Improving Selection for Lung Cancer Screening in Socio-economically Disadvantaged Communities

A thesis submitted to the University of
Manchester for the degree of Doctor of
Philosophy in the Faculty of Biology, Medicine
and Health

2022

Mikey B. Lebrett

School of Biological Sciences

Division of Infection, Immunity and Respiratory
Medicine

Table of Contents

List of Figures **7**

List of Tables **10**

List of Abbreviations **12**

Abstract **14**

Declaration **15**

Copyright Statement **15**

Acknowledgements **16**

Preface **17**

Chapter One - Introduction **18**

1.1 Epidemiology of Lung Cancer 18

1.2 Lung Cancer Subtypes 19

1.3 Prognosis and Diagnosis 20

1.4 Treatment 21

1.5 Risk Factors 21

1.5.1 Smoking 21

1.5.2 Age 23

1.5.3 Asbestos, Occupational Hazards and Radon 23

1.5.4 Socio-economic Deprivation 24

1.5.5 Sex 25

1.5.6 Other Health Conditions 25

1.5.7 Family History 26

1.5.8 Genetic Risk Factors 27

1.5.8.1 Monogenic Variants 27

1.5.8.2 Polygenic Variants 28

1.6 Lung Cancer Screening 30

1.6.1 Principles of Cancer Screening 30

1.6.2 Assessing Screening Performance 31

1.6.3 Harms of Screening 34

1.6.4 Evidence for Lung Cancer Screening 35

1.6.5 Harms from Lung Cancer Screening 38

1.6.6 Selection Methods 40

1.6.6.1 Generalised Eligibility Criteria 40

1.6.6.2 Risk-Based Selection 41

1.6.6.3 Benefit-Based Selection 43

1.6.6.4 Polygenic Risk Scores 44
1.7 Lung Cancer Screening Implementation 46
1.7.1 Community-Based Screening 47
1.7.1.1 The Manchester Lung Health Check Pilot 47
1.7.1.2 Targeted Lung Health Check Programme 49
1.7.2 Risk Perception and Risk Communication..49
1.8 Aims and Objectives..53
Chapter Two - Methods and Protocols.. 54
2.1 Manchester LHC Pilot Cohort..54
2.2 Manchester LHC Pilot Follow-up Data Collection 55
2.3 Life-Gained Calculations 55
2.4 Case-Control Cohort Recruitment and Sample Acquisition..55
2.5 Case-Control DNA Extraction 56
2.5.1 Case Samples..56
2.5.2 Control Samples 57
2.6 DNA Quality Control, Quantification and Normalisation..57
2.7 Genotyping..58
2.8 Data Quality Control 61
2.9 Genomic Imputation 62
2.10 PRS Validation 62
2.11 GWAS and PRS Development 63
2.12 Manchester LHC Programme Cohort and Recruitment..63
Chapter Three - Risk Prediction Model Performance: Study of Manchester LHC Pilot.......... 65
3.1 Introduction 65
3.2 Methods..66
3.2.1 Study Cohort and Risk Prediction 66
3.2.2 Outcome Measures..66
3.2.3 Detection Rate Projections 67
3.2.4 Statistical Analysis..67
3.3 Results..67
3.3.1 Study Cohort Recruitment and Demographics 67
3.3.2 Lung Cancer Detection..68
3.3.3 Risk Prediction Model Calibration..71
3.3.4 Impact of Selection Method on Screening Performance..71
3.3.4.1 Screening Performance in Manchester LHC Pilot 71
3.3.4.2 Manchester LHC Pilot vs. NLST Detection Rates..74
3.3.5 Prevalence of Cardiovascular and Respiratory Comorbidity 75

3.4 Discussion 78
3.4.1 Calibration 78
3.4.2 Detection 80
3.4.3 Comorbidity 82
3.4.4 Conclusion 82
Chapter Four - 'Life-Gained' Screening Selection: Study of Manchester LHC Pilot 83
4.1 Introduction 83
4.2 Methods 84
4.2.1 Study Cohort and Data Collection 84
4.2.2 Analysis Approach 84
4.2.3 Life-Gained Calculations in LHC Pilot Cohort 85
4.2.4 Manchester LHC Pilot Follow-up 85
4.2.5 Model Calibration Calculations 85
4.2.6 Analysis 86
4.3 Results 86
4.3.1 Overview and Comparison of LYFS-CT and $PLCO_{M2012}$ 86
4.3.2 Equivalent Eligibility Thresholds 91
4.3.3 Model Performance within Screened Participants 94
4.3.4 LYFS-CT and $PLCO_{M2012}$ Calibration 100
4.4 Discussion 101
Chapter Five - Genetic Risk Factors for Lung Cancer: Case-Control Study 104
5.1 Introduction 104
5.2 Dataset Processing and Methods Development 105
5.2.1 Cohort Demographic Data Analysis 105
5.2.2 DNA Processing and Quality Control 105
5.2.3 Genotype Data Quality Control 108
5.2.4 Genomic Imputation 114
5.3 Results 115
5.3.1 Cohort Demographics 115
5.3.1.1 Control Cohort 115
5.3.1.2 Case Cohort 116
5.3.1.3 Comparison of Case and Control Cohorts 117
5.3.2 Validation of Previously Published Polygenic Risk Scores – Primary Analysis 118
5.3.2.1 Base clinical model 118
5.3.2.2 PRS Selection 120
5.3.2.3 Overall PRS performance 121
5.3.2.4 Subgroup analysis 129

5.3.2.5 PRS Scores and other Clinical Variables in the Control Cohort..........131
5.3.2.6 Combined PRS..........133
5.3.3 Novel SNPs in the Manchester Cohort..........134
5.3.3.1 Associated SNPs and Genes..........134
5.3.3.2 Polygenic Risk Score Development..........139
5.3.4 Secondary Analysis..........143
5.4 Discussion..........146
5.4.1 Validation of Published PRSs..........146
5.4.2 Novel PRS Development..........148
5.4.3 Strengths and Limitations..........150
5.4.4 Implementation and Next Steps..........152
Chapter Six - Lung Cancer Risk Perception: Study of Manchester LHC Programme 154
6.1 Introduction..........154
6.2 Questionnaire Development and Methods..........155
6.2.1 Measures..........155
6.2.1.1 Risk Perception..........155
6.2.1.2 Disease Knowledge..........155
6.2.1.3 Disease Specific Worry..........156
6.2.2 Patient and Public Involvement and Engagement..........156
6.2.3 Analysis..........158
6.3 Results..........159
6.3.1 Study Cohort Characteristics..........159
6.3.2 Questionnaire Responses..........162
6.3.2.1 Risk Perception..........162
6.3.2.2 Lung Cancer-Specific Worry..........162
6.3.2.3 Lung Cancer Knowledge..........162
6.3.2.4 PHQ-4 Mental Health Score..........164
6.3.3 Questionnaire Responses Stratified by Risk Score and Screening Eligibility..........165
6.3.4 Screening Adherence..........165
6.3.5 Demographic Factors Associated with Risk Perception..........168
6.3.5.1 Perceived Absolute Risk..........168
6.3.5.2 Perceived Comparative Risk..........168
6.3.6 Demographic Factors Associated with Lung Cancer Worry..........169
6.3.7 Demographic Factors Associated with Mental Health..........169
6.4 Discussion..........172
Chapter Seven – Discussion and Future Work 177
7.1 Project Overview..........177

7.2 Future Work 179
References 182
Appendices 210
8.1 Appendix 1 – Comorbidity Classification for Life-Gained Study Follow-up 210
8.2 Appendix 2 - SNP-level Data for PRSs Validated in Manchester Cohort 211
8.3 Appendix 2 – Full GWAS Results for Significant SNPs in Manchester Cohort 215
8.4 Appendix 3 – Lung Cancer Risk Survey 230

Wordcount: 53,070

List of Figures

Figure 1. UK lung cancer incidence by year of diagnosis (1993-2018), overall and stratified by sex. Created using data sourced from CRUK [3]. 18

Figure 2. UK lung cancer mortality by year of death (1971-2018), overall and stratified by sex. Created using data sourced from CRUK [6]. 19

Figure 3. Biases that affect screening trials. A = pathway of a screening participant. Demonstrates lead-time bias if non-screened pathway would be B. Demonstrates overdiagnosis bias if non-screen pathway would be C. 32

Figure 4. Lung cancer incidence (A) and deaths (B) in NLST cohort. Reproduced with permission from [193], Copyright: Massachusetts Medical Society. 36

Figure 5. Lung cancer incidence (A) and mortality (B) among male participants of the NELSON trial. Reproduced with permission from [194], Copyright: Massachusetts Medical Society. 37

Figure 6. Participant journey through LHC programme. 48

Figure 7. Images from the Manchester LHC pilot [251]. 48

Figure 8. Illumina Infinium HTS assay manual workflow protocol. Used under license from Illumina, Inc. All Rights Reserved. 60

Figure 9. Image of the mobile LHC unit from the Manchester Lung Health Check programme [315]. 64

Figure 10. Number needed to screen (NNS) to detect one lung cancer in Manchester LHC pilot, stratified by sex and lung cancer risk group. 70

Figure 11. Projected cumulative number of individuals with lung cancer detected in the Manchester LHC pilot, with 6-years of follow-up. Adapted from [318], reuse permitted under CC BY-BC. 71

Figure 12. Participants of the Manchester LHC pilot eligible for screening by each set of eligibility criteria. Percentage in () = lung cancer detection rate within segment. Number in [] = number of participants eligible for screening. Shaded = eligible in LHC pilot. Reproduced from [318], reuse permitted under CC BY-NC. 73

Figure 13. Distribution of screen-detected lung cancers in the Manchester LHC pilot according to risk score ($\text{PLCO}_{\text{M2012}}$ and LLP_{V2} RPMs). Reproduced from [318], reuse permitted under CC BY-NC. 74

Figure 14. FEV_1/FVC ratio stratified by lung cancer risk group ($\text{PLCO}_{\text{M2012}}$). Horizontal line represents 70% threshold indicative of airflow obstruction. Reproduced from [318], reuse permitted under CC BY-NC. 77

Figure 15. Distribution of LYFS-CT scores for all attendees of the Manchester LHC pilot. LYFS-CT calculates estimated days of life gained from lung cancer screening based on life expectancy, lung cancer risk, and screening mortality benefit. 86

Figure 16. $\text{PLCO}_{\text{M2012}}$ scores vs. LYFS-CT scores for complete LHC pilot cohort. $\text{PLCO}_{\text{M2012}}$ scores presented in proportion form as this is how the lcmodels package generates the scores. These were converted to percentages in the descriptions to retain consistency with other chapters. 90

Figure 17. LYFS-CT score distribution stratified by screening status in the LHC pilot. 1 = had LDCT scan in LHC pilot. 0 = did not have LDCT scan in LHC pilot. LYFS-CT calculates estimated days of life gained from lung cancer screening based on life expectancy, lung cancer risk, and screening mortality benefit. 90

Figure 18. Distribution of LHC pilot participants by LYFS-CT and $\text{PLCO}_{\text{M2012}}$ scores. Vertical line = $\text{PLCO}_{\text{M2012}}$ 1.51% threshold. Horizontal Line = LYFS-CT 12.2 days of life gained threshold. Only participants with $\text{PLCO}_{\text{M2012}}$ score ≤5% displayed in figure to highlight differing selections. 92

Figure 19. Mean number of comorbidities among screened individuals in the LHC pilot, selected (in top 75% of scores) by life-gained-benefit but not risk vs. selected by risk but not life-gained. .98

Figure 20. Sample processing and quality control flowchart for case-control study. 107

Figure 21. Heterozygosity and call rate quality control plot. Red dashed lines indicate the thresholds for sample exclusion. 108
Figure 22. Genetic ancestry of the Manchester cohort dataset vs. the HM3 reference panel. Grey = excluded due to deviation from CEU cluster. CEU: Utah residents with Northern and Western European ancestry from the CEPH collection. CHB: Han Chinese in Beijing, China. JPT: Japanese in Tokyo, Japan. YRI: Yoruba in Ibadan, Nigeria. 109
Figure 23. Aberrant plot excluding 18 outliers from the primary analysis cohort. Dashed line = 99% confidence interval of inferred distribution of 'normal' samples. Normal individuals are coloured from black to grey, with darker colouring indicating sample density. Outliers are coloured from orange to red, with redder colours indicating higher posterior probability of being an outlier.... 110
Figure 24. PC1 and PC2 scatterplot for the final primary analysis cohort. Light blue = case; dark blue = control. 110
Figure 25. Scree plot for primary analysis. Dots represent eigenvalues plotted against PCs. 111
Figure 26. PC1 and PC2 scatterplot for the secondary analysis cohort. Light blue = case; dark blue = control. Outcropping of controls with low PC1 values indicates unexpected abberant genetic variability. These samples were removed from primary analysis but retained in secondary. 112
Figure 27. Scatterplots for first ten principal components in the control cohort of the secondary analysis dataset. Light blue = aberrant plate; dark blue = remaining controls. 113
Figure 28. Scree plot for secondary analysis. 114
Figure 29. Correlation of densities of allele frequencies between the study samples and imputation reference panel. 114
Figure 30. ROC curve for overall base clinical model, clinical model excl. FEV_1/FVC ratio, and individual clinical factors, in primary analysis cohort. 119
Figure 31. ROC curves for all PRSs validated, showing the range of AUC increases conferred by addition of a PRS (Young PRS not plotted as it does not add AUC). 125
Figure 32. Density plots for each of the PRS tools validated, stratified by case/control grouping. Mean score for overall cohort is normalised to 0. Dashed lines = mean scores for case or control subgroups. 126
Figure 33. ROC curves for all PRSs, displayed independently and added to base clinical model. . 128
Figure 34. Reported previous cancer types in the 70 LHC pilot participants comprising the primary analysis dataset. 132
Figure 35. Functional Mapping and Annotation of Genome-Wide Association Studies (FUMA) generated quantile-quantile (Q-Q) plot for Primary Analysis. 134
Figure 36. Functional Mapping and Annotation of Genome-Wide Association Studies (FUMA) generated Manhattan plot of Primary Analysis. Dashed line = threshold for genome-wide significance. 135
Figure 37. Regional plot for Manchester GWAS SNPs in the MAGI2 gene. Displays the genomic environment of the SNPs in chromosome 7 that were significantly associated with lung cancer in my association analysis. Purple dot = top lead SNP. Red/orange/yellow dots = independent significant SNPs, linkage disequilibrium with top lead SNP. Red line = mapped gene. 136
Figure 38. Regional plot for Manchester GWAS SNPs in the DAPK1 gene. Displays the genomic environment of the SNPs in chromosome 9 that were significantly associated with lung cancer in my association analysis. Purple dot = top lead SNP. Red/orange/yellow dots = independent significant SNPs, linkage disequilibrium with top lead SNP. Red line = mapped gene. 136
Figure 39. Chart showing the model fit of various PRSice-2 generated PRSs, with increasingly lenient P-value thresholds (and more SNPs included). 142
Figure 40. ROC curves for PRS52 in the Manchester case-control cohort. 142
Figure 41. Study participant estimates of UK lung cancer incidence. Shaded bar = correct response. 163
Figure 42. Study participant mean survival estimates vs. actual mean survival. 163

Figure 43. Lung cancer risk factor identification in study participants. Reused from [407], permitted under CC BY 4.0. 164
Figure 44. Actual and perceived absolute lung cancer risk score stratified by sex. Reused from [407], permitted under CC BY 4.0 166
Figure 45. Perceived comparative risk vs. actual calculated risk ($PLCO_{m2012}$). Reused from [407], permitted under CC BY-4.0 166

List of Tables

Table 1. Risk factors included in PLCO$_{m2012}$ and LLP$_{v2}$ risk prediction models. Below delineation = factors unique to RPM. 42

Table 2. Demographic and clinical variables of complete Manchester LHC pilot cohort and stratified by PLCO$_{M2012}$ risk category. 69

Table 3. Univariable and multivariable analysis of the risk of screen detected lung cancer stratified by clinical variables and PLCO$_{M2012}$ score 76

Table 4. Comparison of the variables included in PLCO$_{M2012}$ and LYFS-CT. Below delineation = variables unique to each model. 87

Table 5. Demographic and clinical characteristics of the complete Manchester LHC Pilot, stratified by LYFS-CT quintile. 88

Table 6. Demographics and calculated risk and benefit scores for the LHC pilot cohort, using the highest 1430 scores for PLCO$_{M2012}$ and LYFS-CT, amongst the 2541 participants of the LHC. Test for median included non-integer imputed values and could therefore be non-integers. Both median and mean values and P-values provided for comorbidities as both are considered in analysis. 93

Table 7. Demographics and follow-up outcomes of LHC pilot screened participants, stratified by those selected by risk vs. life-gained selection in 1409 scanned participants. 96

Table 8. Outcomes of LHC pilot participants followed-up at 5-years post-LHC. 99

Table 9. Outcomes for those with screen-detected lung cancer in the LHC pilot. 100

Table 10. Comparison of control samples included in the study vs. remaining eligible control population. 115

Table 11. Demographic and clinical characteristics of controls and comparison final primary control cohort vs. failed samples 116

Table 12. Demographic and clinical characteristics of cases and comparison final case cohort vs. failed samples. 117

Table 13. Comparison of demographic variables between post-QC cases and controls for primary analysis. 118

Table 14. Associations between clinical factors and lung cancer from logistic regression in primary analysis cohort. 119

Table 15. Summary of PRSs validated in the Manchester cohort, data sourced from GRS Catalog. 120

Table 16. Proxy SNPs used in lieu of published SNPs absent in this dataset. Magnitude of linkage disequilibrium between published and proxy SNPs represented by R^2 and D'**.** D' is a basic measure of the difference between observed and expected frequencies of a haplotype. R^2 also takes allele frequency into account, expressing the correlation between a pair of loci 121

Table 17. Performance metrics for the nine PRS validations in the Manchester case-control cohort, as well as a combined PRS of most frequently included SNPs. 124

Table 18. Case proportion across the PRS quintiles in the case-control cohort. 126

Table 19. Subgroup analyses for Jia PRS validation. 130

Table 20. Distribution of participants with previous cancer diagnosis across PRS quintiles. 132

Table 21. Combined PRS, SNPs shared by at least four of the validated PRSs 133

Table 22. SNPs displaying high levels of significance from GWAS analysis of the primary analysis dataset. Shaded = lone SNP at that locus. 137

Table 23. Fifty-two SNP PRS developed in the Manchester cohort. 140

Table 24. Comparison of secondary analysis-only controls and primary & secondary analysis controls. 143

Table 25. Comparison of case and control demographics in secondary analysis. 144

Table 26. Comparison of PRS performance in primary and secondary analyses. 145

Table 27. Changes made to study survey after PPIE consultation. 157

Table 28. Study questionnaire responses for all respondents and stratified by consent to linked clinical data. 160
Table 29. Demographic characteristics for the identified subset of the study participants. 161
Table 30. Questionnaire responses for identified subset and stratified by screening eligibility, based on $PLCO_{M2012}$ calculated risk score. 167
Table 31. Binary logistic regression modelling (model 'A') for factors associated with risk perception, lung cancer worry, depression and anxiety (PHQ-4). 170
Table 32. Binary logistic regression modelling (model 'B') for factors associated with risk perception, lung cancer worry, depression and anxiety (PHQ-4). 171

List of Abbreviations

ANOVA	One-way analysis of variance
AUC	Area under the curve
BMI	Body mass index
CAC	Coronary artery calcification
CI	Confidence interval
CMS	Centers for Medicare & Medicaid Services
COPD	Chronic obstructive pulmonary disease
CRUK	Cancer research UK
CT	Computed tomography
CVD	Cardiovascular disease
CXR	Chest x-ray
DNP	Dinitrophenyl
FEV1	Forced expiratory volume in 1 second
FPRP	False positive report probability
FUMA	Functional Mapping and Annotation of Genome-Wide Association
GWAS	Studies
FVC	Forced vital capacity
gnomAD	The Genome Aggregation Database
GOLD	Global Initiative for Chronic Obstructive Lung Disease
GWAS	Genome Wide Association Study
HRC	Haplotype reference panel
HTS	High-throughput screening
HWE	Hardy-Weinberg equilibrium
IBD	Identity by descent
I-ELCAP	International Early Lung Cancer Action Project
ILST	International Lung Screening Trial
IMD	Index of multiple deprivation
IQR	Interquartile range
LCDRAT	Lung Cancer Death Risk Assessment Tool
LDCT	Low-dose computed tomography
LHC	Lung health check
LLP	Liverpool lung project
LSUT	Lung screening uptake trial
LYFS-CT	Life Years Gained from Screening-CT
MAF	Minor allele frequency
MCRC	Manchester Cancer Research Centre
NELSON	Dutch-Belgian Randomised Lung Cancer Screening Trial
NHIS	National Health Interview Study
NHS	National Health Service
NICE	National Institute for Health and Care Institute
NLST	National Lung Screening Trial
NNS	Number needed to screen
NPV	Negative predictive value
NRI	Net reclassification improvement
NSCLC	Non-small cell lung cancer
OR	Odds ratio
PC	Principal component
PCA	Principal component analysis
PLCO	Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial

PPV	Positive predictive value
PRS	Polygenic risk score
QC	Quality control
RCT	Randomised controlled trial
ROC	Receiver operator curve
RPM	Risk prediction model
SABR	Stereotactic ablative radiotherapy
SCLC	Small cell lung cancer
SD	Standard deviation
SNP	Single nucleotide polymorphism
TLHC	Targeted Lung Health Check
TNM	Tumour, Node, Metastasis
UKLS	UK Lung Screening Trial
UK NSC	UK National Screening Committee
USPSTF	United States Preventative Services Task Force
WHO	World Health Organisation
YLST	Yorkshire Lung Screening Trial

Abstract

Background

Lung cancer (LC) is the leading cause of cancer mortality worldwide. Poor survival is driven by late onset of non-specific symptoms, resulting in advanced stage diagnoses. Evidence for the efficacy of low-dose CT (LDCT) screening in detecting cancer earlier, thereby reducing lung-cancer specific mortality, is now well established. Attention has turned to developing and implementing screening programmes in the population. A key aspect of an effective screening programme is the successful selection of participants; this ensures a favourable benefit-to-harm ratio for participants and an efficient and cost-effective programme. This thesis aims to improve screening selection in socio-economically disadvantaged populations by identifying areas of sub-optimal performance and considering strategies for further improvement. The focus on socio-economically deprived populations is of particular importance, as lung cancer risk is often higher in these subgroups, positioning them to be an ideal target population for LDCT screening.

Methods

I) A retrospective study of the Manchester Lung Health Check (LHC) pilot, a community-based LC screening programme, comparing the selection performance and calibration of National Lung Screening Trial (NLST) criteria and two risk prediction models (RPMs) ($PLCO_{M2012}$ and LLP_{V2}), as well as the comorbidity profile of the screening cohort. II) Retrospective modelling of a benefit-based selection approach (LYFS-CT) in the LHC pilot, comparing performance with a risk-based approach and examining the characteristics and outcomes of the screening cohort. III) A Manchester-based case-control study validating nine published polygenic risk score (PRS) tools and assessing if they could improve risk prediction. IV) A cross-sectional questionnaire study of LHC programme participants, examining risk perception, worry and disease knowledge.

