number of methylene groups per double bond	0	0.01	13532	7.54E-01
Threonine	0	0	5687	7.55E-01
CD4:%Act(DR+38+)	-0.01	0.02	456	7.57E-01
Palmitoleate (16:1n7)	0	0	7327	7.58E-01
CD56 on NK eff	0.01	0.02	459	7.59E-01
3-methoxytyrosine	0	0	5656	7.60E-01
Levulinate (4-oxovalerate)	0	0	6535	7.62E-01
X-11593–O-methylascorbate*	0	0	7344	7.63E-01
Margarate (17:0)	0	0	7349	7.64E-01
Alzheimer's disease	0	0.02	54162	7.65E-01
NKearly:%337+335+R7-	0.01	0.02	457	7.66E-01
Hippocampus volume	-1.46	4.96	13163	7.69E-01
Undecanoate (11:0)	0	0	7081	7.69E-01
X-11317	0	0	7362	7.69E-01
X-12261	0	0.01	1027	7.70E-01
Uridine	0	0	7354	7.70E-01
NKearly:%337+158b+	-0.01	0.02	458	7.71E-01
Triglycerides	0	0	91013	7.72E-01
X-11843	0	0.01	3740	7.72E-01
X-06126	0	0.01	7337	7.74E-01
Phenylacetate	0	0	4538	7.75E-01
Description of average fatty acid chain length, not actual carbon number	0	0.01	13476	7.77E-01
Hyodeoxycholate	0	0.01	5673	7.78E-01
Linoleate (18:2n6)	0	0	7333	7.78E-01
X-12038	0	0	7315	7.79E-01

X-11845	0	0.01	2616	7.80E-01
NKeff:%Kir+ (2)	0.01	0.02	457	7.82E-01
X-11437	0	0.01	6379	7.83E-01
Total lipids in very large HDL	0	0.01	19273	7.86E-01
X-12029	0	0	7118	7.86E-01
X-12450	0	0	6004	7.89E-01
Serum creatinine (eGFRcrea)	0	0	15461	7.90E-01
X-12063	0	0.01	6797	7.91E-01
CD8:%R5+	0.01	0.02	460	7.91E-01
Eicosenoate (20:1n9 or 11)	0	0	7352	7.91E-01
CD4nv:%Th0	0.01	0.02	447	7.96E-01
X-11381	0	0	7308	7.97E-01
Fasting proinsulin	0	0.01	10701	7.97E-01
Sitting height ratio	0.01	0.03	3545	8.00E-01
CD32 on mDC	0.01	0.02	459	8.01E-01
Inspection time	0.01	0.03	2645	8.02E-01
CD4mem:%preTh17 (1)	0.01	0.02	447	8.02E-01
7-alpha-hydroxy-3-oxo-4-cholestenoate (7-Hoca)	0	0	7335	8.03E-01
CD8mem:%Th22	-0.01	0.02	443	8.03E-01
Gout	0.01	0.04	67730	8.06E-01
10-heptadecenoate (17:1n7)	0	0	7349	8.06E-01
X-10346	0	0.01	3962	8.06E-01
Total cholesterol in medium HDL	0	0.01	21558	8.07E-01
2-hydroxyisobutyrate	0	0	6180	8.07E-01
X-14374	0	0	7184	8.07E-01
X-11852	0	0.01	2860	8.11E-01
Cyclo(leu-pro)	0	0.01	4520	8.11E-01
X-11849	0	0.01	5064	8.12E-01
CD4:%CD244+ Nave	0	0.02	453	8.13E-01
4-androsten-3beta,17beta-diol disulfate 1*	0	0.01	7355	8.17E-01

N-Butyl Oleate	0	0	4317	8.17E-01
Gallbladder cancer	0.05	0.24	899	8.19E-01
CD4mem:%preTh17 (2)	0	0.02	448	8.21E-01
Taurolithocholate 3-sulfate	0	0.01	6511	8.22E-01
Ratio of bisallylic groups to total fatty acids	0	0.01	13171	8.23E-01
CD4:%Act(25+127+)	0	0.02	457	8.23E-01
Oleate (18:1n9)	0	0	7323	8.23E-01
X-08766	0	0	5676	8.25E-01
CD4:%RTE	0	0.02	454	8.25E-01
Gamma-tocopherol	0	0	5822	8.26E-01
Aspartate	0	0	7281	8.29E-01
Amygdala volume	-0.54	2.49	13160	8.29E-01
Chronic kidney disease	0	0.02	118145	8.30E-01
X-12189	0	0.01	470	8.31E-01
Asthma	0	0.02	26475	8.31E-01
X-12680	0	0	2278	8.31E-01
Neo-openness to experience	0.02	0.08	17375	8.32E-01
Adiponectin	0	0	29347	8.33E-01
PD1 on CD4mem	0	0.02	451	8.33E-01
1-stearoylglycerol (1-monostearin)	0	0	6556	8.34E-01
CD4nv:%TFH (2)	0	0.02	445	8.34E-01
Betaine	0	0	7357	8.35E-01
X-11858	0	0.01	3221	8.38E-01
X-11859	0	0	2683	8.38E-01
Average number of double bonds in a fatty acid chain	0	0.01	13501	8.39E-01
4-methyl-2-oxopentanoate	0	0	7329	8.41E-01
1-eicosatrienoylglycerophosphocholine*	0	0	7362	8.42E-01
X-13069	0	0	6381	8.42E-01
Urate	0	0	7371	8.48E-01
X-13859	0	0	6604	8.49E-01

Total cholesterol	0	0.01	94595	8.51E-01
X-07765	0	0.01	2091	8.51E-01
Mean diameter for VLDL particles	0	0.01	19273	8.54E-01
X-11529	0	0.01	6250	8.55E-01
Mean cell haemoglobin concentration	0	0	28821	8.56E-01
Taurodeoxycholate	0	0.01	1543	8.58E-01
X-13183–stearamide	0	0.01	2474	8.59E-01
Concentration of very large HDL particles	0	0.01	19273	8.61E-01
Zinc	0.01	0.03	2603	8.61E-01
NK:%Term	0	0.02	444	8.64E-01
X-14473	0	0	6486	8.64E-01
Average number of methylene groups in a fatty acid chain	0	0.01	16794	8.66E-01
X-11315	0	0	7337	8.70E-01
Red blood cell count	0	0	30718	8.70E-01
X-10500	0	0	7303	8.71E-01
CD8mem:%"pre-Th17" (1)	0	0.02	445	8.77E-01
1-linoleoylglycerophosphocholine	0	0	7348	8.77E-01
Alpha-ketoglutarate	0	0	5717	8.77E-01
PGC cross-disorder traits	0	0.01	61220	8.77E-01
Estrone 3-sulfate	0	0.01	917	8.78E-01
X-12230	0	0.01	5328	8.79E-01
Age at menopause	0	0.02	69360	8.80E-01
Height	0	0	251418	8.80E-01
Acetoacetate	0	0.01	19262	8.80E-01
X-03003	0	0	7258	8.80E-01
Alcohol dependence	-0.01	0.05	3829	8.81E-01
Concentration of very large VLDL particles	0	0.01	18252	8.81E-01
NK:%Eff	0	0.02	458	8.84E-01
Epiandrosterone sulfate	0	0.01	7326	8.86E-01
NKeff:%Kir (1)	0	0.02	457	8.87E-01

N-acetylthreonine	0	0	6502	8.87E-01
CD4mem:%Th1* (4)	0	0.02	444	8.88E-01
CD4mem:%cFTH (1)	0	0.02	447	8.88E-01
T:%Vg9+Vd2-low	0	0.02	458	8.90E-01
1c+mDC:%32+	0	0.02	460	8.91E-01
X-12405	0	0	6327	8.91E-01
Glycocholate	0	0.01	5624	8.92E-01
NKT:%TM	0	0.02	412	8.93E-01
Neo-agreeableness	0.01	0.07	17375	8.94E-01
Total cholesterol in large VLDL	0	0.01	21235	8.96E-01
Cholesterol esters in medium HDL	0	0.01	19273	8.97E-01
Lung adenocarcinoma	0	0.03	18336	8.97E-01
Free cholesterol in large VLDL	0	0.01	21238	8.97E-01
Phospholipids in very large HDL	0	0.01	19273	8.97E-01
3-methyl-2-oxovalerate	0	0	7331	8.99E-01
CD4nv:%R6+R4-	0	0.02	449	9.01E-01
X-11412	0	0	6497	9.01E-01
Body fat	0	0.01	74337	9.02E-01
5-dodecenoate (12:1n7)	0	0	7322	9.03E-01
NKeff:%Kir+ (3)	0	0.02	455	9.04E-01
Copper	0	0.03	2603	9.04E-01
Birth length	0	0.01	22140	9.06E-01
Obesity class 2	0	0.02	71978	9.10E-01
Waist-to-hip ratio	0	0	144591	9.10E-01
Phospholipids in medium HDL	0	0.01	21558	9.11E-01
CD4mem:%Th9	0	0.02	450	9.13E-01
Depressive symptoms	0	0	161460	9.14E-01
Arginine	0	0	7104	9.15E-01
Childhood obesity	0	0.03	13848	9.16E-01
N-acetylglycine	0	0	6715	9.17E-01

Total lipids in very large VLDL	0	0.01	19273	9.18E-01
Lymph:%T	0	0.02	460	9.21E-01
1-heptadecanoylglycerophosphocholine	0	0	6998	9.23E-01
X-12726	0	0	5081	9.24E-01
CD4mem:%Th1* (1)	0	0.02	453	9.26E-01
Phospholipids in very large VLDL	0	0.01	21237	9.27E-01
Serotonin (5HT)	0	0	5791	9.27E-01
Overweight	0	0.01	158848	9.30E-01
Glycerol 3-phosphate (G3P)	0	0	7337	9.33E-01
X-13658	0	0.01	1383	9.34E-01
N-(2-furoyl)glycine	0	0.01	522	9.34E-01
Total lipids in medium HDL	0	0.01	19273	9.34E-01
Bulimia nervosa	0	0.01	2442	9.35E-01
X-12847	0	0.01	3813	9.36E-01
CD4mem:%Th1* (5)	0	0.02	447	9.37E-01
Neo-conscientiousness	0.01	0.08	17375	9.37E-01
Celiac disease	0	0.03	15283	9.38E-01
CD4mem:%R6+	0	0.02	445	9.39E-01
Concentration of large VLDL particles	0	0.01	18960	9.40E-01
1-linoleoylglycerol (1-monolinolein)	0	0.01	2759	9.41E-01
X-14632	0	0.01	1636	9.43E-01
Phenol sulfate	0	0.01	7361	9.43E-01
Former vs current smoker	0	0.02	70675	9.44E-01
X-11441	0	0	6642	9.44E-01
CD161 on CD4mem	0	0.02	450	9.44E-01
Bradykinin, des-arg(9)	0	0.01	4452	9.50E-01
Putamen volume	0.4	6.32	13145	9.50E-01
CD4mem:%cFTH (3)	0	0.02	445	9.54E-01
4-acetamidobutanoate	0	0	6523	9.55E-01
1,5-anhydroglucitol (1,5-AG)	0	0	7301	9.56E-01

CD3 on DN T	0	0.02	397	9.58E-01
NKeff:%2-158a+	0	0.02	460	9.58E-01
X-09026	0	0	7184	9.60E-01
X-14057	0	0	4529	9.61E-01
X-06307	0	0	6385	9.61E-01
B:%Mature	0	0.02	456	9.63E-01
Melanoma	0	0.06	2829	9.65E-01
N1-methyladenosine	0	0	7365	9.65E-01
X-11818	0	0	6404	9.66E-01
Crohn's disease	0	0.04	5409	9.66E-01
X-02269	0	0.01	7256	9.67E-01
NKearly:%335+314-	0	0.02	461	9.69E-01
GdT:%CM (1)	0	0.02	454	9.70E-01
X-11795	0	0	7319	9.74E-01
Immunoglobulin G index levels	-6.81	207.43	938	9.74E-01
Dehydroisoandrosterone sulfate (DHEA-S)	0	0.01	7346	9.74E-01
Concentration of medium HDL particles	0	0.01	19273	9.74E-01
X-12524	0	0	7362	9.76E-01
X-12039	0	0.01	5185	9.77E-01
Inflammatory bowel disease	0	0.02	34652	9.77E-01
Scyllo-inositol	0	0	6119	9.77E-01
CD4mem:%Th1* (2)	0	0.02	443	9.79E-01
X-14662	0	0.01	3977	9.80E-01
Dimethylarginine (SDMA + ADMA)	0	0	6940	9.80E-01
X-12833	0	0.02	241	9.81E-01
Mean diameter for HDL particles	0	0.01	19273	9.81E-01
X-11530	0	0	6967	9.85E-01
1-oleoylglycerophosphoethanolamine	0	0	7302	9.87E-01
X-18601	0	0.01	7226	9.87E-01
CD32 on pDC	0	0.02	457	9.88E-01

X-13741	0	0.01	5035	9.88E-01
Concentration of chylomicrons and largest VLDL particles	0	0.01	18960	9.88E-01
Laurylcarnitine	0	0	4925	9.89E-01
Systemic lupus erythematosus	-0.08	0.05	3094	9.92E-01
NKT:%4-8-	0	0.02	459	9.93E-01
X-03056–N-[3-(2-Oxopyrrolidin-1-yl)propyl]acetamide	0	0	7363	9.95E-01
Androsterone sulfate	0	0.01	7338	9.96E-01
X-09789	0	0.01	7356	9.97E-01
Symbol search	0	0.04	991	9.97E-01
Difference in height between adolescence and adulthood	0	0.02	4942	9.98E-01
X-11478	0	0	6169	9.99E-01
Birth weight	0	0.01	26836	1.00E+00
Hirschsprungs disease	-0.23	0.13	788	NA

Table A.2: 23 loci reaching tentitive genome-wide significance at P < 10^{-5} for association with dairy milk intake in the UK Biobank

Chromosome	SNP	Allele A	Allele B	Additive Beta	P	LD with lead SNP[a]
2						
	rs150080038	G	A	-0.237595	2.67E-07	Lead
	rs34128006	A	AT	-0.0472842	2.20E-06	Lead
	rs7570971	C	A	-0.046688	8.37E-06	0.173
	rs34252323	G	A	-0.0618038	9.83E-06	0.003
	rs1987446	T	C	-0.0580089	5.52E-06	Lead
	rs964743	G	C	-0.0575868	6.12E-06	1
	rs901613	A	G	-0.0575201	6.25E-06	0.959
	rs6434156	A	G	-0.0571497	7.22E-06	0.959
	rs2046590	A	G	-0.0571434	7.23E-06	0.919
	rs1487358	G	T	-0.0571432	7.23E-06	0.959
	rs371039847;rs75653894	A	AT	-0.0571458	7.26E-06	NA
	rs6717680	C	T	-0.0567405	7.32E-06	0.959
	rs76127353	C	A	0.0570073	7.71E-06	0.959
	rs77335208	C	T	0.0565661	7.81E-06	0.959
	rs1487356	A	T	-0.0569668	8.09E-06	0.959
	rs4666700	A	G	-0.0560145	8.87E-06	0.959
	rs4622676	G	A	0.055915	8.90E-06	0.959
	rs78602588	G	A	0.0558158	9.12E-06	0.959
	rs77494430	C	T	0.0561676	9.21E-06	0.959
	rs139737117	AAC	A	0.0557872	9.49E-06	0.959
	rs725728	G	A	0.0558703	9.73E-06	NA
	rs6747595	G	A	-0.0555214	9.94E-06	0.959
	rs112820548	A	C	0.0555453	9.97E-06	0.959

3						
	rs9820500	C	T	-0.064445	3.36E-06	Lead
	rs9831341	C	T	-0.0645087	3.50E-06	1
	rs9880266	C	A	-0.0643159	3.64E-06	1
	rs9865486	C	A	-0.0630324	4.15E-06	1
	rs9830889	C	T	-0.0639214	4.17E-06	1
	rs9830549	C	A	-0.0638378	4.23E-06	1
	rs9878173	A	G	-0.0635828	4.28E-06	1
	rs9873768	C	G	-0.0635757	4.32E-06	1
	rs9877918	A	G	-0.0636585	4.52E-06	1
	rs9833973	A	G	-0.0635072	4.57E-06	1
	rs9823349	T	A	-0.062024	5.80E-06	1
	rs17007173	C	T	-0.0647467	6.90E-06	1
	rs73040193	A	G	0.242878	6.99E-06	Lead
	rs73040197	G	A	0.242878	6.99E-06	0.739
	rs114342505	C	T	0.164091	5.77E-06	0.497
	rs72558061	T	G	0.155833	4.69E-06	Lead
	rs74482595	A	G	0.152584	7.50E-06	1
	rs201866958	T	TA	0.152612	7.63E-06	1
6						
	rs1316538	C	G	-0.0507008	1.12E-07	Lead
	rs9483145	C	T	-0.0506869	1.18E-07	0.965
	rs9321241	A	G	-0.0505301	1.30E-07	1
	rs9321238	T	G	-0.050472	1.30E-07	1
	rs9321239	A	G	-0.0504686	1.30E-07	1
	rs9321235	C	T	-0.0504335	1.31E-07	1

rs4895887	T	G	-0.0504261	1.35E-07	1
rs9492649	C	A	-0.0503347	1.38E-07	0.966
rs9388806	A	G	-0.0502656	1.45E-07	1
rs11287987	GA	G	-0.0502704	1.46E-07	1
rs9321234	T	A	-0.0502348	1.46E-07	1
rs9321236	C	G	-0.0502194	1.48E-07	1
rs1316493	A	G	-0.0501963	1.50E-07	1
rs1316537	A	G	-0.0501919	1.50E-07	1
rs7740475	C	T	-0.0501886	1.50E-07	1
rs9402248	T	C	-0.0501874	1.50E-07	1
rs9372967	C	A	-0.050146	1.55E-07	1
rs6569690	A	G	-0.0501507	1.55E-07	1
rs9402247	T	C	-0.0501349	1.56E-07	1
rs9402246	T	G	-0.050129	1.56E-07	0.965
rs4895886	C	T	-0.0501194	1.56E-07	0.965
rs1933757	A	G	-0.0501592	1.58E-07	1
rs9372968	G	A	-0.0500997	1.59E-07	1
rs9375741	G	T	-0.0500421	1.63E-07	1
rs9375740	C	T	-0.0500097	1.68E-07	0.965
rs1317232	G	A	-0.0499927	1.69E-07	1
rs7744825	A	C	-0.0498861	1.79E-07	1
rs6932512	G	T	-0.0499234	1.84E-07	1
rs56109346	T	TAA	-0.0498447	1.88E-07	1
rs4897437	A	G	-0.0497796	1.89E-07	1
rs9372966	C	A	-0.0496872	2.02E-07	0.965
rs9375742	A	G	-0.0501798	2.47E-07	0.932
rs9388808	T	C	-0.0494083	3.17E-07	0.932
rs1203341	C	T	0.0488945	3.40E-07	0.863
rs1320549	G	A	-0.0485057	4.01E-07	0.865
rs9402249	G	A	-0.0483538	4.77E-07	0.959

rs1203343	G	A	0.0479606	5.30E-07	0.897
rs11307732	GT	G	-0.0478763	5.65E-07	0.899
rs10682764	C	CCTT	-0.0477727	5.95E-07	0.899
rs4897432	G	A	-0.0477752	6.03E-07	0.865
rs6940599	A	C	-0.0477722	6.05E-07	0.805
rs12526873	A	T	-0.047692	6.10E-07	0.897
rs6569688	T	C	-0.0476707	6.28E-07	0.897
rs4897430	G	A	-0.0477682	6.31E-07	0.897
rs1203347	G	A	0.0475824	6.40E-07	0.897
rs6939930	A	C	-0.0475624	6.62E-07	0.897
rs1203346	T	C	0.0475145	6.62E-07	0.897
rs2876084	C	T	-0.0475452	6.64E-07	0.897
rs766967	C	T	0.0475105	6.64E-07	0.897
rs2326914	A	T	-0.0475391	6.72E-07	0.865
rs1933759	G	A	-0.0475164	6.81E-07	0.863
rs4897433	A	G	-0.0474733	6.88E-07	0.865
rs6934880	A	G	-0.0474732	6.88E-07	0.897
rs1203348	G	C	0.0474164	6.98E-07	0.897
rs67527325	T	TA	0.0476532	7.01E-07	0.838
rs34595276	T	G	-0.047433	7.04E-07	0.897
rs766966	A	C	0.0473361	7.31E-07	0.897
rs1203349	A	G	0.047352	7.38E-07	0.897
rs11307734	AC	A	-0.0473529	7.39E-07	0.899
rs1203344	T	C	0.0472953	7.58E-07	0.863
rs4897431	T	C	-0.0474422	7.59E-07	0.897
rs1203345	T	C	0.0472613	7.68E-07	0.897
rs1203340	G	A	0.0472409	8.11E-07	0.863
rs9385547	T	C	-0.0472794	8.23E-07	0.939
rs9385548	T	C	-0.0470263	9.40E-07	0.979
rs6941016	C	A	-0.0469011	9.46E-07	0.897