Results

There were significant differences in screening selection performance based on the method of selection used. RPMs contributed to increased screening efficiency compared to NLST, but under-estimated LC risk in this population and selected a screening cohort with high levels of comorbidity. Inclusion of spirometry (FEV_1/FVC ratio) or coronary artery calcification in RPMs may improve risk prediction but would further increase participant comorbidity. LYFS-CT selected significantly younger and less comorbid participants but also directed screening away from the most socio-economically disadvantaged. Eight PRS tools were successfully validated in the Manchester cohort and two novel genetic loci were identified for possible inclusion in a future PRS. Participants' comparative risk perception was more accurate than absolute risk perception. Women and those at high LC risk were more likely to have adverse psychological indicators.

Conclusion

Risk-based selection leads to high screening efficiency, but RPMs are not well calibrated for use in socio-economically deprived populations and the optimal RPM and risk threshold strategy is unclear. Benefit-based selection may be an important tool for maximising the screening benefit provided to participants. Prospective studies are required to further elucidate the most advantageous selection strategy. Inclusion of genetic risk factors in RPMs may improve both risk- and benefit-based screening selection. Comparative-based language and decision aids should be employed for communicating risk to screening participants and ensuring effective shared decision making.

Declaration

No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright Statement

I. The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the “Copyright”) and s/he has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.

II. Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made only in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made.

III. The ownership of certain Copyright, patents, designs, trademarks and other intellectual property (the “Intellectual Property”) and any reproductions of copyright works in the thesis, for example graphs and tables (“Reproductions”), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

IV. Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see http://documents.manchester.ac.uk/DocuInfo.aspx?DocID=24420), in any relevant Thesis restriction declarations deposited in the University Library, The University Library’s regulations (see http://www.library.manchester.ac.uk/about/regulations/) and in The University’s policy on Presentation of Theses.

Acknowledgements

It has been a remarkable privilege to have pursued this PhD over the past three years. For this, I am hugely thankful to Dr Phil Crosbie for having selected me for this position, as well as for his professional, expert, and thoughtful supervision throughout my PhD journey. I am incredibly grateful to the rest of my supervisory team, Professor Emma Crosbie, Dr Miriam Smith, and Professor Gareth Evans, for their equally inspiring and supportive supervision. It has been an unsurpassed honour to learn from all four of them.

So many colleagues and staff members at the University of Manchester have been instrumental in training me and supporting my research throughout my PhD. At risk of inadvertent omission, thank you to Helen Byers and Debbie Burt for your extensive assistance in the laboratory. Thank you to John Bowes, Katie Sadler, Cemsel Bafligil, and Li Cheung for guiding me so patiently through the bioinformatics. Thank you to Elaine Harkness for helping me find my statistical feet in the early days of my PhD journey. Thank you to Haval Balata for sharing his detailed knowledge of the Manchester LHC database with me. My risk perception study would not have been possible without the assistance of nursing and trial staff on the LHC screening truck. A special thank you to all my fellow PhD students on both the 5th and 6th floors of St Mary's Hospital, all of whom have contributed to an enjoyable and supportive work environment.

None of this work would have been possible without the financial support of the NIHR Manchester Biomedical Centre. Nor would it have been possible without the selfless provision of samples and data by lung cancer patients and screening participants in Manchester.

I count myself exceptionally blessed to have such a wonderful group of family and friends. To my remarkable and inspiring parents (without whom none of this would have been possible for so many reasons), grandparents, siblings, children (Yonatan and Meira), extended family, and dear friends, thank you for cheering me on throughout this endeavour. To my wonderful wife Rivka, you are the best antidote to imposter syndrome I have ever encountered, as well as an extraordinary partner, mother, and doctor. Thank you for everything.

Preface

This PhD thesis has been formatted in the 'Traditional Style' and is composed of seven chapters in total. Elements of this thesis have been published in peer-reviewed journals.

I)

Lebrett MB, Crosbie EJ, Smith MJ, Woodward ER, Evans DG, Crosbie PAJ.

Targeting lung cancer screening to individuals at greatest risk: the role of genetic factors.

Journal of Medical Genetics 2021;58:217-226.

[Contains elements from Chapter One and Chapter Five's discussion. My role: drafting of paper as first author and finalising submission.]

II)

Lebrett MB, Balata H, Evison M, Colligan D, Duerden R, Elton P, Greaves M, Howells J, Irion K, Karunaratne D, Lyons J, Mellor S, Myerscough A, Netwon T, Sharman A, Smith E, Taylor B, Taylor S, Walsham A, Whittaker J, Barber PV, Tonge J, Booton R, Crosbie PAJ.

Analysis of lung cancer risk model (PLCO$_{M2012}$ and LLP$_{v2}$) performance in a community-based lung cancer screening programme.

Thorax 2020;75:661-668.

[A truncated version of Chapter Three. My role: data analysis and presentation, drafting of paper as first author and finalising submission.]

III)

Lebrett MB, Crosbie EJ, Yorke J, Hewitt K, Rowlands A, Badrick E, Evans DG, Balata H, Booton R, Crosbie PAJ.

Risk Perception and Disease Knowledge in Attendees of a Community-Based Lung Cancer Screening Programme.

Lung Cancer 2022;168:1-9.

[A truncated version of Chapter Six. My role: conceptualisation, data curation, investigation, analysis, presentation, drafting of paper as first author and finalising submission.]

Chapter One - Introduction

1.1 Epidemiology of Lung Cancer

Lung cancer is the second most common cancer worldwide, with an estimated 2.2 million new cases in 2020, representing approximately 11% of total cancer diagnoses. It is the most common cancer in men and third most common cancer in women. Lung cancer is the leading cause of cancer mortality, with 1.8 million people dying of the disease in 2020, representing 18% of total cancer deaths [1]. In the UK, approximately 48,000 new lung cancer cases are diagnosed annually, making it the third most common cancer (13% of all cancer cases). One-in-fifteen UK women and one-in-thirteen UK men will be diagnosed with lung cancer in their lifetime. Overall incidence rates decreased by 9% between 1993 and 2018, primarily as a result of reductions in risk factor exposure [2]. However, there is considerable divergence between the sexes; UK female age-standardised incidence rates increased by 32% between 1993 and 2018, compared to a decrease of 34% for males over the same period [3] (Figure 1). This may be partially driven by sex differences in historic smoking patterns, although other factors have also been implicated. For example, women are at higher risk of developing adenocarcinoma, a specific histological subtype of lung cancer. It is possible that the popularisation of filtered cigarettes in the 1970s (which increase the risk of adenocarcinoma over other lung cancer subtypes) coincided with large increases in female smoking rates, resulting in accelerated incidence among women [4].

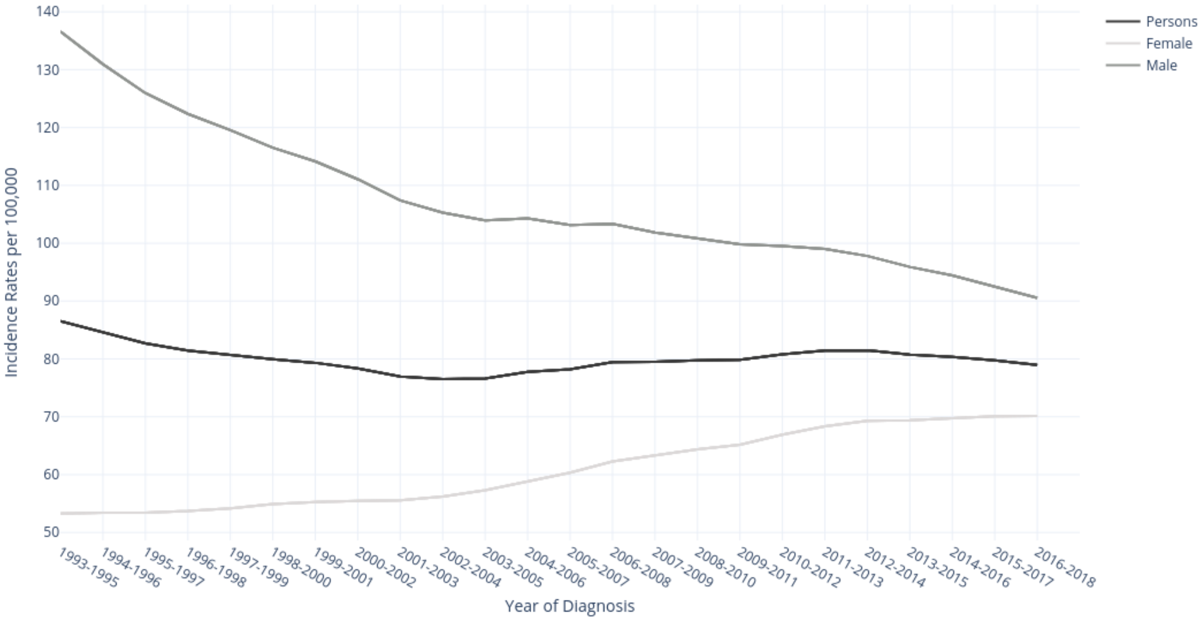


Figure 1. UK lung cancer incidence by year of diagnosis (1993-2018), overall and stratified by sex. Created using data sourced from CRUK [3].

In the UK, lung cancer is responsible for more than 20% of cancer deaths, making it the leading cause of cancer death for British men and women; almost 100 people die from the disease daily [5]. Since 1971, age-standardised mortality rates have decreased by 29%. Like for incidence rates, this statistic masks a marked difference between the sexes; male mortality decreased by 58% over this period, whilst female mortality increased by 81%. However, in the most recent decade for which there is complete data, mortality rates have decreased for both sexes (male: 22%; female: 5%) (Figure 2) [6]. Age-stratified mortality trends show that rates peaked and fell in younger age groups before older age groups. Most mortality reduction is caused by lower rates of smoking uptake, rather than higher quit rates. Consequently, it takes longer for mortality rates to drop in older age groups who have much higher rates of ever-smoking (even if a proportion are former smokers), compared to younger age groups who are less likely to start smoking initially [6]. Mortality is projected to decrease further in the next 15 years, with a 28% reduction forecasted between 2014 and 2035 [7].

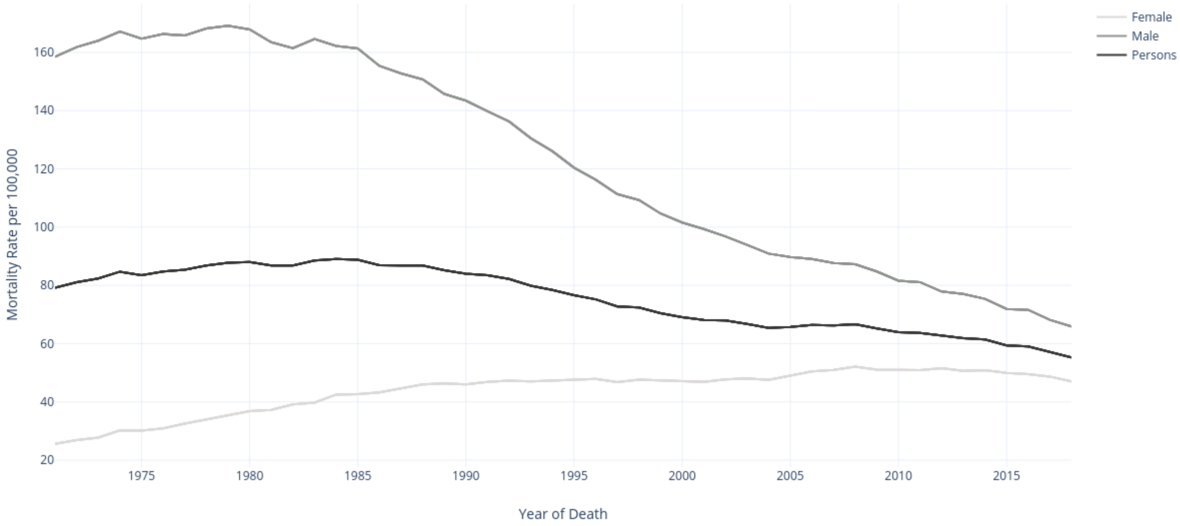


Figure 2. UK lung cancer mortality by year of death (1971-2018), overall and stratified by sex. Created using data sourced from CRUK [6].

1.2 Lung Cancer Subtypes

Cancer is characterised by cells that have acquired key biological abilities (‘hallmarks’), facilitating their development into malignant tumours. Six of these hallmarks were originally defined by *Hanahan* and *Weinberg* in a paper published in 2000: self-sufficiency in growth signals, insensitivity to anti-growth signals, tissue invasion and metastasis, limitless replicative potential, sustained angiogenesis, and evasion of apoptosis [8]. In 2011, *Hanahan* and *Weinberg* proposed two additional hallmarks: reprogramming of energy metabolism and evading immune destruction, as well as two enabling characteristics: genome instability and mutation, and tumour-promoting inflammation [9]. Lung cancer occurs when cells in the lungs or airways acquire these biological abilities, transform, and develop into malignant tumours.

The most common pathological subtype of lung cancer is non-small-cell lung cancer (NSCLC), comprising approximately 80-85% of cases. NSCLC can be further subcategorised histologically into adenocarcinomas, squamous cell carcinomas, and large-cell carcinomas. Adenocarcinomas and squamous cell carcinomas are the most commonly diagnosed subtypes of NSCLC (~40% and ~30% of diagnoses respectively), with squamous cell carcinomas strongly associated with tobacco smoke exposure and adenocarcinomas being the dominant subtype in non-smokers with lung cancer [10]. Small-cell lung cancer (SCLC) is less common but more aggressive than NSCLC, with faster doubling times and a greater propensity to metastasise early [11]. The genetic diversity and clinical differences between SCLC, adenocarcinomas and squamous cell carcinomas have led some to suggest that these subtypes should be treated as distinct diseases [12].

1.3 Prognosis and Diagnosis

Lung cancer prognosis is poor. In England and Wales, only 40% of patients survive for a year following diagnosis. This proportion drops to approximately 16% survival over five years and 10% survival over ten years [5]. In four decades, five-year age-standardised survival only improved by 7% for women and 4% for men [13]; this stands in stark contrast to the doubling of overall cancer survival in the UK in that time [14].

Poor lung cancer survival is driven by late diagnosis in the majority of patients. NSCLC progression is classified using the TNM (tumour, node, metastasis) staging system, describing the size of the tumour and extent of spread to the lymph nodes and rest of the body. The TNM system can be used to define the overall stage of the disease described by the number staging system; stage I and stage II refer to tumours that are ‘early stage’ (small and contained within the lungs), stage III and stage IV refer to advanced or ‘late stage’ cancers that have grown and spread to other tissues and organs [15]. Most lung cancer cases are diagnosed at a late stage, with approximately half diagnosed at stage IV, at which point the cancer has metastasised and is incurable [5]. The one-year survival rate of stage IV disease is 19%, compared to 88% for those diagnosed at stage I [16]. Even within stage I, tumour diameter is predictive of survival. Five-year survival decreases by 5% for each 1 cm that tumour diameter increases; this emphasises the importance of early detection, even at the earliest stage of lung cancer development [17].

Late clinical presentation of lung cancer is caused by the asymptomatic nature of early-stage disease; even when symptoms do appear, they are largely non-specific. The National Institute for Health and Care Excellence (NICE) advises doctors to activate the lung cancer referral pathway if a patient aged over 40 suffers from any two of: fatigue, cough, shortness of breath, chest pain, weight loss or appetite loss. If the patient is an ever-smoker, they are referred with any one of these symptoms. Any patient is referred immediately if they suffer from one of: finger clubbing, persistent or recurrent chest infections, supraclavicular lymphadenopathy or thrombocytosis [18].

Many of these symptoms are indicators of other smoking related conditions, such as COPD and emphysema. This can result in lung cancer remaining undetected even after the presentation of symptoms.

1.4 Treatment

Treatment options for lung cancer depends on the stage of the disease, tumour subtype, as well as the condition of the patient. Treatment success rates are always superior when the cancer is detected at an early stage. Early-stage NSCLC is often most successfully treated through surgery, with a lobectomy (partial lung removal) or pneumonectomy (full lung removal) providing the most favourable survival statistics for patients with stage I and stage II diseases. Resection rates in the UK doubled from 9% to 18% between 2006 and 2018 [19].

Radiotherapy or chemotherapy may be offered in addition to surgery to reduce the chance of recurrence, or instead of surgery for high-risk patients or patients with unresectable tumours [20]. Stereotactic ablative radiotherapy (SABR) is a form of radiotherapy which uses multiple thin beams of radiation directed from multiple angles to target doses more accurately at the tumour; there is evidence that it provides more favourable outcomes than standard radiotherapy in early-stage NSCLC [21,22], and is recommended by NICE as the preferential treatment modality for patients who cannot undergo surgery [23].

Patients with stage III NSCLC may be offered adjuvant chemotherapy followed by surgery (or vice versa), chemoradiation (a combination of radiotherapy and chemotherapy), or external radiotherapy. Stage IV NSCLC is usually treated palliatively, with chemotherapy, radiotherapy, combination therapies and sometimes surgery being employed to ease symptoms [20]. Immunotherapy and targeted biological therapies may also be used to treat specific NSCLC types, depending on the specific genetic profile of the tumour [23].

1.5 Risk Factors

1.5.1 Smoking

Approximately 80% of lung cancer cases can be attributed to modifiable risk factors, one of the highest rates of all cancer types [24]. Smoking constitutes the bulk of this modifiable risk [25]. It is estimated that between 72% and 86% of lung cancer cases in the UK are caused by smoking [24,26,27]. The relationship between smoking and lung cancer began emerging in the first half of the 20th century, with evidence accumulating from a large variety of population, animal, and cellular studies [28]. Prospective evidence relating to the link between lung cancer and smoking was gathered by *Doll, Hill* and *Peto et. al.* in a paradigmatic cohort study which began in 1951 and ran for 50-years, terminating in 2001. The study regularly surveyed the smoking habits of tens of thousands of British doctors and collected detailed data relating to rates and causes of mortality

[29,30]. The first results published in 1954 showed that men who smoked an average of >25 g of tobacco a day had a death rate of 1.14 per 1000, compared to a rate of 0 in non-smokers and 0.66 in all subjects [29]. The final publication from the study in 2005 demonstrated that smokers born between 1900-1930 lost an average of 10 years of life compared to non-smokers; lung cancer (and chronic obstructive lung disease) accounted for a quarter of the excess mortality among smokers [30].

Considerable research has confirmed this link as causal, with cigarette smoke found to contain more than 70 carcinogenic compounds [31]. These compounds are catalysed into forms that covalently bond with DNA, producing DNA adducts. A high volume of DNA adducts can overwhelm DNA repair pathways and result in persistent damage, which increases the likelihood of somatic mutations [32]. Mutations caused by adducts on the *KRAS* oncogene are common in lung cancer, present in between 16% and 40% of NSCLC cases. These mutations occur with particularly high frequency in adenocarcinomas and in ever-smokers, although can also occur in squamous cell carcinomas and in never-smokers [33]. Several clinical trials are underway with the objective of developing targeted therapies for the treatment of *KRAS*-mutated lung cancers [34].

Increasing either smoking duration or intensity contributes to greater lung cancer risk, with duration having a larger effect; smoking at higher intensity for a shorter period is less deleterious than smoking at lower intensity for longer periods [35]. Use of other tobacco delivery systems such as shisha, cigars and pipes have also been shown to be robustly associated with increased lung cancer risk [36–38].

Passive smokers (individuals who are exposed to second-hand environmental tobacco smoke) may have an increased lung cancer risk of 20-30% [39]. This association was established as causal through large meta-analyses of epidemiological studies [40]. In the UK, 1% of lung cancer cases are attributed to second-hand smoke. This proportion increases to 15% of lung cancers diagnosed in patients who have never smoked [24]. However, a large prospective study of 76,000 women contended that there is no clear link between passive smoking and lung cancer, positing that recall bias may contribute to an increased association between the two factors in other studies [41]; the findings of this study were in turn questioned [42]. The extent of the association between second-hand smoke and lung cancer remains controversial.

Smoking cessation is the most effective strategy to reduce lung cancer risk. The final paper from the prospective study by *Doll et.al.* reported that those who quit smoking by age 50 halved their smoking-derived mortality hazard, and those quitting at age 30 suffering no loss in life expectancy at all [30]. Another prospective study of 8,907 subjects with lengthy follow-up reported that ever-smokers with a median pack-year exposure of at least 21.3 who had quit smoking in the last five years reduced their lung cancer risk by 39.1% when compared to current smokers [43]. A prospective study of over one million UK women found that smoking cessation before the age of 40 protects the individual from more than 90% of the excess mortality associated with continued smoking [44].

1.5.2 Age

As with most cancer types, lung cancer risk increases with age. This is due to a confluence of factors including accumulated risk factor exposure, genetic and epigenetic mutations, and cellular or biochemical changes that promote carcinogenesis [45,46]. In the UK, 44% of cases occur in people aged over 75. The highest rate of lung cancer occurs between the ages of 80-84 years in women and 85-89 years in men. An average of 1,140 lung cancer cases are diagnosed in people below the age of 50 per year, less than 3% of the total incidence [5]. Consequently, screening eligibility criteria usually limit lung cancer screening to those in older age brackets, with the US Preventive Services Task Force (USPSTF) recommending screening to those aged 50-80 [47], and the NHS Targeted Lung Health Check (TLHC) Programme inviting participants aged 55-74 [48]. Age also negatively impacts prognosis [49]. In the UK, five-year net survival decreases progressively from 42% in 15-39 year-olds to 6% in 80-99 year-olds in men, and from 48% to 7% in women [5].

1.5.3 Asbestos, Occupational Hazards and Radon

Occupational exposure to carcinogenic agents is estimated to be responsible for approximately 13% of lung cancer cases in the UK [24,50]. Asbestos exposure is the main contributing factor; it is estimated that between 6% and 8% of annual UK lung cancer deaths are linked to this exposure [51]. Asbestos is a generic term referring to six naturally occurring silicate minerals which are fireproof, insulating, and soundproof, properties which positioned it to be used extensively in industry and construction, particularly in the late-19th and 20th centuries [52]. All forms of asbestos are carcinogenic. When asbestos is inhaled, the microscopic fibres instigate programmed cell necrosis as well as the release of HMGB1, a protein that triggers a chronic inflammatory response, which in turn promotes cellular transformation and carcinogenesis [53]. It appears that the association between asbestos exposure and lung cancer can be characterised as a linear dose-response relationship [54]. It has also been reported that the risk of lung cancer begins to increase at very low levels of asbestos exposure and excess risk can persist for more than 40 years after exposure [55]. There is evidence that asbestos and tobacco smoke work synergistically to increase lung cancer risk [54–56].