	rs7768049	A	G	-0.0469797	9.66E-07	0.829
	rs11307733	CA	C	-0.047024	9.67E-07	0.899
	rs1320550	G	A	-0.0467197	1.05E-06	0.897
	rs4144216	A	C	-0.0466733	1.11E-06	0.863
	rs10872372	T	C	-0.0466768	1.11E-06	0.863
	rs1203350	A	T	0.0464028	1.28E-06	0.897
	rs9402250	G	A	-0.0463978	1.35E-06	0.939
	rs1203342	C	A	0.0460896	1.48E-06	0.863
	rs6932344	G	A	-0.0455175	1.77E-06	0.871
	rs17801702	C	T	-0.0499654	5.68E-06	0.121
	rs138412515	G	A	0.209479	5.90E-06	NA
	rs6932172	G	A	-0.04306	6.46E-06	0.842
	rs6899678	T	C	-0.0433357	8.01E-06	0.805
7						
	rs62469670	T	G	0.0715656	2.46E-07	Lead
	rs62469669	T	C	0.0637225	2.83E-07	0.861
	rs10282350	G	A	0.0631591	3.66E-07	0.72
	rs34720120	CATT	C	0.0623507	5.25E-07	NA
	rs11978376	A	C	0.0665371	6.68E-07	0.938
	rs5887549	CACTT	C	0.0615924	7.22E-07	NA
	rs10273512	G	A	0.0612477	7.56E-07	0.678
	rs10246281	G	A	0.0593224	8.32E-07	0.64
	rs10275602	A	T	0.0592114	9.21E-07	0.64
	rs12536824	G	A	0.0653031	1.05E-06	0.752
	rs10237663	G	C	0.0586826	1.20E-06	0.605
	rs6976275	G	A	0.0580288	1.70E-06	0.64
	rs2175858	G	A	0.0581637	1.86E-06	0.576
	rs6944280	A	G	-0.0575589	1.87E-06	0.804
	rs1033171	C	A	0.0579532	1.90E-06	0.64
	rs10232652	C	T	0.0574611	1.92E-06	0.605

	rs2175859	A	G	0.0578189	1.97E-06	0.605
	rs9641922	A	G	-0.057438	1.98E-06	0.64
	rs12537969	C	T	0.0577046	2.01E-06	0.64
	rs10216052	C	A	0.0573737	2.11E-06	0.605
	rs10232734	G	A	0.0570502	2.26E-06	0.64
	rs13438632	T	C	0.0572288	2.34E-06	0.64
	rs10249435	C	T	0.0570528	2.48E-06	0.64
	rs4731842	T	C	-0.0567947	2.54E-06	0.64
	rs11284200	GT	G	-0.0561209	5.10E-06	0.777
9						
	rs7029703	C	T	-0.0425765	6.84E-06	Lead
	rs4745564	G	A	0.0423501	7.76E-06	0.934
	rs2031908	G	A	0.0439018	9.68E-06	0.269
12						
	rs12367733	T	C	-0.143609	2.59E-06	Lead
	rs12368986	C	T	-0.158181	3.24E-06	1
	rs55990161	C	G	-0.144978	4.99E-06	1
	rs11053206	A	G	-0.144072	5.01E-06	1
	rs116910420	G	A	-0.144072	5.01E-06	1
	rs11053148	T	C	-0.140757	5.11E-06	1
	rs12367560	C	T	-0.143928	5.13E-06	1
	rs11513528	C	G	-0.147056	5.56E-06	1
	rs541856930	C	G	-0.144844	5.85E-06	1
	rs137981389	T	A	-0.148399	6.24E-06	1
	rs11053169	A	G	-0.146692	6.37E-06	1
	rs145769471	G	A	-0.144502	6.69E-06	1
	rs199787760	A	AT	-0.141554	6.95E-06	1
	rs111813337	G	A	-0.143033	7.16E-06	1
	rs11053270	G	A	-0.143882	7.73E-06	1
	rs148178498	C	T	-0.137526	8.27E-06	1

	rs11053219	C	T	-0.138988	9.09E-06	1
	rs11180909	C	T	-0.151124	1.22E-06	Lead
	rs11514054	C	T	-0.149353	1.41E-06	1
	rs11495697	G	A	-0.14852	1.55E-06	0.661
	rs11514015	C	T	-0.15451	1.73E-06	0.661
	rs11495460	G	T	-0.154999	2.80E-06	1
	rs142166599	T	G	-0.142709	3.06E-06	1
	rs117320563	G	A	-0.144096	3.16E-06	1
	rs148991213	C	T	-0.143756	3.23E-06	1
	rs148422305	G	A	-0.156584	3.26E-06	0.663
	rs532879136	T	C	-0.148081	3.54E-06	NA
	rs11520053	C	A	-0.148081	3.54E-06	0.594
	rs11495596	A	G	-0.150416	3.55E-06	1
	rs11182825	A	G	-0.142708	3.55E-06	1
	rs117742570	T	C	-0.143111	3.60E-06	1
	rs12368917	A	G	-0.151452	3.67E-06	1
	rs182128654	C	A	-0.153522	4.07E-06	0.257
	rs11520300	T	A	-0.142026	4.16E-06	1
	rs11182241	C	G	-0.143025	4.33E-06	1
	rs117129580	A	C	-0.146458	5.23E-06	0.661
	rs55847803	G	A	-0.148419	5.56E-06	1
	rs143900000	C	A	-0.158851	6.83E-06	1
	rs11520109	G	T	-0.141539	6.96E-06	1
	rs190561272	G	T	-0.144662	7.82E-06	1
14						
	rs146106168	C	T	0.195123	9.40E-06	Lead
16						
	rs72778535	G	A	0.0804969	1.18E-06	Lead

rs4647869	C	A	0.058601	3.50E-07	Lead
rs4647863	C	T	0.0576396	8.13E-07	0.876
rs4648330	A	G	0.0570008	9.09E-07	0.783
rs55803651	C	T	0.0562153	1.21E-06	0.973
rs4647861	C	T	0.0559103	1.39E-06	0.973
rs12373152	G	A	0.0554761	1.61E-06	0.84
rs111782763	T	C	0.0548826	3.53E-06	0.921
rs4354969	A	T	0.0538334	3.85E-06	0.792
rs150284072	TTTTG	T	0.0545475	4.03E-06	NA
rs4792352	T	C	0.0434281	4.21E-06	Lead
rs1003833	C	G	-0.0480521	4.68E-06	0.251
rs9898604	T	G	0.0431503	4.86E-06	1
rs7215589	G	C	0.0431044	4.93E-06	1
rs12602364	A	G	0.042722	5.30E-06	0.772
rs28759190	G	A	0.0430104	5.35E-06	1
rs12450092	G	C	0.0428166	5.68E-06	0.96
rs73296883	G	A	0.042731	5.95E-06	0.96
rs16947606	T	C	0.0427098	5.97E-06	0.96
rs7216140	C	T	0.0425534	6.43E-06	0.96
rs8073006	G	A	0.0424363	6.94E-06	0.96
rs6502250	C	T	0.0422429	7.55E-06	0.96
rs5009607	T	C	0.0421016	8.07E-06	0.96
rs4444379	T	G	0.0421077	8.16E-06	0.96
rs4792351	G	A	0.0419692	8.87E-06	0.98
rs114417133	A	G	0.0419086	9.04E-06	0.98
rs58488326	GA	G	0.0419047	9.08E-06	0.98
rs7208389	A	T	0.0418907	9.13E-06	0.98
rs9916579	G	T	0.0418903	9.13E-06	0.98

	rs146908369	ATTACAGGCATGAGCCACTGTC	A	0.0418388	9.62E-06	0.96
20						
	rs77410568	A	G	0.204402	7.05E-06	Lead
21						
	rs71317650	G	A	-0.233565	2.32E-06	Lead
	rs148462362	TC	T	0.110927	6.71E-06	Lead
	rs77409573	A	G	0.155487	4.39E-06	Lead
	rs62230080	C	T	-0.121612	2.05E-06	Lead
	rs62227539	C	A	-0.113876	4.36E-06	1
	rs35040005	C	T	-0.121953	5.17E-06	0.781
	rs73369341	A	G	-0.116588	5.76E-06	1
	rs62227544	A	G	-0.116034	5.96E-06	1
	rs139044928	C	G	-0.118554	8.86E-06	0.781
22						
	rs114463484	A	G	-0.101943	6.08E-06	Lead
	rs148735502	T	C	-0.107446	7.30E-06	1
	rs201947651	T	TACAC	-0.101283	7.89E-06	0.245

[a] Linkage disequilibrium for each SNP at a given loci with the lead SNP for that loci

Appendix B

EPIC Baseline Questionnaire

CONFIDENTIAL

PATIENT'S QUESTIONNAIRE

This questionnaire asks for some background information about you, especially about what you eat. Please fill it in at home and bring it with you to the surgery.

Please answer every question. If you are uncertain about how to answer a question then do the best you can, but please do not leave a question blank. If you have any problems with the questions please ask the nurse to help when you come for your appointment.

Your answers will be treated as strictly confidential and will be used only for medical research.

Surname:

Forename(s):

Address:

Postcode:

Please complete this section before going to question 1.

Date of birth: ☐ day ☐ month 19 ☐ year

Are you male or female? Male ☐ Female ☐

How tall are you? ☐ feet and ☐ inches **or** ☐ centimetres

How much do you weigh? ☐ stones and ☐ pounds **or** ☐ kilogrammes

How old were you when you left school? ☐ years old

Do you eat any meat (including bacon, ham, poultry, game, meat pies, sausages)? Yes ☐ No ☐

If no, how old were you when you last ate meat? ☐ years old

Do you eat any fish? Yes ☐ No ☐

If no, how old were you when you last ate fish? ☐ years old

Do you eat any dairy products (including milk, cheese, butter, yogurt)? Yes ☐ No ☐

If no, how old were you when you last ate dairy products? ☐ years old

Do you eat any eggs (including eggs in cakes and other baked foods)? Yes ☐ No ☐

If no, how old were you when you last ate eggs? ☐ years old

Listed below are 130 food items divided into sections according to food type. For each food there is an amount shown, either a "medium serving" or a common household unit such as a slice or teaspoon. Please put a tick (✓) in the box to indicate how often, **on average**, you have eaten the specified amount of each food **during the last 12 months**.

EXAMPLES:

For white bread the amount is one slice, so if you ate 4 or 5 slices a day, you should put a tick in the column headed "4-5 per day".

FOODS AND AMOUNTS	AVERAGE USE IN LAST 12 MONTHS								
BREAD AND SAVOURY BISCUITS (one slice or biscuit)	Never or less than once/month	1-3 per month	Once a week	2-4 per week	5-6 per week	Once a day	2-3 per day	4-5 per day	6+ per day
White bread and rolls								✓	

For chips, the amount is a "medium serving", so if you had a helping of chips twice a week you should put a tick in the column headed "2-4 per week".

FOODS AND AMOUNTS	AVERAGE USE IN LAST 12 MONTHS								
POTATOES, RICE AND PASTA (medium serving)	Never or less than once/month	1-3 per month	Once a week	2-4 per week	5-6 per week	Once a day	2-3 per day	4-5 per day	6+ per day
Chips				✓					

For very seasonal fruits such as strawberries and raspberries you should estimate your average use when the fruits are in season, so if you ate strawberries or raspberries about once a week when they were in season you should put a tick in the column headed "once a week"

FOODS AND AMOUNTS	AVERAGE USE IN LAST 12 MONTHS								
FRUIT (1 fruit or medium serving)	Never or less than once/month	1-3 per month	Once a week	2-4 per week	5-6 per week	Once a day	2-3 per day	4-5 per day	6+ per day
Strawberries, raspberries, kiwi fruit			✓						

1. Please estimate your average food use as best you can, and please answer every question.

MEAT AND FISH

Did you eat any meat or fish in the last 12 months? Yes ☐ No ☐

If no, please go to next page

If yes , please fill in this page

PLEASE PUT A TICK (✓) ON EVERY LINE

FOODS AND AMOUNTS	AVERAGE USE IN LAST 12 MONTHS								
MEAT AND FISH (medium serving)	Never or less than once/month	1-3 per month	Once a week	2-4 per week	5-6 per week	Once a day	2-3 per day	4-5 per day	6+ per day
Beef: roast, steak, mince, stew or casserole									
Beefburgers									
Pork: roast, chops, stew or slices									
Lamb: roast, chops or stew									
Chicken or other poultry e.g. turkey									
Bacon									
Ham									
Corned beef, Spam, luncheon meats									
Sausages									
Savoury pies, e.g. meat pie, pork pie, pasties, steak & kidney pie, sausage rolls									
Liver, liver paté, liver sausage									
Fried fish in batter, as in fish and chips									
Fish fingers, fish cakes									
Other white fish, fresh or frozen, e.g. cod, haddock, plaice, sole, halibut									
Oily fish, fresh or canned, e.g. mackerel, kippers, tuna, salmon, sardines, herring									
Shellfish, e.g. crab, prawns, mussels									
Fish roe, taramasalata									
	Never or less than once/month	1-3 per month	Once a week	2-4 per week	5-6 per week	Once a day	2-3 per day	4-5 per day	6+ per day

What did you do with the visible fat on your meat?

Ate most of the fat ☐ Ate as little as possible ☐

Ate some of the fat ☐ Did not eat meat ☐

How often did you eat grilled or roast meat? ☐☐ times a week

How well cooked did you usually have grilled or roast meat?

Well done /dark brown ☐ Lightly cooked/rare ☐

Medium ☐ Did not eat meat ☐

PLEASE PUT A TICK (✓) ON EVERY LINE

FOODS AND AMOUNTS	AVERAGE USE IN LAST 12 MONTHS								
BREAD AND SAVOURY BISCUITS (one slice or biscuit)	Never or less than once/month	1-3 per month	Once a week	2-4 per week	5-6 per week	Once a day	2-3 per day	4-5 per day	6+ per day
White bread and rolls									
Brown bread and rolls									
Wholemeal bread and rolls									
Cream crackers, cheese biscuits									
Crispbread, e.g. Ryvita									
CEREALS (one bowl)									
Porridge, Readybrek									
Breakfast cereal such as cornflakes, muesli etc.									
POTATOES, RICE AND PASTA (medium serving)									
Boiled, mashed, instant or jacket potatoes									
Chips									
Roast potatoes									
Potato salad									
White rice									
Brown rice									
White or green pasta, e.g. spaghetti, macaroni, noodles									
Wholemeal pasta									
Lasagne, moussaka									
Pizza									
	Never or less than once/month	1-3 per month	Once a week	2-4 per week	5-6 per week	Once a day	2-3 per day	4-5 per day	6+ per day

Please check that you have a tick (✓) on EVERY line

PLEASE PUT A TICK (✓) ON EVERY LINE

FOODS AND AMOUNTS	AVERAGE USE IN LAST 12 MONTHS								
DAIRY PRODUCTS AND FATS	Never or less than once/month	1-3 per month	Once a week	2-4 per week	5-6 per week	Once a day	2-3 per day	4-5 per day	6+ per day
Single or sour cream (tablespoon)									
Double or clotted cream (tablespoon)									
Low fat yogurt, fromage frais (125g carton)									
Full fat or Greek yogurt (125g carton)									
Dairy desserts (125g carton)									
Cheese, e.g. Cheddar, Brie, Edam (medium serving)									
Cottage cheese, low fat soft cheese (medium serving)									
Eggs as boiled, fried, scrambled, etc. (one)									
Quiche (medium serving)									
Low calorie, low fat salad cream(tablespoon)									
Salad cream, mayonnaise (tablespoon)									
French dressing (tablespoon)									
Other salad dressing (tablespoon)									
The following on bread or vegetables									
Butter (teaspoon)									
Block margarine, wrapped, NOT tub, e.g. Stork,Krona (teaspoon)									
Polyunsaturated margarine, in tub, e.g. Flora, sunflower (teaspoon)									
Other soft margarine, dairy spreads, in tub, e.g. Blue Band, Clover (teaspoon)									
Low fat spread, in tub, e.g. Outline, Gold (teaspoon)									
Very low fat spread, in tub (teaspoon)									
	Never or less than once/month	1-3 per month	Once a week	2-4 per week	5-6 per week	Once a day	2-3 per day	4-5 per day	6+ per day

Please check that you have a tick (✓) on EVERY line

PLEASE PUT A TICK (✓) ON EVERY LINE

FOODS AND AMOUNTS	AVERAGE USE IN LAST 12 MONTHS								
SWEETS AND SNACKS (medium serving)	Never or less than once/month	1-3 per month	Once a week	2-4 per week	5-6 per week	Once a day	2-3 per day	4-5 per day	6+ per day
Sweet biscuits, chocolate , e.g. digestive (one)									
Sweet biscuits, plain, e.g. Nice, ginger (one)									
Cakes e.g. fruit, sponge, home baked									
Cakes e.g. fruit, sponge, ready made									
Buns, pastries e.g. scones, flapjacks, home baked									
Buns, pastries e.g. croissants, doughnuts, ready made									
Fruit pies, tarts, crumbles, home baked									
Fruit pies, tarts, crumbles, ready made									
Sponge puddings, home baked									
Sponge puddings, ready made									
Milk puddings, e.g. rice, custard, trifle									
Ice cream, choc ices									
Chocolates, single or squares									
Chocolate snack bars e.g. Mars, Crunchie									
Sweets, toffees, mints									
Sugar added to tea, coffee, cereal (teaspoon)									
Crisps or other packet snacks, e.g. Wotsits									
Peanuts or other nuts									
SOUPS, SAUCES, AND SPREADS									
Vegetable soups (bowl)									
Meat soups (bowl)									
Sauces, e.g. white sauce, cheese sauce, gravy (tablespoon)									
Tomato ketchup (tablespoon)									
Pickles, chutney (tablespoon)									
Marmite, Bovril (teaspoon)									
Jam, marmalade, honey (teaspoon)									
Peanut butter (teaspoon)									
	Never or less than once/month	1-3 per month	Once a week	2-4 per week	5-6 per week	Once a day	2-3 per day	4-5 per day	6+ per day

Please check that you have a tick (✓) on EVERY line

PLEASE PUT A TICK (✓) ON EVERY LINE

FOODS AND AMOUNTS	AVERAGE USE IN LAST 12 MONTHS								
DRINKS	Never or less than once/month	1-3 per month	Once a week	2-4 per week	5-6 per week	Once a day	2-3 per day	4-5 per day	6+ per day
Tea (cup)									
Coffee, instant or ground (cup)									
Coffee, decaffeinated (cup)									
Coffee whitener, e.g. Coffee-mate (teaspoon)									
Cocoa, hot chocolate (cup)									
Horlicks, Ovaltine (cup)									
Wine (glass)									
Beer, lager or cider (half pint)									
Port, sherry, vermouth, liqueurs (glass)									
Spirits, e.g. gin, brandy, whisky, vodka (single)									
Low calorie or diet fizzy soft drinks (glass)									
Fizzy soft drinks, e.g. Coca cola, lemonade (glass)									
Pure fruit juice (100%) e.g. orange, apple juice (glass)									
Fruit squash or cordial (glass)									
FRUIT (1 fruit or medium serving) **For very seasonal fruits such as strawberries, please estimate your average use when the fruit is in season**									
Apples									
Pears									
Oranges, satsumas, mandarins									
Grapefruit									
Bananas									
Grapes									
Melon									
Peaches, plums, apricots									
Strawberries, raspberries, kiwi fruit									
Tinned fruit									
Dried fruit, e.g. raisins, prunes									
	Never or less than once/month	1-3 per month	Once a week	2-4 per week	5-6 per week	Once a day	2-3 per day	4-5 per day	6+ per day

Please check that you have a tick (✓) on EVERY line

PLEASE PUT A TICK (✓) ON EVERY LINE

FOODS AND AMOUNTS	AVERAGE USE IN LAST 12 MONTHS								
VEGETABLES Fresh, frozen or tinned (medium serving)	Never or less than once/month	1-3 per month	Once a week	2-4 per week	5-6 per week	Once a day	2-3 per day	4-5 per day	6+ per day
Carrots									
Spinach									
Broccoli, spring greens, kale									
Brussels sprouts									
Cabbage									
Peas									
Green beans, broad beans, runner beans									
Marrow, courgettes									
Cauliflower									
Parsnips, turnips, swedes									
Leeks									
Onions									
Garlic									
Mushrooms									
Sweet peppers									
Beansprouts									
Green salad, lettuce, cucumber, celery									
Watercress									
Tomatoes									
Sweetcorn									
Beetroot									
Coleslaw									
Avocado									
Baked beans									
Dried lentils, beans, peas									
Tofu , soya meat, TVP, Vegeburger									
	Never or less than once/month	1-3 per month	Once a week	2-4 per week	5-6 per week	Once a day	2-3 per day	4-5 per day	6+ per day

Please check that you have a tick (✓) on EVERY line

Your diet last year, continued

2. Are there any **other** foods which you ate more than once a week? Yes ☐ No ☐

If yes, please list below

Food	Usual serving size	Number of times eaten each week

3. What type of milk did you most often use?

Select one only

Full cream, silver ☐ Semi-skimmed, red/white ☐
Skimmed/fat free ☐ Channel Islands, gold ☐
Dried milk ☐ Soya ☐
Other ☐ specify ☐ None ☐

If you used soya milk, please describe brand and type ☐

4. How much milk did you drink each day, including milk with tea, coffee, cereals etc?

None ☐ Three quarters of a pint ☐
Quarter of a pint ☐ One pint ☐
Half a pint ☐ More than one pint ☐

5. Did you usually eat breakfast cereal, excluding porridge and Ready Brek mentioned earlier? Yes ☐ No ☐

If yes, which brand and type of breakfast cereal, including muesli, did you usually eat?