Air pollution, specifically particulate matter smaller than 2.5µm ($PM_{2.5}$), is considered by the World Health Organisation (WHO) to be an extremely significant source of environmental carcinogens [57]. Several recent meta-analyses demonstrated a significant link between $PM_{2.5}$ exposure and lung cancer incidence and mortality [58–60]. Diesel exhaust is an important contributor to $PM_{2.5}$ pollution and increases lung cancer risk [61]. Silica dust exposure may also increase lung cancer risk [62].

Radon is a colourless, odourless, tasteless gas emitted by naturally occurring uranium and thorium as they decay. Unusually high concentrations of the gas can accumulate in indoor spaces. Radon can also contaminate building materials and drinking water [63]. When radon penetrates the lungs, ionising alpha radiation that is emitted can cause DNA damage. This can contribute to the development of lung cancer [64]. Radon exposure is considered the second leading cause of lung cancer by the World Health Organisation (WHO), with between 3% and 14% of lung cancer cases linked to the gas, although risk of exposure is heavily dependent on geography [65].

1.5.4 Socio-economic Deprivation

Socio-economic deprivation is associated with higher lung cancer risk and poorer survival [66–68]. In England, socio-economic status is measured using the index of multiple deprivation (IMD). This is a metric that ranks small geographical areas of England by relative poverty, from 1 (most deprived) to 32,844 (least deprived), by combining and weighting data relating to income, employment, education, health, crime, housing and living environment [69]. Lung cancer incidence rates are 170% higher in the most deprived IMD quintile compared to the least deprived; if the whole UK population had the age-specific crude incidence rates of the least deprived quintile, it is estimated that there would be 14,000 fewer annual lung cancer cases [70].

A large multi-national meta-analysis found that men with lower educational attainment have a higher lung cancer mortality rate than those with higher attainment in ten separate populations, across all age groups [71]. Similar differences exist between manual and non-manual workers [72], and between women with varying numbers of factors associated with socio-economic deprivation [73]. This disparity is principally due to higher exposure to tobacco smoke in more deprived communities [74]. In addition to higher primary smoking rates, second-hand smoke exposure and prenatal tobacco exposure are also higher in more deprived communities (53,54). Furthermore, lower socio-economic status is associated with lower smoking quit rates. One study of over 900 British female smokers found that those who stayed in school past the age of 16 had a smoking cessation rate of 22% in one year, compared to 11% amongst women who left school before the age of 16. Previous studies have shown that only 15% of adult smokers with the lowest socio-economic status were able to successfully quit smoking, compared to 60% amongst the most affluent [77].

There is some evidence that deprivation may be independently associated with lung cancer, even when smoke exposure is controlled for. One study reported an independent association between deprivation and both lung cancer incidence and mortality, even when adjusted for tobacco abuse (a binary variable comprising of a loosely linked collection of smoking related disorders, used here as a surrogate for detailed smoking data. The binary and poorly defined nature of this variable may weaken the conclusions of this study) [78]. Another study demonstrated that even when smoking history is controlled for, people living in more deprived areas suffer from poorer lung function [79]. Several factors may drive this association, including unequal access to healthcare, isolation from health-promoting environments, and increased air pollution [67,68,78].

1.5.5 Sex

Lung cancer incidence in the UK is currently higher amongst men than women. This is primarily caused by historic smoking patterns; whilst the difference has been shrinking in recent decades, smoking prevalence has always been higher in men than women [80]. When smoking history is controlled for, there is evidence that women are at higher risk of lung cancer than men. A case-control study from 1993 reported that when only examining subjects with a 40-pack-year history, women have approximately three times the likelihood of developing lung cancer compared to men [81]. Similar patterns, albeit with lower magnitudes of differences, were reported in other studies [82]. A Swiss study demonstrated that women with lung cancer are significantly younger and have smoked considerably less than their equivalent male counterparts, a pattern consistently found in other populations [83]. Furthermore, lung cancer risk is higher amongst female never-smokers than male never-smokers [84]. A recent American study concluded that the higher incidence of lung cancer amongst women cannot be explained by smoking behaviour alone [85], leading some studies to establish female sex as an independent risk factor for lung cancer [82,86–88]. However, the large Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial (PLCO) study did not find an independent link between female sex and lung cancer risk [89].

The mechanism responsible for possible increased susceptibility to lung cancer amongst women is not clear. It is likely that there are biological, hormonal, and genetic contributing factors [90]. For example, oestrogen may play an important role in the development of lung cancer by promoting angiogenesis and the formation of DNA adducts [91,92]. Additionally, studies have shown that women are more likely to carry tumour promoting *KRAS* and *p53* mutations which may work in cooperation with oestrogen to encourage cancer growth [93,94].

1.5.6 Other Health Conditions

Lung cancer risk is increased in individuals with other respiratory diseases. It is often difficult to establish if this relationship is causal due to detection bias, reverse causality and misdiagnosis. Chronic obstructive pulmonary disease (COPD) is an umbrella term for a range of conditions including chronic bronchitis and emphysema. Symptoms associated with COPD include persistent coughing, breathlessness, wheezing and chest infections. COPD is diagnosed by measuring airflow obstruction using a spirometer. Forced expiratory volume in the first second of forced breath

(FEV_1) and forced vital capacity (FVC) – the total volume of air in a forced exhalation – are measured; the ratio between the two values is referred to as the FEV_1/FVC ratio. A ratio below 0.7 is indicative of airflow obstruction and the possibility of COPD [95]. COPD develops gradually as the body responds to inhaled irritants, launching a chronic inflammatory response which eventually leads to scarring, the breakdown of lung tissue and the narrowing of airways [96]. It is this inflammation that may catalyse the development of lung cancer [97]. A large meta-analysis found that COPD confers an increased risk of lung cancer, even when adjusting for smoking history [98]. Further studies confirm that COPD is an independent risk factor for lung cancer [99,100].

Pneumonia is an inflammatory condition usually caused by a viral or bacterial infection. A meta-analysis of twenty-two studies found strong evidence that an individual with a previous diagnosis of pneumonia is at increased risk of lung cancer [98]. However, the strongest association between pneumonia and lung cancer comes when there is a very short time period between the diagnosis of the two diseases, indicating that misdiagnosis and reverse causality may be sources of bias [101]. Furthermore, some studies have claimed that pneumonia actually has a protective effect against lung cancer and an individual's risk of lung cancer decreases with an increase in pneumonia diagnoses [102]. Previous diagnosis of tuberculosis, a bacteria borne respiratory disease, is associated with increased lung cancer risk, although the nature of the causality is not clearly established [103].

Previous diagnosis of a malignant tumour increases the risk of development of second primary lung cancer. Contributors to this association include exposure to radiation during cancer treatment [104], genetic predisposition, and exposure to shared risk factors [105]. Patients treated for head and neck tumours are particularly susceptible to lung cancer [106,107]. There is evidence that being overweight or obese is protective against lung cancer, even amongst never-smokers. Research to determine the biological mechanism behind this association is ongoing [108,109].

1.5.7 Family History

Family history of lung cancer is an important predictor of personal risk. Having a first-degree relative diagnosed with lung cancer, multiple relatives diagnosed, or relatives with early-onset disease, all increase an individual's risk [110,111]. There is evidence that this increased risk is conferred by both shared environment as well as genetic factors. A pooled analysis of over 24,000 lung cancer cases found that those with a family history of lung cancer had an increased risk of approximately 50%, even after controlling for smoking and other confounding environmental factors. The highest increased risk was conferred to individuals with a sibling diagnosed with lung cancer, even after controlling for tobacco exposure [112]. It should be noted that sampling bias may affect the case-control studies included in this pooled analysis. It is unclear whether the magnitude of the effect reported would be reproduced at a population level. Despite this limitation, the implication of a significant familial element to increased lung cancer risk, with a potential genetic contribution, remains strongly supported by this and other studies.

A large Icelandic study found that spouses of lung cancer patients have a 1.75-fold increased risk of lung cancer, indicating that shared environment is an important factor in the development of lung cancer. The same study demonstrated that first-degree relatives had a greater risk, up to a 3.5-fold increase. The exact nature of the interaction between environmental and genetic mechanisms to increase risk is not elucidated [113]. A multicentre study found that the risk of lung cancer increases with family history of the disease even amongst non-smoking women [114]. Similarly, another study showed that non-smoking relatives of never-smoker lung cancer patients have a higher risk of contracting the disease when compared to controls, even though tobacco smoke did not contribute [115]. In summary, whilst there is certainly an environmental aspect to increased lung cancer risk derived from family history of the disease, there is considerable evidence of an important genetic contribution as well.

1.5.8 Genetic Risk Factors

A large prospective twin based study estimated that the overall heritability of lung cancer is 18% [116]. Heritability refers to the limit of genetic risk stratification on a population level and individuals may have a much higher level of genetically conferred lung cancer risk [117]. Genetic risk may be of particular importance in early onset lung cancer [118], as well as for individuals with multiple primary lung cancers [119]. A study of 230 never-smokers with lung cancer found that 18% had family history of the disease, and a large proportion had specific genetic pathogenic variants that increase an individual’s susceptibility to developing lung cancer [120]. Considerable research has taken place in recent decades to establish the nature of the genetic component of lung cancer risk.

1.5.8.1 Monogenic Variants

There is limited evidence that pathogenic variants in a single gene confer risk for lung cancer. Li Fraumeni Syndrome, a disorder arising from germline *TP53* variants, is a notable exception which increases risk for several cancers, including lung [121–123]. An association between lung cancer risk and several rare inherited *EGFR* variants has also been reported [124]; these germline variants might cause genetic instability which predisposes cells to somatic mutations and tumorigenesis [125]. For example, the T790M variant is both a germline mutation and an important somatic variant with implications for therapy [126,127].

Other potential risk alleles have been identified through segregation analysis of families with high lung cancer incidence [128–130]. A linkage analysis of 52 high-risk families identified chr6q as a locus containing an inherited high-penetrance allele significantly associated with lung cancer risk. A further study published in 2010 confirmed the association of this chromosomal region with increased lung cancer risk, even in never smokers [131]; fine mapping identified gene *RGS17* as a candidate for familial lung cancer susceptibility [132]. Whilst *RGS17* overexpression has been shown to aid tumour cell proliferation, it has not been convincingly proven as a lung cancer susceptibility gene [133]. A study published in 2015 demonstrated that a high-penetrance missense mutation in the *YAP1* oncogene significantly increases the risk of lung cancer [134]. Another reported association was with the c.823C>T (p.Arg275Trp) missense variant in *PARK2* [135]. However, given its low allele frequency in gnomAD (<0.002) and its lack of subsequent validation, it appears unlikely to be a high-risk allele [136].

1.5.8.2 Polygenic Variants

Considering the limited evidence for a monogenic inheritance pattern for lung cancer risk, research into high-frequency, low-penetrance risk alleles has become a more promising endeavour. Genome-wide association studies (GWAS), in which millions of single nucleotide polymorphisms (SNPs) are genotyped in a very large case-control cohort, have yielded hundreds of variants potentially associated with lung cancer risk. In a GWAS, genome wide significance for an allele is usually established with a P-value of less than 5×10^{-8}; odds ratios (OR) can then be calculated for the identified SNPs, indicating the increased or decreased likelihood of lung cancer with the presence of that particular variant.

Since 2008, through many thousands of GWAS, more than 45 genetic loci have been associated with lung cancer risk, although the strength of evidence varies in each case [137]. A large 2017 meta-analysis synthesised evidence from more than 1000 publications published until 2015, examining 246 SNPs from 138 loci. It concluded that 22 variants in 21 genes showed significant association with lung cancer with robust cumulative epidemiological evidence. Epidemiological evidence was graded by the Venice Criteria, a score which considers the amount of evidence (based on sample size and number of studies), replication of association (homogeneity of effect and magnitude across multiple studies), and protection from bias (based on several common areas of bias in genetic studies) [138]. It also reported significant heterogeneity between the SNPs associated with various subgroups, including ethnicity, lung cancer histology and smoking status [139]. A large number of similar meta-analyses have been published in recent years [140].

A 2017 review aimed to assess and summarise the evidence for lung cancer associated SNPs from more than 200 separate GWAS and meta-analyses, all published up to 2016 with at least 1000 cases [140]. The study concluded that 137 variants were associated with lung cancer, 80 of which were statistically significant. SNPs derived from meta-analyses were graded for strength of evidence using Venice Criteria and false positive report probability (FPRP) [141]; of the variants derived from the meta-analyses, 15 SNPs were graded as ‘strong’ for evidence of association and 19 SNPs were graded as ‘moderate’. This review did not weigh and synthesise the evidence for each SNP as a formal meta-analysis would have; when there was conflicting evidence from different studies, the evidence from the largest study was treated as authoritative. Nevertheless, this study serves as an important summary of the SNPs likely to exhibit robust association with lung cancer.

A further large case-control study was published in 2017, in which an aggregated dataset of 20,266 cases and 56,450 controls was analysed, resulting in the discovery of 10 novel SNPs associated with lung cancer and the confirmation of 8 SNPs previously reported. The study claims to identify the SNPs responsible for 12.3% of the additional familial relative risk of lung cancer [142].

SNPs associated with traits or disease risk are often ethnicity specific. In European populations, a significant volume of lung cancer risk variants are localised to several gene clusters:

***CHRNA*:** Expression of this gene, located in the 15q25 chromosomal region, has been found to contribute to cancer cell signalling, proliferation, angiogenesis and inhibition of apoptosis [143]. Additionally, studies have identified *CHRNA5* as having a role in nicotine addiction and dependency [144]. Several variants linked to lung cancer are located on this gene cluster. For example, AA risk genotype at rs16969968 in *CHRNA5* is associated with both an increased risk and earlier diagnosis of lung cancer [145]. Several studies have demonstrated that increased lung cancer risk is an independent association related to SNPs in this gene [146–148]. SNP rs1051730 in the *CHRNA* gene is a variant with significantly robust association with lung cancer risk in European populations [140].

***CLPTM1L*:** The *CLPTM1L* gene is located in the 5p15 chromosomal region. Two variants (rs401681 and rs402710) on this gene are particularly strongly associated with increased lung cancer risk [149]. The gene segment containing these polymorphisms may regulate telomerase reverse transcriptase (TERT) expression, allowing cells to resist apoptosis and become malignant [150].

***BAT3*:** The *BAT3* gene is located in the 6p21 chromosomal region. The protein product of this gene cluster has been shown to be crucial in p53 acetylation during the repair or apoptosis of damaged, potentially malignant, cells. BAT3 may also be released in response to stress signals, engaging natural killer cells to target tumour cells [151].

A polygenic risk score (PRS) is a measure of an individual’s genetic risk of developing a specific disease, based on the combination of multiple low-penetrance SNPs with known association to that disease. Whilst each SNP has minimal impact on disease risk in isolation, when combined they can alter risk significantly. There are several methodological approaches to building a PRS tool; fundamentally, the sum of the number of risk alleles present in a particular individual is calculated and weighted according to the variant's strength of association with the condition in question, as measured by the odds ratio (OR) [152].

PRS tools have been shown to be useful in risk prediction in a wide variety of cancer types and other disease areas. There is evidence that a PRS can be used to reduce overdiagnosis in prostate cancer screening programmes [153], as well as facilitate the stratification of colorectal cancer screening by risk [154]. Use of a PRS has also been proposed for the identification of individuals at increased risk of cardiovascular disease [155] and Alzheimer's disease [156]. A study of more than 81,000 individuals published in November 2019 demonstrated that polygenic and monogenic risk factors interact with each other to modify risk in breast cancer, coronary artery disease and colon cancer [157]. A selection of polygenic variants can influence the level of penetrance of the monogenic risk factor; consequently, a PRS can be used to predict the level of increased risk conferred by the monogenic risk variant carried by the individual. The development of a validated PRS for breast cancer is particularly advanced, with a 313 SNP PRS demonstrated to be efficacious in breast cancer risk prediction. In prospective validation, the PRS area under the curve (AUC) was 0.63 and it was well calibrated; those in the top 1% of PRS scores had between 2.78 and 4.37-fold increased risk compared to the middle quintile, and those in the bottom 1% had 0.16 to 0.27-fold risk when compared to the middle group [158].

1.6 Lung Cancer Screening

1.6.1 Principles of Cancer Screening

A crucial strategy for diagnosing cancer earlier, thereby increasing the chances of curative treatment and improving patient outcomes, is systematically screening asymptomatic people at risk of the disease. Cancer screening can be population based, in which all individuals of a certain sex and age in a population are offered screening (such as in the NHS cervical, breast and colon cancer screening programmes [159]), or targeted, in which more precise tools are used to direct screening to those most at risk of the disease (such as NHS England's Targeted Lung Health Check (TLHC) programme for lung cancer screening [48]).

In 1968, *Wilson* and *Jungner* published ten guiding principles to establish the appropriateness of pursuing organised screening for a condition (Box 1) [160]. In subsequent years, attempts have been made to refine and modernise the principles to account for knowledge gained from half a century of further screening research [161,162], although the original principles remain remarkably enduring. One systematic review and consensus study developed twelve screening principles, eight of which aligned with and further developed principles from *Wilson* and *Jungner's* original set, with the novel principles focussing on the benefits, harms, quality, management and ethics of the screening programme as a clinical system (rather than just the core test or intervention) [163].

Box 1. Wilson and Jungner's Principles of Screening [160]

1. The condition sought should be an important health problem.

2. There should be an accepted treatment for patients with recognized disease.

3. Facilities for diagnosis and treatment should be available.

4. There should be a recognizable latent or early symptomatic stage.

5. There should be a suitable test or examination.

6. The test should be acceptable to the population.

7. The natural history of the condition, including development from latent to declared disease, should be adequately understood.

8. There should be an agreed policy on whom to treat as patients.

9. The cost of case-finding (including diagnosis and treatment of patients diagnosed) should be economically balanced in relation to possible expenditure on medical care as a whole.

10. Case-finding should be a continuing process and not a 'once and for all' project.

1.6.2 Assessing Screening Performance

Despite the relatively straightforward theory behind cancer screening, it can be challenging to establish whether systematically applying a specific screening test to a population would provide the necessary benefits to justify its implementation.

First, test validity needs to be established. There are several possible outcomes of a screening test. A true positive or true negative refers to the test accurately detecting the presence or absence of cancer respectively. A false positive means that the test indicated a possibility of cancer, which upon further investigation is not found to be concerning. Conversely, a false negative means that the test reported no concerning findings and missed the presence of cancer. The performance of a screening test can be assessed by examining its sensitivity (proportion of true positives who are screen positive) and specificity (proportion of true negatives who are screen negative). These metrics can be combined into a receiver-operating characteristic (ROC) curve; the area under the curve (AUC) indicates the test's overall ability to discriminate successfully between cases and non-cases (an AUC of 0.5 means that the test is no better than random at discriminating between cases and non-cases; 1 means that the test can discriminate perfectly, with no false-negatives or false-positives) [164]. Positive predictive value (PPV) and negative predictive value (NPV), which indicate the proportion of screen positives/negatives who were true positives/negatives, are also useful performance metrics [165].

Once test validity is established, actual effectiveness of applying the test in a population must be assessed. This can be challenging, as several sources of bias can affect screening, obscuring the true impact of the intervention. Lead-time bias occurs when screening diagnoses cancer early, but the patient still dies from that disease and does not live longer than they would have in the absence of screening. In such a situation, it can appear that screening has increased survival time, but this is only because the disease was identified earlier (Figure 3). Length-time bias occurs as a result of the increased chance of screening detecting cancers that are not causing symptoms and are therefore less aggressive and less likely to cause death. This bias also makes it seem that screening has increased survival time, when it is in-fact a result of the less aggressive nature of the tumours detected. Overdiagnosis bias refers to a situation in which the cancer detected by screening would never have been diagnosed in the patient’s lifetime, due to low aggressiveness of the cancer, competing comorbid conditions, or short baseline life expectancy (Figure 3) [166].

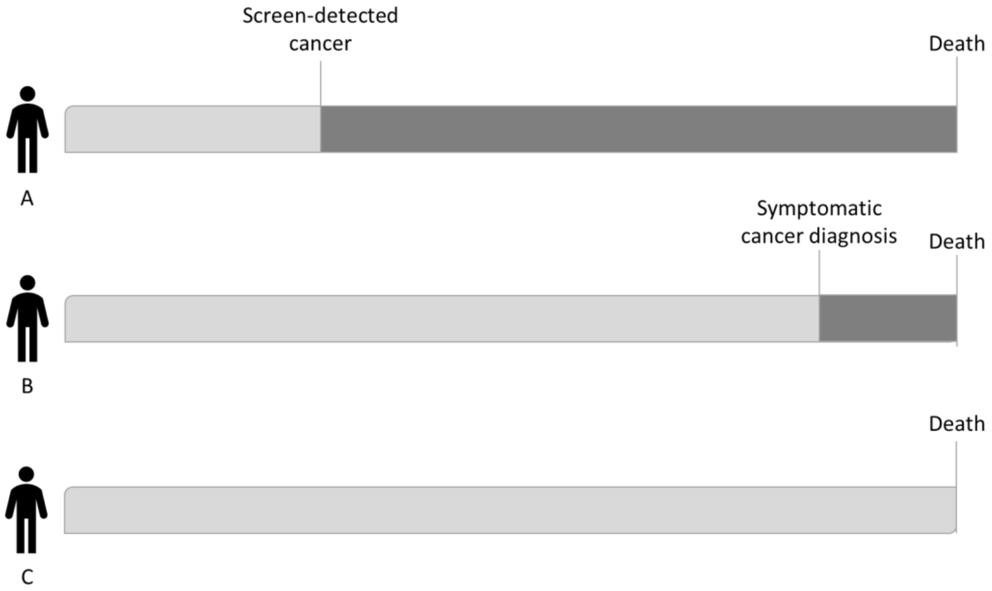


Figure 3. Biases that affect screening trials. A = pathway of a screening participant. Demonstrates lead-time bias if non-screened pathway would be B. Demonstrates overdiagnosis bias if non-screen pathway would be C.

To investigate the performance of screening tests in the population, single arm observational studies can be conducted. In addition to the measures of test validity described above, population-based studies can assess stage of cancer at diagnosis and survival time following detection to indicate whether cancer diagnoses are being made at an earlier stage than would be expected under regular conditions. However, this study design is particularly susceptible to lead-time, length-time and overdiagnosis biases [167]. Consequently, whilst results from single-arm observational studies can provide useful indications of a potentially efficacious screening intervention, it does not provide evidence as to whether a real-world screening programme would provide actual benefit to the population.

To provide more robust evidence, prospective randomised controlled trials (RCT) are necessary. In RCTs, participants are randomised into intervention and control groups, with a comparison of outcomes between the two groups indicating the impact of the intervention. The primary outcome of interest in screening RCTs is usually cancer-specific mortality; if there is a lower rate of cancer death in the intervention group after extended follow-up, it is presumed that this mortality benefit can be attributed to screening [165,167]. This metric reflects the fundamental objective of implementing screening: reducing the number of people who die prematurely from the disease of interest. However, this outcome measure is liable to bias relating to lead-time and overdiagnosis; it cannot demonstrate if overall life expectancy increases due to the screening intervention [165,167,168].