List the one or two types most often used

Brand	Type

6. What kind of fat did you most often use for frying, roasting, grilling etc?

Select one only

Butter ☐ Solid white vegetable fat ☐
Lard/dripping ☐ Margarine ☐
Vegetable oil ☐ None ☐

If you used vegetable oil, please give type e.g. corn, sunflower ☐

7. What kind of fat did you most often use for baking cakes etc?

Select one only

Butter ☐ Solid white vegetable fat ☐
Lard/dripping ☐ Margarine ☐
Vegetable oil ☐ None ☐

If you used margarine, please give type e.g. Flora, Stork ☐

8. How often did you eat food that was fried at home?

Daily ☐ Less than once a week ☐
4-6 times a week ☐ Never ☐
1-3 times a week ☐

9. How often did you eat fried food away from home?

Daily ☐ Less than once a week ☐
4-6 times a week ☐ Never ☐
1-3 times a week ☐

10. How often did you add salt to food while cooking?

Always ☐ Rarely ☐
Usually ☐ Never ☐
Sometimes ☐

11. How often did you add salt to any food at the table?

Always ☐ Rarely ☐
Usually ☐ Never ☐
Sometimes ☐

12. Did you regularly use a salt substitute (e.g. LoSalt)? Yes ☐ No ☐

If yes, which brand? _____

13. Have you regularly taken any vitamins, minerals, fish oils, fibre or other food supplements during the last 12 months? Yes ☐ No ☐

If yes, list brand and daily dose

Name and brand of supplements	Daily dose

14. In the last 12 months, have you eaten a modified diet for any of these reasons?

Tick more than one box if applicable

High blood pressure ☐ High blood cholesterol ☐
Stomach problems (e.g. ulcer or gastritis) ☐ Overweight/obesity ☐
Bowel problems (e.g. irritable bowel or diverticulitis) ☐ Diabetes ☐
Allergies (e.g. skin rash) ☐ Concern over eating a healthy diet ☐
Concern over a family history of illness ☐ Not modified my diet ☐
Other ☐ specify _____

15. When you were aged 20, about how many alcoholic drinks did you have each week?

Put "0" if none, "occ" if occasional but less than one drink a week

Please answer EACH line

Beer or cider	pints each week
Wine	glasses each week
Sherry or other fortified wine	glasses each week
Spirits	glasses (singles) each week

16. When you were aged 30, about how many alcoholic drinks did you have each week?

Put "0" if none, "occ" if occasional but less than one drink a week

Please answer EACH line

Beer or cider	pints each week
Wine	glasses each week
Sherry or other fortified wine	glasses each week
Spirits	glasses (singles) each week

Not yet aged 30

17. Have you ever smoked as much as one cigarette a day for as long as a year? Yes No

If no, please go to question 18

If yes, how old were you when you started smoking cigarettes regularly? years old

Did you smoke at the following ages? If so, how many cigarettes did you smoke and were they usually filter cigarettes?

Age 20	cigs per day	Filter	No filter	Non smoker
Age 30	cigs per day	Filter	No filter	Non smoker
Age 40	cigs per day	Filter	No filter	Non smoker
Age 50	cigs per day	Filter	No filter	Non smoker

Do you smoke cigarettes now? Yes No

If yes, how many cigarettes do you smoke each day? cigarettes

Do you usually smoke filter cigarettes? Yes No

Do you usually smoke low tar cigarettes? Yes No

Which brand do you normally smoke?

How deeply do you inhale? Deeply into the lungs A little Not at all

If you have stopped smoking, how old were you when you last smoked ? years old

18. Do you currently smoke cigars? Yes No

19. Do you currently smoke a pipe? Yes No

20. Approximately how much did you weigh when you were 20 years old?

stones lbs **or** kg

21. What is your present waist size? inches **or** centimetres

22. What is your present hip size? inches **or** centimetres

23. In a typical week during the last 12 months, how many hours did you spend on each of the following activities? **Put "0" if none**

Housework, such as cleaning, washing, cooking, child care		☐☐	hours per week
Do-it-yourself		☐☐	hours per week
Gardening	in summer	☐☐	hours per week
	in winter	☐☐	hours per week
Walking, including walking to work, shopping and leisure	in summer	☐☐	hours per week
	in winter	☐☐	hours per week
Cycling, including cycling to work and leisure	in summer	☐☐	hours per week
	in winter	☐☐	hours per week
Other physical exercise, such as keep-fit, aerobics, swimming, jogging, tennis	in summer	☐☐	hours per week
	in winter	☐☐	hours per week

24. In a typical week during the last 12 months, did you practise any of these activities vigorously enough to cause sweating or a faster heartbeat? Yes ☐ No ☐

If yes, for how many hours each week did you practise such vigorous physical activity?

☐☐ hours per week

25. In a typical day during the last 12 months, how many floors of stairs did you climb up?

Put "0" if none ☐☐ floors per day

26. Have you ever been told by a doctor that you have, or had, any of the following conditions?

Please tick all which apply and give the age at which each condition was first diagnosed.

Heart attack, coronary thrombosis, myocardial infarction	Yes ☐	at age	☐☐	yrs old	No ☐
Angina	Yes ☐	at age	☐☐	yrs old	No ☐
Stroke	Yes ☐	at age	☐☐	yrs old	No ☐
High blood pressure (hypertension)	Yes ☐	at age	☐☐	yrs old	No ☐
High blood cholesterol, hyperlipidaemia	Yes ☐	at age	☐☐	yrs old	No ☐
Diabetes	Yes ☐	at age	☐☐	yrs old	No ☐
Gallstones	Yes ☐	at age	☐☐	yrs old	No ☐
Polyps in the large intestine	Yes ☐	at age	☐☐	yrs old	No ☐
Cancer	Yes ☐	at age	☐☐	yrs old	No ☐

If yes, what type of cancer? _____

Any other illnesses or operations?

Do not include hysterectomy or breast surgery. These are covered in the women's section later in the questionnaire.

	Age first diagnosed
_____	☐☐ yrs old
_____	☐☐ yrs old
_____	☐☐ yrs old
_____	☐☐ yrs old

27. Are you currently receiving long-term treatment for any illness or condition? Yes ☐ No ☐

If yes, please give details:

Illness or condition	Treatment	Dose	Frequency

28. Please give details of the ages of your mother and father, and whether they have ever had cancer or a heart attack. If you are adopted or if your parents remarried, please give details of your blood relatives only.

	Age now	OR	Age at death	Details of any cancer and/or heart attacks: Disease type	Age first diagnosed
Mother	☐☐ yrs		☐☐ yrs		☐☐ yrs
					☐☐ yrs
Father	☐☐ yrs		☐☐ yrs		☐☐ yrs
					☐☐ yrs

29. Do you have any brothers or sisters ? Yes ☐ No ☐

If yes, please give their ages, whether they are full or half brothers or sisters, and whether they have ever had cancer or a heart attack.

Brother OR Sister	Full OR Half	Age now	OR	Age at death	Details of any cancer and/or heart attacks: Disease type	Age first diagnosed
☐☐	☐☐	☐☐ yrs		☐☐ yrs		☐☐ yrs
						☐☐ yrs
☐☐	☐☐	☐☐ yrs		☐☐ yrs		☐☐ yrs
						☐☐ yrs
☐☐	☐☐	☐☐ yrs		☐☐ yrs		☐☐ yrs
						☐☐ yrs
☐☐	☐☐	☐☐ yrs		☐☐ yrs		☐☐ yrs
						☐☐ yrs
☐☐	☐☐	☐☐ yrs		☐☐ yrs		☐☐ yrs
						☐☐ yrs
☐☐	☐☐	☐☐ yrs		☐☐ yrs		☐☐ yrs
						☐☐ yrs
☐☐	☐☐	☐☐ yrs		☐☐ yrs		☐☐ yrs
						☐☐ yrs
☐☐	☐☐	☐☐ yrs		☐☐ yrs		☐☐ yrs
						☐☐ yrs
☐☐	☐☐	☐☐ yrs		☐☐ yrs		☐☐ yrs
						☐☐ yrs

30. How old were you when you finished full time education? ☐☐ years old

Not yet finished ☐

31. Do you have any of the following qualifications? **Tick all applicable**

CSE ☐ GCE "O" level ☐ "A" level, Highers ☐

Teaching diploma, HNC ☐ Degree ☐ None of these ☐

Other ☐ describe ☐

32. Have you ever had a paid job ? Yes ☐ No ☐

If yes, please answer for you current or most recent job

What is/was your job title? ☐

What do/did you do in your job? ☐

What does/did the organization you work for make or do? ☐

How many hours do/did you work each week? ☐☐ hours

Are/were you a Manager? ☐ Foreman/woman? ☐ Supervisor? ☐ None of these? ☐

Are/were you self-employed? Yes ☐ No ☐

In this job, which of the following best describes your physical activity. **Tick one only**

Sedentary occupation. You spend most of your time sitting (such as in an office). ☐

Standing occupation. You spend most of your time standing or walking. However, your work does not require intense physical effort (e.g. shop assistant, hairdresser, guard). ☐

Manual work. This involves some physical effort including handling of heavy objects and use of tools (e.g. plumber, electrician, carpenter). ☐

Heavy manual work. This involves very vigorous physical activity including handling very heavy objects (e.g. docker, miner, bricklayer, construction worker). ☐

Do you have a paid job at present ? Yes ☐ No ☐

If no, how would you describe yourself?

Housewife/husband ☐ Unemployed ☐

Retired ☐ Student ☐

Other ☐ describe ☐

When did you last work? 1 9 ☐☐ year Never ☐

33. What is your marital status?

Married or living as married ☐ Widowed ☐ Separated ☐
Divorced ☐ Single ☐

If you are not married or living as married, please go to question 34

If married or living as married, has your partner ever had a paid job? Yes ☐ No ☐

If yes, please answer for your partner's current or most recent job.

What is/was your partner's job title? ☐

What does/did your partner do in this job? ☐

What does/did the organization your partner works for make or do? ☐

Is/was your partner a Manager? ☐ Foreman/woman? ☐ Supervisor? ☐ None of these? ☐

Is/was your partner self-employed? Yes ☐ No ☐

Does your partner have a paid job at present? Yes ☐ No ☐

If no, how would you describe your partner?

Housewife/husband ☐ Unemployed ☐ Retired ☐ Student ☐
Other ☐ describe ☐

When did your partner last work? 1 9 ☐☐ year Never ☐

34. To which of these groups do you consider you belong?

White ☐ Indian ☐ Pakistani ☐
Bangladeshi ☐ Chinese ☐ Black - Caribbean ☐
Black - other ☐ describe ☐
Other ☐ describe ☐

Question 35 is for men only. Women please go to question 36

35. Have you had a vasectomy? Yes ☐ No ☐

If yes, at what age? ☐☐ years old

Now please go to question 52 on page 18

Questions 36 to 51 are for women only.

36. How old were you when you had your first menstrual period? ☐☐ years old

37. When you were aged between thirty and forty, how many days were there between the start of one menstrual period and the start of the next? **Ignore times when you were pregnant, breastfeeding or taking an oral contraceptive (the pill)**

24 days or less ☐ 25 to 26 days ☐ 27 to 29 days ☐
30 to 31 days ☐ 32 or more days ☐ Irregular ☐
No menstrual cycles ☐ Used the pill continuously ☐ Don't know ☐
Not yet aged 30 ☐

38. Have you ever been pregnant? Yes ☐ No ☐

If yes, please go to question 39

If no, have you ever tried to become pregnant? Yes ☐ No ☐

Please go to question 43

39. Have you had any children? Yes ☐ No ☐

If yes, fill in one line for each child you have had.
If twins or triplets, fill in one line per child.

	Date of birth (day/month/year)	Boy (tick as applicable)	Girl (tick as applicable)	Number of weeks breastfed even if only occasional (put "0" if none, "1 week" if 1 to 6 days)
1	/ / 19	☐	☐	weeks
2	/ / 19	☐	☐	weeks
3	/ / 19	☐	☐	weeks
4	/ / 19	☐	☐	weeks
5	/ / 19	☐	☐	weeks
6	/ / 19	☐	☐	weeks
7	/ / 19	☐	☐	weeks
8	/ / 19	☐	☐	weeks
9	/ / 19	☐	☐	weeks
10	/ / 19	☐	☐	weeks

40. Have you had any stillbirths? Yes ☐ No ☐

If yes, please record the year(s) 19☐☐ 19☐☐ 19☐☐

41. Have you had any miscarriages? Yes ☐ No ☐

If yes, please record the year(s) 19☐☐ 19☐☐ 19☐☐
19☐☐ 19☐☐ 19☐☐

42. Have you had any **other** pregnancies that lasted less than 28 weeks? Yes ☐ No ☐

If yes, please record the year(s) 19☐☐ 19☐☐ 19☐☐

43. Have you ever seen a doctor because of fertility problems? Yes ☐ No ☐

If yes, has a doctor ever told you that you were infertile? Yes ☐ No ☐

Have you ever had surgery for infertility? Yes ☐ No ☐

Have you ever been treated with drugs for infertility? Yes ☐ No ☐

44. Have you ever used oral contraceptives (the pill)? Yes ☐ No ☐

If yes, how old were you when you first used the pill? ☐☐ years old

For how long altogether did you use the pill? ☐☐ years

Are you currently using the pill? Yes ☐ No ☐

If no, how old were you when you last used it? ☐☐ years old

45. Have you ever used a coil or intra-uterine device (IUD)? Yes ☐ No ☐

If yes, do you have a coil or IUD at present? Yes ☐ No ☐

46. How many "natural" menstrual periods have you had in the last 12 months?
Do not count bleeding while using the pill or HRT (hormone replacement therapy)
None ☐ 1 to 3 ☐ 4 to 5 ☐ 6 to 9 ☐ 10 or more ☐
Not applicable because using the pill or HRT ☐

47. What was the date of the start of your last "natural" menstrual period? **Do not count bleeding while using the pill or HRT (hormone replacement therapy). Record as fully as possible**
Date ☐☐/☐☐/19☐☐ or age ☐☐ years old Don't know ☐

48. Have you ever used HRT (hormone replacement therapy for menopause)? Yes ☐ No ☐
If yes, how old were you when you first used HRT? ☐☐ years old
For how long altogether have you used HRT? years ☐☐ and months ☐☐
Are you currently using HRT? Yes ☐ No ☐
If no, how old were you when you last used HRT? ☐☐ years old
In what form do/did you take HRT? **Tick all which apply**
By mouth (pill form) ☐ By injection ☐
By implantation under the skin ☐ By cream (vaginal or skin) ☐
By adhesive patches on the skin ☐ By pessary (vaginal) ☐
Other ☐ describe ☐

What brand name are you currently using or did you last use?
Cyclo-Progynova ☐ Harmogen ☐ Prempak-C ☐
Estracombi ☐ Livial ☐ Progynova ☐
Estraderm ☐ Nuvelle ☐ Trisequens ☐
Estrapak ☐ Premarin ☐ don't know ☐
Other ☐ describe ☐

Do/did you have periods or bleeding while taking HRT?
Not at all ☐ Some spotting ☐ Light bleeding ☐ Heavy bleeding ☐

49. Have you had a hysterectomy (womb removed)? Yes ☐ No ☐
If yes, how old were you when you had your hysterectomy? ☐☐ years old

50. Have you had an operation to remove one or both ovaries? Yes ☐ No ☐ Don't know ☐
If yes, how old were you? ☐☐ years old
Were one or both ovaries removed? One ☐ Both ☐ Don't know ☐

51. Have you ever had a breast biopsy (minor surgery to remove tissue from your breast for diagnostic purposes)? Yes ☐ No ☐ Don't know ☐
If yes, how old were you (first occurrence)? ☐☐ years old

Questions 52 is for men and women.

52. If we have any queries about your answers to this questionnaire, would you be happy for us to contact you? Yes ☐ No ☐

If yes, please give your telephone number

Telephone: Dialling code ____ Number ____ Daytime ☐ Evening ☐ Anytime ☐

We would like to write to you again to tell you about the progress of EPIC and to find out whether your diet has changed.

In case you change your address and we lose contact with you, could you give us the name and address of a friend or relative who would know your new address? Please inform them that you have done this.

Contact name

Contact address

Postcode

Please go back and check that you have answered all the questions, then bring this questionnaire with you for your appointment with the nurse.
If you have any questions about the study you can telephone us on 0865 516329

Thank you for your help

Section to be completed with the nurse during your appointment - *please leave blank*

Patient's consent

I agree to participate in the study and:

i) give my doctor permission to provide clinical information from my medical records;

ii) understand that personal details will be used only for research;

iii) agree to provide a blood sample for research purposes.

Signature _____ Date __/__/19__

NHS No. _____

Blood pressure 1st reading ____/____ mmHg

2nd reading ____/____ mmHg

Resting pulse ____ per minute

Height without shoes ____ cm

Weight in light indoor clothing ____.__ kg

Waist circumference ____ cm

Hip circumference ____ cm

Date of blood sample Date __/__/19__

Time of blood sample (24hr clock) __ __ hours __ __ minutes

Number of cigarettes smoked in last 24 hours ____ cigarettes

Put 'nil' if none

Were any prescription medicines, over the counter medicines or nutritional supplements taken today or yesterday? Yes ☐ No ☐

If yes, record

Name of prescription	Dose	Tick if taken Today	Tick if taken Yesterday
		☐	☐
		☐	☐
		☐	☐
		☐	☐
		☐	☐

For women only

Please check through the answers to questions 46, 47, 48, 49, 50 and update the date of the start of the last "natural' menstrual period if applicable.

EPIC is supported by:

British Heart Foundation

Cancer Research Campaign

Department of Health

Europe Against Cancer Programme
Commission of the European Communities

Imperial Cancer Research Fund

MRC

Medical Research Council

Ministry of Agriculture Fisheries & Food

World Health Organization

Bibliography

[1] Lindsey A Torre, Rebecca L Siegel, Elizabeth M Ward, and Ahmedin Jemal. Global cancer incidence and mortality rates and trendsan update. *Cancer Epidemiology Biomarkers & Prevention*, 25(1):16–27, 2016.

[2] Martin CS Wong, William B Goggins, Harry HX Wang, Franklin DH Fung, Colette Leung, Samuel YS Wong, Chi Fai Ng, and Joseph JY Sung. Global incidence and mortality for prostate cancer: Analysis of temporal patterns and trends in 36 countries. *European Urology*, 2016.

[3] Ahmedin Jemal, Stacey A Fedewa, Jiemin Ma, Rebecca Siegel, Chun Chieh Lin, Otis Brawley, and Elizabeth M Ward. Prostate cancer incidence and psa testing patterns in relation to uspstf screening recommendations. *Jama*, 314(19):2054–2061, 2015.

[4] Ann W Hsing, Lilian Tsao, and Susan S Devesa. International trends and patterns of prostate cancer incidence and mortality. *International journal of cancer*, 85(1):60–67, 2000.

[5] Jacques Ferlay, Hai-Rim Shin, Freddie Bray, David Forman, Colin Mathers, and Donald Maxwell Parkin. Estimates of worldwide burden of cancer in 2008: Globocan 2008. *International journal of cancer*, 127(12):2893–2917, 2010.

[6] Ahmedin Jemal, Freddie Bray, Melissa M Center, Jacques Ferlay, Elizabeth Ward, and David Forman. Global cancer statistics. *CA: a cancer journal for clinicians*, 61(2):69–90, 2011.

[7] Frederick Albright, Robert A Stephenson, Neeraj Agarwal, Craig C Teerlink, William T Lowrance, James M Farnham, and Lisa A Cannon Albright. Prostate cancer risk prediction based on complete prostate cancer family history. *The Prostate*, 75(4):390–398, 2015.

[8] Ola Bratt, Linda Drevin, Olof Akre, Hans Garmo, and Pär Stattin. Family history and probability of prostate cancer, differentiated by risk category: A nationwide population-based study. *Journal of the National Cancer Institute*, 108(10):djw110, 2016.

[9] Rebecca L Siegel, Kimberly D Miller, and Ahmedin Jemal. Cancer statistics, 2015. *CA: a cancer journal for clinicians*, 65(1):5–29, 2015.

[10] Yoav Ben-Shlomo, Simon Evans, Fowzia Ibrahim, Biral Patel, Ken Anson, Frank Chinegwundoh, Cathy Corbishley, Danny Dorling, Bethan Thomas, David Gillatt, et al. The risk of prostate cancer amongst black men in the united kingdom: the process cohort study. *European urology*, 53(1):99–105, 2008.

[11] Rosalind A Eeles, Zsofia Kote-Jarai, Ali Amin Al Olama, Graham G Giles, Michelle Guy, Gianluca Severi, Kenneth Muir, John L Hopper, Brian E Henderson, Christopher A Haiman, et al. Identification of seven new prostate cancer susceptibility loci through a genome-wide association study. *Nature genetics*, 41(10):1116–1121, 2009.

[12] Rosalind A Eeles, Ali Amin Al Olama, Sara Benlloch, Edward J Saunders, Daniel A Leongamornlert, Malgorzata Tymrakiewicz, Maya Ghoussaini, Craig Luccarini, Joe Dennis, Sarah Jugurnauth-Little, et al. Identification of 23 new prostate cancer susceptibility loci using the icogs custom genotyping array. *Nature genetics*, 45(4):385–391, 2013.

[13] Rosalind A Eeles, Zsofia Kote-Jarai, Graham G Giles, Ali Amin Al Olama, Michelle Guy, Sarah K Jugurnauth, Shani Mulholland, Daniel A Leongamornlert, Stephen M Edwards, Jonathan Morrison, et al. Multiple newly identified loci associated with prostate cancer susceptibility. *Nature genetics*, 40(3):316–321, 2008.