Some very large RCTs may be able to demonstrate changes in all-cause (rather than cancer-specific) mortality. This metric is not biased by subjective cause of death classification, takes harms caused by the intervention into account, and shows that the cancer-specific mortality reduction due to screening is not wholly eliminated by competing causes of death. However, a modelling study found that RCTs need to be extremely large (40,000-600,000 participants per arm) and long running (11-20 years) to have the possibility of demonstrating significant reduction in all-cause mortality [168]. Consequently, some consider all-cause mortality too stringent an outcome measure for the purposes of assessing cancer screening utility [169,170].

Whilst RCT design minimises the effect of the biases which affect single-arm studies, all trial designs, (including RCTs) are vulnerable to ‘selection bias’ (also known as ‘healthy volunteer bias’) in which the trial participants are healthier and more affluent than the general population [171]. For example, an analysis of a large lung cancer screening RCT found that those who had cancer diagnosed were less comorbid and had significantly better post-operative outcomes than would be expected in the general population [172]. The expertise of the healthcare staff running the trial may also be more specialised than the equivalent practitioners in the wider community. This can lead to overly optimistic study results, highlighting the importance of real-world studies of screening in the target population to ensure benefits are also observable in a non-trial setting.

1.6.3 Harms of Screening

It is important to consider that screening, as with any health intervention, has the potential to cause harm to participants. Even if a screening test demonstrates good performance, an analysis of the harms and benefits must be conducted to ensure that implementing a screening programme will provide net benefit to the population.

There are several potential harms associated with screening. At the most fundamental level, participants may be exposed to potential harm from the screening test itself, whether radiation from CT scans [173], x-rays or mammography [174], or adverse reactions to blood, cervical, or faecal sampling. Investigations for suspicious screening findings such as physical examinations, biopsies, colonoscopies, ultrasounds and colposcopies can also cause harm [175].

Overdiagnosis occurs when disease that has no clinical significance for the patient is diagnosed through screening. This can result in the patient undergoing invasive and unnecessary treatment to cure a cancer that would not have shortened their life or impaired its quality [176]. The mechanics of screening may exacerbate the issue of overdiagnosis, as by definition, checking asymptomatic people for cancer at semi-regular intervals is more likely to detect slow growing, less aggressive cancers. Aggressive, fast growing cancers are more likely to cause symptoms and be diagnosed in a clinical setting [177]. In some cases, cancer treatment can result in complications that shorten the patient's life, potentially more significantly than the cancer would have [178]. Overdiagnosis can be quantified and monitored through extended follow-up of RCTs, modelling studies, pathological and imaging studies, or cohort studies. Each method has strengths and limitations [179]. To mitigate the potential harm of overdiagnosis, evidence for screening benefit should be robust before screening is implemented. Once screening is implemented, methods to further reduce harm from overdiagnosis include selecting an optimum screening frequency, targeting screening to appropriate populations, using multiple screening modalities and additional biomarkers, and implementing carefully considered diagnostic and treatment pathways (including options such as watchful waiting and active surveillance) [180].

Another screening related harm is false-positive results, in which a screening attendee is referred for further investigation which does not result in an eventual cancer diagnosis. This can cause anxiety and other negative psychosocial impacts [181–183] and lead to invasive and unnecessary investigations that may cause further harm. Conversely, false-negative results can give screening participants undue confidence that leads them to ignore symptoms of the disease, reduce their engagement with healthcare professionals and make negative health-related choices. This can ultimately lead to poorer prognosis when the disease is correctly diagnosed [184]. Screening tests can also result in indeterminate results, in which repeat screening, follow-up screening, or further investigations are necessary to confirm if there is cause for concern. There is some evidence of psychosocial harm in participants who receive an indeterminate screening result [185,186].

1.6.4 Evidence for Lung Cancer Screening

The difficulty in diagnosing lung cancer at an early stage, coupled with the high mortality rate that results from late diagnosis, positions lung cancer screening to be a crucial strategy in reducing deaths from lung cancer. If individuals with lung cancer who have not yet developed symptoms can be identified, it could provide them with the best possible chance of having curative treatment at an early stage.

Two methods of lung cancer screening trialled in the latter half of the 20th century were chest x-ray (CXR) and sputum cytology [187,188]. The PLCO study (an RCT that ran from 1993 to 2001) randomised 154,901 men and women, age 55-74, into an intervention arm that received four annual rounds of CXR, and a control arm. It confirmed that CXR conferred no lung cancer mortality reduction nor any evident stage shift in cancers after 13 years of follow-up [189]. A Cochrane meta-analysis published in 2004 concluded that neither sputum cytology or CXR resulted in benefit to patients [190].

Low dose computed tomography (LDCT) began emerging as a superior alternative in the 1990s. In contrast to a traditional CXR which produces a single plane, two-dimensional image, LDCT employs a rotating x-ray tube to take a series of images from different angles, which are then combined to produce a detailed three-dimensional representation of the lungs and surrounding tissue. A Japanese study of 1,369 high-risk individuals produced the first evidence for the effectiveness of LDCT for lung cancer detection, identifying 15 cases of lung cancer, 11 of which had been missed by CXR [191]. Further evidence accrued with the International Early Lung Cancer Action Project (I-ELCAP), an international multi-centre LDCT programme running until 2005 in which 31,567 high-risk individuals were screened with LDCT (27,456 had repeat screening one year after baseline). The programme identified 484 lung cancers, 85% of which were diagnosed at stage I [192].

The first conclusive evidence that LDCT screening reduces lung cancer mortality was provided by the National Lung Screening Trial (NLST). This trial, based in the United States, randomised 53,454 current or former (within 15 years) smokers, age 55-74 with at least 30 pack-year smoking history, to either LDCT or CXR screening for three annual rounds over two years. CXR was selected as the intervention in the control arm to complement the PLCO trial that was already comparing CXR vs. standard of care (no screening). After six years of follow-up, the LDCT arm had a 20% reduction in lung cancer-specific mortality (Figure 4) and a 6.7% reduction in all-cause mortality compared to the CXR arm. Half of the lung cancer diagnosed in the LDCT arm were stage I, compared to 31% in the CXR arm [193].

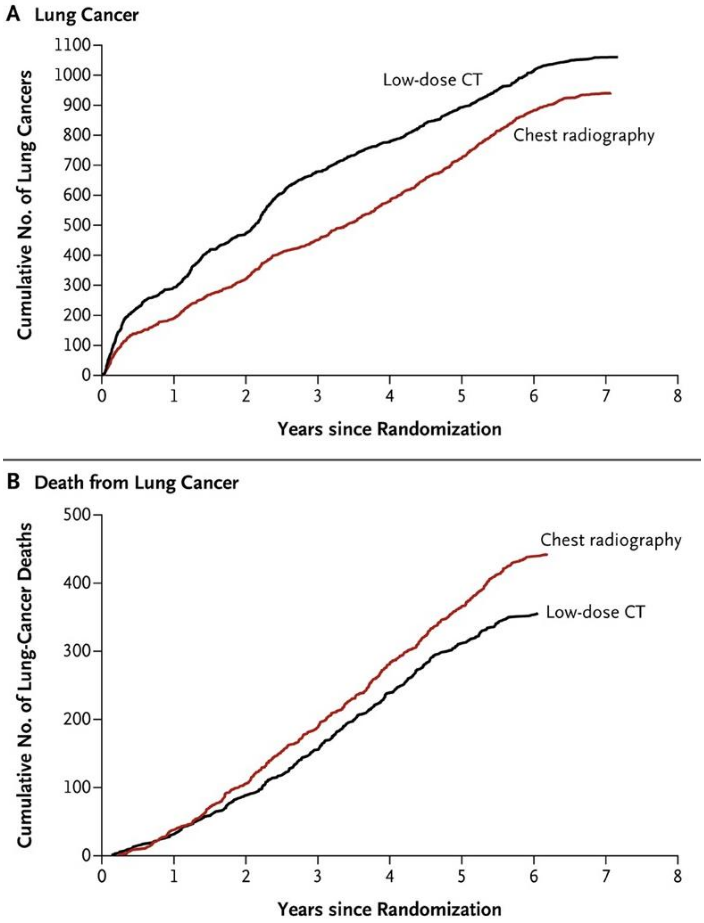


Figure 4. Lung cancer incidence (A) and deaths (B) in NLST cohort. Reproduced with permission from [193], Copyright: Massachusetts Medical Society.

The Dutch-Belgian Randomised Lung Cancer Screening Trial (NELSON) confirmed the findings from NLST in a European population. 13,195 men and 2,594 women were randomised into two arms, one of which received four rounds of LDCT over 5.5 years with intervals of increasing length (1, 2 and 2.5 years), the other of which received no screening. The primary outcome was lung cancer mortality reduction in men. A smaller, sub-analysis was conducted in the female subset; only 16% of NELSON participants were women, ostensibly due to lower female smoking prevalence and thus lower eligibility at the time of trial recruitment. Therefore, a small sample of high-risk women were proactively approached to participate. All participants were current or former smokers who had smoked within 10 years and had a tobacco exposure of either ≥15 cigarettes per day for 25 years or ≥10 cigarettes per day for 30 years. After ten years of follow-up, the trial reported a 26% reduction in lung cancer-specific mortality in men and a 33% reduction in women due to LDCT screening (Figure 5). There is lower confidence around the female mortality reduction due to the smaller sample size in that subgroup. Almost 60% of screen-detected lung cancers were early-stage, compared to less than 15% of non-screen detected lung cancers [194].

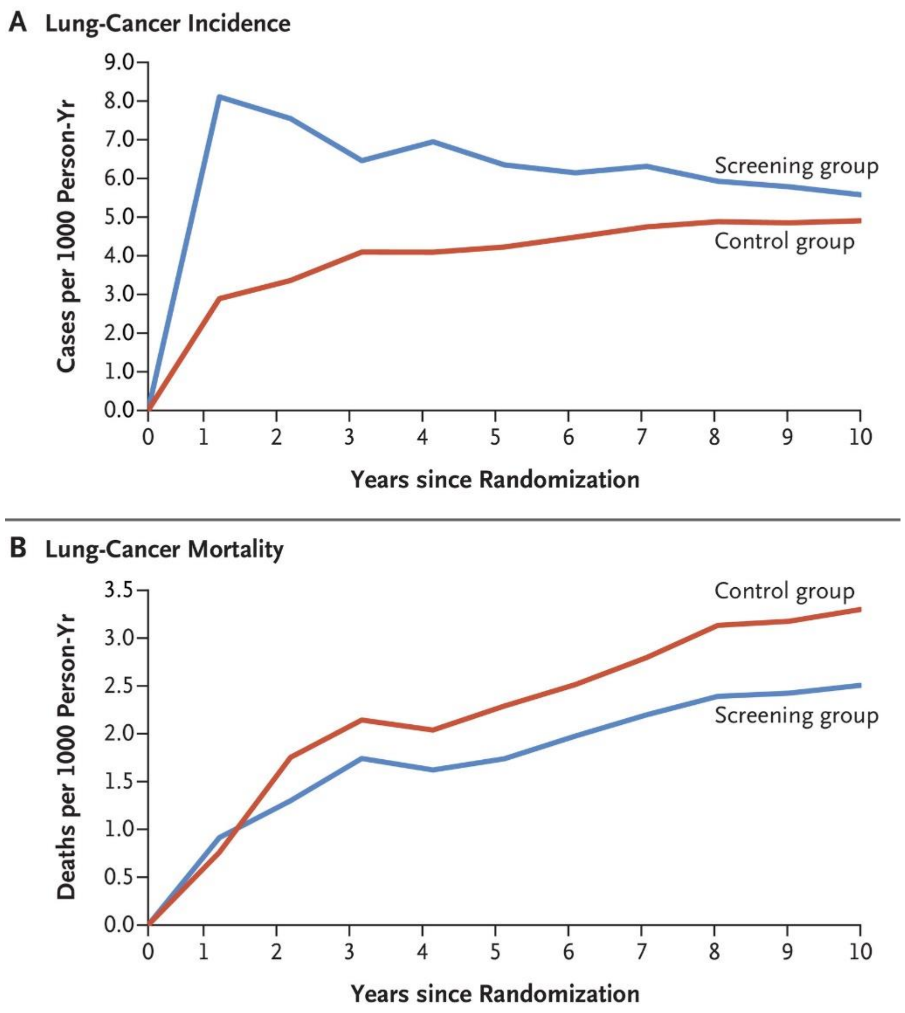


Figure 5. Lung cancer incidence (A) and mortality (B) among male participants of the NELSON trial. Reproduced with permission from [194], Copyright: Massachusetts Medical Society.

Other, smaller, studies have reported similar results, although these were often not powered to find statistically significant differences in mortality when viewed independently. The ITALUNG trial randomised 3206 ever-smoking (minimum 20 pack-years in last 10 years) participants aged 55-69 into an intervention arm that received four rounds of annual LDCT screening and a control arm with standard care. The study reported borderline significant reductions in lung cancer (30%; p=0.07) and overall mortality (17%; p=0.08) due to LDCT screening [195].The Danish Lung Cancer Screening Trial, which randomised 4104 participants (age 50-70, minimum 20 pack years) to have five annual LDCT scans or no screening, also reported nonsignificant trends, although post-hoc analysis in the high-risk subgroup showed results that aligned with the NLST [196]. The German LUSI trial randomised 4052 participants to five annual rounds of LDCT screening or standard care, and reported a statistically significant reduction in mortality in women, but not in men [197]. The Multicentric Italian Lung Detection (MILD) trial reported a 39% reduction in lung cancer mortality due to LDCT screening over 10-years [198]. A meta-analysis of nine RCTs (several of which were not powered to show statistically significant mortality reductions in isolation) found a 16%

relative reduction in lung cancer mortality and a small reduction in overall mortality in the screening arm [199].

1.6.5 Harms from Lung Cancer Screening

Whilst evidence for the benefits of LDCT screening for lung cancer continues to accumulate, it is important to consider how the various categories of screening related harms may apply in this context.

Lung cancer screening trials have been significantly affected by overdiagnosis in the past; one historic CXR and sputum cytology trial had a likely overdiagnosis rate of 51% [200]. However, more recent LDCT trials have shown smaller, albeit still notable, overdiagnosis rates. The overdiagnosis rate for NLST was estimated by dividing the difference in lung cancer cases between the arms by the number of screen-detected cases at several follow-up points after the final screening round. Using this method, the initial estimation of overdiagnosis rate (calculated 4.5 years after the final screening round) was 18%. However, with an extended follow-up period of 12 years, the estimated rate reduced to 3% as the large divergence in lung cancer incidence initially observed between the two arms narrowed considerably [201]. The NELSON trial reported an estimated overdiagnosis rate of 19.7% at 4.5 years after the final screening round, reducing to 8.9% with a single additional year of follow-up [194]. The Danish Lung Cancer Screening Trial reported a much higher estimated overdiagnosis rate of 67% at five years after the final screening round [202]. This elevated rate may be explained by the higher baseline lung cancer risk in the intervention group than in the control group, the high frequency of LDCT screening in this trial, or the standard of care approach applied in the control group (compared to CXR in NLST) [203]. This stands in stark contrast to the 0% overdiagnosis rate reported by the ITALUNG trial in the same follow-up period [195]. Treatment for lung cancer after detection by screening is not risk free; for example, of 250 individuals who had lung cancer detected by screening in a variety of UK programmes, five (2%) had treatments or investigations that resulted in major complications, of which two were post-operative deaths [204]. Even if no major complications occur, treatments can cause a reduction in quality of life [205]. The higher the overdiagnosis rate, the more likely it is that patients are undergoing potentially harmful treatment for a tumour that would not have impacted their life had it not been detected through screening.

False positive results are also a concern in lung cancer screening programmes. NLST reported that 23% of the LDCT scans were false positive, equivalent to 96% of the total positive LDCT results [193]. NELSON reported a much lower false positive rate, at 1.2% of total LDCT scans or 57% of all positive LDCT results. The addition of an ‘indeterminate’ classification and surveillance screening in NELSON (before findings were deemed positive) was a significant contributing factor to this reduction in false positives [206]; this is because borderline findings could be monitored and treated based on growth and volume, rather than a single time-point decision based on nodule diameter. As well as the psychological harms caused by false-positive results, clinical investigations assessing positive LDCT screening results have the potential to cause physical harm. One meta-analysis reported that CT guided core biopsies, a common diagnostic procedure performed to investigate potential lung cancer tumours, resulted in an almost 40% complication rate and 5.6% major complication rate [207]. Adverse events such as these are particularly significant when occurring in individuals who are found not to have cancer. They can significantly reduce the net benefit provided to that patient by undergoing screening.

There are several secondary potential screening related harms in lung cancer screening. Radiation exposure from LDCT scans must be considered; the increased risk is minor in most instances, although there may be cause for concern in populations at lower risk of lung cancer and with repeated scanning [208]. There may also be psychological impacts of participating in screening, whatever the eventual screening result, although this has been shown to be of limited clinical significance [209,210]. The number of pulmonary nodules defined as ‘indeterminate’ in screening can also be of concern, as these require additional LDCT scans (and therefore exposure to potential harm) and may result in anxiety in the interim period between scans. In NELSON, 9.2% of scans had an indeterminate result [194].

Several approaches can be taken to mitigate screening-related harms. One approach is to implement robust nodule management processes, carefully considered informed decision making, and appropriate screening intervals. These approaches seek to lower the risk of harm, whatever the characteristics of the screening cohort. An alternative (and complementary) approach is to shift the harm-to-benefit ratio of the cohort overall by selecting screening participants who are most at risk of lung cancer and/or most likely to benefit from screening, thereby reducing the weight of screening-related harms when considered in combination with the potential benefits. This was demonstrated in a recent study that examined screening related harms by aggregating data from five UK lung cancer screening programmes, representing more than 11,000 screened individuals. It found an overall false positive rate of 2% (n=219/10,898), comparable to NELSON, and much lower than NLST. More than 50% of those with a positive screening result went on to have lung cancer confirmed, much higher than both NLST (4%) and NELSON (38%). Of attendees who were not diagnosed with lung cancer, 0.6% underwent an invasive diagnostic test, although this did not lead to any complications or deaths. Approximately 5% of surgeries undertaken were for benign lesions, lower than the 24.4% seen in NLST. Risk-based screening selection in the UK programmes was identified as a crucial factor responsible for the lower level of harm observed when compared to NELSON and NLST [204].

The improvement in treatments and the development of our understanding of lung cancer in recent decades, as well as the ongoing accumulation of evidence relating to the efficacy of LDCT screening for early detection, would appear to satisfy criteria 1-5 and 7 in *Wilson* and *Jungner's* principles of screening (Box 1). The remaining principles relate to screening implementation, ensuring that the intervention is acceptable, cost-effective, and appropriately targeted to minimise harms while maximising benefits. This is a crucial area of ongoing research to actualise the potential of LDCT screening for lung cancer. Developing the optimal method for selecting individuals for screening is a key facet in ensuring that lung cancer screening provides benefit and meets all the screening principles.

1.6.6 Selection Methods

1.6.6.1 Generalised Eligibility Criteria

In NLST and NELSON, selection of participants for screening was achieved by employing generalised eligibility criteria, a basic threshold of age and smoke exposure (the two main risk factors for lung cancer). NLST enrolled participants aged 55-74, with a smoking history of ≥30 pack years who had smoked within 15 years of study entry. NELSON offered screening to those who had smoked within 10 years and had an exposure of either ≥15 cigarettes per day for 25 years or ≥10 cigarettes per day for 30 years. The 2013 screening guidelines from the US Preventive Services Task Force (USPSTF) broadly matched the NLST eligibility criteria [211]; in 2020, updated and expanded USPSTF guidelines were published, recommending annual lung cancer screening be offered to those aged 50-80 with a ≥20 pack-year history, who have smoked in the past 15 years [47].

Whilst aiming to select a screening cohort that has a high-risk of lung cancer, thereby ensuring a favourable lung cancer detection rate and benefit-to-harm ratio for participants, generalised eligibility criteria have limitations. The rudimentary nature of the criteria results in a cohort with a heterogenous mix of risk profiles. One study of NLST found that when the trial population was stratified by individualised lung cancer risk, 5276 participants were screened to prevent one lung cancer death in the lowest risk group, compared to 161 in the highest. Overall, only 1% of lung cancer deaths were prevented in the lowest risk quintile of the trial [212]. Furthermore, it is estimated that up to 45% of ex-smokers who develop lung cancer do so more than 15 years after quitting [213]. A recent meta-analysis confirmed that ever-smokers maintain an increased level of lung cancer risk well after 15 years since quitting [214]; other studies have demonstrated increased lung cancer incidence even 25 years after quitting [43]. Former smokers of more than 15 years are currently excluded from screening by most guidelines. Substituting generalised eligibility criteria with individualised risk prediction could be an effective method for increasing the risk profile of the screening cohort to maximise programme efficiency and benefit to participants [215,216].

1.6.6.2 Risk-Based Selection

Risk prediction models (RPMs) are tools that use multiple risk factors to calculate an individual's specific, personal risk of developing lung cancer over a particular period. In addition to increasing the overall risk-profile of the screening cohort, by taking several lung cancer risk factors into account, RPMs avoid assigning screening eligibility based solely on smoke exposure and age. Whilst age is an accurate predictor of lung cancer, actual lung cancer risk varies considerably due to other risk factors [217]. Similarly, individuals who are deemed ineligible for screening by generalised criteria due to extended quit-periods may still be at significant increased risk of lung cancer [43,214].

At least 20 RPMs have been developed for lung cancer thus far [218]. Two important RPMs in the context of lung cancer screening in the UK are $PLCO_{M2012}$ and LLP_{V2}. $PLCO_{M2012}$ is a logistic regression model predicting 6-year lung cancer risk [89]. It was built using the disease incidence data of more than 80,000 smokers and has been externally validated in several large trials [219] (see Table 1 for factors included in the RPM). One study showed that in a group of 37,332 ever-smokers, the $PLCO_{M2012}$ RPM selected 81 more people with lung cancer for screening than the NLST inclusion criteria would have [89]. When compared to the $USPSTF_{2013}$ generalised eligibility criteria, $PLCO_{M2012}$ identified 12.4% more lung cancers whilst screening 8.8% fewer individuals [220].

The Liverpool Lung Project (LLP) RPM predicts lung cancer risk over 5-years [221]. LLP was developed in a case-control study of 579 lung cancer cases and 1157 controls, all resident in Liverpool; a 10-fold cross validation within the study resulted in an AUC of 0.7 [87]. External validation in a further three case-control cohorts confirmed the efficacy of the RPM [222]. Observations in the UK Lung Screening Trial (UKLS), in which the RPM was applied prospectively, resulted in the adaptation of the RPM to LLP_{V2} [223] (see Table 1 for factors included in the RPM), with a further model (LLP_{V3}) published in 2020, having been recalibrated to country-wide cancer incidence data [224].