[14] Ali Amin Al Olama, Zsofia Kote-Jarai, Sonja I Berndt, David V Conti, Fredrick Schumacher, Ying Han, Sara Benlloch, Dennis J Hazelett, Zhaoming Wang, Ed Saunders, et al. A meta-analysis of 87,040 individuals identifies 23 new susceptibility loci for prostate cancer. *Nature genetics*, 46(10):1103–1109, 2014.

[15] Ruth C Travis, Paul N Appleby, Richard M Martin, Jeff MP Holly, Demetrius Albanes, Amanda Black, H Bas Bueno-de Mesquita, June M Chan, Chu Chen, Maria-Dolores Chirlaque, et al. A meta-analysis of individual participant data reveals an association between circulating levels of igf-i and prostate cancer risk. *Cancer research*, 76(8):2288–2300, 2016.

[16] Francesca L Crowe, Timothy J Key, Naomi E Allen, Paul N Appleby, Andrew Roddam, Kim Overvad, Henning Grønbæk, Anne Tjønneland, Jutte Halkjær, Laure Dossus, et al. The association between diet and serum concentrations of igf-i, igfbp-1, igfbp-2, and igfbp-3 in the european prospective investigation into cancer and nutrition. *Cancer Epidemiology Biomarkers & Prevention*, 18(5):1333–1340, 2009.

[17] T Norat, L Dossus, S Rinaldi, Kim Overvad, Henning Grønbæk, A Tjønneland, A Olsen, F Clavel-Chapelon, MC Boutron-Ruault, H Boeing, et al. Diet, serum insulin-like growth factor-i and igf-binding protein-3 in european women. *European journal of clinical nutrition*, 61(1):91–98, 2007.

[18] Dagfinn Aune, Deborah A Navarro Rosenblatt, Doris SM Chan, Ana Rita Vieira, Rui Vieira, Darren C Greenwood, Lars J Vatten, and Teresa Norat. Dairy products, calcium, and prostate cancer risk: a systematic review and meta-analysis of cohort studies. *The American journal of clinical nutrition*, pages ajcn–067157, 2015.

[19] Kana Wu, Donna Spiegelman, Tao Hou, Demetrius Albanes, Naomi E Allen, Sonja I Berndt, Piet A Van Den Brandt, Graham G Giles, Edward Giovannucci, R Alexandra Goldbohm, et al. Associations between unprocessed red and processed meat, poultry, seafood and egg intake and the risk of prostate cancer: A pooled analysis of 15 prospective cohort studies. *International journal of cancer*, 138(10):2368–2382, 2016.

[20] World Cancer Research Fund and American Institute for Cancer Research. *Food, nutrition, physical activity, and the prevention of cancer: a global perspective*, volume 1. Amer Inst for Cancer Research, 2007.

[21] WCRF/AICR. World cancer research fund/american institute for cancer research continuous update project: Diet, nutrition, physical activity, and prostate cancer. 2014.

[22] Neil M Davies, Tom R Gaunt, Sarah J Lewis, Jeff Holly, Jenny L Donovan, Freddie C Hamdy, John P Kemp, Rosalind Eeles, Doug Easton, Zsofia Kote-Jarai, et al. The effects of height and bmi on prostate cancer incidence and mortality: a mendelian randomization study in 20,848 cases and 20,214 controls from the practical consortium. *Cancer Causes & Control*, 26:1603–1616, 2015.

[23] Mohummad Minhaj Siddiqui, Kathryn M Wilson, Mara M Epstein, Jennifer R Rider, Neil E Martin, Meir J Stampfer, Edward L Giovannucci, and Lorelei A Mucci. Vasectomy and risk of aggressive prostate cancer: a 24-year follow-up study. *Journal of Clinical Oncology*, 32(27):3033–3038, 2014.

[24] Madhur Nayan, Robert J Hamilton, Erin M Macdonald, Qing Li, Muhammad M Mamdani, Craig C Earle, Girish S Kulkarni, Keith A Jarvi, and David N Juurlink. Vasectomy and risk of prostate cancer: population based matched cohort study. *bmj*, 355:i5546, 2016.

[25] Leslie K Dennis and Deborah V Dawson. Meta-analysis of measures of sexual activity and prostate cancer. *Epidemiology*, 13(1):72–79, 2002.

[26] Saverio Caini, Sara Gandini, Maria Dudas, Viviane Bremer, Ettore Severi, and Alin Gherasim. Sexually transmitted infections and prostate cancer risk: a systematic review and meta-analysis. *Cancer epidemiology*, 38(4):329–338, 2014.

[27] Christopher A Haiman, Daniel O Stram, Andrew J Vickers, Lynne R Wilkens, Katharina Braun, Camilla Valtonen-André, Mari Peltola, Kim Pettersson, Kevin M Waters, Loic Le Marchand, et al. Levels of beta-microseminoprotein in blood and risk of prostate cancer in multiple populations. *Journal of the National Cancer Institute*, page djs486, 2012.

[28] Andrew J Vickers, David Ulmert, Angel M Serio, Thomas Björk, Peter T Scardino, James A Eastham, Göran Berglund, and Hans Lilja.

The predictive value of prostate cancer biomarkers depends on age and time to diagnosis: Towards a biologically-based screening strategy. *International journal of cancer*, 121(10):2212–2217, 2007.

[29] David Ulmert, Angel M Cronin, Thomas Björk, Matthew F O'Brien, Peter T Scardino, James A Eastham, Charlotte Becker, Göran Berglund, Andrew J Vickers, and Hans Lilja. Prostate-specific antigen at or before age 50 as a predictor of advanced prostate cancer diagnosed up to 25 years later: a case-control study. *BMC medicine*, 6(1):6, 2008.

[30] Andrew Gelman and John Carlin. Beyond power calculations: Assessing type s (sign) and type m (magnitude) errors. *Perspectives on Psychological Science*, 9(6):641–651, 2014.

[31] Katherine S Button, John PA Ioannidis, Claire Mokrysz, Brian A Nosek, Jonathan Flint, Emma SJ Robinson, and Marcus R Munafò. Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14(5):365–376, 2013.

[32] Laurence S Freedman, Arthur Schatzkin, Douglas Midthune, and Victor Kipnis. Dealing with dietary measurement error in nutritional cohort studies. *Journal of the National Cancer Institute*, 103(14):1086–1092, 2011.

[33] LK Dennis, DV Dawson, and MI Resnick. Vasectomy and the risk of prostate cancer: a meta-analysis examining vasectomy status, age at vasectomy, and time since vasectomy. *Prostate cancer and prostatic diseases*, 5(3):193–203, 2001.

[34] Eric J Jacobs, Rebecca L Anderson, Victoria L Stevens, Christina C Newton, Ted Gansler, and Susan M Gapstur. Vasectomy and prostate cancer incidence and mortality in a large us cohort. *Journal of Clinical Oncology*, page JCO662361, 2016.

[35] SS Howards. Possible biological mechanisms for a relationship between vasectomy and prostatic cancer. *European journal of Cancer*, 29(7):1060–1062, 1993.

[36] Zsofia Kote-Jarai, Douglas F Easton, Janet L Stanford, Elaine A Ostrander, Johanna Schleutker, Sue A Ingles, Daniel Schaid, Stephen Thibodeau, Thilo Dörk, David Neal, et al. Multiple novel prostate cancer predisposition loci confirmed by an international study: the practical consortium. *Cancer Epidemiology Biomarkers & Prevention*, 17(8):2052–2061, 2008.

[37] Christopher I Amos, Joe Dennis, Zhaoming Wang, Jinyoung Byun, Fredrick R Schumacher, Simon A Gayther, Graham Casey, David J Hunter, Thomas A Sellers, Stephen B Gruber, et al. The oncoarray consortium: a network for understanding the genetic architecture of common cancers. *Cancer Epidemiology and Prevention Biomarkers*, 26(1):126–135, 2017.

[38] DJ Hunter, E Riboli, CA Haiman, D Albanes, D Altshuler, SJ Chanock, RB Haynes, BE Henderson, R Kaaks, DO Stram, et al. A candidate gene approach to searching for low-penetrance breast and prostate cancer genes. *Nature reviews. Cancer*, 5(12):977–985, 2005.

[39] H Lilja and P-A Abrahamsson. Three predominant proteins secreted by the human prostate gland. *The Prostate*, 12(1):29–38, 1988.

[40] Charlotte Becker, Timo Piironen, Kim Pettersson, Jonas Hugosson, and Hans Lilja. Testing in serum for human glandular kallikrein 2, and free and total prostate specific antigen in biannual screening for prostate cancer. *The Journal of urology*, 170(4):1169–1174, 2003.

[41] Charlotte Becker, Timo Piironen, Kim Pettersson, Jonas Hugosson, and Hans Lilja. Clinical value of human glandular kallikrein 2 and free and total prostate-specific antigen in serum from a population of men with prostate-specific antigen levels 3.0 ng/ml or greater. *Urology*, 55(5):694–699, 2000.

[42] Amine Benchikh, Caroline Savage, Angel Cronin, Gilles Salama, Arnauld Villers, Hans Lilja, and Andrew Vickers. A panel of kallikrein markers can predict outcome of prostate biopsy following clinical workup: an independent validation study from the european randomized study of prostate cancer screening, france. *BMC cancer*, 10(1):635, 2010.

[43] Andrew J Vickers, Angel M Cronin, Monique J Roobol, Caroline J Savage, Mari Peltola, Kim Pettersson, Peter T Scardino, Fritz H Schröder, and Hans Lilja. A four-kallikrein panel predicts prostate cancer in men with recent screening: data from the european randomized study of screening for prostate cancer, rotterdam. *Clinical Cancer Research*, 16(12):3232–3239, 2010.

[44] Andrew Vickers, Angel Cronin, Monique Roobol, Caroline Savage, Mari Peltola, Kim Pettersson, Peter T Scardino, Fritz Schröder, and Hans Lilja. Reducing unnecessary biopsy during prostate cancer screening using a four-kallikrein panel: an independent replication. *Journal of Clinical Oncology*, 28(15):2493–2498, 2010.

[45] Andrew J Vickers, Tineke Wolters, Caroline J Savage, Angel M Cronin, M Frank OBrien, Kim Pettersson, Monique J Roobol, Gunnar Aus, Peter T Scardino, Jonas Hugosson, et al. Prostate-specific antigen velocity for early detection of prostate cancer: result from a large, representative, population-based cohort. *European urology*, 56(5):753–760, 2009.

[46] Andrew J Vickers, Angel M Cronin, Gunnar Aus, Carl-Gustav Pihl, Charlotte Becker, Kim Pettersson, Peter T Scardino, Jonas Hugosson, and Hans Lilja. A panel of kallikrein markers can reduce unnecessary biopsy for prostate cancer: data from the european randomized study of prostate cancer screening in göteborg, sweden. *BMC medicine*, 6(1):19, 2008.

[47] Charlotte Becker, Timo Piironen, Kim Pettersson, Thomas Björk, Kirk J Wojno, Joseph E Oesterling, and Hans Lilja. Discrimination of men with prostate cancer from those with benign disease by measurements of human glandular kallikrein 2 (hk2) in serum. *The Journal of urology*, 163(1):311–316, 2000.

[48] Franz Recker, Maciej K Kwiatkowski, Timo Piironen, Kim Pettersson, Andreas Huber, Gerd Lümmen, and Reto Tscholl. Human glandular kallikrein as a tool to improve discrimination of poorly differentiated and non-organ-confined prostate cancer compared with prostate-specific antigen. *Urology*, 55(4):481–485, 2000.

[49] Alexander Haese, Charlotte Becker, Joachim Noldus, Markus Graefen, Edith Huland, Hartwig Huland, and Hans Lilja. Human glandular kallikrein 2: a potential serum marker for predicting the organ confined versus nonorgan confined growth of prostate cancer. *The Journal of urology*, 163(5):1491–1497, 2000.

[50] Alexander Haese, Markus Graefen, Thomas Steuber, Charlotte Becker, Kim Pettersson, Timo Piironen, Joachim Noldus, Hartwig Huland, and Hans Lilja. Human glandular kallikrein 2 levels in serum for discrimination of pathologically organ-confined from locally-advanced prostate cancer in total psa-levels below 10 ng/ml. *The Prostate*, 49(2):101–109, 2001.

[51] Thomas Steuber, Andrew J Vickers, Angel M Serio, Ville Vaisanen, Alexander Haese, Kim Pettersson, James A Eastham, Peter T Scardino, Hartwig Huland, and Hans Lilja. Comparison of free and total forms of serum human kallikrein 2 and prostate-specific antigen for prediction of locally advanced and recurrent prostate cancer. *Clinical chemistry*, 53(2):233–240, 2007.

[52] Alexander Haese, Ville Vaisanen, Hans Lilja, Michael W Kattan, Harry G Rittenhouse, Kim Pettersson, Daniel W Chan, Hartwig Huland, Lori J Sokoll, and Alan W Partin. Comparison of predictive accuracy for pathologically organ confined clinical stage t1c prostate cancer using human glandular kallikrein 2 and prostate specific antigen combined with clinical stage and gleason grade. *The Journal of urology*, 173(3):752–756, 2005.

[53] René Raaijmakers, Stijn H de Vries, Bert G Blijenberg, Mark F Wildhagen, Renske Postma, Chris H Bangma, Claude Darte, and Fritz H Schröder. hk2 and free psa, a prognostic combination in predicting minimal prostate cancer in screen-detected men within the psa range 4–10 ng/ml. *european urology*, 52(5):1358–1364, 2007.

[54] Leena Peltonen, Nabil Enattah, Irma Jarvela, Timo Sahi, Erkki Savilahti, and Joseph Terwilliger. Identification of a dna variant associated with adult type hypolactasia, August 28 2012. US Patent 8,252,537.

[55] Ruth C Travis, Paul N Appleby, Afshan Siddiq, Naomi E Allen, Rudolf Kaaks, Federico Canzian, Silke Feller, Anne Tjønneland, Nina Føns Johnsen, Kim Overvad, et al. Genetic variation in the lactase gene, dairy product intake and risk for prostate cancer in the european prospective investigation into cancer and nutrition. *International Journal of Cancer*, 132(8):1901–1910, 2013.

[56] Suvi Torniainen, Maria Hedelin, Ville Autio, Heli Rasinperä, Katarina Augustsson Bälter, Åsa Klint, Rino Bellocco, Fredrik Wiklund, Pär Stattin, Tarja Ikonen, et al. Lactase persistence, dietary intake of milk, and the risk for prostate cancer in sweden and finland. *Cancer Epidemiology Biomarkers & Prevention*, 16(5):956–961, 2007.

[57] Fernando Pires Hartwig, Bernardo Lessa Horta, George Davey Smith, Christian Loret de Mola, and Cesar Gomes Victora. Association of lactase persistence genotype with milk consumption, obesity and blood pressure: a mendelian randomization study in the 1982 pelotas (brazil) birth cohort, with a systematic review and meta-analysis. *International journal of epidemiology*, 45(5):1573–1587, 2016.

[58] Adil J Malek, Yann C Klimentidis, Kenneth P Kell, and José R Fernández. Associations of the lactase persistence allele and lactose intake with body composition among multiethnic children. *Genes & nutrition*, 8(5):487–494, 2013.

[59] Ming Ding, Tao Huang, Helle KM Bergholdt, Børge G Nordestgaard, Christina Ellervik, and Lu Qi. Dairy consumption, systolic blood pressure, and risk of hypertension: Mendelian randomization study. *bmj*, 356:j1000, 2017.

[60] Uk family food survey 2015. `https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/597667/Family_Food_2015-09mar17.pdf`. Accessed: 2017-06-15.

[61] Gibran Hemani, Jie Zheng, Kaitlin H Wade, Charles Laurin, Benjamin Elsworth, Stephen Burgess, Jack Bowden, Ryan Langdon, Vanessa Tan, James Yarmolinsky, et al. Mr-base: a platform for systematic causal inference across the phenome using billions of genetic associations. *bioRxiv*, page 078972, 2016.

[62] Sebastien Antoni, Isabelle Soerjomataram, Bjørn Møller, Freddie Bray, and Jacques Ferlay. An assessment of globocan methods for deriving national estimates of cancer incidence. *Methods*, 3:4, 2016.

[63] Lindsey A Torre, Freddie Bray, Rebecca L Siegel, Jacques Ferlay, Joannie Lortet-Tieulent, and Ahmedin Jemal. Global cancer statistics, 2012. *CA: a cancer journal for clinicians*, 65(2):87–108, 2015.

[64] Ruth Etzioni, David F Penson, Julie M Legler, Dante Di Tommaso, Rob Boer, Peter H Gann, and Eric J Feuer. Overdiagnosis due to prostate-specific antigen screening: lessons from us prostate cancer incidence trends. *Journal of the National Cancer Institute*, 94(13):981–990, 2002.

[65] H Gilbert Welch and Peter C Albertsen. Prostate cancer diagnosis and treatment after the introduction of prostate-specific antigen screening: 1986–2005. *Journal of the National Cancer Institute*, 101(19):1325–1329, 2009.

[66] Gurdarshan S Sandhu and Gerald L Andriole. Overdiagnosis of prostate cancer. *Journal of the National Cancer Institute. Monographs*, 2012(45):146, 2012.

[67] Benjamin F Hankey, Eric J Feuer, Limin X Clegg, Richard B Hayes, Julie M Legler, Phillip C Prorok, Lynn A Ries, Ray M Merrill, and Richard S Kaplan. Cancer surveillance series: interpreting trends in prostate cancerpart i: evidence of the effects of screening in recent prostate cancer incidence, mortality, and survival rates. *Journal of the National Cancer Institute*, 91(12):1017–1024, 1999.

[68] Simon M Collin, Richard M Martin, Chris Metcalfe, David Gunnell, Peter C Albertsen, David Neal, Freddie Hamdy, Peter Stephens, J Athene Lane, Rollo Moore, et al. Prostate-cancer mortality in the usa and uk in 1975–2004: an ecological study. *The lancet oncology*, 9(5):445–452, 2008.

[69] Vickie L Shavers and Martin L Brown. Racial and ethnic disparities in the receipt of cancer treatment. *Journal of the National Cancer Institute*, 94(5):334–357, 2002.

[70] Elizabeth Ward, Ahmedin Jemal, Vilma Cokkinides, Gopal K Singh, Cheryll Cardinez, Asma Ghafoor, and Michael Thun. Cancer disparities by race/ethnicity and socioeconomic status. *CA: a cancer journal for clinicians*, 54(2):78–93, 2004.

[71] S Moss, J Melia, J Sutton, C Mathews, and M Kirby. Prostate-specific antigen testing rates and referral patterns from general practice data in england. *International journal of clinical practice*, 70(4):312–318, 2016.

[72] Pär Stattin, Andrew J Vickers, Daniel D Sjoberg, Robert Johansson, Torvald Granfors, Mattias Johansson, Kim Pettersson, Peter T Scardino, Göran Hallmans, and Hans Lilja. Improving the specificity of screening for lethal prostate cancer using prostate-specific antigen and a panel of kallikrein markers: a nested case–control study. *European urology*, 2015.

[73] Andrew J Vickers, Angel M Cronin, Thomas Björk, Jonas Manjer, Peter M Nilsson, Anders Dahlin, Anders Bjartell, Peter T Scardino, David Ulmert, and Hans Lilja. Prostate specific antigen concentration at age 60 and death or metastasis from prostate cancer: case-control study. *Bmj*, 341:c4521, 2010.

[74] Mark A Preston, Julie L Batista, Kathryn M Wilson, Sigrid V Carlsson, Travis Gerke, Daniel D Sjoberg, Douglas M Dahl, Howard D Sesso, Adam S Feldman, Peter H Gann, et al. Baseline prostate-specific antigen levels in midlife predict lethal prostate cancer. *Journal of Clinical Oncology*, 34(23):2705–2711, 2016.

[75] WJ Catalona, JP Richie, Frederick R Ahmann, MA Hudson, PT Scardino, RC Flanigan, JB Dekernion, TL Ratliff, LR Kavoussi, and Bruce L Dalkin. Comparison of digital rectal examination and serum prostate specific antigen in the early detection of prostate cancer: results of a multicenter clinical trial of 6,630 men. *The Journal of urology*, 151(5):1283–1290, 1994.

[76] Fritz H Schröder, Jonas Hugosson, Monique J Roobol, Teuvo LJ Tammela, Stefano Ciatto, Vera Nelen, Maciej Kwiatkowski, Marcos Lujan,

Hans Lilja, Marco Zappa, et al. Screening and prostate-cancer mortality in a randomized european study. *New England Journal of Medicine*, 360(13):1320–1328, 2009.

[77] Jonas Hugosson, Sigrid Carlsson, Gunnar Aus, Svante Bergdahl, Ali Khatami, Pär Lodding, Carl-Gustaf Pihl, Johan Stranne, Erik Holmberg, and Hans Lilja. Mortality results from the göteborg randomised population-based prostate-cancer screening trial. *The lancet oncology*, 11(8):725–732, 2010.

[78] Gerald L Andriole, E David Crawford, Robert L Grubb III, Saundra S Buys, David Chia, Timothy R Church, Mona N Fouad, Edward P Gelmann, Paul A Kvale, Douglas J Reding, et al. Mortality results from a randomized prostate-cancer screening trial. *New England Journal of Medicine*, 360(13):1310–1319, 2009.

[79] Jonathan E Shoag, Sameer Mittal, and Jim C Hu. Reevaluating psa testing rates in the plco trial. *New England Journal of Medicine*, 374(18):1795–1796, 2016.

[80] Peter C Albertsen, James A Hanley, and Judith Fine. 20-year outcomes following conservative management of clinically localized prostate cancer. *Jama*, 293(17):2095–2101, 2005.

[81] N Mottet, J Bellmunt, E Briers, et al. European association of urology. guidelines on prostate cancer, 2015.

[82] Andrew Wolf, Richard C Wender, Ruth B Etzioni, Ian M Thompson, Anthony V D'Amico, Robert J Volk, Durado D Brooks, Chiranjeev Dash, Idris Guessous, Kimberly Andrews, et al. American cancer society guideline for the early detection of prostate cancer: update 2010. *CA: a cancer journal for clinicians*, 60(2):70–98, 2010.

[83] Michał Kiciński, Jaco Vangronsveld, and Tim S Nawrot. An epidemiological reappraisal of the familial aggregation of prostate cancer: a meta-analysis. *PloS one*, 6(10):e27130, 2011.

[84] H Ballentine Carter, Peter C Albertsen, Michael J Barry, Ruth Etzioni, Stephen J Freedland, Kirsten Lynn Greene, Lars Holmberg, Philip Kantoff, Badrinath R Konety, Mohammad Hassan Murad, et al. Early

detection of prostate cancer: Aua guideline. *The Journal of urology*, 190(2):419–426, 2013.

[85] Andrew J Vickers, Daniel D Sjoberg, David Ulmert, Emily Vertosick, Monique J Roobol, Ian Thompson, Eveline AM Heijnsdijk, Harry De Koning, Coral Atoria-Swartz, Peter T Scardino, et al. Empirical estimates of prostate cancer overdiagnosis by age and prostate-specific antigen. *BMC medicine*, 12(1):1, 2014.

[86] MJ Resnick, DJ Canter, TJ Guzzo, BM Brucker, M Bergey, SS Sonnad, AJ Wein, and SB Malkowicz. Does race affect postoperative outcomes in patients with low-risk prostate cancer who undergo radical prostatectomy? *Urology*, 73(3):620–623, 2009.

[87] Folakemi T Odedina, Titilola O Akinremi, Frank Chinegwundoh, Robin Roberts, Daohai Yu, R Renee Reams, Matthew L Freedman, Brian Rivers, B Lee Green, and Nagi Kumar. Prostate cancer disparities in black men of african descent: a comparative literature review of prostate cancer burden among black men in the united states, caribbean, united kingdom, and west africa. *Infectious Agents and Cancer*, 4(1):1, 2009.

[88] Yu-Hsuan Shao, Kitaw Demissie, Weichung Shih, Amit R Mehta, Mark N Stein, Calpurnyia B Roberts, Robert S DiPaola, and Grace L Lu-Yao. Contemporary risk profile of prostate cancer in the united states. *Journal of the National Cancer Institute*, 101(18):1280–1283, 2009.

[89] H Nicole Tran, Yan Li, Natalia Udaltsova, Mary Anne Armstrong, Gary D Friedman, and Arthur L Klatsky. Risk of cancer in asian americans: a kaiser permanente cohort study. *Cancer Causes & Control*, 27(10):1197–1207, 2016.

[90] Ruth Etzioni, Kristin M Berry, Julie M Legler, and Pamela Shaw. Prostate-specific antigen testing in black and white men: an analysis of medicare claims from 1991–1998. *Urology*, 59(2):251–255, 2002.

[91] Angela B Mariotto, Ruth Etzioni, Martin Krapcho, and Eric J Feuer. Reconstructing psa testing patterns between black and white men in

the us from medicare claims and the national health interview survey. *Cancer*, 109(9):1877–1886, 2007.

[92] TJ Littlejohns, RC Travis, TJ Key, and NE Allen. Op15 characteristics of men who have had a prostate-specific antigen test: cross-sectional findings for 212,039 men from uk biobank. *Journal of Epidemiology and Community Health*, 69(Suppl 1):A15–A15, 2015.

[93] Michel P Coleman, Manuela Quaresma, Franco Berrino, Jean-Michel Lutz, Roberta De Angelis, Riccardo Capocaccia, Paolo Baili, Bernard Rachet, Gemma Gatta, Timo Hakulinen, et al. Cancer survival in five continents: a worldwide population-based study (concord). *The lancet oncology*, 9(8):730–756, 2008.

[94] Song-Yi Park, Christopher A Haiman, Iona Cheng, Sungshim Lani Park, Lynne R Wilkens, Laurence N Kolonel, Loïc Le Marchand, and Brian E Henderson. Racial/ethnic differences in lifestyle-related factors and prostate cancer risk: the multiethnic cohort study. *Cancer Causes & Control*, 26(10):1507–1515, 2015.

[95] Haitao Chen, Hongjie Yu, Jianqing Wang, Zheng Zhang, Zhengrong Gao, Zhuo Chen, Yulan Lu, Wennuan Liu, Deke Jiang, S Lilly Zheng, et al. Systematic enrichment analysis of potentially functional regions for 103 prostate cancer risk-associated loci. *The Prostate*, 2015.

[96] John I Jones and David R Clemmons. Insulin-like growth factors and their binding proteins: Biological actions*. *Endocrine reviews*, 16(1):3–34, 1995.

[97] Adda Grimberg. Mechanisms by which igf-i may promote cancer. *Cancer biology & therapy*, 2(6):630–635, 2003.

[98] NE Allen, TJ Key, PN Appleby, RC Travis, AW Roddam, A Tjønneland, NF Johnsen, K Overvad, J Linseisen, S Rohrmann, et al. Animal foods, protein, calcium and prostate cancer risk: the european prospective investigation into cancer and nutrition. *British journal of cancer*, 98(9):1574–1581, 2008.

[99] Teodora Ribarska, Klaus-Marius Bastian, Annemarie Koch, and Wolfgang A Schulz. Specific changes in the expression of imprinted genes in prostate cancer–implications for cancer progression and epigenetic regulation. *Asian J Androl*, 14(3):436–450, 2012.

[100] Andreas Hoeflich, Raquel Reisinger, Harald Lahm, Wieland Kiess, Werner F Blum, Helmut J Kolb, Matthias M Weber, and Eckhard Wolf. Insulin-like growth factor-binding protein 2 in tumorigenesis protector or promoter? *Cancer research*, 61(24):8601–8610, 2001.

[101] R Mehrian-Shai, CD Chen, T Shi, S Horvath, SF Nelson, JKV Reichardt, and CL Sawyers. Insulin growth factor-binding protein 2 is a candidate biomarker for pten status and pi3k/akt pathway activation in glioblastoma and prostate cancer. *Proceedings of the National Academy of Sciences*, 104(13):5563–5568, 2007.

[102] Mari-Anne Rowlands, Jeff MP Holly, David Gunnell, Rebecca Gilbert, Jenny Donovan, J Athene Lane, Gemma Marsden, Simon M Collin, Freddie Hamdy, David E Neal, et al. The relation between adiposity throughout the life course and variation in igfs and igfbps: evidence from the protect (prostate testing for cancer and treatment) study. *Cancer Causes & Control*, 21(11):1829–1842, 2010.

[103] Stephen B Wheatcroft, Mark T Kearney, Ajay M Shah, Vivienne A Ezzat, John R Miell, Michael Modo, Stephen CR Williams, Will P Cawthorn, Gema Medina-Gomez, Antonio Vidal-Puig, et al. Igf-binding protein-2 protects against the development of obesity and insulin resistance. *Diabetes*, 56(2):285–294, 2007.

[104] Bruce Armstrong and Richard Doll. Environmental factors and cancer incidence and mortality in different countries, with special reference to dietary practices. *International journal of cancer*, 15(4):617–631, 1975.

[105] William Haenszel and Minoru Kurihara. Studies of japanese migrants. i. mortality from cancer and other diseases among japanese in the united states. *Journal of the National Cancer Institute*, 40(1):43–68, 1968.

[106] Alice S Whittemore, Laurence N Kolonel, Anna H Wu, Esther M John, Richard P Gallagher, Geoffrey R Howe, J David Burch, Jean Hankin, Darlene M Dreon, Dee W West, et al. Prostate cancer in relation to diet, physical activity, and body size in blacks, whites, and asians in the united states and canada. *Journal of the National Cancer Institute*, 87(9):652–661, 1995.

[107] H Shimizu, RK Ross, L Bernstein, R Yatani, BE Henderson, and TM Mack. Cancers of the prostate and breast among japanese and white immigrants in los angeles county. *British journal of cancer*, 63(6):963, 1991.

[108] Motoki Iwasaki, Cecilia Polidoro Mameri, Gerson Shigueaki Hamada, and Shoichiro Tsugane. Cancer mortality among japanese immigrants and their descendants in the state of sao paulo, brazil, 1999–2001. *Japanese journal of clinical oncology*, 34(11):673–680, 2004.

[109] Richard Doll and Richard Peto. The causes of cancer: quantitative estimates of avoidable risks of cancer in the united states today. *Journal of the National Cancer Institute*, 66(6):1192–1308, 1981.

[110] CS Muir, J Nectoux, and J Staszewski. The epidemiology of prostatic cancer: geographical distribution and time-trends. *Acta oncologica*, 30(2):133–140, 1991.

[111] Michael Marmot, T Atinmo, T Byers, J Chen, T Hirohata, A Jackson, W James, L Kolonel, S Kumanyika, C Leitzmann, et al. Food, nutrition, physical activity, and the prevention of cancer: a global perspective. 2007.

[112] Timothy J Key, Naomi E Allen, Elizabeth A Spencer, and Ruth C Travis. The effect of diet on risk of cancer. *The Lancet*, 360(9336):861–868, 2002.

[113] Alison J Price, Ruth C Travis, Paul N Appleby, Demetrius Albanes, Aurelio Barricarte Gurrea, Tone Bjørge, H Bas Bueno-de Mesquita, Chu Chen, Jenny Donovan, Randi Gislefoss, et al. Circulating folate and vitamin b 12 and risk of prostate cancer: A collaborative analysis of individual participant data from six cohorts including 6875 cases and 8104 controls. *European urology*, 2016.

[114] Michael Fenech. The role of folic acid and vitamin b12 in genomic stability of human cells. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 475(1):57–67, 2001.

[115] Yan Song, Jorge E Chavarro, Yin Cao, Weiliang Qiu, Lorelei Mucci, Howard D Sesso, Meir J Stampfer, Edward Giovannucci, Michael Pollak, Simin Liu, et al. Whole milk intake is associated with prostate cancer-specific mortality among us male physicians. *The Journal of nutrition*, pages jn–112, 2013.

[116] PC Dagnelie, AG Schuurman, RA Goldbohm, and PA Van den Brandt. Diet, anthropometric measures and prostate cancer risk: a review of prospective cohort and intervention studies. *BJU international*, 93(8):1139–1150, 2004.

[117] Dominik D Alexander, Pamela J Mink, Colleen A Cushing, and Bonnie Sceurman. A review and meta-analysis of prospective studies of red and processed meat intake and prostate cancer. *Nutrition journal*, 9(1):1, 2010.

[118] Erin L Richman, Meir J Stampfer, Alan Paciorek, Jeanette M Broering, Peter R Carroll, and June M Chan. Intakes of meat, fish, poultry, and eggs and risk of prostate cancer progression[1], [2]. *American journal of clinical nutrition AJN*, 2010.

[119] Naomi E Allen, Ruth C Travis, Paul N Appleby, Demetrius Albanes, Matt J Barnett, Amanda Black, H Bas Bueno-de Mesquita, Mélanie Deschasaux, Pilar Galan, Gary E Goodman, et al. Selenium and prostate cancer: Analysis of individual participant data from fifteen prospective studies. *Journal of the National Cancer Institute*, 108(11):djw153, 2016.

[120] Timothy J Key, Paul N Appleby, Ruth C Travis, Demetrius Albanes, Anthony J Alberg, Aurelio Barricarte, Amanda Black, Heiner Boeing, H Bas Bueno-de Mesquita, June M Chan, et al. Carotenoids, retinol, tocopherols, and prostate cancer risk: pooled analysis of 15 studies–3. *The American journal of clinical nutrition*, 102(5):1142–1157, 2015.

[121] A Perez-Cornago, PN Appleby, T Pischon, KK Tsilidis, A Tjønneland, A Olsen, K Overvad, R Kaaks, T Kühn, H Boeing, et al. Tall height and obesity are associated with an increased risk of aggressive prostate cancer: results from the epic cohort study. 2017.

[122] Chi Gao, Chirag J Patel, Kyriaki Michailidou, Ulrike Peters, Jian Gong, Joellen Schildkraut, Fredrick R Schumacher, Wei Zheng, Paolo Boffetta, Isabelle Stucker, et al. Mendelian randomization study of adiposity-related traits and risk of breast, ovarian, prostate, lung and colorectal cancer. *International Journal of Epidemiology*, page dyw129, 2016.

[123] D Bansal, A Bhansali, G Kapil, K Undela, and P Tiwari. Type 2 diabetes and risk of prostate cancer: a meta-analysis of observational studies. *Prostate cancer and prostatic diseases*, 16(2):151–158, 2013.

[124] Konstantinos K Tsilidis, Naomi E Allen, Paul N Appleby, Sabine Rohrmann, Ute Nöthlings, Larraitz Arriola, Marc J Gunter, Veronique Chajes, Sabina Rinaldi, Isabelle Romieu, et al. Diabetes mellitus and risk of prostate cancer in the european prospective investigation into cancer and nutrition. *International Journal of Cancer*, 136(2):372–381, 2015.

[125] Eric L Ding, Yiqing Song, Vasanti S Malik, and Simin Liu. Sex differences of endogenous sex hormones and risk of type 2 diabetes: a systematic review and meta-analysis. *Jama*, 295(11):1288–1299, 2006.

[126] David Spiegelhalter. *Sex by numbers: what statistics can tell us about sexual behaviour*. Profile Books, 2015.

[127] Mark H Kawachi, Robert R Bahnson, Michael Barry, J Erik Busby, Peter R Carroll, H Ballentine Carter, William J Catalona, Michael S Cookson, Jonathan I Epstein, Ruth B Etzioni, et al. Prostate cancer early detection. *Journal of the National Comprehensive Cancer Network*, 8(2):240–262, 2010.

[128] Emily A Vertosick, Bing Ying Poon, and Andrew J Vickers. Relative value of race, family history and prostate specific antigen as indications for early initiation of prostate cancer screening. *The Journal of urology*, 192(3):724–729, 2014.

[129] Andrew J Vickers, David Ulmert, Daniel D Sjoberg, Caroline J Bennette, Thomas Björk, Axel Gerdtsson, Jonas Manjer, Peter M Nilsson, Anders Dahlin, Anders Bjartell, et al. Strategy for detection of prostate cancer based on relation between prostate specific antigen at age 40-55 and long term risk of metastasis: case-control study. *Bmj*, 346:f2023, 2013.

[130] Richard J Bryant, Daniel D Sjoberg, Andrew J Vickers, Mary C Robinson, Rajeev Kumar, Luke Marsden, Michael Davis, Peter T Scardino, Jenny Donovan, David E Neal, et al. Predicting high-grade cancer at ten-core prostate biopsy using four kallikrein markers measured in blood in the protect study. *Journal of the National Cancer Institute*, 107(7):djv095, 2015.

[131] Katharina Braun, Daniel D Sjoberg, Andrew J Vickers, Hans Lilja, and Anders S Bjartell. A four-kallikrein panel predicts high-grade cancer on biopsy: Independent validation in a community cohort. *European urology*, 2015.

[132] Andrew Vickers, Emily A Vertosick, Daniel D Sjoberg, Monique J Roobol, Freddie Hamdy, David Neal, Anders Bjartell, Jonas Hugosson, Jenny L Donovan, Arnauld Villers, et al. Properties of the four kallikrein panel outside the diagnostic grey zone: meta-analysis of patients with positive digital rectal exam or prostate-specific antigen 10 ng/ml and above. *The Journal of Urology*, 2016.

[133] FH Schröder, P Hermanek, L Denis, WR Fair, MK Gospodarowicz, and M Pavone-Macaluso. The tnm classification of prostate cancer. *The Prostate*, 21(S4):129–138, 1992.

[134] Donald F Gleason. Histologic grading of prostate cancer: a perspective. *Human pathology*, 23(3):273–279, 1992.

[135] Joseph C Presti Jr. Prostate biopsy: current status and limitations. *Reviews in urology*, 9(3):93, 2007.

[136] Klaus Eichler, Susanne Hempel, Jennifer Wilby, Lindsey Myers, Lucas M Bachmann, and Jos Kleijnen. Diagnostic value of systematic biopsy methods in the investigation of prostate cancer: a systematic review. *The Journal of urology*, 175(5):1605–1612, 2006.

[137] Peter C Albertsen, James A Hanley, George H Barrows, David F Penson, Pam DH Kowalczyk, M Melinda Sanders, and Judith Fine. Prostate cancer and the will rogers phenomenon. *Journal of the National Cancer Institute*, 97(17):1248–1253, 2005.

[138] Jonathan I Epstein, William C Allsbrook Jr, Mahul B Amin, Lars L Egevad, ISUP Grading Committee, et al. The 2005 international society of urological pathology (isup) consensus conference on gleason grading of prostatic carcinoma. *The American journal of surgical pathology*, 29(9):1228–1242, 2005.

[139] Charlie Schmidt. Gleason scoring system faces change and debate. *Journal of the National Cancer Institute*, 101(9):622–629, 2009.

[140] E Riboli, KJ Hunt, N Slimani, P Ferrari, T Norat, M Fahey, UR Charrondiere, B Hemon, C Casagrande, J Vignat, et al. European prospective investigation into cancer and nutrition (epic): study populations and data collection. *Public health nutrition*, 5(6b):1113–1124, 2002.

[141] Uk biobank: Protocol for a large-scale prospective epidemiological resource. `http://www.ukbiobank.ac.uk/wp-content/uploads/2011/11/UK-Biobank-Protocol.pdf?phpMyAdmin=trmKQlYdjjnQIgJ\%2CfAzikMhEnx6`. Accessed: 2017-03-12.

[142] Shengxu Li, Jing Hua Zhao, Jian'an Luan, Ulf Ekelund, Robert N Luben, Kay-Tee Khaw, Nicholas J Wareham, and Ruth JF Loos. Physical activity attenuates the genetic predisposition to obesity in 20,000 men and women from epic-norfolk prospective population study. *PLoS Med*, 7(8):e1000332, 2010.

[143] Francesca L Crowe, Timothy J Key, Naomi E Allen, Paul N Appleby, Kim Overvad, Henning Grønbæk, Anne Tjønneland, Jytte Halkjær, Laure Dossus, Heiner Boeing, et al. A cross-sectional analysis of the associations between adult height, bmi and serum concentrations of igf-i and igfbp-1-2 and-3 in the european prospective investigation into cancer and nutrition (epic). *Annals of human biology*, 38(2):194–202, 2011.

[144] Julie A Schmidt, Sabina Rinaldi, Augustin Scalbert, Pietro Ferrari, David Achaintre, Marc J Gunter, Paul N Appleby, Timothy J Key, and Ruth C Travis. Plasma concentrations and intakes of amino acids in male meat-eaters, fish-eaters, vegetarians and vegans: a cross-sectional analysis in the epic-oxford cohort. *European journal of clinical nutrition*, 70(3):306–312, 2016.

[145] Heiner Boeing, A Korfmann, and MM Bergmann. Recruitment procedures of epic-germany. *Annals of nutrition and metabolism*, 43(4):205–215, 1999.

[146] Antonio Agudo, Laia Cabrera, Pilar Amiano, Eva Ardanaz, Aurelio Barricarte, Toni Berenguer, María D Chirlaque, Miren Dorronsoro, Paula Jakszyn, Nerea Larranaga, et al. Fruit and vegetable intakes, dietary antioxidant nutrients, and total mortality in spanish adults: findings from the spanish cohort of the european prospective investigation into cancer and nutrition (epic-spain). *The American journal of clinical nutrition*, 85(6):1634–1642, 2007.

[147] Anne Tjønneland, Anja Olsen, Katja Boll, Connie Stripp, Jane Christensen, Gerda Engholm, and Kim Overvad. Study design, exposure variables, and socioeconomic determinants of participation in diet, cancer and health: a population-based prospective cohort study of 57,053 men and women in denmark. *Scandinavian journal of public health*, 35(4):432–441, 2007.

[148] N Day, S Oakes, R Luben, K-T Khaw, S Bingham, A Welch, and N Wareham. Epic-norfolk: study design and characteristics of the cohort. 1999.

[149] B Margetts and Pirjo Pietinen. European prospective investigation into cancer and nutrition: validity studies on dietary assessment methods. *International Journal of Epidemiology*, 26(90001):1–5, 1997.

[150] Nadia Slimani, Geneviève Deharveng, Ruth U Charrondière, Anne Linda van Kappel, Marga C Ocké, Ailsa Welch, Areti Lagiou, Marti van Liere, Antonio Agudo, Valeria Pala, et al. Structure of the standardized computerized 24-h diet recall interview used as reference

method in the 22 centers participating in the epic project. *Computer methods and programs in biomedicine*, 58(3):251–266, 1999.

[151] Anja Kroke, Kerstin Klipstein-Grobusch, Susanne Voss, Jutta Möseneder, Frank Thielecke, Rudolf Noack, and Heiner Boeing. Validation of a self-administered food-frequency questionnaire administered in the european prospective investigation into cancer and nutrition (epic) study: comparison of energy, protein, and macronutrient intakes estimated with the doubly labeled water, urinary nitrogen, and repeated 24-h dietary recall methods. *The American journal of clinical nutrition*, 70(4):439–447, 1999.