Table 1. Risk factors included in PLCO$_{m2012}$ and LLP$_{v2}$ risk prediction models. Below delineation = factors unique to RPM.

PLCO$_{M2012}$	**LLP$_{v2}$**
Age	Age
Smoking duration	Smoking duration
Previous cancer diagnosis	Previous cancer diagnosis
Family history of lung cancer	Family history of lung cancer
COPD diagnosis	Pneumonia/ emphysema/bronchitis/tuberculosis/COPD diagnosis
Smoking status (current vs. former)	Sex
Cigs per day	Asbestos exposure
Smoking quit time	Family history of early onset lung cancer
Ethnicity	
Education	
BMI	

An important study published in 2021 evaluated eight RPMs in three large UK cohorts (UK Biobank, EPIC-UK, and Generations Study), totalling more than 270,000 participants in which there were 1474 incident lung cancer cases, and 826 lung cancer deaths [225]. The RPMs evaluated were: PLCO$_{M2012}$ [89], LLP [221], LLP$_{V2}$ [223], LLP$_{V3}$ [224], Bach [226], Hoggart [227], LCRAT [228], and LCDRAT [228] (which predicts risk of death from lung cancer). In the study, all the RPMs overestimated the number of lung cancer cases or deaths in these cohorts, with LLP$_{V3}$ and PLCO$_{M2012}$ having the best calibration, and LLP$_{V2}$ having the worst. AUC ranged from 0.77 for LLP$_{V2}$ in the UK Biobank cohort, to 0.84 for Bach and LCDRAT in the EPIC-UK and Generations cohorts respectively [225].

The study found that had USPSTF$_{2013}$ guidelines been used to select participants for screening, 51% of lung cancer cases would have been eligible. At their respective recommended thresholds, LCDRAT and LCRAT would have selected 61% of cases, PLCO$_{M2012}$ and Bach would have selected 58%, LLP$_{V3}$ would have selected 57%, and LLP$_{V2}$ would have selected 54%. Had the expanded USPSTF$_{2020}$ criteria been used to determine screening eligibility, it would have selected 26% of the total cohort, including 66% of lung cancer cases. The 26% highest risk individuals by each RPM would have contained the following proportion of total cases: LCDRAT and Bach – 77%, PLCO$_{M2012}$ and LCRAT – 75%. LLP$_{V3}$ – 71%, and LLP$_{V2}$ – 70%. All RPM thresholds resulting in these figures were below 1%, apart from LLP_{V2} which was 1.3% [225]. The research cohorts examined were less socio-economically disadvantaged than the general UK population, indicating that model calibration may be better when applied in a real-world setting. However, this also implies that less deprived individuals are more likely to have higher lung cancer risk scores (and thereby be more likely to be deemed eligible for screening) when compared to more deprived participants; this could result in the exacerbation of health inequalities [229,230].

There is a considerable need for further retrospective and prospective analyses of lung cancer RPMs in actual screening programmes to examine how employing different models impacts on the overall performance of the programme. Selecting the most effective RPM, as well as ensuring it is calibrated appropriately for the population it is being used in, are key elements in ensuring the efficiency and success of the screening programme, as well as limiting potential harms to participants.

Whilst RPM-based selection may have considerable advantages over generalised eligibility criteria, there are limitations. One study demonstrated that whilst RPMs may have prevented more deaths from lung cancer than generalised eligibility criteria, they yielded fewer life-years per-death prevented and contributed to higher overdiagnosis rates [231]. This is primarily because they selected older and more comorbid participants for screening, who were more at risk of lung-cancer, but had less to gain from having cancer diagnosed early. The International Lung Screening Trial (ILST) was the first study to publish a prospective comparison of the performance of $USPSTF_{2013}$ criteria and $PLCO_{M2012}$ for screening selection in almost 6000 participants. Interim analysis found that $PLCO_{M2012}$ was a more efficient selection method, resulting in a higher lung cancer detection rate than the generalised eligibility criteria (92-95% vs. 76-79% of cancers detected depending on threshold; $p=0.0001$. 3.6% vs. 3% positive predictive value; $p=0.11$). However, it also selected older and more comorbid individuals with shorter life expectancy estimations (based on several predictors including age, sex, BMI, comorbidities, and smoking history). It should be noted that cumulative estimated life expectancy amongst those diagnosed with lung cancer was higher in the $PLCO_{M2012}$ group (2249 vs. 2001 years; $p=0.015$) [232].

1.6.6.3 Benefit-Based Selection

Benefit-based screening selection is an alternative method of selecting participants for screening. It aims to capitalise on the increased screening efficiency derived from RPMs, without selecting an older, more comorbid cohort likely to receive limited benefit from early cancer detection. Life-gained prediction frameworks seek to incorporate individually calculated lung cancer risk, as well as projected benefits from screening, into a single metric. This metric can be used to select a screening cohort with an adequate level of lung cancer risk to warrant screening, as well as a long enough life expectancy to derive benefit from the intervention.

The Life Years Gained from Screening-CT (LYFS-CT) model generates a score indicating an individual's projected days-of-life gained from screening. It does so by calculating life expectancy using an overall mortality model developed and validated in two subsets of ever-smokers in the National Health Interview Survey (NHIS) cohort (>100,000 subjects in total). Predictors of overall mortality include age, ethnicity, gender, BMI, smoking history, and co-morbid medical conditions such as diabetes, stroke and hypertension. Expected mortality benefit from screening (20.4% based on NLST) is then used to adjust life expectancy, with the difference between life expectancy with 3-annual rounds of LDCT screening vs. without constituting the LYFS-CT score [233]. It is important to note that the 'days of life gained' metric 'spreads' the expected screening benefit across the population. For example, an individual with 2% chance of developing lung cancer based on an RPM calculation, and 20 days of life-expectancy gained from undergoing screening based on LYFS-CT, has a 98% chance of gaining no days of life (if lung cancer is not detected) and a 2% chance of gaining 1000 days of life (if lung cancer is detected).

When retrospectively applied to a test cohort of 28,458 ever-smokers (a representative NHIS sample of the approximately 60,712,710 ever-smokers resident in the USA [234]), life-gained based selection resulted in fewer lung cancer deaths prevented, but more life-years gained from screening in the population (overall, per-detected case and per-prevented death), when compared to RPM based selection. An individual selected by LYFS-CT was more likely to be younger, female, African-American, a current smoker and have fewer comorbidities, with a moderately high risk of developing lung cancer. Interestingly, USPSTF generalised eligibility criteria outperformed both life-gained and RPM-based selection for life years gained per-prevented death, although it underperformed when it came to total number of deaths averted and life years gained counted independently [233]. The American College of Chest Physicians recommends the use of the LYFS-CT model to inform screening selection [235].

To our knowledge, no study has tested benefit-based selection in an actual lung cancer screening cohort. Considering the impact benefit-based selection may have on reducing harms and increasing benefits for screening participants, further research is required to compare the performance of benefit-based and risk-based screening selection in real-world screening programme settings.

1.6.6.4 Polygenic Risk Scores

Whilst some RPMs and life-gained selection tools consider family history of lung cancer when predicting risk, no currently used model includes a direct measure of genetic risk. Despite this, there is evidence that a PRS of low-penetrance SNPs could aid in lung cancer risk prediction and screening selection.

A 2009 study demonstrated that integrating a 20 SNP PRS into an RPM for lung cancer increased its predictive ability when compared to standard risk factors alone [236]. However, the model was not externally validated in an independent population and had certain non-standardised study design elements [237]. Two other studies demonstrated that the inclusion of specific genetic markers resulted in modest improvements in predictive ability [238–240]. All three of these studies preceded the large GWAS and meta-analyses published in the last few years that have provided the best evidence for which SNPs are most robustly associated with lung cancer risk. Therefore, whilst serving as an important proof of concept, they are of limited clinical utility. More recent case-control studies again demonstrated improvement in risk prediction conferred by the SNPs, albeit modest. The limited magnitude of effect was likely a product of the small SNP panels tested; successful PRS systems rely on the combination of a very large number of independent SNPs from a range of loci [241].

An important demonstration of an effective PRS for lung cancer risk prediction came in 2019, with the publication of a large Chinese study which reported the development of a 19 SNP PRS which had been prospectively validated in a cohort of more than 95,000 individuals. The study compared the PRS tool's risk prediction when compared to age and pack-year history alone. The top 10% of the cohort for genetic risk were 2.96 times more likely to develop lung cancer than the bottom 10%. Heavy smokers with intermediate genetic risk had similar levels of risk to light smokers with high genetic risk. Light smokers with low genetic risk had similar overall risk to non-smokers [242]. It should be noted that the PRS developed is specific to a Chinese population. It also did not test the PRS in an actual screening programme, nor did it compare the risk prediction against RPMs, which have superior risk prediction to generalised eligibility criteria (and it is therefore more challenging to improve risk prediction further with the addition of a PRS).

A 2020 study published in *Nature Communications* systematically assessed the value of adding a PRS to RPMs for risk prediction across 16 cancer types. It tested the tools in a cohort of 413,870 UK Biobank subjects, 22,755 of which were incident cancers. The study reported that the RPM for lung cancer resulted in exceptionally high levels of calibration, discrimination and predictive ability, based on conventional risk factors alone without inclusion of a PRS. Adding the PRS did result in a modest improvement in predictive value, but this was limited by the success of the standard RPM. However, non-event net reclassification index (the proportion of non-cases with risk classified downwards due to the PRS) was very high, indicating that a PRS might have utility in directing screening away from those who are less likely to develop cancer. Lung cancer risk prediction in non-smokers could be much improved by use of a PRS, as RPMs rely on risk factors that non-smokers are not exposed to [243].

A further study published in March 2021 used a machine-learning approach to construct a PRS for lung cancer based on the OncoArray data of more than 23,000 individuals (approximately 13,000 cases and 10,000 controls) which was then validated in more than 335,000 individuals sourced from the UK Biobank (of which 1,768 were cases). The PRS was associated with lung cancer risk in a dose-response relationship, although it had limited impact on improving discrimination over existing risk models using standard risk factors. However, the study did find that integrating a PRS into risk prediction could have a significant impact on the age at which an individual reaches the absolute risk threshold to be eligible for screening. For example, the average age at which ever-smokers with no family history of lung cancer would be eligible for screening was 61, whilst those who were in the top 1% of PRS would become eligible at age 56. Current smokers with a family history of lung cancer, and who were in the top 10% for genetic risk, reached eligibility before the age of 50, considerably earlier than current screening guidelines would deem them eligible. Smoking cessation reduced lung cancer risk significantly, irrespective of PRS decile. Never-smokers did not hit the screening threshold, irrespective of genetic risk [244].

Further research is required to develop, validate, and optimise PRS tools for lung cancer risk prediction. Validating PRSs in actual screening cohorts, in a variety of populations, and integrating them into existing RPMs, are areas of particular importance in investigating whether these tools have the potential to improve screening selection in a real-world lung cancer screening programme.

1.7 Lung Cancer Screening Implementation

Since the publication of the NLST and NELSON trial results confirming the lung cancer mortality benefit derived from screening, attention has turned to developing strategies for the effective implementation of lung cancer screening. There are several practical prerequisites for an effective screening programme relating to infrastructure, organisation, funding, and governance. In addition to these general areas of concern, four challenges have been identified specifically relevant to the implementation of lung cancer screening: developing individually tailored invitations to screening, optimising individualised lung cancer risk assessment, investigating individualised screening intervals, and integrating comorbidity reducing strategies into screening programmes. Each category has several subsidiary questions that must be addressed to ensure a successful screening programme [245].

A small RCT that had significant implications for screening implementation was the UK Lung Screening Trial (UKLS), which took place in 2011. It randomised 4055 participants into screening and control arms. Among the 1994 scanned participants, 42 lung cancers were detected, 86% were stage I or stage II at diagnosis [223]. Long-term follow-up found a lung cancer mortality reduction in the screening arm of the trial (relative rate 0.65, 95%CI 0.41-1.02; p=0.062) [199].

UKLS was the first screening trial to prospectively employ an RPM for determining participant eligibility; it used the LLP_{V2} RPM to stratify individuals, those with a score of ≥5% were eligible to participate. The study demonstrated that lung cancer screening, employing an RPM for cohort selection, could be cost-effective in the NHS at £8466 per quality adjusted life year (approximately one-fifth of the equivalent cost in NLST). However, the invitation strategy (postal questionnaires) had limited effectiveness; positive response rate was 31%, with only 3.5% of those respondents deemed eligible for screening based on risk calculation [223].

Results from trials such as UKLS have several important implications for screening implementation. However, a trial setting is fundamentally distinct from a clinical setting, particularly in relation to how representative the trial cohort is of actual individuals at high-risk of lung cancer [246,247]. In order to develop a model of lung cancer screening that could be practically implemented in real-world populations, community-based screening pilots and programmes are required.

1.7.1 Community-Based Screening

1.7.1.1 The Manchester Lung Health Check Pilot

The Manchester Lung Health Check (LHC) pilot was a pioneering, community-based screening programme with significant implications for widescale screening implementation. Much of the research in this thesis utilises the Manchester LHC pilot cohort. The LHC pilot ran in 2016-2017 and was designed as a 'one-stop' Lung Health Check located in easily accessible community locations in three socio-economically deprived areas of Manchester. This approach aimed to increase convenience for attendees and reduce barriers to participation amongst those most at risk of lung cancer [248]. As well as ensuring the overall success of the screening programme, reducing barriers to participation (particularly amongst high-risk deprived populations) is crucial to ensuring screening programmes do not exacerbate health inequalities. Screening has the potential to worsen health inequalities by improving outcomes in less disadvantaged people who are more likely to attend screening, whilst neglecting high-risk, more disadvantaged people who do not attend [229,230].

A schematic of the LHC process is presented in Figure 6. In the LHC pilot, 'ever-smokers' aged 55-74 from three deprived areas of Manchester were invited to attend mobile LHC units in easily accessible community locations such as supermarket and stadia carparks (Figure 7). Through nurse-guided completion of a questionnaire, each participant's 6-year risk of lung cancer was calculated using the $PLCO_{M2012}$ RPM; those with a score of ≥1.51% were offered annual screening over two rounds, with the first LDCT scan taking place immediately in an adjacent unit. LDCT images were examined by consultant radiologists and were classed as 'negative', 'positive' or 'indeterminate'. Those with a positive scan were assessed in a lung cancer clinic and managed in accordance with national guidelines. Participants with an indeterminate scan returned after three months for a surveillance scan, whilst people with negative scans returned a year later for a second round of screening. In line with the holistic approach of the LHC model, participants also underwent spirometry to assess airflow obstruction and detect undiagnosed COPD, had QRISK scores calculated (a model used to quantify cardiovascular risk [249]), and were offered smoking cessation advice if they were a current smoker [248,250].

Overall, 2541 participants attended the LHC pilot, most of whom were in the lowest deprivation quintile in England. 1429 (56.2%) were eligible for screening, of whom 4.4% were diagnosed with lung cancer, a significantly higher detection rate than the 1.7% seen in the first two rounds of NLST and 1.6% in the first two rounds of NELSON. More than 80% of the detected cancers were early stage; 63% underwent surgical resection [248,250].

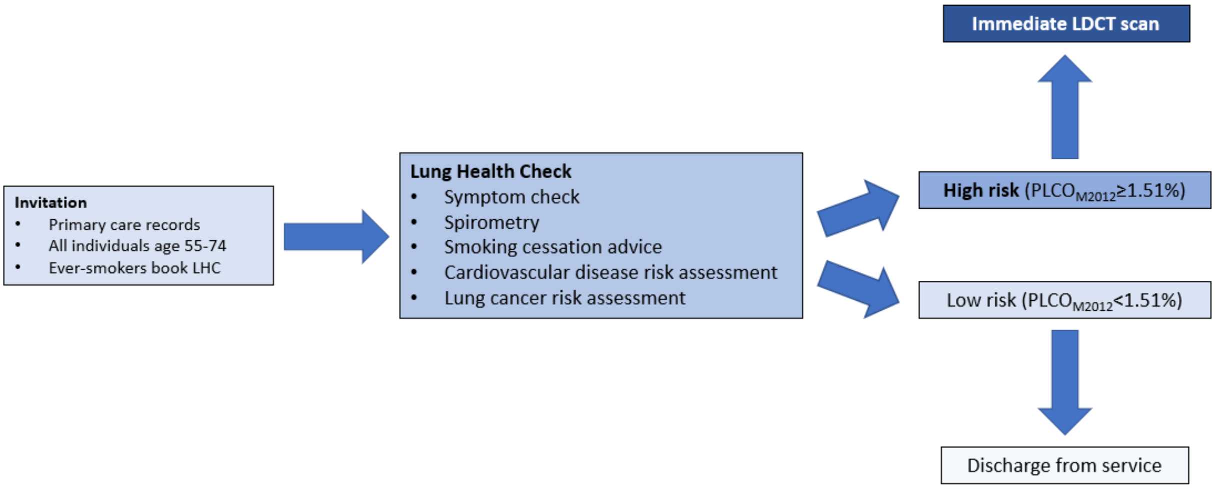

Immediate LDCT scan
Invitation
Primary care records
All individuals age 55-74
Ever-smokers book LHC
Lung Health Check
Symptom check
Spirometry
Smoking cessation advice
Cardiovascular disease risk assessment
Lung cancer risk assessment
High risk ($\text{PLCO}_{M2012} \geq 1.51\%$)
Low risk ($\text{PLCO}_{M2012} < 1.51\%$)
Discharge from service

Figure 6. Participant journey through LHC programme.

Alliance Medical
Mobile CT Scanner Unit
Optima

Figure 7. Images from the Manchester LHC pilot [251].

1.7.1.2 Targeted Lung Health Check Programme

The success of the Manchester LHC pilot led to the development of the NHS Targeted Lung Health Check (TLHC) programme, in which the LHC model is being rolled out to dozens of further sites [48]. It is currently operating at 23 sites in England, with a further expansion expected in the near-future [252]. In March 2022, the UK National Screening Committee (UK NSC) published an interim report proposing that screening high-risk individuals for lung cancer, using the TLHC model, should be recommended nationwide and is likely to be cost-effective for the NHS [253].

Further research into optimising screening implementation is taking place embedded in, or parallel to, the TLHC programme. The Yorkshire Lung Screening Trial (YLST), which is operating within a community based LHC setting, will be the first trial to prospectively evaluate the performance of three screening selection methods (USPSTF, $PLCO_{2012}$, and LLP_{V2}). A cohort-based study previously compared these three approaches and found $PLCO_{M2012}$ to be best performing in several respects, but analysis of real-world performance is essential [225]. YLST will also investigate the effectiveness of fully integrated smoking cessation services within the LHC and the cost-effectiveness of the screening programme [254].

Another implementation trial which took place within the context of an LHC programme is the Lung Screen Uptake Trial (LSUT). It tested whether an innovative leaflet seeking to address psychological barriers for screening and increase uptake would be an effective intervention. The trial found that the intervention did not improve uptake, but confirmed that the LHC approach resulted in much better participation than previous trials and studies, particularly in socio-economically disadvantaged communities [255].

1.7.2 Risk Perception and Risk Communication

With the continued development of lung cancer screening programmes as clinical services, and the likelihood that personalised risk calculation will be used to assess screening eligibility in these services, understanding participants’ perception of lung cancer risk (whether it be population, personal, or factors that increase risk) is becoming a crucial nascent area of implementation research.

Shared decision making is a process in which the clinician provides the patient with the necessary evidence and treatment options available to them and supports the patient to make informed decisions for their care. This protects patient autonomy, self-determination, and agency [256]. It has been argued that shared decision making provides an extension to the ethical framework of informed consent [256], although some conceptual and practical challenges remain in its implementation [257,258].

A crucial step in facilitating shared decision making is seeking to understand what the patient already knows about their medical situation, correcting incorrect assumptions, and providing high quality information to allow for informed deliberations and decisions. This is particularly important in the context of an intervention such as risk-based lung cancer screening, which may be preference-sensitive in patients who are older or have significant comorbidities [259]. USPSTF and Centers for Medicare & Medicaid Services (CMS) both emphasise the importance of shared decision making in lung cancer screening implementation [211,260].

Before communicating risk and explaining options for appropriate interventions, it is important to understand the patient's baseline risk perception for the disease in question. This creates an opportunity to correct inaccurate risk perception, ensuring the participant has an appropriate framework for considering the information relating to their actual risk and interventions. Several factors have been implicated in risk perception formation. These include numeracy, beliefs about the preventability and severity of the disease in question, general mental health and worry, previous exposure to the disease, and demographic factors such as age, sex, education and BMI [261–264].

Inaccurate risk perception is common in all disease areas. A 1993 study of women in breast cancer family history clinics found that only 44% of women estimated their personal risk to within 50% of their actual risk. The study also found that women at high risk of the disease had an inaccurate perception of the population risk of breast cancer, with only 11% estimating this correctly [265]. A further study in 2016 found that 66% of low-risk women overestimate their risk of breast cancer, and the majority of both high and low risk women estimated the population risk of the disease incorrectly [266].

A 2008 systematic review of 61 studies from across disease areas concluded that the majority of people (particularly those with family history of the disease) overestimate their risk of cancer. There was considerable evidence for poor population risk recall, and mixed evidence as to the association between anxiety and risk perception [267]. A study of coronary heart disease, breast cancer, colorectal cancer and diabetes mellitus risk perception found that high-risk patients usually underestimate risk, whilst low-risk patients overestimate risk. The study also found that BMI was the demographic factor most commonly associated with inaccurate risk perception [268].

Several studies have examined lung cancer risk perception. One study evaluated risk perception in a subset of the NLST screening cohort. It found that personal risk estimates were higher amongst current smokers than in former smokers. Unlike in other studies, they did not find an optimistic risk perception bias amongst smokers (possibly due to those seeking out screening being less likely to underestimate personal risk). Worry was also linked to risk perception [269]. Further studies have reported that lung cancer risk perception is influenced by cigarette smoking and family history of the disease [270,271]. There is also evidence that both smokers and non-smokers overestimate lung cancer 5-year survival rates [272].

A study published in 2019 examined risk perceptions in a small cohort of 70 NLST eligible screening participants. The study found that there was a particularly pessimistic risk perception amongst the participants, with the mean lifetime risk estimate being 52.1%. Following shared decision making counselling in which their actual calculated risk scores were provided, this was reduced to a mean of 32.8%. Those at higher objective risk of lung cancer had a smaller reduction in risk perception post-counselling. Interestingly, whilst risk perception decreased, interest in screening increased post-counselling [273].

A cross-sectional study published in 2021 was the first to compare actual lung cancer risk and perceived risk in attendees of a screening programme. The study surveyed 2,514 participants of the PanCan study, all current or former smokers with ≥2% risk of developing lung cancer over 6-years. It reported that actual and perceived comparative risk perception were positively aligned. Younger age, higher pack-year history, being a former smoker, having a family history of lung cancer, respiratory symptoms and history of COPD were also associated with higher risk perception. No clear relationship between lung cancer-specific worry and perceived risk was found [274].