[152] K Overvad, A Tjønneland, J Haraldsdöttir, S Bang, M Ewertz, and O Møller-Jensen. Development of a semi-quantitative food frequency questionnaire to assess food, energy and nutrient intake in denmark. *International Journal of Epidemiology*, 20:906–912, 1991.

[153] Elio Riboli, SÖLVE Elmståhl, Rodolfo Saracci, Bo Gullberg, and FOLKE Lindgärde. The malmö food study: validity of two dietary assessment methods for measuring nutrient intake. *International journal of epidemiology*, 26(suppl 1):S161, 1997.

[154] SA Bingham, C Gill, A Welch, K Day, A Cassidy, KT Khaw, MJ Sneyd, TJA Key, L Roe, and NE Day. Comparison of dietary assessment methods in nutritional epidemiology: weighed records v. 24 h recalls, food-frequency questionnaires and estimated-diet records. *British Journal of Nutrition*, 72(04):619–643, 1994.

[155] A Fry, T Littlejohns, C Sudlow, N Doherty, L Adamska, T Sprosen, R Collins, and NE Allen. Comparison of sociodemographic and health-related characteristics of uk biobank participants with the general population.

[156] Uk biobank touchscreen questionnaire. `https://www.ukbiobank.ac.uk/wp-content/uploads/2011/06/Touch_screen_questionnaire.pdf?phpMyAdmin=trmKQlYdjjnQIgJ\%2CfAzikMhEnx6`. Accessed: 2017-05-9.

[157] 24-hour dietary recall questionnaire. `https://biobank.ctsu.ox.ac.uk/crystal/docs/DietWebQ.pdf`. Accessed: 2017-05-9.

[158] Lynne Henderson, Karen Irving, Jan Gregory, Christopher J Bates, A Prentice, J Perks, G Swan, and M Farron. The national diet & nutrition survey: adults aged 19 to 64 years. 2003.

[159] Daniel Dorling and Bethan Thomas. *People and places: A 2001 census atlas of the UK*. Policy Press, 2004.

[160] Rachel Craig and Nicola Shelton. *The Health Survey for England 2007*. TSO, 2008.

[161] Deborah Jarvis, S Chinn, C Luczynska, and P Burney. Association of respiratory symptoms and lung function in young adults with use of domestic gas appliances. *The Lancet*, 347(8999):426–431, 1996.

[162] David Coggon, Peter Croft, Samantha Kellingray, DS Barrett, Magnus McLaren, and Cyrus Cooper. Occupational physical activities and osteoarthritis of the knee. *Arthritis and rheumatism*, 43(7):1443–1449, 2000.

[163] Michael L Booth, BARBARA E Ainsworth, MICHAEL Pratt, U Ekelund, AGNETA Yngve, JAMES F Sallis, and PEKKA Oja. International physical activity questionnaire: 12-country reliability and validity. *Med sci sports Exerc*, 195(9131/03):3508–1381, 2003.

[164] Frank B Hu, Tricia Y Li, Graham A Colditz, Walter C Willett, and JoAnn E Manson. Television watching and other sedentary behaviors in relation to risk of obesity and type 2 diabetes mellitus in women. *Jama*, 289(14):1785–1791, 2003.

[165] RW Jakes, NE Day, KT Khaw, R Luben, S Oakes, A Welch, S Bingham, and NJ Wareham. Television viewing and low participation in vigorous recreation are independently associated with obesity and markers of cardiovascular disease risk: Epic-norfolk population-based study. *European journal of clinical nutrition*, 57(9):1089–1096, 2003.

[166] Aiden Doherty, Dan Jackson, Nils Hammerla, Thomas Plötz, Patrick Olivier, Malcolm H Granat, Tom White, Vincent T van Hees, Michael I Trenell, Christoper G Owen, et al. Large scale population assessment of physical activity using wrist worn accelerometers: The uk biobank study. *PLoS One*, 12(2):e0169649, 2017.

[167] C Brayne, N Day, and C Gill. Methodological issues in screening for dementia. *Neuroepidemiology*, 11(Suppl. 1):88–93, 1992.

[168] Celeste De Jager, Andrew D Blackwell, Marc M Budge, and Barbara J Sahakian. Predicting cognitive decline in healthy older adults. *The American Journal of Geriatric Psychiatry*, 13(8):735–740, 2005.

[169] Uk biobank/data management sharing plan. `http://www.ukbiobank.ac.uk/wp-content/uploads/2013/10/ukbiobank-data-management.pdf`. Accessed: 2017-03-12.

[170] A description of the prostate cancer association group to investigate cancer associated alterations in the genome (practical). `http://practical.ccge.medschl.cam.ac.uk`. Accessed: 2017-03-12.

[171] Supplementary information for: Identification of 23 new prostate cancer susceptibility loci using the icogs custom genotyping array. `http://www.nature.com/ng/journal/v45/n4/extref/ng.2560-S1.pdf`. Accessed: 2017-03-12.

[172] A description of the available phenotypic information within the prostate cancer association group to investigate cancer associated alterations in the genome (practical). `http://practical.ccge.medschl.cam.ac.uk/members/data/data-dictionary/`. Accessed: 2017-03-12.

[173] World Health Organization, World Health Organization, et al. Icd-10. *International statistical classification of diseases and related health problems (10th edition). Geneva, SR: World Health Organization (WHO)*, 1992.

[174] Diem Nguyen Bentzon, Mette Nyegaard, Anders Børglum, Torben Ørntoft, Michael Borre, and Karina Dalsgaard Sørensen. Replication of prostate cancer risk variants in a danish case-control association study. *Open Journal of Urology*, 2(02):45, 2012.

[175] Rajeev Mahajan, Aaron Blair, Joseph Coble, Charles F Lynch, Jane A Hoppin, Dale P Sandler, and Michael CR Alavanja. Carbaryl exposure and incident cancer in the agricultural health study. *International journal of cancer*, 121(8):1799–1805, 2007.

[176] ATBC Cancer Prevention Study Group et al. The alpha-tocopherol, beta-carotene lung cancer prevention study: design, methods, participant characteristics, and compliance. *Annals of epidemiology*, 4(1):1–10, 1994.

[177] A Discacciati, N Orsini, S-O Andersson, Ove Andrén, J-E Johansson, CS Mantzoros, and A Wolk. Coffee consumption and risk of localized, advanced and fatal prostate cancer: a population-based prospective study. *Annals of oncology*, page mdt105, 2013.

[178] Maren Weischer, Børge G Nordestgaard, Richard M Cawthon, Jacob J Freiberg, Anne Tybjærg-Hansen, and Stig E Bojesen. Short telomere length, cancer survival, and cancer risk in 47102 individuals. *Journal of the National Cancer Institute*, page djt016, 2013.

[179] Børge G Nordestgaard, Tom M Palmer, Marianne Benn, Jeppe Zacho, Anne Tybjærg-Hansen, George Davey Smith, and Nicholas J Timpson. The effect of elevated body mass index on ischemic heart disease risk: causal estimates from a mendelian randomisation approach. *PLoS Med*, 9(5):e1001212, 2012.

[180] Heiko Müller, Elke Raum, Dietrich Rothenbacher, Christa Stegmaier, and Hermann Brenner. Association of diabetes and body mass index with levels of prostate-specific antigen: implications for correction of prostate-specific antigen cutoff values? *Cancer Epidemiology and Prevention Biomarkers*, 18(5):1350–1356, 2009.

[181] Ilir Agalliu, Claudia A Salinas, Philip D Hansten, Elaine A Ostrander, and Janet L Stanford. Statin use and risk of prostate cancer: results from a population-based epidemiologic study. *American journal of epidemiology*, 168(3):250–260, 2008.

[182] Sarah C Markt, Erin E Flynn-Evans, Unnur Valdimarsdottir, Lara G Sigurdardottir, Rulla M Tamimi, Julie L Batista, Sebastien Haneuse, Steven W Lockley, Meir Stampfer, Kathryn M Wilson, et al. Sleep duration and disruption and prostate cancer risk: a 23-year prospective study. *Cancer Epidemiology and Prevention Biomarkers*, pages cebp–1274, 2015.

[183] Anita V Mitra, Elizabeth K Bancroft, and Rosalind A Eeles. A review of targeted screening for prostate cancer: introducing the impact study. *BJU international*, 99(6):1350–1355, 2007.

[184] Sofia Maia, Marta Cardoso, Pedro Pinto, Manuela Pinheiro, Catarina Santos, Ana Peixoto, Maria José Bento, Jorge Oliveira, Rui Henrique, Carmen Jerónimo, et al. Identification of two novel hoxb13 germline mutations in portuguese prostate cancer patients. *PloS one*, 10(7):e0132728, 2015.

[185] KU Leuven, department of cellular and molecular medicine. `https://gbiomed.kuleuven.be/english/research/50000618/50753346`. Accessed: 2017-02-22.

[186] Gemma Castaño-Vinyals, Nuria Aragonés, Beatriz Pérez-Gómez, Vicente Martín, Javier Llorca, Victor Moreno, Jone M Altzibar, Eva Ardanaz, Sílvia De Sanjosé, José Juan Jiménez-Moleón, et al. Population-based multicase-control study in common tumors in spain (mcc-spain): rationale and study design. *Gaceta Sanitaria*, 29(4):308–315, 2015.

[187] GG Giles and DR English. The melbourne collaborative cohort study. *IARC scientific publications*, pages 69–70, 2003.

[188] Hui-Yi Lin, Hyun Y Park, Selina Radlein, Nupam P Mahajan, Thomas A Sellers, Babu Zachariah, Julio Pow-Sang, Domenico Coppola, Vadivel Ganapathy, and Jong Y Park. Protein expressions and genetic variations of slc5a8 in prostate cancer risk and aggressiveness. *Urology*, 78(4):971–e1, 2011.

[189] CH Hennekens. Final report on the aspirin component of the ongoing physicians health study. *New England Journal of Medicine*, 321(3):129–135, 1989.

[190] Robert Szulkin, Erik Holmberg, Pär Stattin, Jianfeng Xu, Sigun Zheng, Juni Palmgren, Henrik Grönberg, and Fredrik Wiklund. Prostate cancer risk variants are not associated with disease progression. *The Prostate*, 72(1):30–39, 2012.

[191] Laura Fachal, Antonio Gómez-Caamaño, Gillian C Barnett, Paula Peleteiro, Ana M Carballo, Patricia Calvo-Crespo, Sarah L Kerns, Manuel Sánchez-García, Ramón Lobato-Busto, Leila Dorling, et al. A three-stage genome-wide association study identifies a susceptibility locus for late radiotherapy toxicity at 2q24. 1. *Nature genetics*, 46(8):891–894, 2014.

[192] Maria Vias, Charlie E Massie, Philip East, Helen Scott, Anne Warren, Zongxiang Zhou, Alexander Yu Nikitin, David E Neal, and Ian G Mills. Pro-neural transcription factors as cancer markers. *BMC medical genomics*, 1(1):17, 2008.

[193] Felicity Lose, Srilakshmi Srinivasan, Tracy OMara, Louise Marquart, Suzanne Chambers, Robert A Gardiner, Joanne F Aitken, Amanda B Spurdle, Jyotsna Batra, Judith A Clements, et al. Genetic association of the klk4 locus with risk of prostate cancer. *PloS one*, 7(9):e44520, 2012.

[194] Anglian Breast Cancer Study Group et al. Prevalence and penetrance of brca1 and brca2 mutations in a population-based series of breast cancer cases. *British Journal of Cancer*, 83(10):1301, 2000.

[195] Esther M John, Gary G Schwartz, Jocelyn Koo, David Van Den Berg, and Sue A Ingles. Sun exposure, vitamin d receptor gene polymorphisms, and risk of advanced prostate cancer. *Cancer research*, 65(12):5470–5479, 2005.

[196] Tobias Nordström, Markus Aly, Martin Eklund, Lars Egevad, and Henrik Grönberg. A genetic score can identify men at high risk for prostate cancer among men with prostate-specific antigen of 1–3 ng/ml. *European urology*, 65(6):1184–1190, 2014.

[197] Bimal Bhindi, Jennifer Locke, Shabbir MH Alibhai, Girish S Kulkarni, David S Margel, Robert J Hamilton, Antonio Finelli, John Trachtenberg, Alexandre R Zlotta, Ants Toi, et al. Dissecting the association between metabolic syndrome and prostate cancer risk: analysis of a large clinical cohort. *European urology*, 67(1):64–70, 2015.

[198] Uk genetic prostate cancer study. http://www.icr.ac.uk/our-research/research-divisions/division-of-genetics-and-epidemiology/oncogenetics/research-projects/prostate-cancer-research/ukgpcs. Accessed: 2017-02-22.

[199] Manuel Luedeke, Carmen M Linnert, Matthias D Hofer, Harald M Surowy, Antje E Rinckleb, Josef Hoegel, Rainer Kuefer, Mark A Rubin, Walther Vogel, and Christiane Maier. Predisposition for tmprss2-erg fusion in prostate cancer by variants in dna repair genes. *Cancer Epidemiology and Prevention Biomarkers*, 18(11):3030–3035, 2009.

[200] Sabine Rohrmann, Dina N Paltoo, Elizabeth A Platz, Sandra C Hoffman, George W Comstock, and Kathy J Helzlsouer. Association of vasectomy and prostate cancer among men in a maryland cohort. *Cancer Causes & Control*, 16(10):1189–1194, 2005.

[201] Edward Giovannucci, Tor D Tosteson, Frank E Speizer, Alberto Ascherio, Martin P Vessey, and Graham A Colditz. A retrospective cohort study of vasectomy and prostate cancer in us men. *Jama*, 269(7):878–882, 1993.

[202] Stephen Sidney, Charles P Quesenberry Jr, Marianne C Sadler, Harry A Guess, Eva G Lydick, and Eugene V Cattolica. Vasectomy and the risk of prostate cancer in a cohort of multiphasic health-checkup examinees: second report. *Cancer Causes & Control*, 2(2):113–116, 1991.

[203] Henriet Nienhuis, Michael Goldacre, Valerie Seagroatt, Leicester Gill, and Martin Vessey. Incidence of disease after vasectomy: a record linkage retrospective cohort study. *Bmj*, 304(6829):743–746, 1992.

[204] Henrik Moller, Lisbeth B Knudsen, and Elsebeth Lynge. Risk of testicular cancer after vasectomy: cohort study of over 73 000 men. *Bmj*, 309(6950):295–299, 1994.

[205] Lynley A Cook, Asha Pun, Maria F Gallo, Laureen M Lopez, and HAAM Van Vliet. Scalpel versus no-scalpel incision for vasectomy. *Cochrane Database Syst Rev*, 3, 2014.

[206] Pamela J Schwingl and Harry A Guess. Safety and effectiveness of vasectomy. *Fertility and Sterility*, 73(5):923–936, 2000.

[207] P Christensen, OA Al-Aqidi, FS Jensen, and T Dørflinger. [vasectomy. a prospective, randomized trial of vasectomy with bilateral incision versus the li vasectomy]. *Ugeskrift for laeger*, 164(18):2390–2394, 2002.

[208] David Sokal, Susan McMullen, Deb Gates, Rosalie Dominik, and The Male Sterilization Investigator Team. A comparative study of the no scalpel and standard incision approaches to vasectomy in 5 countries. *The Journal of urology*, 162(5):1621–1625, 1999.

[209] Brian Cox, Mary J Sneyd, Charlotte Paul, Brett Delahunt, and David CG Skegg. Vasectomy and risk of prostate cancer. *Jama*, 287(23):3110–3115, 2002.

[210] Janet L Stanford, Kristine G Wicklund, Barbara McKnight, Janet R Daling, and Michael K Brawer. Vasectomy and risk of prostate cancer. *Cancer Epidemiology Biomarkers & Prevention*, 8(10):881–886, 1999.

[211] Stuart S Howards and Herbert B Peterson. Vasectomy and prostate cancer: chance, bias, or a causal relationship? *JAMA*, 269(7):913–914, 1993.

[212] S Pereira, M Martinez, FE Martinez, and W Mello Júnior. Repercussions of castration and vasectomy on the ductal system of the rat ventral prostate. *Cell biology international*, 30(2):169–174, 2006.

[213] Zeng-Nan Mo, Xun Huang, Shi-Chun Zhang, and Jin-Rui Yang. Early and late long-term effects of vasectomy on serum testosterone, dihydrotestosterone, luteinizing hormone and follicle-stimulating hormone levels. *The Journal of urology*, 154(6):2065–2069, 1995.

[214] RK Naz, J Deutsch, TM Phillips, AC Menge, and H Fisch. Sperm antibodies in vasectomized men and their effects on fertilization. *Biology of reproduction*, 41(1):163–173, 1989.

[215] Lars Linnet. Clinical immunology of vasectomy and vasovasostomy. *Urology*, 22(2):101–114, 1983.

[216] Maarten C Bosland. Sex steroids and prostate carcinogenesis. *Annals of the New York Academy of Sciences*, 1089(1):168–176, 2006.

[217] DC Skegg, JD Mathews, J Guillebaud, MP Vessey, S Biswas, KM Ferguson, Y Kitchin, MD Mansfield, and IF Sommerville. Hormonal assessment before and after vasectomy. *Br Med J*, 1(6010):621–622, 1976.

[218] Peng Xian-sheng Li Fu-de, Miao Zhong-rei Wong Yong Hu, Xiao-zhong Zeng Fu-xiong Wu, and Xiou-qing Lan Jun. Long-term effects of vasectomy on the pituitary-gonadal axis. *Reproduction & Contraception*, 1:004, 1986.

[219] Ihor Batruch, Irene Lecker, Daniel Kagedan, Christopher R Smith, Brendan J Mullen, Ethan Grober, Kirk C Lo, Eleftherios P Diamandis, and Keith A Jarvi. Proteomic analysis of seminal plasma from normal volunteers and post-vasectomy patients identifies over 2000 proteins and candidate biomarkers of the urogenital system. *Journal of proteome research*, 10(3):941–953, 2011.

[220] Jennifer L Kelsey. *Methods in observational epidemiology*, volume 26. Monographs in Epidemiology and, 1996.

[221] Knud Juel and Karin Helweg-Larsen. The danish registers of causes of death. *Danish medical bulletin*, 46(4):354–357, 1999.

[222] Hans H Storm, Elli Vera Michelsen, Inge Haunstrup Clemmensen, and Jesper Pihl. The danish cancer registry–history, content, quality and use. *Danish medical bulletin*, 44(5):535–539, 1997.

[223] Beatriz Pérez-Gómez, Nuria Aragonés, Marina Pollán, Berta Suárez, Virginia Lope, Alicia Llácer, and Gonzalo López-Abente. Accuracy of cancer death certificates in spain: a summary of available information. *Gaceta Sanitaria*, 20:42–51, 2006.

[224] Emma L Turner, Chris Metcalfe, Jenny L Donovan, Sian Noble, Jonathan AC Sterne, J Athene Lane, Eleanor I Walsh, Elizabeth M Hill, Liz Down, Yoav Ben-Shlomo, et al. Contemporary accuracy of death certificates for coding prostate cancer as a cause of death: Is reliance on death certification good enough? a comparison with blinded

review by an independent cause of death evaluation committee. *British journal of cancer*, 2016.

[225] Ruth C Travis, Francesca L Crowe, Naomi E Allen, Paul N Appleby, Andrew W Roddam, Anne Tjønneland, Anja Olsen, Jakob Linseisen, Rudolf Kaaks, Heiner Boeing, et al. Serum vitamin d and risk of prostate cancer in a case-control analysis nested within the european prospective investigation into cancer and nutrition (epic). *American journal of epidemiology*, 169(10):1223–1232, 2009.

[226] Katja Mitrunen, Kim Pettersson, Timo Piironen, Thomas Björk, Hans Lilja, and Timo Lövgren. Dual-label one-step immunoassay for simultaneous measurement of free and total prostate-specific antigen concentrations and ratios in serum. *Clinical Chemistry*, 41(8):1115–1120, 1995.

[227] Pauliina Nurmikko, Kim Pettersson, Timo Piironen, Jonas Hugosson, and Hans Lilja. Discrimination of prostate cancer from benign disease by plasma measurement of intact, free prostate-specific antigen lacking an internal cleavage site at lys145-lys146. *Clinical chemistry*, 47(8):1415–1423, 2001.

[228] T Piironen, Janita Lövgren, M Karp, R Eerola, Å Lundwall, B Dowell, T Lövgren, Hans Lilja, and K Pettersson. Immunofluorometric assay for sensitive and specific measurement of human prostatic glandular kallikrein (hk2) in serum. *Clinical chemistry*, 42(7):1034–1041, 1996.

[229] Per-Anders Abrahamsson, C Andersson, Thomas Björk, Per Fernlund, Hans Lilja, A Murne, and Håkan Weiber. Radioimmunoassay of beta-microseminoprotein, a prostatic-secreted protein present in sera of both men and women. *Clinical chemistry*, 35(7):1497–1503, 1989.

[230] Camilla Valtonen-André, Charlotta Sävblom, Per Fernlund, Hans Lilja, Aleksander Giwercman, and Åke Lundwall. Beta-microseminoprotein in serum correlates with the levels in seminal plasma of young, healthy males. *Journal of andrology*, 29(3):330–337, 2008.

[231] Naomi E Allen, Timothy J Key, Paul N Appleby, Ruth C Travis, Andrew W Roddam, Sabina Rinaldi, Lars Egevad, Sabine Rohrmann,

Jakob Linseisen, Tobias Pischon, et al. Serum insulin-like growth factor (igf)-i and igf-binding protein-3 concentrations and prostate cancer risk: results from the european prospective investigation into cancer and nutrition. *Cancer Epidemiology Biomarkers & Prevention*, 16(6):1121–1127, 2007.