In addition to its importance in shared decision making, risk perception may be a determinant of health behaviours [275]. For example, there is some evidence that risk perception and lung cancer worry were positively associated with intention to quit smoking [274]. Risk perception can also mediate feelings of anxiety and general quality of life [268]. There is evidence that risk perception is an important factor in individual’s willingness to attend screening [276–278]. Being aware of the characteristics of those with an overly optimistic or pessimistic risk perception could facilitate the targeting of screening related messaging to the subset of the population who would benefit most [279].

There are several frameworks for risk conceptualisation, useful for understanding a participant’s perceived risk and for facilitating effective risk communication. Two primary approaches are to use absolute risk, a numerical estimation of an individual’s risk of a disease in a particular timeframe, or comparative risk, where you present an individual’s risk in the context of a larger population or group [280]. There is mixed evidence as to which approach is most effective for risk communication [281]. Visual aids representing risk have been shown to assist participants in their understanding of risk information [281]. There has been limited development of visual aids, standardised communication protocols, and other communication tools for interacting with lung cancer screening participants [282]. Two infographics were created and published presenting baseline outcomes and screening related harms for people undergoing lung cancer screening in the UK [204], but this is not directly applicable to communicating a person’s individual lung cancer risk and screening eligibility. An online tool is available that allows individuals to calculate their personal lung cancer risk; it presents risk information in a variety of formats, including in absolute and comparative terms [283]. More research is required to investigate how socio-economically disadvantaged participants of community-based LHC programmes perceive their risk, and which tools are effective for risk communication and shared decision making.

1.8 Aims and Objectives

Aim

The overall aim of this project is to improve the selection of individuals in socio-economically disadvantaged communities for lung cancer screening.

Hypotheses

- Current lung cancer screening selection strategies perform sub-optimally in socio-economically disadvantaged populations.
- Benefit-based screening selection reduces the comorbidity profile of the cohort selected for screening when compared to risk-based selection.
- Assessing the genetic profile of ever-smokers improves prediction of lung cancer risk.
- Risk perception accuracy is low in socio-economically disadvantaged lung cancer screening participants.

Objectives

I. To analyse the performance of several risk-based methods of cohort selection in a community-based screening programme, including generalised eligibility criteria and two RPMs.

II. To investigate the impact of applying a life-gained selection strategy in a community-based lung cancer screening programme, assessing whether this approach may contribute to improving lung cancer screening selection in this population.

III. To validate previously published polygenic risk score tools, investigating whether reported genetic risk factors for lung cancer are present in the Manchester population and whether they could be used to improve lung cancer risk prediction.

IV. To explore novel genetic signals emergent from the Manchester case-control dataset.

V. To assess lung cancer risk perception, disease knowledge and lung cancer-specific worry in attendees of a community-based lung cancer screening programme, with implications for effective risk communication and shared decision making in this setting.

Chapter Two - Methods and Protocols

2.1 Manchester LHC Pilot Cohort

The results and analyses described in Chapters 3-5 are derived from the Manchester LHC pilot cohort (introduced in section 1.7.1.1) [248,250]. The Manchester LHC pilot took place in 2016-2017, located in Harpurhey, Gorton and Wythenshawe, three socio-economically deprived areas of Manchester. Participants were invited through their registered GP practices. All 96 GP practices in North, Central and South Manchester Clinical Commissioning Groups (CCG) were approached to participate, 20 practices expressed an interest, and 15 were selected based on proximity to the LHC pilot sites (one practice withdrew prior to pilot commencement) [284]. Ever-smokers aged 55-74 were invited to attend the LHC, which took place in mobile units at easily accessible community locations.

Participants completed a questionnaire under the guidance of a specialist nurse, facilitating the calculation of an individualised 6-year risk estimate of lung cancer using the $PLCO_{M2012}$ RPM. Those with a score of ≥1.51% were offered annual screening over two rounds. LDCT images were examined by consultant radiologists and were classed as 'negative', 'positive' or 'indeterminate' for lung cancer. Pulmonary nodules were managed in accordance with British Thoracic Society guidelines [285]. The radiologist also subjectively classified the level of coronary artery calcification (CAC) for each scanned individual (none, mild, moderate or severe). The baseline LDCT images were also used to determine the presence or absence of emphysema.

In addition to lung cancer risk calculation, all participants of the LHC pilot underwent spirometry (pre-bronchodilator) using a desktop spirometer (Vitalograph® ALPHA). Forced Expiratory Volume in one second (FEV_1), Forced Vital Capacity (FVC), respective predictive values, and FEV_1/FVC ratio were recorded. Participants completed a questionnaire related to respiratory symptoms. Respondents were classed as 'symptomatic' if they reported one or more of: breathlessness (≥2 Medical Research Council (MRS) dyspnoea scale [286]), sputum production (≥teaspoon per day), and/or cough (≥6 weeks in duration), in accordance with the Global Initiative for Chronic Obstructive Lung Disease (GOLD) recommendations [95]. Cardiovascular disease history (myocardial infarction, stroke, angina or transient ischaemic attack) was collected in attendees of the second screening round.

The $PLCO_{M2012}$ RPM predicts lung cancer risk based on an array of risk factor and demographic variables (Table 1) [89]. In the LHC pilot, UK educational levels were adapted from the US reference in accordance with the International Standard Classification of Education [287]. Additional data including asbestos exposure, early-onset familial lung cancer, history of respiratory disease (pneumonia, tuberculosis, bronchitis and COPD) were collected to calculate 5-

year lung cancer risk using the LLP_{v2} RPM (Table 1) [221], although this score was not used to determine screening eligibility in the pilot. LLP_{V2} and $PLCO_{M2012}$ have been shown to perform differently in validation studies [225]. Consequently, it is important to note that having prospective data based solely on $PLCO_{M2012}$ somewhat limits the post-hoc analysis contained in this thesis; specific impacts for each analysis are discussed in the individual results chapters.

LHC pilot participants provided written informed consent to participate in research (REC Ref: 17/EE/0092). All data were stored on an ethically approved Microsoft Access database (REC Ref: 16/NW/0013).

2.2 Manchester LHC Pilot Follow-up Data Collection

For the analysis in Chapter Four, I collected data on outcomes, mortality, and comorbidity diagnoses for all LHC pilot participants who had screen-detected lung cancer, as well as for a subset of cancer-free individuals, at a fixed time-point approximately 5-years after the pilot took place. GP, hospital and NHS records were used to ascertain alive/dead status, cause of death, age of death or current age, details of cancer staging, treatment, complications and recurrence (for those diagnosed with lung cancer), and details of newly developed comorbidities (in a subset of the follow-up cohort).

2.3 Life-Gained Calculations

For the analysis in Chapter Four, I used the *lcmodels* R package (v.4.0.3) [288] to recalculate $PLCO_{M2012}$ risk scores and calculate LYFS-CT life-gained scores (and associated life expectancy projections) for the whole LHC pilot cohort (LYFS-CT introduced in section 1.6.6.3). LYFS-CT uses 23 variables in its calculation of expected life-gained from screening. Eleven are the same as that of the Lung Cancer Death Risk Assessment Tool (LCDRAT) model: age, gender, ethnicity, education, BMI, years smoked, years quit, cigarettes per day, pack-years, lung disease, pneumonia, family history of lung cancer (binary and numeric) [228]. The remaining variables are additional comorbidities: hypertension, coronary heart disease, angina, heart attack, other heart disease, stroke, diabetes, chronic bronchitis, kidney failure, liver disease, health condition requiring special equipment, and year of assessment.

2.4 Case-Control Cohort Recruitment and Sample Acquisition

Samples used for the case-control study described in Chapter Five were sourced from two Manchester-based resources. The control cohort comprised of whole blood samples provided by attendees of the second round of the Manchester LHC pilot (described in section 2.1) who were screen-negative for lung cancer. Blood was aseptically extracted by research nurses on the screening truck and stored at -80 °C within 48 hours. Written, informed consent was provided by all participants.

The case cohort comprised of whole blood samples from patients with histologically confirmed NSCLC, sourced from the Manchester Cancer Research Centre (MCRC) Biobank. All case samples originated from patients who had undergone surgical resection of lung cancer between 2010 and 2018. Blood was aseptically extracted by biobank technicians and immediately collected in EDTA tubes. After extraction, samples were stored temporarily on dry ice and transferred to -80 °C freezers in the biobank no later than 48 hours after extraction. Samples were transferred from the biobank to the University of Manchester laboratories on dry ice and then stored in -80 °C freezers until DNA extraction was performed. Alongside blood samples, the biobank provided clinical data relating to the cases including age, sex, smoking history, cancer histology, tumour stage, date of surgery and spirometry. Written, informed consent was provided by all participants (REC Ref: 18/NW/0092).

2.5 Case-Control DNA Extraction

2.5.1 Case Samples

QIAGEN (Hilden, Germany) Gentra® PureGene® and QIAGEN FlexiGene® kits were used for DNA extraction from the case samples. A detailed QIAGEN Gentra® PureGene® protocol is available [289]. In brief, 9 ml of red blood cell lysis solution was added to 3 ml of whole blood, releasing the cell nuclei and mitochondria. The samples were incubated at room temperature for 5 minutes, following which they were centrifuged for 2 minutes at 2000 x g. The supernatant containing the cell fragments was discarded, leaving a white blood cell pellet which was then resuspended in the residual liquid by vortex mixing. A further 3 ml of cell lysis solution was added to the samples, followed by 10 seconds of vortex mixing. Next, 1 ml of protein precipitation solution was added to the samples, which were then vortex mixed for 20 seconds and centrifuged at 2000 x g for 5 minutes. The resultant supernatant was transferred into 3 ml of isopropanol and inverted 50 times until the DNA pellet was visible. The DNA was centrifuged at 2000 x g for 3 minutes, drained by inversion onto absorbent paper, and cleaned with 3 ml of 70% ethanol. The samples were centrifuged again, inverted onto absorbent paper, air-dried for 5 minutes, rehydrated in 300 µl of DNA rehydration solution, and briefly mixed. The suspended DNA samples were then incubated for 1 hour at 65 °C and overnight at room temperature.

A detailed QIAGEN FlexiGene® protocol is available [290]. In brief, for each patient sample, 3 ml of blood was added to 7.5 ml of lysis buffer, mixed briefly, and centrifuged at 2000 x g for 5 minutes. The supernatant was discarded, and the pellet drained onto absorbent paper for 2 minutes. 1.5 ml of a protease buffer was added, and the mixture vortex mixed until homogenised. The samples were then incubated at 65 °C for 10 minutes, at which point the solution changed colour from red to green indicating protein digestion. 1.5 ml of 100% isopropanol was added, and the samples were centrifuged at 2000 x g for 3 minutes to precipitate the DNA. The DNA pellets were drained and washed with 1.5 ml of 70% ethanol. They were then re-centrifuged, drained, air-dried and rehydrated in 300 µl of buffer, after which the suspended DNA samples were incubated for 1 hour at 65 °C.

2.5.2 Control Samples

DNA from control blood samples was extracted using the Qiagen QIAamp® DNA Blood Midi Kit. A detailed protocol for this procedure is available [291]. In brief, up to 2 ml of blood from each sample was added to 200 µl protease and mixed. 2 ml of lysis buffer was added, and the tubes were inverted 15 times and mixed vigorously for 1 minute. The samples were incubated at 70 °C for 10 minutes. Ethanol (100%) was added, and the samples mixed. The solution was transferred to QIAamp® Midi (Qiagen, Hilden, Germany) columns placed in 15 ml collection tubes and centrifuged at 1850 x g for 3 minutes to bind the DNA; DNA adsorbs to a silica-based membrane inside the column, whilst impurities and contaminants pass through. The filtrate was discarded, and 2 ml volumes of two wash buffers were added to the columns in sequence; the columns were centrifuged at 4500 x g for 1 minute and 15 minutes respectively. Finally, 300 µl of elution buffer was added to the columns, incubated at room temperature for 5 minutes, and centrifuged at 4500 x g for 2 minutes. To increase DNA concentration, the eluate was reloaded into the columns to undergo a second round of incubation and centrifugation.

2.6 DNA Quality Control, Quantification and Normalisation

Extracted DNA samples were tested for purity and concentration using Thermo Fisher NanoDrop™ (Waltham, Massachusetts, United States) either immediately (for the QIAGEN FlexiGene® and QiaAmp DNA Blood Midi Kit) or 24 hours after extraction (for the QIAGEN Gentra® PureGene®). The NanoDrop™ was calibrated using injection water, blanked using the rehydration buffer provided with the DNA extraction kit, after which a volume of 2 µl DNA of each sample was used for quantification. A concentration of ≥100 ng/µl was considered adequate for downstream application. A 260/280 absorbance ratio of ≥1.8 and 260/230 absorbance ratio of ≥2.0 was considered pure; low ratios indicate the presence of residual reagents from the extraction process which may interfere with downstream processes [292].

Samples with inadequate concentration or purity were either re-purified, concentrated, or re-extracted. Samples with low purity or concentration were purified or concentrated using Zymo Research Genomic Clean and Concentrator™ (Irvine, California, USA). A detailed protocol is available [293]. In brief, one volume of DNA was added to double its volume of DNA binding buffer. The mixture was transferred to a Zymo-Spin™ Column housed in a collection tube and centrifuged for 30 seconds at 15,000 x g. DNA binds to the silica-based membrane in the column. DNA wash buffer was then added to the column which was centrifuged for 1 minute. The wash step was repeated. Finally, DNA elution buffer (volume dependant on concentration of original

DNA sample) was added to the column, incubated at room temperature for 1 minute, and centrifuged in a clean collection tube for 30 seconds. Samples with extremely low DNA concentrations (indicating a possible loss of genomic material during extraction) were re-extracted from residual blood where available.

All DNA samples were quantified using the Qubit™ dsDNA BR Assay Kit (Thermo Fisher, Waltham, Massachusetts, United States), to facilitate accurate normalisation in preparation for genotyping. A detailed protocol is available [294]. In brief, 1 µl of each DNA sample was combined with 199 µl of prepared Qubit™ working solution in Qubit™ assay tubes. The two standards were made up with 10 µl of standard stock and 190 µl of working solution. Each mixture was vortex mixed. The fluorometer was blanked and calibrated before each run using freshly prepared standards. The concentration of each DNA sample was then measured. The value provided by the fluorometer was multiplied by 200 to ascertain the DNA concentration of the original sample. The DNA samples were normalised to 50 ng/µl in 20 µl aliquots and dispensed into 96-well plates, using RNase and DNase-free sterile injection water for dilutions. DNA samples were stored at -20 °C until genotyping.

2.7 Genotyping

Genotyping of DNA samples was performed on the Illumina iScan™ System (San Diego, California, United States), employing the Infinium OncoArray-500K for high-throughput screening (HTS). This is a BeadChip designed by the OncoArray consortium for the purpose of studying cancer risk and predisposition by examining approximately 500,000 SNPs related to common cancers, including lung cancer [295].

A detailed protocol for performing Infinium HTS is available [296]. A schematic overview is presented in Figure 8. A brief description follows, with several mix and centrifuge steps omitted for brevity. First, 4 µl of each DNA sample was added to 20 µl of buffer and denatured with 4 µl of 0.1N NaOH. The mixtures were then neutralised with 34 µl of neutralisation buffer. 38 µl of an amplification master mix was added to each sample following which they were amplified by incubation in a 37 °C hybridisation oven overnight. This results in uniformly amplified DNA samples in sufficient quantity for genotyping.

The DNA was then enzymatically fragmented with 25 µl of a fragmentation buffer and incubated on 37 °C heat block for one hour. This cleaves DNA into 300-600 base pair fragments, the optimal length for hybridisation to the OncoArray BeadChip. The fragmented DNA samples were coloured with 50 µl of a visualisation reagent, precipitated with 155 µl of 100% 2-propanol, mixed by inversion, incubated at 4 °C for 30 minutes, and centrifuged at 3000 x g and 4 °C for 20 minutes. Following centrifugation, the plate was drained onto an absorbent pad, leaving the blue-coloured

DNA pellets in the plate. The samples were then resuspended in 23 µl of buffer, the plate was heat sealed, and incubated in the hybridisation oven at 48 °C for 1 hour

Next, 14 µl of each DNA sample was transferred onto the OncoArray 500K BeadChip, hybridising to complementary oligonucleotide probes which cover the silica beads on the chip during an overnight incubation period (48 °C). Each bead is coated with DNA oligonucleotide probes specific to one locus of interest. DNA fragments only bind to the complementary probes; specificity is ensured by the high stringency buffer conditions, elevated temperatures, and length of probes.

Following overnight incubation, the BeadChip covers were removed, and excess reagents were removed by submerging the chips in wash buffer. The BeadChips were integrated into specialised flow-through chambers and placed vertically onto a water circulator set to 44 °C. The BeadChips then underwent an extensive manual washing and staining procedure. A reagent was added to wash away unhybridized excess DNA. Two buffers were then added to condition the BeadChip surface for DNA extension. Next, DNA polymerase was added to extend the DNA fragments with a single hapten-labelled base for each probe. G and C nucleotides are labelled with biotin, A and T nucleotides are labelled with dinitrophenyl (DNP). Following extension, the probes were stained with green-fluorescent streptavidin (which binds to biotin) and red-fluorescent anti-DNP antibodies. Successive rounds of staining and washing amplify the fluorescent signal. After staining, the BeadChips were washed to remove excess reagents, submerged in a coating solution, and dried in a vacuum desiccator for at least one hour.

Once the BeadChips were coated and dried, they were loaded into the Illumina iScan™ System. The scanner exposed the BeadChip to high-performance red and green lasers, which excite the fluorophores bound to the DNA samples, resulting in the display of a red or green signal depending on the variant present for a particular SNP. For example, if the signal produced by a particular bead is red, it can be assumed that there is a homozygous TT allele at that locus. A green signal would indicate homozygous CC, whilst a yellow signal indicates a heterozygous allele. The BeadChip includes a series of built-in internal control probes, designed to confirm the success of each genotyping step, thereby ensuring reproducibility and robustness of results [297]. I added cross-plate sample duplicates to each genotyping run to serve as positive controls.

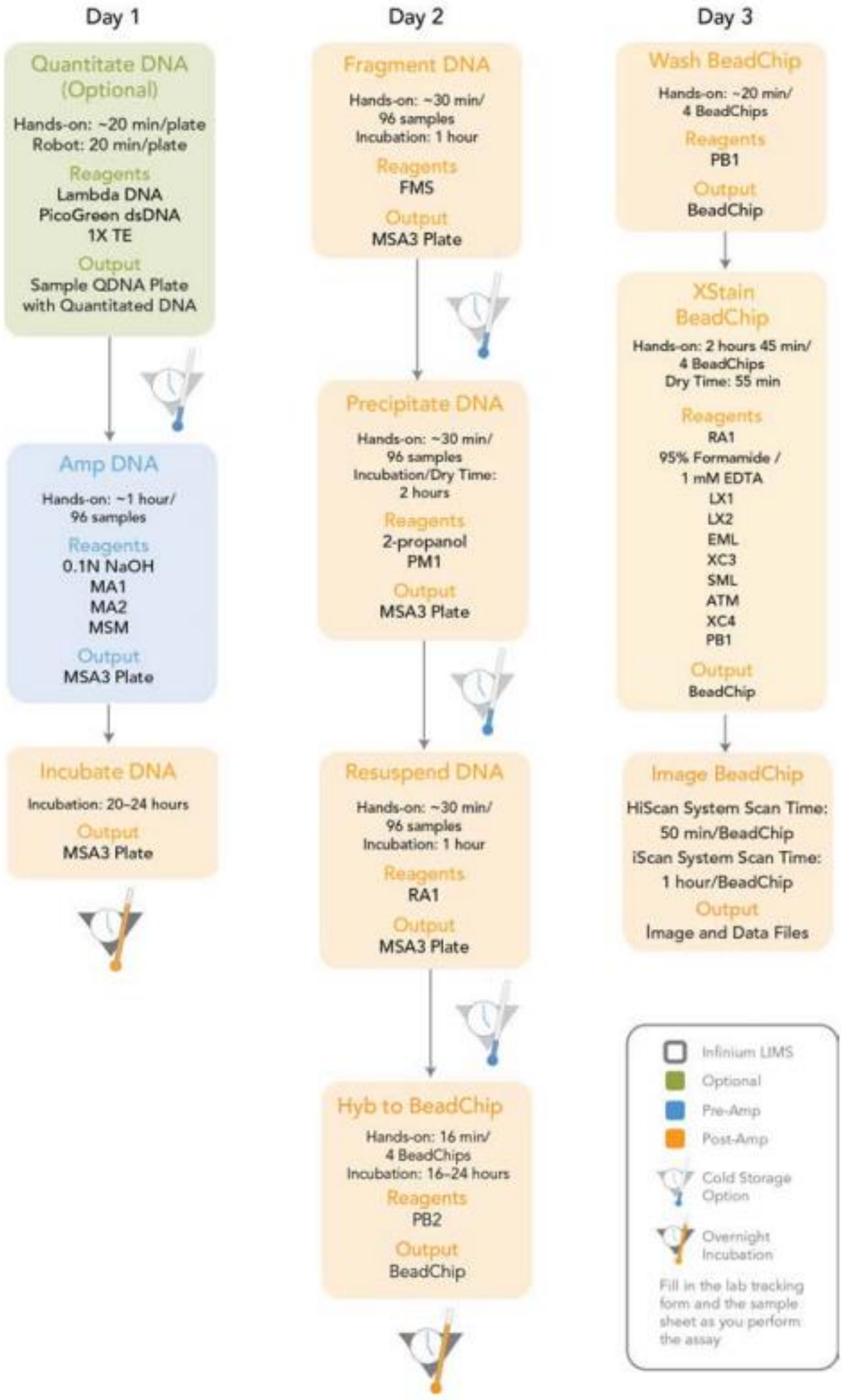

Day 1
Quantitate DNA (Optional)
Hands-on: ~20 min/plate
Robot: 20 min/plate
Reagents
Lambda DNA
PicoGreen dsDNA
1X TE
Output
Sample QDNA Plate with Quantitated DNA
Amp DNA
Hands-on: ~1 hour/ 96 samples
Reagents
0.1N NaOH
MA1
MA2
MSM
Output
MSA3 Plate
Incubate DNA
Incubation: 20-24 hours
Output
MSA3 Plate
Day 2
Fragment DNA
Hands-on: ~30 min/ 96 samples
Incubation: 1 hour
Reagents
FMS
Output
MSA3 Plate
Precipitate DNA
Hands-on: ~30 min/ 96 samples
Incubation/Dry Time: 2 hours
Reagents
2-propanol
PM1
Output
MSA3 Plate
Resuspend DNA
Hands-on: ~30 min/ 96 samples
Incubation: 1 hour
Reagents
RA1
Output
MSA3 Plate
Hyb to BeadChip
Hands-on: 16 min/ 4 BeadChips
Incubation: 16-24 hours
Reagents
PB2
Output
BeadChip
Day 3
Wash BeadChip
Hands-on: ~20 min/ 4 BeadChips
Reagents
PB1
Output
BeadChip
XStain BeadChip
Hands-on: 2 hours 45 min/ 4 BeadChips
Dry Time: 55 min
Reagents
RA1
95% Formamide / 1 mM EDTA
LX1
LX2
EML
XC3
SML
ATM
XC4
PB1
Output
BeadChip
Image BeadChip
HiScan System Scan Time: 50 min/BeadChip
iScan System Scan Time: 1 hour/BeadChip
Output
Image and Data Files
☐ Infinium LIMS
🟩 Optional
🟦 Pre-Amp
🟧 Post-Amp
Cold Storage Option
Overnight Incubation
Fill in the lab tracking form and the sample sheet as you perform the assay

Figure 8. Illumina Infinium HTS assay manual workflow protocol. Used under license from Illumina, Inc. All Rights Reserved.