[232] Alison J Price, Naomi E Allen, Paul N Appleby, Francesca L Crowe, Ruth C Travis, Sarah J Tipper, Kim Overvad, Henning Grønbæk, Anne Tjønneland, Nina Føns Johnsen, et al. Insulin-like growth factor-i concentration and risk of prostate cancer: results from the european prospective investigation into cancer and nutrition. *Cancer Epidemiology Biomarkers & Prevention*, 21(9):1531–1541, 2012.

[233] Anne E Cust, Naomi E Allen, Sabina Rinaldi, Laure Dossus, Christine Friedenreich, Anja Olsen, Anne Tjønneland, Kim Overvad, Françoise Clavel-Chapelon, Marie-Christine Boutron-Ruault, et al. Serum levels of c-peptide, igfbp-1 and igfbp-2 and endometrial cancer risk; results from the european prospective investigation into cancer and nutrition. *International journal of cancer*, 120(12):2656–2664, 2007.

[234] Andrew W Roddam, Naomi E Allen, Paul Appleby, Timothy J Key, Luigi Ferrucci, H Ballentine Carter, E Jeffrey Metter, Chu Chen, Noel S Weiss, Annette Fitzpatrick, et al. Insulin-like growth factors, their binding proteins, and prostate cancer risk: analysis of individual patient data from 12 prospective studies. *Annals of internal medicine*, 149(7):461, 2008.

[235] Naomi E Allen, Timothy J Key, Laure Dossus, Sabina Rinaldi, Anne Cust, Annekatrin Lukanova, Petra H Peeters, N Charlotte Onland-Moret, Petra H Lahmann, Franco Berrino, et al. Endogenous sex hormones and endometrial cancer risk in women in the european prospective investigation into cancer and nutrition (epic). *Endocrine-related cancer*, 15(2):485–497, 2008.

[236] Mary Lunn and Don McNeil. Applying cox regression to competing risks. *Biometrics*, pages 524–532, 1995.

[237] Charles J Flickinger, Leigh Ann Bush, Mollie V Williams, Soren Naaby-Hansen, Stuart S Howards, and John C Herr. Post-obstruction rat

sperm autoantigens identified by two-dimensional gel electrophoresis and western blotting. *Journal of reproductive immunology*, 43(1):35–53, 1999.

[238] N Anahí Franchi, Conrado Avendaño, Rosa I Molina, Andrea D Tissera, Cristina A Maldonado, Sergio Oehninger, and Carlos E Coronel. β-microseminoprotein in human spermatozoa and its potential role in male fertility. *Reproduction*, 136(2):157–166, 2008.

[239] Kok Onn Lee, YOUNGMAN Oh, Linda C Giudice, Pinchas Cohen, DONNA M Peehl, and RG Rosenfeld. Identification of insulin-like growth factor-binding protein-3 (igfbp-3) fragments and igfbp-5 proteolytic activity in human seminal plasma: a comparison of normal and vasectomized patients. *The Journal of Clinical Endocrinology & Metabolism*, 79(5):1367–1372, 1994.

[240] Pinhas Cohen, Donna M Peehl, Thomas A Stamey, Kristin F Wilson, David R Clemmons, and Ron G Rosenfeld. Elevated levels of insulin-like growth factor-binding protein-2 in the serum of prostate cancer patients. *The Journal of Clinical Endocrinology & Metabolism*, 76(4):1031–1035, 1993.

[241] Sue M Firth and Robert C Baxter. Cellular actions of the insulin-like growth factor binding proteins. *Endocrine reviews*, 23(6):824–854, 2002.

[242] WJ Azar, S Zivkovic, GA Werther, and VC Russo. Igfbp-2 nuclear translocation is mediated by a functional nls sequence and is essential for its pro-tumorigenic actions in cancer cells. *Oncogene*, 33(5):578–588, 2014.

[243] Herbert Yu, Michael R Nicar, Runhua Shi, Hans J Berkel, Robert Nam, John Trachtenberg, and Eleftherios P Diamandis. Levels of insulin-like growth factor i (igf-i) and igf binding proteins 2 and 3 in serial postoperative serum samples and risk of prostate cancer recurrence. *Urology*, 57(3):471–475, 2001.

[244] Shahrokh F Shariat, Dolores J Lamb, Michael W Kattan, Cuong Nguyen, JaHong Kim, Josie Beck, Thomas M Wheeler, and Kevin M

Slawin. Association of preoperative plasma levels of insulin-like growth factor i and insulin-like growth factor binding proteins-2 and-3 with prostate cancer invasion, progression, and metastasis. *Journal of Clinical Oncology*, 20(3):833–841, 2002.

[245] Matthew Fankhauser, Yuen Tan, Geoff Macintyre, Izhak Haviv, Matthew KH Hong, Anne Nguyen, John S Pedersen, Anthony J Costello, Christopher M Hovens, and Niall M Corcoran. Canonical androstenedione reduction is the predominant source of signaling androgens in hormone-refractory prostate cancer. *Clinical Cancer Research*, 20(21):5547–5557, 2014.

[246] Ruth C Travis, Timothy J Key, Naomi E Allen, Paul N Appleby, Andrew W Roddam, Sabina Rinaldi, Lars Egevad, Peter H Gann, Sabine Rohrmann, Jakob Linseisen, et al. Serum androgens and prostate cancer among 643 cases and 643 controls in the european prospective investigation into cancer and nutrition. *International journal of cancer*, 121(6):1331–1338, 2007.

[247] Karen E Knudsen. Hormone whodunit: clues for solving the case of intratumor androgen production. *Clinical Cancer Research*, 20(21):5343–5345, 2014.

[248] K Purvis, SK Saksena, Z Cekan, E Diczfalusy, and J Giner. Endocrine effects of vasectomy. *Clinical endocrinology*, 5(3):263–272, 1976.

[249] Fritz H Schröder, Jonas Hugosson, Monique J Roobol, Teuvo LJ Tammela, Marco Zappa, Vera Nelen, Maciej Kwiatkowski, Marcos Lujan, Liisa Määttänen, Hans Lilja, et al. Screening and prostate cancer mortality: results of the european randomised study of screening for prostate cancer (erspc) at 13 years of follow-up. *The Lancet*, 384(9959):2027–2035, 2014.

[250] Xing Xu, Camilla Valtonen-André, Charlotta Sävblom, Christer Halldén, Hans Lilja, and Robert J Klein. Polymorphisms at the microseminoprotein-β locus associated with physiologic variation in β-microseminoprotein and prostate-specific antigen levels. *Cancer Epidemiology Biomarkers & Prevention*, 19(8):2035–2042, 2010.

[251] Bao-Li Chang, Elaine Spangler, Stephen Gallagher, Christopher A Haiman, Brian Henderson, William Isaacs, Marnita L Benford, LaCreis R Kidd, Kathleen Cooney, Sara Strom, et al. Validation of genome-wide prostate cancer associations in men of african descent. *Cancer Epidemiology Biomarkers & Prevention*, 20(1):23–32, 2011.

[252] Bao-Li Chang, Scott D Cramer, Fredrik Wiklund, Sarah D Isaacs, Victoria L Stevens, Jielin Sun, Shelly Smith, Kristen Pruett, Lina M Romero, Kathleen E Wiley, et al. Fine mapping association study and functional analysis implicate a snp in msmb at 10q11 as a causal variant for prostate cancer risk. *Human molecular genetics*, 18(7):1368–1375, 2009.

[253] Hong Lou, Meredith Yeager, Hongchuan Li, Jesus Gonzalez Bosquet, Richard B Hayes, Nick Orr, Kai Yu, Amy Hutchinson, Kevin B Jacobs, Peter Kraft, et al. Fine mapping and functional analysis of a common variant in msmb on chromosome 10q11. 2 associated with prostate cancer susceptibility. *Proceedings of the National Academy of Sciences*, 106(19):7933–7938, 2009.

[254] H Weiber, C Andersson, A Murne, G Rannevik, C Lindström, H Lilja, and P Fernlund. Beta microseminoprotein is not a prostate-specific protein. its identification in mucous glands and secretions. *The American journal of pathology*, 137(3):593, 1990.

[255] M Baijal-Gupta, MW Clarke, MA Finkelman, CM McLachlin, and VK Han. Prostatic secretory protein (psp94) expression in human female reproductive tissues, breast and in endometrial cancer cell lines. *Journal of endocrinology*, 165(2):425–433, 2000.

[256] Seema Garde, Jennifer E Fraser, Najib Nematpoor, Rebecca Pollex, Catherine Morin, André Forté, Shafaat Rabbani, Chandra Panchal, and Madhulika B Gupta. Cloning, expression, purification and functional characterization of recombinant human psp94. *Protein expression and purification*, 54(2):193–203, 2007.

[257] Seema Garde, Anil Sheth, Arthur T Porter, and Kenneth J Pienta. Effect of prostatic inhibin peptide (pip) on prostate cancer cell growth in vitro and in vivo. *The Prostate*, 22(3):225–233, 1993.

[258] Seema V Garde, Vathsala S Basrur, Li Li, Malcolm A Finkelman, Awtar Krishan, Larry Wellham, Edgar Ben-Josef, Maher Haddad, John D Taylor, Arthur T Porter, et al. Prostate secretory protein (psp94) suppresses the growth of androgen-independent prostate cancer cell line (pc3) and xenografts by inducing apoptosis. *The Prostate*, 38(2):118–125, 1999.

[259] Balakrishna L Lokeshwar, Kalpana S Hurkadli, Anil R Sheth, and Norman L Block. Human prostatic inhibin suppresses tumor growth and inhibits clonogenic cell survival of a model prostatic adenocarcinoma, the dunning r3327g rat tumor. *Cancer research*, 53(20):4855–4859, 1993.

[260] Nicholas Shukeir, Ani Arakelian, Gaoping Chen, Seema Garde, Marcia Ruiz, Chandra Panchal, and Shafaat A Rabbani. A synthetic 15-mer peptide (pck3145) derived from prostate secretory protein can reduce tumor growth, experimental skeletal metastases, and malignancy-associated hypercalcemia. *Cancer research*, 64(15):5370–5377, 2004.

[261] Nicholas Shukeir, Ani Arakelian, Salam Kadhim, Seema Garde, and Shafaat A Rabbani. Prostate secretory protein psp-94 decreases tumor growth and hypercalcemia of malignancy in a syngenic in vivo model of prostate cancer. *Cancer research*, 63(9):2072–2078, 2003.

[262] Anneli ML Edström Hägerwall, Victoria Rydengård, Per Fernlund, Matthias Mörgelin, Maria Baumgarten, Alexander M Cole, Martin Malmsten, Birthe B Kragelund, and Ole E Sørensen. β-microseminoprotein endows post coital seminal plasma with potent candidacidal activity by a calcium-and ph-dependent mechanism. *PLoS Pathog*, 8(4):e1002625, 2012.

[263] Kazuko Akiyama, Yasuyuki Yoshioka, Karl Schmid, Gwynneth D Offner, Robert F Troxler, Ryouichi Tsuda, and Mitsuwo Hara. The amino acid sequence of human β-microseminoprotein. *Biochimica et Biophysica Acta (BBA)-Protein Structure and Molecular Enzymology*, 829(2):288–294, 1985.

[264] AE Tilley, MR Staudt, J Fuller, BP De, RG Crystal, and N Qadir. Smoking-induced up-regulation of microseminoprotein beta gene expression in the human airway. *Am J Respir Crit Care Med*, 185:A6052, 2012.

[265] Robert K Nam, Eleftherios P Diamandis, Ants Toi, John Trachtenberg, Angeliki Magklara, Andreas Scorilas, Panayotis A Papnastasiou, Michael AS Jewett, and Steven A Narod. Serum human glandular kallikrein-2 protease levels predict the presence of prostate cancer among men with elevated prostate-specific antigen. *Journal of Clinical Oncology*, 18(5):1036–1036, 2000.

[266] Hayley C Whitaker, Zsofia Kote-Jarai, Helen Ross-Adams, Anne Y Warren, Johanna Burge, Anne George, Elizabeth Bancroft, Sameer Jhavar, Daniel Leongamornlert, Malgorzata Tymrakiewicz, et al. The rs10993994 risk allele for prostate cancer results in clinically relevant changes in microseminoprotein-beta expression in tissue and urine. *PloS one*, 5(10):e13363, 2010.

[267] Andrew R Girvan, Peter Chang, Isaac van Huizen, Madeleine Moussa, Jim W Xuan, Larry Stitt, Joseph L Chin, Yasuto Yamasaki, and Jonathan I Izawa. Increased intratumoral expression of prostate secretory protein of 94 amino acids predicts for worse disease recurrence and progression after radical prostatectomy in patients with prostate cancer. *Urology*, 65(4):719–723, 2005.

[268] Yushi Imasato, Jim W Xuan, Hideki Sakai, Jonathon I Izawa, Yutaka Saito, Joseph L Chin, and Madeleine Moussa. Psp94 expression after androgen deprivation therapy: a comparative study with prostate specific antigen in benign prostate and prostate cancer. *The Journal of urology*, 164(5):1819–1824, 2000.

[269] Hideki Sakai, Toshifumi Tsurusaki, Shigeru Kanda, Takehiko Koji, Jim W Xuan, and Yutaka Saito. Prognostic significance of β-microseminoprotein mrna expression in prostate cancer. *The Prostate*, 38(4):278–284, 1999.

[270] Isaac Van Huizen, Guojun Wu, Madeleine Moussa, Joseph L Chin, Aaron Fenster, James C Lacefield, Hideki Sakai, Norman M Greenberg, and Jim W Xuan. Establishment of a serum tumor marker for preclinical trials of mouse prostate cancer models. *Clinical Cancer Research*, 11(21):7911–7919, 2005.

[271] Liisa Sjöblom, Outi Saramäki, Matti Annala, Katri Leinonen, Janika Nättinen, Teemu Tolonen, Tiina Wahlfors, Matti Nykter, G Steven Bova, Johanna Schleutker, et al. Microseminoprotein-beta expression in different stages of prostate cancer. *PloS one*, 11(3):e0150241, 2016.

[272] Anders S Bjartell, Hikmat Al-Ahmadie, Angel M Serio, James A Eastham, Scott E Eggener, Samson W Fine, Lene Udby, William L Gerald, Andrew J Vickers, Hans Lilja, et al. Association of cysteine-rich secretory protein 3 and β-microseminoprotein with outcome after radical prostatectomy. *Clinical cancer research*, 13(14):4130–4138, 2007.

[273] Hiroyuki Hyakutake, Hideki Sakai, Yasuo Yogi, Ryouichi Tsuda, Yuzo Minami, Yoshiaki Yushita, Hiroshi Kanetake, Ichiro Nakazono, and Yutaka Saito. Beta-microseminoprotein immunoreactivity as a new prognostic indicator of prostatic carcinoma. *The prostate*, 22(4):347–355, 1993.

[274] Toshifumi Tsurusaki, Takehiko Koji, Hideki Sakai, Hiroshi Kanetake, Paul K Nakane, and Yutaka Saito. Cellular expression of beta-microseminoprotein (β-msp) mrna and its protein in untreated prostate cancer. *The Prostate*, 35(2):109–116, 1998.

[275] Magnus Ulvsbäck, Clas Lindström, Håkan Weiber, Per-Anders Abrahamsson, Hans Lilja, and Åke Lundwall. Molecular cloning of a small prostate protein, known as β-microsemenoprotein, psp 94 or β-inhibin, and demonstration of transcripts in non-genital tissues. *Biochemical and biophysical research communications*, 164(3):1310–1315, 1989.

[276] Mark M Pomerantz, Yashaswi Shrestha, Richard J Flavin, Meredith M Regan, Kathryn L Penney, Lorelei A Mucci, Meir J Stampfer, David J Hunter, Stephen J Chanock, Eric J Schafer, et al. Analysis of the 10q11 cancer risk locus implicates msmb and ncoa4 in human prostate tumorigenesis. *PLoS Genet*, 6(11):e1001204, 2010.

[277] Liesel M FitzGerald, Xiaotun Zhang, Suzanne Kolb, Erika M Kwon, Ying Ching Liew, Antonio Hurtado-Coll, Beatrice S Knudsen, Elaine A Ostrander, and Janet L Stanford. Investigation of the relationship between prostate cancer and msmb and ncoa4 genetic variants and protein expression. *Human mutation*, 34(1):149–156, 2013.

[278] Siobhan Sutcliffe, Angelo M De Marzo, Karen S Sfanos, and Martin Laurence. Msmb variation and prostate cancer risk: clues towards a possible fungal etiology. *The Prostate*, 74(6):569–578, 2014.

[279] Daniel I Swerdlow, Karoline B Kuchenbaecker, Sonia Shah, Reecha Sofat, Michael V Holmes, Jon White, Jennifer S Mindell, Mika Kivimaki, Eric J Brunner, John C Whittaker, et al. Selecting instruments for mendelian randomization in the wake of genome-wide association studies. *International journal of epidemiology*, 45(5):1600–1616, 2016.

[280] Sonja A Swanson and Miguel A Hernán. The challenging interpretation of instrumental variable estimates under monotonicity. *International Journal of Epidemiology*, page dyx038, 2017.

[281] Felix R Day, Po-Ru Loh, Robert A Scott, Ken K Ong, and John RB Perry. A robust example of collider bias in a genetic association study. *American journal of human genetics*, 98(2):392, 2016.

[282] Andreas Scorilas, Mario Plebani, Saverio Mazza, Daniela Basso, Antoninus R Soosaipillai, Nikos Katsaros, Francesco Pagano, and Eleftherios P Diamandis. Serum human glandular kallikrein (hk2) and insulin-like growth factor 1 (igf-1) improve the discrimination between prostate cancer and benign prostatic hyperplasia in combination with total and% free psa. *The Prostate*, 54(3):220–229, 2003.

[283] Alexander Haese, Markus Graefen, Charlotte Becker, Joachin Noldus, Jared Katz, Ilias Cagiannos, Michael Kattan, Peter T Scardino, Edith Huland, Hartwig Huland, et al. The role of human glandular kallikrein 2 for prediction of pathologically organ confined prostate cancer. *The Prostate*, 54(3):181–186, 2003.

[284] Thomas Steuber, Andrew J Vickers, Alexander Haese, Charlotte Becker, Kim Pettersson, Felix K-H Chun, Michael W Kattan, James A

Eastham, Peter T Scardino, Hartwig Huland, et al. Risk assessment for biochemical recurrence prior to radical prostatectomy: Significant enhancement contributed by human glandular kallikrein 2 (hk2) and free prostate specific antigen (psa) in men with moderate psa-elevation in serum. *International journal of cancer*, 118(5):1234–1240, 2006.

[285] Carla A Borgono, Iacovos P Michael, and Eleftherios P Diamandis. Human tissue kallikreins: physiologic roles and applications in cancer. *Molecular cancer research*, 2(5):257–280, 2004.

[286] Janita Lövgren, Camilla Valtonen-André, Karel Marsal, Hans Liua, and Åke Lundwall. Measurement of prostate-specific antigen and human glandular kallikrein 2 in different body fluids. *Journal of andrology*, 20(3):348–355, 1999.

[287] Steven R Potter and Alan W Partin. Tumor markers: An update on human kallikrein 2. *Reviews in urology*, 2(4):221, 2000.

[288] Shahrokh F Shariat, Axel Semjonow, Hans Lilja, Caroline Savage, Andrew J Vickers, and Anders Bjartell. Tumor markers in prostate cancer i: blood-based markers. *Acta oncologica*, 50(sup1):61–75, 2011.

[289] George M Yousef and Eleftherios P Diamandis. The new human tissue kallikrein gene family: structure, function, and association to disease 1. *Endocrine reviews*, 22(2):184–204, 2001.

[290] Abhay Kumar, Stephen D Mikolajczyk, Amita S Goel, Lisa S Millar, and Mohammad S Saedi. Expression of pro form of prostate-specific antigen by mammalian cells and its conversion to mature, active form by human kallikrein 2. *Cancer research*, 57(15):3111–3114, 1997.

[291] Roland R Tremblay, David Deperthes, Bernard Tetu, and Jean Y Dube. Immunohistochemical study suggesting a complementary role of kallikreins hk2 and hk3 (prostate-specific antigen) in the functional analysis of human prostate tumors. *The American journal of pathology*, 150(2):455, 1997.

[292] Zhiqun Shang, Yuanjie Niu, Qiliang Cai, Jing Chen, Jing Tian, Shuyuan Yeh, Kuo-Pao Lai, and Chawnshang Chang. Human kallikrein

2 (klk2) promotes prostate cancer cell growth via function as a modulator to promote the ara70-enhanced androgen receptor transactivation. *Tumor Biology*, 35(3):1881–1890, 2014.

[293] Thomas K Takayama, Corey A Carter, and Ta Deng. Activation of prostate-specific antigen precursor (pro-psa) by prostin, a novel human prostatic serine protease identified by degenerate pcr. *Biochemistry*, 40(6):1679–1687, 2001.

[294] Janita Lövgren, Kristiina Rajakoski, Matti Karp, Åke Lundwall, and Hans Lilja. Activation of the zymogen form of prostate-specific antigen by human glandular kallikrein 2. *Biochemical and biophysical research communications*, 238(2):549–555, 1997.

[295] Mitchell G Lawrence, John Lai, and Judith A Clements. Kallikreins on steroids: structure, function, and hormonal regulation of prostate-specific antigen and the extended kallikrein locus. *Endocrine reviews*, 31(4):407–446, 2010.

[296] Micheal F Darson, Anna Pacelli, Patrick Roche, Harry G Rittenhouse, Robert L Wolfert, Charles YF Young, George G Klee, Donald J Tindall, and David G Bostwick. Human glandular kallikrein 2 (hk2) expression in prostatic intraepithelial neoplasia and adenocarcinoma: a novel prostate cancer marker. *Urology*, 49(6):857–862, 1997.

[297] Micheal F Darson, Anna Pacelli, Patrick Roche, Harry G Rittenhouse, Robert L Wolfert, Mohammad S Saeid, Charles YF Young, George G Klee, Donald J Tindall, and David G Bostwick. Human glandular kallikrein 2 expression in prostate adenocarcinoma and lymph node metastases. *Urology*, 53(5):939–944, 1999.

[298] Angeliki Magklara, Andreas Scorilas, William J Catalona, and Eleftherios P Diamandis. The combination of human glandular kallikrein and free prostate-specific antigen (psa) enhances discrimination between prostate cancer and benign prostatic hyperplasia in patients with moderately increased total psa. *Clinical chemistry*, 45(11):1960–1966, 1999.

[299] Alexander Haese, Markus Graefen, Thomas Steuber, Charlotte Becker, Joachim Noldus, Andreas Erbersdobler, Edith Huland, Hartwig Huland, and Hans Lilja. Total and gleason grade 4/5 cancer volumes are major contributors of human kallikrein 2, whereas free prostate specific antigen is largely contributed by benign gland volume in serum from patients with prostate cancer or benign prostatic biopsies. *The Journal of urology*, 170(6):2269–2273, 2003.

[300] Katharina Braun, Daniel D Sjoberg, Andrew J Vickers, Hans Lilja, and Anders S Bjartell. A four-kallikrein panel predicts high-grade cancer on biopsy: independent validation in a community cohort. *European urology*, 69(3):505–511, 2016.

[301] Dipen J Parekh, Sanoj Punnen, Daniel D Sjoberg, Scott W Asroff, James L Bailen, James S Cochran, Raoul Concepcion, Richard D David, Kenneth B Deck, Igor Dumbadze, et al. A multi-institutional prospective trial in the usa confirms that the 4kscore accurately identifies men with high-grade prostate cancer. *European urology*, 68(3):464–470, 2015.

[302] Henrik Grönberg, Jan Adolfsson, Markus Aly, Tobias Nordström, Peter Wiklund, Yvonne Brandberg, James Thompson, Fredrik Wiklund, Johan Lindberg, Mark Clements, et al. Prostate cancer screening in men aged 50–69 years (sthlm3): a prospective population-based diagnostic study. *The lancet oncology*, 16(16):1667–1676, 2015.

[303] Adrian E Raftery. Bayesian model selection in social research. *Sociological methodology*, pages 111–163, 1995.

[304] Wessel N van Wieringen. Lecture notes on ridge regression. *arXiv preprint arXiv:1509.09169*, 2015.

[305] Andrew Gelman, Aleks Jakulin, Maria Grazia Pittau, and Yu-Sung Su. A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, pages 1360–1383, 2008.

[306] Elizabeth R DeLong, David M DeLong, and Daniel L Clarke-Pearson. Comparing the areas under two or more correlated receiver operating

characteristic curves: a nonparametric approach. *Biometrics*, pages 837–845, 1988.

[307] Margaret Sullivan Pepe, Jing Fan, and Christopher W Seymour. Estimating the receiver operating characteristic curve in studies that match controls to cases on covariates. *Academic radiology*, 20(7):863–873, 2013.

[308] Margaret Sullivan Pepe, Jing Fan, Christopher W Seymour, Christopher Li, Ying Huang, and Ziding Feng. Biases introduced by choosing controls to match risk factors of cases in biomarker research. *Clinical chemistry*, 58(8):1242–1251, 2012.

[309] Stephen J Freedland, Elizabeth A Platz, Joseph C Presti, William J Aronson, Christopher L Amling, Christopher J Kane, and Martha K Terris. Obesity, serum prostate specific antigen and prostate size: implications for prostate cancer detection. *The Journal of urology*, 175(2):500–504, 2006.

[310] Eleanor L Watts, Paul N Appleby, Demetrius Albanes, Amanda Black, June M Chan, Chu Chen, Piera M Cirillo, Barbara A Cohn, Michael B Cook, Jenny L Donovan, et al. Circulating sex hormones in relation to anthropometric, sociodemographic and behavioural factors in an international dataset of 12,300 men. *PloS one*, 12(12):e0187741, 2017.

[311] Lionel L Bañez, Robert J Hamilton, Alan W Partin, Robin T Vollmer, Leon Sun, Carmen Rodriguez, Yiting Wang, Martha K Terris, William J Aronson, Joseph C Presti, et al. Obesity-related plasma hemodilution and psa concentration among men with prostate cancer. *Jama*, 298(19):2275–2280, 2007.

[312] Jiyoung Ahn, Demetrius Albanes, Ulrike Peters, Arthur Schatzkin, Unhee Lim, Michal Freedman, Nilanjan Chatterjee, Gerald L Andriole, Michael F Leitzmann, and Richard B Hayes. Dairy products, calcium intake, and risk of prostate cancer in the prostate, lung, colorectal, and ovarian cancer screening trial. *Cancer Epidemiology Biomarkers & Prevention*, 16(12):2623–2630, 2007.

[313] Xiang Gao, Michael P LaValley, and Katherine L Tucker. Prospective studies of dairy product and calcium intakes and prostate cancer risk: a meta-analysis. *Journal of the National Cancer Institute*, 97(23):1768–1777, 2005.

[314] Albano Beja-Pereira, Gordon Luikart, Phillip R England, Daniel G Bradley, Oliver C Jann, Giorgio Bertorelle, Andrew T Chamberlain, Telmo P Nunes, Stoitcho Metodiev, Nuno Ferrand, et al. Gene-culture coevolution between cattle milk protein genes and human lactase genes. *Nature genetics*, 35(4):311–313, 2003.

[315] Melvin B Heyman et al. Lactose intolerance in infants, children, and adolescents. *Pediatrics*, 118(3):1279–1286, 2006.

[316] Kenneth M Weiss. The unkindest cup. *The Lancet*, 363(9420):1489–1490, 2004.

[317] Lynne C Olds and Eric Sibley. Lactase persistence dna variant enhances lactase promoter activity in vitro: functional role as a cis regulatory element. *Human molecular genetics*, 12(18):2333–2340, 2003.

[318] Rikke H Lewinsky, Tine GK Jensen, Jette Møller, Allan Stensballe, Jørgen Olsen, and Jesper T Troelsen. T- 13910 dna variant associated with lactase persistence interacts with oct-1 and stimulates lactase promoter activity in vitro. *Human molecular genetics*, 14(24):3945–3953, 2005.

[319] Lin Fang, Jong Kun Ahn, Dariusz Wodziak, and Eric Sibley. The human lactase persistence-associated snp- 13910* t enables in vivo functional persistence of lactase promoter–reporter transgene expression. *Human genetics*, 131(7):1153–1159, 2012.

[320] Jesper T Troelsen, Jørgen Olsen, Jette Møller, and Hans SjÖstrÖm. An upstream polymorphism associated with lactase persistence has increased enhancer activity. *Gastroenterology*, 125(6):1686–1694, 2003.

[321] Nabil Sabri Enattah, Mikko Kuokkanen, Carol Forsblom, Sirajedin Natah, Aino Oksanen, Irmä Järvelä, Leena Peltonen, and Erkki Savilahti. Correlation of intestinal disaccharidase activities with the c/t-

13910 variant and age. *World journal of gastroenterology: WJG*, 13(25):3508, 2007.

[322] Edyta Madry, Aleksandra Lisowska, Jarosław Kwiecień, Ryszard Marciniak, Anna Korzon-Burakowska, Sławomira Drzymała-Czyż, Ewa Mojs, and Jaroslaw Walkowiak. Adult-type hypolactasia and lactose malabsorption in poland. *Acta Biochimica Polonica*, 57(4):585–588, 2009.

[323] Zbigniew Dzialanski, Michael Barany, Peter Engfeldt, Anders Magnuson, Lovisa A Olsson, and Torbjörn K Nilsson. Lactase persistence versus lactose intolerance: Is there an intermediate phenotype? *Clinical biochemistry*, 49(3):248–252, 2016.

[324] H Rasinperä, M Kuokkanen, KL Kolho, H Lindahl, NS Enattah, E Savilahti, A Orpana, and I Järvelä. Transcriptional downregulation of the lactase (lct) gene during childhood. *Gut*, 54(11):1660–1661, 2005.

[325] Carina M Schlebusch, Per Sjödin, Pontus Skoglund, and Mattias Jakobsson. Stronger signal of recent selection for lactase persistence in maasai than in europeans. *European Journal of Human Genetics*, 21(5):550–553, 2013.

[326] Todd Bersaglieri, Pardis C Sabeti, Nick Patterson, Trisha Vanderploeg, Steve F Schaffner, Jared A Drake, Matthew Rhodes, David E Reich, and Joel N Hirschhorn. Genetic signatures of strong recent positive selection at the lactase gene. *The American Journal of Human Genetics*, 74(6):1111–1120, 2004.

[327] Andrew Szilagyi. Adult lactose digestion status and effects on disease. *Canadian Journal of Gastroenterology and Hepatology*, 29(3):149–156, 2015.

[328] AA Paul and DAT Southgate. Mccance and widdowsons. *Food Composition Tables in Translation*, page 129, 1978.

[329] W Nadia H Koek, Joyce B van Meurs, Bram CJ van der Eerden, Fernando Rivadeneira, M Carola Zillikens, Albert Hofman, Barbara Obermayer-Pietsch, Paul Lips, Huibert A Pols, André G Uitterlinden, et al. The t-13910c polymorphism in the lactase phlorizin hydrolase

gene is associated with differences in serum calcium levels and calcium intake. *Journal of bone and mineral research*, 25(9):1980–1987, 2010.

[330] Q Yang, SL Lin, SL Au Yeung, MK Kwok, L Xu, GM Leung, and CM Schooling. Genetically predicted milk consumption and bone health, ischemic heart disease and type 2 diabetes: a mendelian randomization study. *European Journal of Clinical Nutrition*, 2017.

[331] Genotyping and quality control of uk biobank, a large-scale, extensively phenotyped prospective resource. `http://www.ukbiobank.ac.uk/wp-content/uploads/2014/04/UKBiobank_genotyping_QC_documentation-web.pdf`. Accessed: 2017-05-9.

[332] Bradley Efron. Better bootstrap confidence intervals. *Journal of the American statistical Association*, 82(397):171–185, 1987.

[333] Anthony Christopher Davison and David Victor Hinkley. *Bootstrap methods and their application*, volume 1. Cambridge university press, 1997.

[334] Vincenzo Verardi and Christophe Croux. Robust regression in stata. 2008.

[335] Carlotta Sacerdote, Simonetta Guarrera, George Davey Smith, Sara Grioni, Vittorio Krogh, Giovanna Masala, Amalia Mattiello, Domenico Palli, Salvatore Panico, Rosario Tumino, et al. Lactase persistence and bitter taste response: instrumental variables and mendelian randomization in epidemiologic studies of dietary factors and cancer risk. *American journal of epidemiology*, 166(5):576–581, 2007.

[336] Helle KM Bergholdt, Børge G Nordestgaard, Anette Varbo, and Christina Ellervik. Milk intake is not associated with ischaemic heart disease in observational or mendelian randomization analyses in 98 529 danish adults. *International journal of epidemiology*, 44(2):587–603, 2015.

[337] Mikko Kuokkanen, Nabil S Enattah, Aino Oksanen, Erkki Savilahti, Arto Orpana, and Irma Järvelä. Transcriptional regulation of the lactase-phlorizin hydrolase gene by polymorphisms associated with adult-type hypolactasia. *Gut*, 52(5):647–652, 2003.

[338] Nick Patterson, Alkes L Price, and David Reich. Population structure and eigenanalysis. *PLoS genetics*, 2(12):e190, 2006.

[339] Catherine JE Ingram, Mohamed F Elamin, Charlotte A Mulcare, Michael E Weale, Ayele Tarekegn, Tamiru Oljira Raga, Endashaw Bekele, Farouk M Elamin, Mark G Thomas, Neil Bradman, et al. A novel polymorphism associated with lactose tolerance in africa: multiple causes for lactase persistence? *Human genetics*, 120(6):779–788, 2007.

[340] Nabil Sabri Enattah, Tine GK Jensen, Mette Nielsen, Rikke Lewinski, Mikko Kuokkanen, Heli Rasinpera, Hatem El-Shanti, Jeong Kee Seo, Michael Alifrangis, Insaf F Khalil, et al. Independent introduction of two lactase-persistence alleles into human populations reflects different history of adaptation to milk culture. *The American Journal of Human Genetics*, 82(1):57–72, 2008.

[341] Catherine JE Ingram, Tamiru Oljira Raga, Ayele Tarekegn, Sarah L Browning, Mohamed F Elamin, Endashaw Bekele, Mark G Thomas, Michael E Weale, Neil Bradman, and Dallas M Swallow. Multiple rare variants as a cause of a common phenotype: several different lactase persistence associated alleles in a single ethnic group. *Journal of molecular evolution*, 69(6):579, 2009.

[342] Alison M Smith and Katrine I Baghurst. Public health implications of dietary differences between social status and occupational category groups. *Journal of Epidemiology & Community Health*, 46(4):409–416, 1992.

[343] AE Perrin, C Simon, G Hedelin, D Arveiler, et al. Ten-year trends of dietary intake in a middle-aged french population: relationship with educational level. *European Journal of Clinical Nutrition*, 56(5):393, 2002.

[344] A Sanchez-Villegas, JA Martinez, R Prattala, E Toledo, et al. A systematic review of socioeconomic differences in food habits in europe: consumption of cheese and milk. *European journal of clinical nutrition*, 57(8):917, 2003.

[345] Nicole Darmon and Adam Drewnowski. Does social class predict diet quality? *The American journal of clinical nutrition*, 87(5):1107–1117, 2008.

[346] Sabine Rohrmann, Jakob Linseisen, Marianne Uhre Jakobsen, Kim Overvad, Ole Raaschou-Nielsen, Anne Tjonneland, Marie Christine Boutron-Ruault, Rudolf Kaaks, Nikolaus Becker, Manuela Bergmann, et al. Consumption of meat and dairy and lymphoma risk in the european prospective investigation into cancer and nutrition. *International journal of cancer*, 128(3):623–634, 2011.

[347] Jimmy Chun Yu Louie, VM Flood, DJ Hector, AM Rangan, and TP Gill. Dairy consumption and overweight and obesity: a systematic review of prospective cohort studies. *Obesity Reviews*, 12(7):e582–e592, 2011.

[348] Dengfeng Gao, Ning Ning, Congxia Wang, Yuhuan Wang, Qing Li, Zhe Meng, Yang Liu, and Qiang Li. Dairy products consumption and risk of type 2 diabetes: systematic review and dose-response meta-analysis. *PloS one*, 8(9):e73965, 2013.

[349] Nita G Forouhi. Association between consumption of dairy products and incident type 2 diabetesinsights from the european prospective investigation into cancer study. *Nutrition reviews*, 73(suppl 1):15–22, 2015.

[350] RA Ralston, JH Lee, H Truby, CE Palermo, and KZ Walker. A systematic review and meta-analysis of elevated blood pressure and consumption of dairy foods. *Journal of human hypertension*, 26(1):3–13, 2012.

[351] Sabita S Soedamah-Muthu, Lisa DM Verberne, Eric L Ding, Mariëlle F Engberink, and Johanna M Geleijnse. Dairy consumption and incidence of hypertensionnovelty and significance. *Hypertension*, 60(5):1131–1137, 2012.

[352] Stephanie De Smet, Nathalie Michels, Carolien Polfliet, Sara DHaese, Inge Roggen, Stefaan De Henauw, and Isabelle Sioen. The influence of

dairy consumption and physical activity on ultrasound bone measurements in flemish children. *Journal of bone and mineral metabolism*, 33(2):192–200, 2015.

[353] Larry A Tucker, Andrea Erickson, James D LeCheminant, and Bruce W Bailey. Dairy consumption and insulin resistance: the role of body fat, physical activity, and energy intake. *Journal of diabetes research*, 2015, 2015.

[354] Jing Ma, Edward Giovannucci, Michael Pollak, June M Chan, J Michael Gaziano, Walter Willett, and Meir J Stampfer. Milk intake, circulating levels of insulin-like growth factor-i, and risk of colorectal cancer in men. *Journal of the National Cancer Institute*, 93(17):1330–1336, 2001.

[355] Li-Qiang Qin, Ka He, and Jia-Ying Xu. Milk consumption and circulating insulin-like growth factor-i level: a systematic literature review. *International journal of food sciences and nutrition*, 60(sup7):330–340, 2009.

[356] Peter C Elwood, Janet E Pickering, D Ian Givens, and John E Gallacher. The consumption of milk and dairy foods and the incidence of vascular disease and diabetes: an overview of the evidence. *Lipids*, 45(10):925–939, 2010.

[357] D Aune, R Lau, DSM Chan, R Vieira, DC Greenwood, E Kampman, and T Norat. Dairy products and colorectal cancer risk: a systematic review and meta-analysis of cohort studies. *Annals of oncology*, 23(1):37–45, 2012.

[358] Zhaoxia Yu and Daniel J Schaid. Methods to impute missing genotypes for population data. *Human genetics*, 122(5):495–504, 2007.

[359] Alexandros Kanterakis, Patrick Deelen, Freerk van Dijk, Heorhiy Byelas, Martijn Dijkstra, and Morris A Swertz. Molgenis-impute: imputation pipeline in a box. *BMC research notes*, 8(1):359, 2015.

[360] Jonathan Marchini, Bryan Howie, Simon Myers, Gil McVean, and Peter Donnelly. A new multipoint method for genome-wide association

studies by imputation of genotypes. *Nature genetics*, 39(7):906–913, 2007.

[361] Orestis A Panagiotou, John PA Ioannidis, and Genome-Wide Significance Project. What should the genome-wide significance threshold be? empirical replication of borderline genetic associations. *International journal of epidemiology*, 41(1):273–286, 2011.

[362] Kevin N Laland, John Odling-Smee, and Sean Myles. How culture shaped the human genome: bringing genetics and the human sciences together. *Nature reviews. Genetics*, 11(2):137, 2010.

[363] Catherine JE Ingram, Charlotte A Mulcare, Yuval Itan, Mark G Thomas, and Dallas M Swallow. Lactose digestion and the evolutionary genetics of lactase persistence. *Human genetics*, 124(6):579–591, 2009.

[364] Joshua M Akey. Constructing genomic maps of positive selection in humans: Where do we go from here? *Genome research*, 19(5):711–722, 2009.

[365] Gil McVean. The structure of linkage disequilibrium around a selective sweep. *Genetics*, 175(3):1395–1406, 2007.

[366] Joseph Lachance. Detecting selection-induced departures from hardy-weinberg proportions. *Genetics Selection Evolution*, 41(1):15, 2009.

[367] Jiemin Liao, Xiang Li, Tien-Yin Wong, Jie Jin Wang, Chiea Chuen Khor, E Shyong Tai, Tin Aung, Yik-Ying Teo, and Ching-Yu Cheng. Impact of measurement error on testing genetic association with quantitative traits. *PloS one*, 9(1):e87044, 2014.

[368] William Barendse. The effect of measurement error of phenotypes on genome wide association studies. *BMC genomics*, 12(1):232, 2011.

[369] Stephen Leslie, Bruce Winney, Garrett Hellenthal, Dan Davison, Abdelhamid Boumertit, Tammy Day, Katarzyna Hutnik, Ellen C Royrvik, Barry Cunliffe, Daniel J Lawson, et al. The fine scale genetic structure of the british population. *Nature*, 519(7543):309, 2015.

[370] InterAct Consortium et al. Design and cohort description of the interact project: an examination of the interaction of genetic and lifestyle factors on the incidence of type 2 diabetes in the epic study. *Diabetologia*, 54(9):2272, 2011.

[371] Emma AD Clifton, Felix R Day, Emanuella De Lucia Rolfe, Nita G Forouhi, Soren Brage, Simon J Griffin, Nicholas J Wareham, and Ken K Ong. Associations between body mass index-related genetic variants and adult body composition: The fenland cohort study. *International journal of obesity (2005)*, 41(4):613, 2017.

[372] Bimal Bhindi, Christopher JD Wallis, Madhur Nayan, Ann M Farrell, Landon W Trost, Robert J Hamilton, Girish S Kulkarni, Antonio Finelli, Neil E Fleshner, Stephen A Boorjian, et al. The association between vasectomy and prostate cancer: A systematic review and meta-analysis. *JAMA Internal Medicine*, 2017.

[373] Elisabetta Rapiti, Gerald Fioretta, Robin Schaffar, Isabel Neyroud-Caspar, Helena M Verkooijen, Franz Schmidlin, Raymond Miralbell, Roberto Zanetti, and Christine Bouchardy. Impact of socioeconomic status on prostate cancer diagnosis, treatment, and prognosis. *Cancer*, 115(23):5556–5565, 2009.

[374] Jonathan I Epstein, Lars Egevad, Mahul B Amin, Brett Delahunt, John R Srigley, Peter A Humphrey, Grading Committee, et al. The 2014 international society of urological pathology (isup) consensus conference on gleason grading of prostatic carcinoma: definition of grading patterns and proposal for a new grading system. *The American journal of surgical pathology*, 40(2):244–252, 2016.