2.8 Data Quality Control

Illumina's GenomeStudio (v.2.0) was used for the first stages of quality control (QC) and data processing. SNPs were initially clustered using the Infinium OncoArray-500K v1.0 (Rev. B) Cluster File provided by Illumina. Samples with a call rate <95% were excluded, and in most cases, were re-genotyped to generate an iteration with an adequate call rate. The final sample-set was then re-clustered independent of pre-defined cluster positions. The PLINK Input Report Plug-in (v2.1.4) was used to generate PLINK compatible PED and MAP files from GenomeStudio. The PED file was converted to binary (BED) format in PLINK (v.1.9) [298]. PLINK was used for all QC steps unless otherwise specified.

Thresholds for SNP QC were established based on previous studies [299]. Y chromosome and mitochondrial SNPs were excluded. SNPs with a missingness rate of >0.02 were excluded, thereby removing SNPs with low genotype call rates. SNPs with a minor allele frequency of <0.01 were excluded, as these are considered rare variants which are more prone to genotyping errors. Furthermore, this study is not powered to establish associations between rare variants and the phenotype of interest. SNPs which deviated from Hardy Weinberg Equilibrium (HWE) (1e-4) were excluded. In such a deviation, the allele frequencies do not conform to the HWE (which assumes constant allele frequencies across generations), indicating genotyping error or evolutionary selection. Symmetric SNPs (i.e. A/T or G/C SNPs) were excluded, as errors can occur in the Illumina "A" and "B" allele SNP designation process in these cases [300]. Following these steps, a checking tool was run to confirm strand alignment and position of SNPs in comparison to the haplotype reference panel (HRC) and 1000 Genome datasets [301].

Thresholds for individual/participant QC were sourced from previous studies [299]. Samples with a call rate of <98% or divergent heterozygosity (>3 SDs), indicating inbreeding or sample contamination, were excluded. Individuals with non-concordant phenotypic and genotypic sex were excluded. KING (v.1.9) software was used to assess Identity by Descent (IBD) [302]. Unexpected duplicates were excluded, as were one member of each pair of 1st or 2nd degree relatives (relatedness can skew association analyses).

The FlashPCA (v.2.0) tool for principal component analysis (PCA) was used to ascertain genetic ancestry and genetic variation in the cohort [303]. To estimate genetic ancestry, I combined the Manchester cohort genotypes with the HapMap3 reference dataset and performed PCA on the combined data [304]. Samples which deviated from the European cluster were excluded. To ensure there were no distinct genetic subgroupings within the cohort which could skew results, samples from within the European ancestry subgroup were also excluded if they had outlying genetic variation based on a scatter plot of PC1 and PC2 values. The Aberrant (v.1.0) R package was used to exclude outliers [305].

The scree plot generated during the PCA was used to determine the number of principal components (PCs) to be used as covariates in the association analysis, with the number of points deviating from the trend providing an indication of the appropriate number of PCs to account for the genetic variance in the cohort.

2.9 Genomic Imputation

I used the Michigan Imputation Server for genomic imputation [306]. The human genome reference build, HRC r.1.1 2016 (GRCh37/hg19) was used as the reference panel and Eagle v.2.4 was used for phasing. The imputed dataset was filtered for duplicated SNPs and SNPs with low imputation confidence (r^2 <0.5).

2.10 PRS Validation

I searched PubMed and the PGS Catalog [307] databases for previously developed lung cancer PRS in order to validate them in the Manchester case-control cohort. Effect alleles listed in the literature were matched with the Manchester cohort genotyped data by comparing minor allele frequency (MAF). The LDproxy tool [308] was used to identify proxy SNPs when SNP data were not available in the study cohort. The proxy-SNP appearing in the Manchester dataset with the highest R^2 was used, with the correlated alleles provided by LDproxy used to identify the effect allele. If no proxy SNP with an R^2 of >0.5 was available, that SNP was excluded from the PRS for the purposes of validation. Genetic load was calculated using the --score function in PLINK (v.1.9) [298]. This function applies a simple linear scoring system, by summing the number of effect alleles for the specified SNPs, multiplied by the allelic weight. Allelic weight was represented by the natural logarithm of the published ORs. Scores were centred around the mean to facilitate comparison between PRSs.

Following calculation of genetic load scores for each PRS, I used the ggplot2 (v.3.3.5) package in RStudio to create density plots of scores stratified by phenotype, in order to visually represent the divergence in PRS between cases and controls [309]. Model discrimination was assessed by calculating area under the curve (AUC) using the pROC (v.1.17.1.0) package in RStudio [310].

I fitted a logistic regression model using the demographic and clinical data available in both cases and controls. Variables included were age, sex, BMI, smoking status and FEV_1/FVC ratio. Using the PredictABEL (v.1.2-4) R package [311], AUC was calculated for the base clinical model, then each of the PRSs were added to the model to examine additional discriminatory value conferred by the genetic score. Statistical significance of any AUC improvement conferred by PRS inclusion was tested using the Likelihood-ratio test. I also used PredictABEL to calculate net reclassification improvement (NRI) index for each model. NRI is a metric that attempts to quantify how well a model reclassifies subjects compared to a previous model. Cases or controls who are correctly reclassified by the new model into a higher or lower risk groups respectively are assigned +1, and the converse are assigned -1. The sum of all individuals in each group is then divided by the total number of subjects in the group and added together to result in the NRI [312].

I selected the best performing PRS from those validated in the Manchester cohort to perform subgroup analysis. I split the cohort by sex, age (above vs. below median), BMI (above vs. below median), and smoking status (current vs. former), as well as by lung cancer histological subtype in the cases (adenocarcinoma vs. other) and NLST eligibility, calculating PRS performance scores in each of the subgroups individually.

SNPs that appeared in multiple published PRSs were also identified and synthesised into a ‘combined PRS’ before testing it in the study data, to explore whether the SNPs most employed in a variety of PRSs would provide superior performance if used in combination.

2.11 GWAS and PRS Development

Association analysis was performed using SNPTEST (v.2.5.2). The analysis was adjusted for PCs generated during PCA, as well as sex, age, smoking status and BMI. The output was uploaded to the Functional Mapping and Annotation of Genome-Wide Association Studies (FUMA GWAS v.1.3.6) platform for GWAS data visualisation and candidate gene exploration [313]. Maximum P-value for independently significant SNPs was set at 1e-5.

PRSice-2 (v.2.3.3; 2020-08-05) was used to prune SNPs in linkage disequilibrium, identify the most predictive PRS, and generate a list of SNPs in the PRS ordered by P-value thresholds [314]. The analysis was adjusted for the same covariates detailed above (section 2.10). The SNPTEST output files were used as the base dataset; raw genotyping data from the case-control dataset was used as the target dataset. The beta from the SNPTEST output was used as the effect magnitude. PRSice-2 was also used to create a plot displaying model fit at each P-value threshold.

2.12 Manchester LHC Programme Cohort and Recruitment

The cross-sectional study investigated lung cancer-specific risk perception and disease knowledge, described in Chapter Six, took place in a subset of first-round participants of the Manchester Lung Health Check Programme (LHC). This programme took place in 2019-2020 (Figure 9) and was an expanded version of the Manchester LHC pilot described in 2.1. The programme invited ever-smokers aged 55-80 resident in deprived areas of North and East Manchester to a community-based LHC, in which lung cancer risk was calculated, respiratory symptoms examined, and smoking cessation advice offered. As in the LHC pilot and the NHS TLHC protocol [48], those deemed at high risk of lung cancer ($PLCO_{M2012}$ 6-year risk ≥1.51%) were offered immediate low-dose computed tomography (LDCT) screening (T0) followed by a second-round LDCT scan one year later (T0+12). Data collected in the LHC included a variety of demographic factors: age, sex, ethnicity, smoking status, smoking history, educational attainment, asbestos exposure, family history of lung cancer, previous cancer diagnosis, BMI and postcode-based index of multiple deprivation (IMD). Some of these factors were used to generate a 6-year estimate of lung cancer risk using the $\text{PLCO}_{\text{M2012}}$ risk model [89].

I recruited participants for the study immediately as they entered the LHC site, prior to undergoing their LHC or any contact with clinical staff. All English-speaking participants who attended the LHC on the days when this study was operating were offered to take part, without any selection based on demographic factors. Participants consented to take part in this study and were asked if they agreed to have their questionnaire responses linked to the clinical and demographic data collected in their LHC. Participants who declined to be identified were still able to complete the questionnaire and their responses remained anonymous.

Manchester Lung Health Check
ENTRANCE
www.mft.nhs.uk/lunghealthcheck
Manchester Lung Health Check
Cobalt
NHS
Manchester University
NHS Foundation Trust
LAMBOO MEDICAL UK

Figure 9. Image of the mobile LHC unit from the Manchester Lung Health Check programme [315].

Chapter Three - Risk Prediction Model Performance: Study of Manchester LHC Pilot

3.1 Introduction

Lung cancer is the leading cause of cancer death worldwide [1]. Poor survival is driven by late clinical presentation, due to early-stage cancer being largely asymptomatic [5]. LDCT screening significantly improves lung cancer outcomes by detecting lung cancer at an earlier stage [193,194]. The selection method used to determine screening eligibility is a crucial determinant of screening programme efficiency and effectiveness. Lung cancer screening trials often employ generalised eligibility criteria for screening selection (also recommended by USPSTF guidelines [47]). Implementation studies and NHS England's TLHC programme favour screening selection based on individual risk prediction (using RPMs), which may improve screening performance compared to generalised eligibility criteria [89,220] (described in detail in sections 1.6.6 and 1.7.1.2).

The Manchester LHC pilot was a community-based screening programme that offered targeted LDCT screening to ever-smokers in socio-economically deprived areas of Manchester. The LHC pilot had a detection rate of 4.4%, significantly higher than comparable screening trials and programmes [248,250]. Screening eligibility in the pilot was determined using the $PLCO_{M2012}$ RPM at a threshold of ≥1.51%. This RPM was originally developed in the PLCO study population (a large, North American, relatively affluent cohort [189]), which may not be representative of real-world lung cancer screening attendees. Consequently, $PLCO_{M2012}$ and other selection tools developed in trial settings (such as LLP_{V2} and NLST criteria), may function unexpectedly or sub-optimally when applied to community-based lung cancer screening programmes. With the ongoing rollout of the TLHC programme in England, understanding the characteristics and deficiencies of the selection methods being widely used in clinical practice is of particular importance [48]. Furthermore, socio-economically disadvantaged communities, which are key targets for lung cancer screening due to the high risk profile of the population, may be particularly vulnerable to sub-standard risk prediction; this could lead to an exacerbation of health inequalities [229,230] (described in detail in section 1.7.1.1).

This study aimed to test the hypothesis that screening selection strategies perform sub-optimally in socio-economically disadvantaged populations (see hypotheses in section 1.8). I retrospectively examined the performance of the $PLCO_{M2012}$ RPM which was used in the Manchester LHC pilot for screening selection. I also modelled various alternative methods of screening selection in this cohort and examined the hypothetical screening performance had these strategies been used in the pilot. Selection performance was assessed by considering RPM calibration, the number of participants selected for screening, and the associated cancer detection rates. Model calibration is important, as RPMs and associated risk-based screening thresholds are frequently developed in populations unrepresentative of the programmes they are used in clinically. A mis-calibrated model may underestimate or overestimate risk in the target population, thereby restricting screening from those who may benefit or providing screening too widely and exposing low-risk participants to harm. The number of participants selected for screening and cancer detection rates are key metrics in judging the overall efficiency of a screening programme.

A further measure of selection performance is the comorbidity profile of the screening cohort. Lung cancer risk, being predominantly driven by age and smoke exposure, is strongly correlated with respiratory and cardiovascular comorbidities [316]. Evidence for a relationship of this nature could be used to augment RPMs with additional clinical risk factors to improve risk prediction. However, it may also indicate that the benefit high-risk individuals derive from screening could be limited by competing causes of mortality [233], which would severely impact on the overall effectiveness of the screening programme.

3.2 Methods

3.2.1 Study Cohort and Risk Prediction

The population analysed in this study was the Manchester LHC pilot cohort (LHC pilot described in section 1.7.1.1). In depth clinical and demographic data were recorded prospectively by the clinical team at the LHC location. Detailed descriptions of cohort recruitment, data collection, and risk prediction calculations are described in section 2.1.

3.2.2 Outcome Measures

The number of participants selected for screening and cancer detection rates were examined by comparing $PLCO_{M2012}$ at a ≥1.51% threshold (used in the LHC pilot for selection) with several alternative selection methods: NLST generalised eligibility criteria (age 55-74, ≥30 pack years, current smoker within 15 years [193]), LLP_{v2} RPM at a ≥5% threshold (as per the UKLS trial [223]), and LLP_{v2} at a ≥2.5% threshold (as per the TLHC protocol [48]). The comparison of $PLCO_{M2012}$ ≥1.51% and LLP_{v2} ≥2.5% is particularly pertinent with important clinical implications, as these criteria are recommended by National Health Service England (NHSE) for determining screening eligibility in the TLHC programme (see section 1.7.1.2) [48].

I examined the presence of several cardiovascular and respiratory comorbidities in the screening cohort, including airflow obstruction, coronary artery calcification, emphysema, and respiratory symptoms. I assessed the association between these comorbid conditions and $PLCO_{M2012}$ calculated lung cancer risk, as well as explored whether these comorbidities are associated with lung cancer detection in the Manchester LHC pilot.

RPM calibration was examined by comparing the projected number of detected lung cancers in the Manchester LHC pilot with the number of cancers predicted by the $\text{PLCO}_{\text{M2012}}$ and LLP_{V2} RPMs.

3.2.3 Detection Rate Projections

In order to assess RPM calibration, I calculated hypothetical detection rates to approximate the number of lung cancers that might have been detected in Manchester if screening had been continued for a total of six years. I used lung cancer incidence data from the NELSON trial as a basis for projections because it was undertaken in a comparable European population, at a large scale and the duration of screening was similar (5.5 years) to the 6-year risk calculated by $\text{PLCO}_{\text{M2012}}$. In the NELSON trial, lung cancer detection rates at each screen were 0.9% at baseline, 0.8% in the second screening round (1-year after baseline), 1.1% in the third (3-years after baseline) and 0.8% in the fourth screening round (5.5-years after baseline) [194]. This is one-third and one half the detection rate seen in Manchester at baseline (3.0%) and the second round (1.6%). The projected detection rates used for analysis were therefore adjusted twofold to 2.2% (3rd round) and 1.6% (4th round). The denominator was the total number eligible for screening minus the number of participants diagnosed with lung cancer prior to that screening round, thus the estimated number of cases assumes complete adherence. Lung cancers diagnosed outside of screening (interval cancers) in the NELSON trial were not included in the analysis, and therefore, it was assumed that no interval cancers would occur. This is consistent with the LHC pilot data which showed no interval cancers between the first and second screening round.

3.2.4 Statistical Analysis

I carried out statistical analysis using IBM SPSS Statistics (v.25). Basic descriptive statistical techniques were used to elucidate the characteristics of screening cohorts as selected by the various eligibility criteria. Comparison of means was conducted by T test or one-way analysis of variance (ANOVA). All P-values for ANOVA were adjusted for multiple testing by post-hoc Bonferroni correction. P-values for trend were calculated by fitting linear or logistic regression models, using a single independent variable with 1 degree of freedom. I tested a number of variables, not included in the $\text{PLCO}_{\text{M2012}}$ model, in a logistic regression analysis to see if they were independently predictive of lung cancer after controlling for $\text{PLCO}_{\text{M2012}}$ risk score, as indicated by a statistically significant P-value for the variable in question (<0.05). These variables included FEV1/FVC ratio, CAC and emphysema.

3.3 Results

3.3.1 Study Cohort Recruitment and Demographics

Detailed results from both screening rounds of the Manchester LHC pilot have been published previously [248,250]. In brief, 2541 ever-smokers attended the LHC pilot. Complete demographic descriptors are presented in Table 2; 51% were female, mean age was 64 years, 37% had airflow obstruction, 35% were current smokers, and the average number of pack-years was 35. Three-quarters of attendees were in the lowest IMD quintile, and two-thirds had no educational qualifications. In total, 1429 (56.2%) attendees were eligible for, and 1384 had, a baseline scan. A small proportion (n=25) had a scan at T1 only (98.6% of those eligible had at least one scan), resulting in 1409 having at least one scan over the two rounds (Table 2).

3.3.2 Lung Cancer Detection

A total of 62 participants were diagnosed with lung cancer in the LHC pilot, 42 at T0 (3%), 19 at T1 (1.6%) and one during post-T1 nodule surveillance [248,250]. The number needed to screen (NNS) to detect one lung cancer was 23. NNS was strongly associated with lung cancer risk, with an NNS of 9 amongst those in the highest risk group ($PLCO_{M2012}$ ≥10%) compared to a range of 66-73 amongst those with a risk score of $PLCO_{M2012}$ 1.51%-3%. The NNS was lower for women than men at every risk threshold until ≥6%, at which point it was similar (Figure 10). At the $PLCO_{M2012}$ ≥1.51% threshold used in the LHC pilot, the NNS was 26 for men and 20 for women (Table 2).

Table 2. Demographic and clinical variables of complete Manchester LHC pilot cohort and stratified by PLCO$_{M2012}$ risk category.

Variable	**All**	**PLCO$_{M2012}$ Score Group**							
		<1.51%	**1.51-2%**	**2-3%**	**3-4%**	**4-6%**	**6-9%**	**≥10%**	**p value for trend**
Number Participants (%)	2541	1112	265 (19)	294 (21)	233 (16)	263 (18)	221 (16)	153 (11)	-
Number Screened (%)	1409 (55.5)	-	263 (19)	291 (21)	230 (16)	256 (18)	218 (16)	151 (11)	-
% Screened with Lung Cancer (n) [% Male/Female]	4.4 (62) [3.9/4.9]	-	1.5 (4) [1.5/1.6]	1.4 (4) [1.5/1.3]	3.0 (7) [2.6/3.4]	5.5 (14) [3.2/ 7.6]	7.8 (17) [6.6/ 9.4]	10.6 (16) [10.8/ 10.4]	<0.0001
NNS to Detect 1 LC in Group (M/F)	22.7 (25.9/ 20.3) ^{n}	-	66 (67/65)	73 (66/80)	33 (38/29)	18 (31/13)	13 (15/11)	9 (9/10)	-
NNS to detect 1 LC at or above risk score (M/F)	-	-	23 (26/20)	20 (23/18)	16 (19/14)	13 (16/11)	11 (12/10)	9 (9/10)	-
Mean Age ±SD	64.1 ±5.5	63.4 ±5.5	62.4 ±5.6	63.4 ±5.4	64.1 ±5.2	65.0 ±4.8	66.1 ±4.5	69.7 ±3.6	<0.0001
% Female (n)	51 (1296)	52 (573)	49 (130)	55 (162)	51 (118)	52 (136)	44 (98)	52 (79)	0.388
% Current Smokers (n)	35 (891)	12 (137)	39 (104)	41 (121)	51 (118)	57 (149)	66 (145)	77 (117)	<0.0001
Mean Age Started Smoking ±SD	16.6 ±5.3	17.6 ±6.3	16.3 ±5.1	16.2 ±4.3	16.0 ±4.0	15.8 ±4.0	15.3 ±4.0	15.0 ±3.1	<0.0001
Mean Years Smoked ±SD	34.6 ±14.7	22.8 ±12.6	37.5 ±8.0	40.2 ±6.8	42.6 ±6.9	45.7 ±5.6	48.5 ±6.0	53.5 ±4.6	<0.0001
Mean Cigs per Day ±SD	20.4 ±13	16 ±12	21 ±10	22 ±11	24 ±12	26 ±17	26 ±12	27 ±14	<0.0001
% Family History$^{\#}$ (n)	21.8 (553)	15 (161)	17 (45)	24 (70)	19 (45)	30 (80)	39 (85)	44 (67)	<0.0001
% MRC Dyspnoea 1 (n)	70.5 (1791)	78 (871)	63 (168)	72 (211)	70 (162)	66 (174)	57 (125)	52 (80)	<0.0001
% Symptomatic* (n)	50.3 (1278)	38 (423)	52 (136)	51 (149)	55 (129)	62 (163)	73 (162)	76 (116)	<0.0001
Mean FEV$_1$ ±SD	2.3 ±0.8	2.6 ±0.8	2.4 ±0.7	2.3 ±0.7	2.2 ±0.7	2.0 ±0.7	2.0 ±0.7	1.7 ±0.6	<0.0001
Mean FVC ±SD	3.3 ±1	3.4 ±1	3.3 ±0.9	3.2 ±0.9	3.2 ±0.9	3.0 ±1.0	3.1 ±1.0	2.9 ±0.9	<0.0001
Mean FEV$_1$/FVC Ratio ±SD	70.8 ±10.6	75 ±9	72 ±9	70 ±10	68 ±10	67 ±11	64 ±11	61 ±12	<0.0001
% Airflow Obstruction (n)	37 (948)	21 (232)	28 (75)	39 (114)	50 (114)	58 (152)	68 (149)	74 (112)	<0.0001
% Lowest IMD Quintile (n)	74.5 (1893)	67.8 (754)	79.6 (211)	78.2 (230)	77.7 (181)	79.1 (208)	81.9 (181)	83.7 (128)	<0.0001

% No Educational Qualifications (n)		61.7 (1567)	51.2 (569)	57.4 (152)	61.9 (182)	71.7 (167)	74.5 (196)	77.8 (172)	84.3 (129)	<0.0001
CVD %Yes (n)		-	-	17.8 (33)	17.0 (35)	24.8 (38)	18.2 (31)	31.1 (42)	30.0 (27)	0.005
Mean QRISK2 score ±SD		-	-	21 ±11	22 ±12	23 ±11	24 ±10	27 ±13	31 ±12	<0.0001
% Emphysema (n)		-	-	54 (142)	60 (171)	64 (146)	65 (164)	69 (146)	77 (114)	<0.0001
% CAC (n)	None	-	-	34 (81)	26 (66)	27 (52)	29 (62)	23 (42)	20 (25)	<0.0001
	Mild	-	-	54 (80)	40 (103)	38 (72)	37 (81)	39 (71)	39 (49)	
	Mod-Severe	-	-	12 (75)	35 (90)	35 (67)	34 (74)	39 (71)	41 (52)	

*LC diagnosed in a first degree relative.

†Defined as the presence of ≥1 symptom at baseline, including: breathlessness (MRC dyspnoea scale ≥2), cough ≥6 weeks in duration, sputum production ≥teaspoon/day.

‡QRISK2 score calculated in the second screening round only in those with no prior history of CVD.

^{n}within screened (n=1409) participants

CAC, coronary artery calcification; CVD, cardiovascular disease; FEV1, forced expiratory volume in 1 second; FVC, forced vital capacity; LC, lung cancer; NNS, number needed to screen.

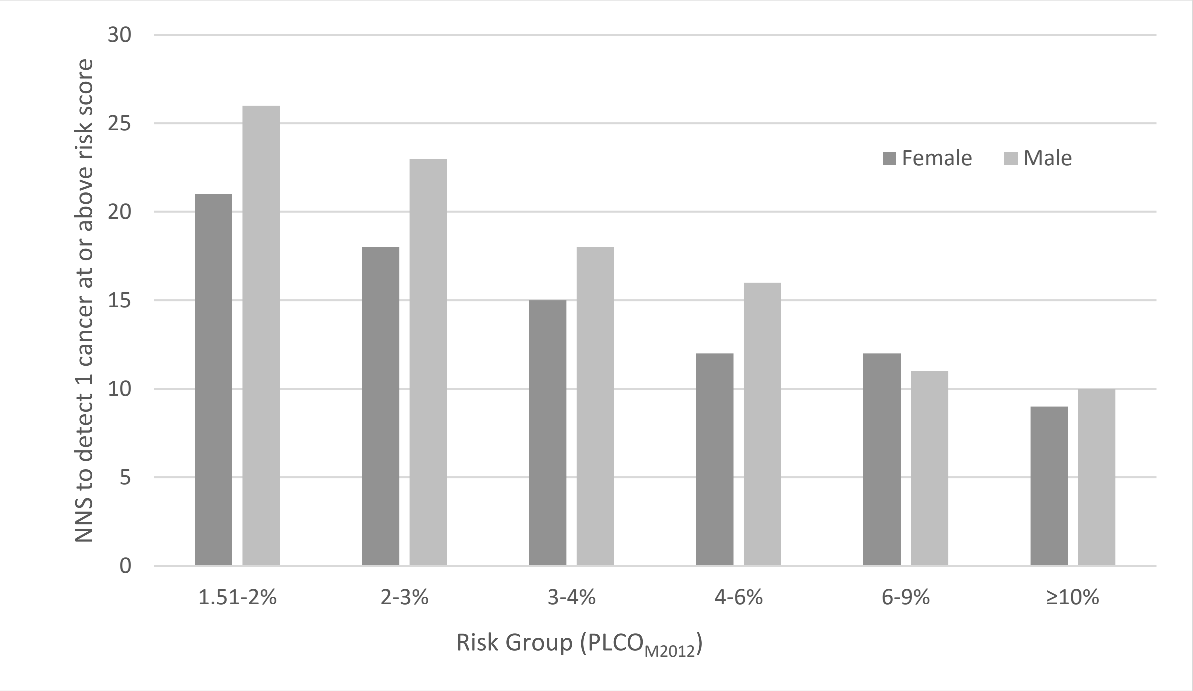


Figure 10. Number needed to screen (NNS) to detect one lung cancer in Manchester LHC pilot, stratified by sex and lung cancer risk group.

3.3.3 Risk Prediction Model Calibration

The mean $PLCO_{M2012}$ score amongst screen-eligible participants (n=1429) was 5% (±4.03). Based on this, the RPM predicts the incidence of 71 lung cancers over six years; 87% (n=62) of these predicted cancers were detected in the 15-months of the pilot. Modelling, based on adjusted NELSON trial detection rates (detailed in section 3.2.2), predicts that 113 lung cancers would be diagnosed over a period of 5.5 years. This is 59% more than predicted by the RPM. Even with a more conservative unadjusted detection rate, 87 lung cancers might be detected during this time, 23% higher than predicted. If annual detection rates were to mirror those seen in the Danish Lung Cancer Screening Trial [317], 168 lung cancer might be detected after 6 years, almost 2.5 times as many as predicted by the $PLCO_{M2012}$ model (Figure 11).

The mean LLP_{V2} score amongst screen-eligible participants was 6.7% (±5.4), higher than $PLCO_{M2012}$. Based on this, LLP_{V2} predicts the incidence of 94 lung cancers over five years, of which 66% had been detected in the 15-months of the LHC pilot. The adjusted NELSON incidence trend would reach this number between three and four years, with a final 5-year projection of 109, 16% higher than the cancers predicted by LLP_{V2} (Figure 11).

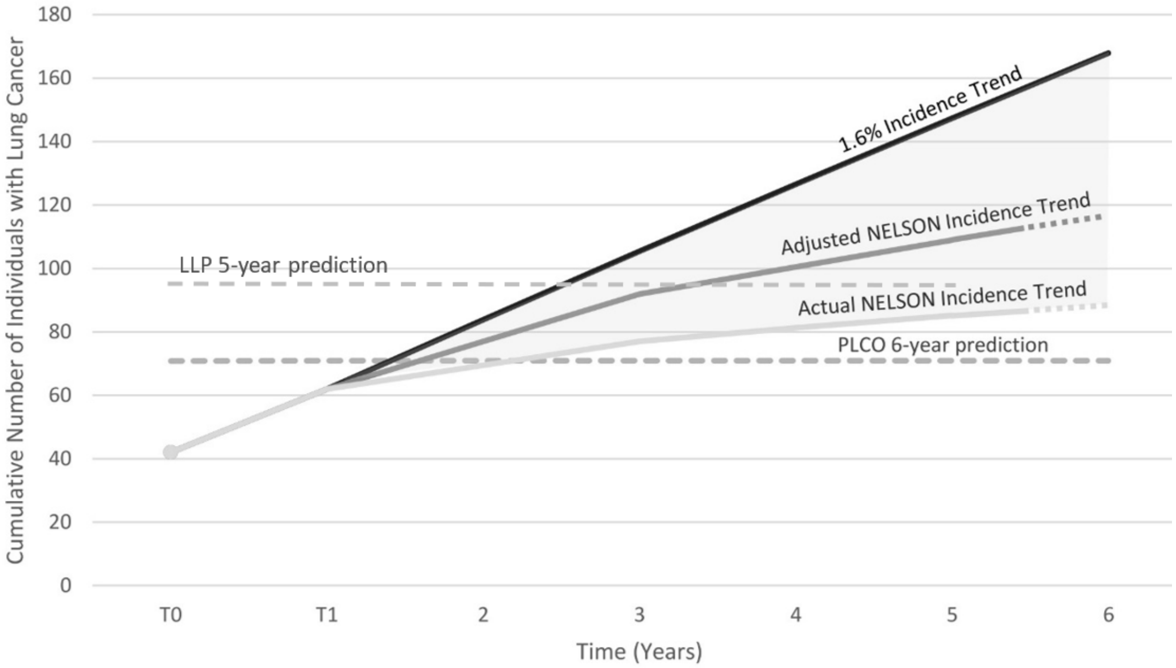


Figure 11. Projected cumulative number of individuals with lung cancer detected in the Manchester LHC pilot, with 6-years of follow-up. Adapted from [318], reuse permitted under CC BY-BC.

3.3.4 Impact of Selection Method on Screening Performance

3.3.4.1 Screening Performance in Manchester LHC Pilot

UKLS employed the LLP_{V2} RPM at a ≥5% threshold to assess eligibility for screening. Had this criterion been used in the LHC pilot, 42% fewer participants (n=826) would have been eligible for screening and 26% of cancers detected in the pilot would have been missed (n=16/62). Approximately half of those eligible by $PLCO_{M2012}$ ≥1.51% criteria would have received screening had LLP_{V2} ≥5% been used for selection (52.1%, n=745/1429), it is unknown how many additional cancers would have been diagnosed in the 81 participants who were only eligible under LLP_{V2} ≥5% criteria (Figure 12). The distribution of screen detected lung cancers by both $PLCO_{M2012}$ and LLP_{v2} is shown in Figure 13, most cancers were detected in the higher risk groups by both RPMs.

At a LLP_{V2} ≥2.5% threshold (as has been recommended alongside $PLCO_{M2012}$ ≥1.51% in the NHS TLHC programme [48]), a similar number of participants would have been eligible for screening (n=1430) as were by $PLCO_{M2012}$ ≥1.51%, although four of the detected cancers would have been missed (6.5%) (Figure 12). Lung cancer outcomes are unknown among the 272 participants who were eligible by LLP_{v2} ≥2.5% but ineligible by $PLCO_{M2012}$ ≥1.51%. This cohort was significantly older (66.7 ±4.9 vs 62.4 ±5.6 years, p<0.0001), had lower smoke exposure (pack-years 19.4 ±13.7 vs 37.8 ±15.5, p<0.0001) and had a lower proportion of current smokers (19% vs 39%, p<0.0001) than the lowest risk screened group in the pilot ($PLCO_{M2012}$ 1.51%–1.99%; n=265). Had both LLP_{V2} ≥2.5% and $PLCO_{M2012}$ ≥1.51% been used to ascertain screening eligibility, almost 20% more attendees (n=272) would have received screening.

Had NLST generalised eligibility criteria (aged 55-74, ≥30 pack-year smoking history, smoker within 15 years) been used for screening selection, 17% fewer participants would have been selected for screening (n=1188) and 18% of the cancers detected in the pilot would have been missed (n=11) (Figure 12). The $PLCO_{M2012}$ ≥1.51% threshold selected 92.7% of those eligible by NLST criteria. It is unknown how many additional cancers would have been diagnosed in those only eligible by NLST criteria (n=94). Had no additional cancers been found in that subset, the detection rate in the NLST eligible pilot cohort would have been 4.3% (n=51/1188) overall, 2.8% at T0 (n=35/1188) and 1.5% at T1 (n=16/1084). This detection rate was more than double that seen in NLST, which saw an overall 1.7% detection rate (n=438/26309), 1% at T0 and 0.7% at T1 (p=0.0001) [193].

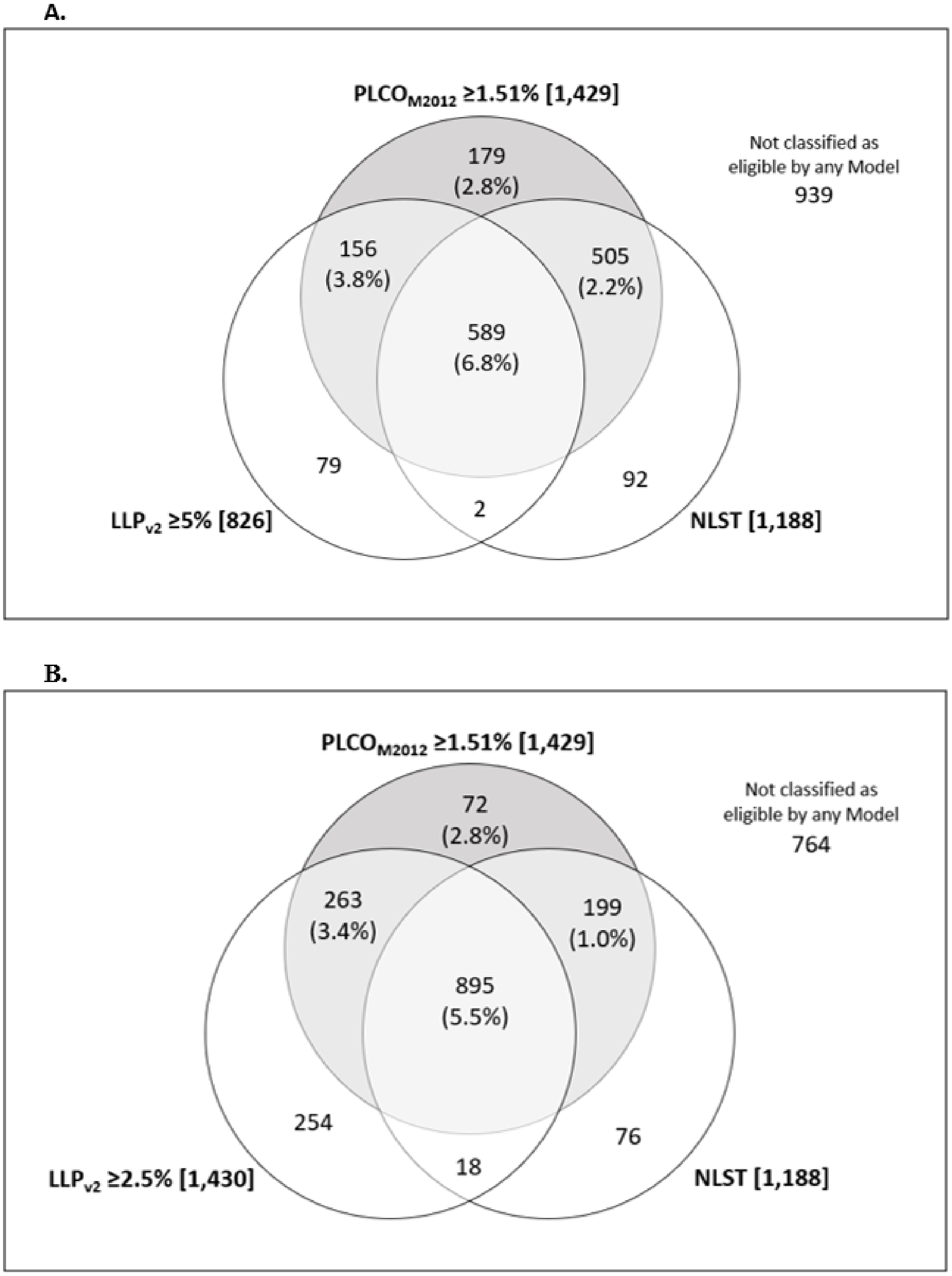

A.
$\text{PLCO}_{\text{M2012}}$ ≥1.51% [1,429]
Not classified as
eligible by any Model
939
179
(2.8%)
156
(3.8%)
505
(2.2%)
589
(6.8%)
79
92
2
LLP_{v2} ≥5% [826]
NLST [1,188]
B.
$\text{PLCO}_{\text{M2012}}$ ≥1.51% [1,429]
Not classified as
eligible by any Model
764
72
(2.8%)
263
(3.4%)
199
(1.0%)
895
(5.5%)
254
76
18
LLP_{v2} ≥2.5% [1,430]
NLST [1,188]

Figure 12. Participants of the Manchester LHC pilot eligible for screening by each set of eligibility criteria. Percentage in () = lung cancer detection rate within segment. Number in [] = number of participants eligible for screening. Shaded = eligible in LHC pilot. Reproduced from [318], reuse permitted under CC BY-NC.

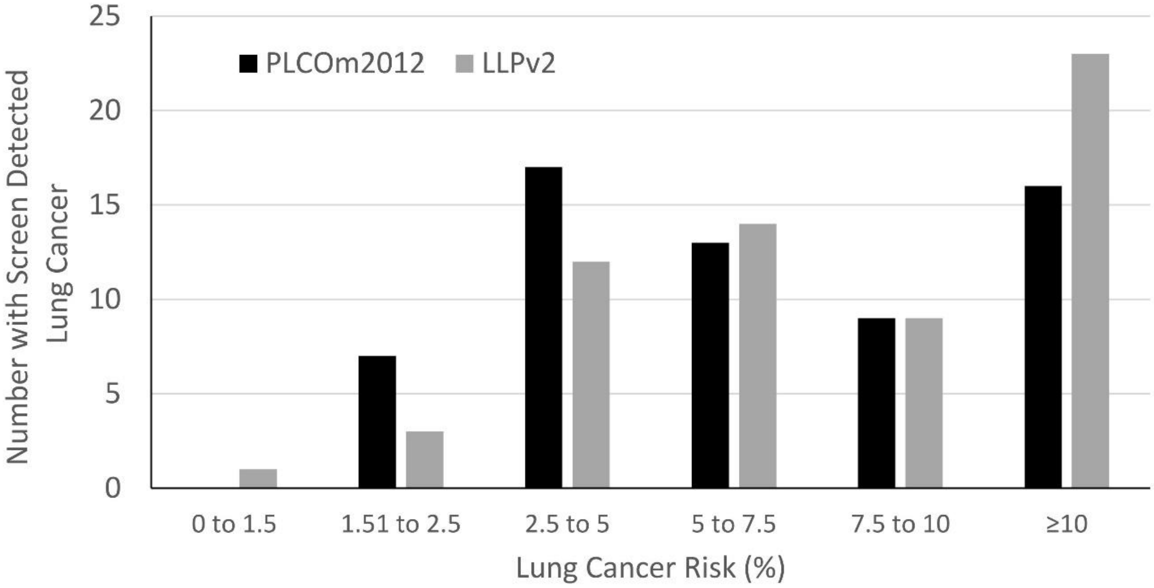


Figure 13. Distribution of screen-detected lung cancers in the Manchester LHC pilot according to risk score (PLCO$_{M2012}$ and LLP$_{V2}$ RPMs). Reproduced from [318], reuse permitted under CC BY-NC.

3.3.4.2 Manchester LHC Pilot vs. NLST Detection Rates

The analysis in section 3.3.4.1 revealed a large detection rate discrepancy between the Manchester LHC pilot and NLST. I contributed to an investigation of the factors responsible for this discrepancy through a joint project with Dr Hillary Robbins *et. al.*

In this study, detection ratios were calculated between several dataset iterations from NLST and the LHC pilot. The NLST dataset was restricted to only lung cancer detections from baseline, interval, and year one screenings to match the LHC pilot; the LHC pilot dataset was restricted to exclude NLST-ineligible participants. This resulted in 1079 LHC pilot participants with 51 detected lung cancers (4.7% detection rate), and 26,268 NLST participants with 454 detected cancers (1.7% detection rate), indicating 2.7-fold higher detection in the LHC pilot compared to NLST (95%CI 1.6-2.8). When NLST participants who did not meet the PLCO$_{M2012}$ ≥1.51% screening threshold were also excluded, the NLST detection rate increased to 2.1%, reducing the detection ratio to 2.2 (95%CI 1.3-2.3).

Even within this further restricted dataset, the risk distribution was significantly different; for example, 4.4% of LHC-eligible NLST participants had a PLCO$_{M2012}$ score >10%, compared to 13.6% of LHC pilot participants. Indirect standardisation was applied to adjust for this difference; detection frequencies at ten categories of baseline risk in NLST were calculated and applied to the same categories in the LHC pilot to determine how many cancers would be expected if there was matching risk-specific detection. Following this analysis, the detection ratio reduced, but remained notable, at 1.6 (95%CI 1.2-2.1).

The residual discrepancy was more pronounced in participants who were older (detection ratio across increasing 5-year age groups from 55-74: 0.8 → 1.5 → 1.7 → 2.1), more socio-economically deprived (IMD rank ≤1500 = 1.9 vs. IMD rank >3500 = 1.4), and who had lower FEV_1/FVC ratios (<60% = 2.3 vs. >70% = 1.3).

3.3.5 Prevalence of Cardiovascular and Respiratory Comorbidity

Table 2 details the demographic and clinical characteristics of the Manchester LHC pilot attendees stratified by lung cancer risk. As expected, variables used as predictors of lung cancer in the $PLCO_{M2012}$ RPM are closely associated with risk score. For example, mean age increases from 62.4 years in the lowest screening-eligible risk group ($PLCO_{M2012}$ score: 1.51%-1.99%) to 69.7 in the highest risk group ($PLCO_{M2012}$ score: ≥10%) ($p<0.001$). Percentage of current smokers increases from 39% in the lowest group, to 77% in the highest group ($p<0.001$). Percentage of attendees with a first-degree relative with lung cancer increases from 17% in the lowest group, to 44% in the highest group ($p<0.001$).

Other clinical factors which are not considered in the $PLCO_{M2012}$ RPM also demonstrated an association with lung cancer risk score in this study cohort. Half (n=1273) of all LHC attendees reported at least one respiratory symptom (breathlessness - MRC dyspnoea scale ≥2, cough ≥6 weeks in duration or sputum production ≥teaspoon/day). In the highest $PLCO_{M2012}$ risk group the prevalence was 76%, double the 38% prevalence in the screening-ineligible <1.51% risk group ($p<0.001$). Respiratory symptoms were predictive of lung cancer diagnosis in univariable analysis (OR 2.37, 95%CI 1.16-4.85; $p=0.02$) (Table 3).

Table 3. Univariable and multivariable analysis of the risk of screen detected lung cancer stratified by clinical variables and PLCO$_{M2012}$ score.

Variable		**Number**	**Risk of Lung Cancer**			
			Univariable		**Multivariable**	
			OR (95% CI)	p value	$_{adj}$OR$^{\#}$ (95% CI)	p value
Emphysema	No	504	1	-	-	-
	Yes	883	1.12 (0.65-1.92)	0.68	-	-
CAC	None	328	1	-	1	-
	Mild	456	2.44 (1.03-5.75)	0.042	2.38 (0.996-5.67)	0.051
	Mod-Sev	429	2.84 (1.21-6.65)	0.016	2.62 (1.11-6.20)	0.029
Symptomatic	No	393	1	-	1	-
	Yes	1007	2.37 (1.16-4.85)	0.018	1.55 (0.74-3.28)	0.25
FEV$_1$/FVC ratio		1,400	0.97 (0.95-0.99)	0.001	0.98 (0.96-1.01)	0.19
PLCO$_{M2012}$ score		1,409	1.13 (1.08-1.17)	<0.0001	1.11 (1.05-1.16)	<0.0001

Poor lung function also showed a robust association with lung cancer risk. Almost three-quarters of those with a risk score of ≥10% had airflow obstruction (defined as FEV$_1$/FVC ratio <0.7), compared to 21% in those at the lowest risk of lung cancer. Overall, 37% of LHC attendees had airflow obstruction. FEV$_1$/FVC ratio was inversely correlated with lung cancer risk score (Table 2, Figure 14; $p<0.001$). FEV$_1$/FVC ratio was predictive of lung cancer diagnosis in univariable analysis (OR 0.97, 95%CI 0.95-0.99; $p=0.001$). When adjusted for PLCO$_{M2012}$ score the predictiveness had borderline statistical significance ($_{adj}$OR 0.98, 95%CI 0.96-1; $p=0.07$); when adjusted for the full multivariable model the predictiveness was not statistically significant (Table 3). The prevalence of scan-detected emphysema also increased across the lung cancer risk groups. Of total scan reports, 64% (n=883/1387) had evidence of emphysema. This ranged from 54% in the lowest screening-eligible risk group, to 77% in the highest ($p<0.001$).

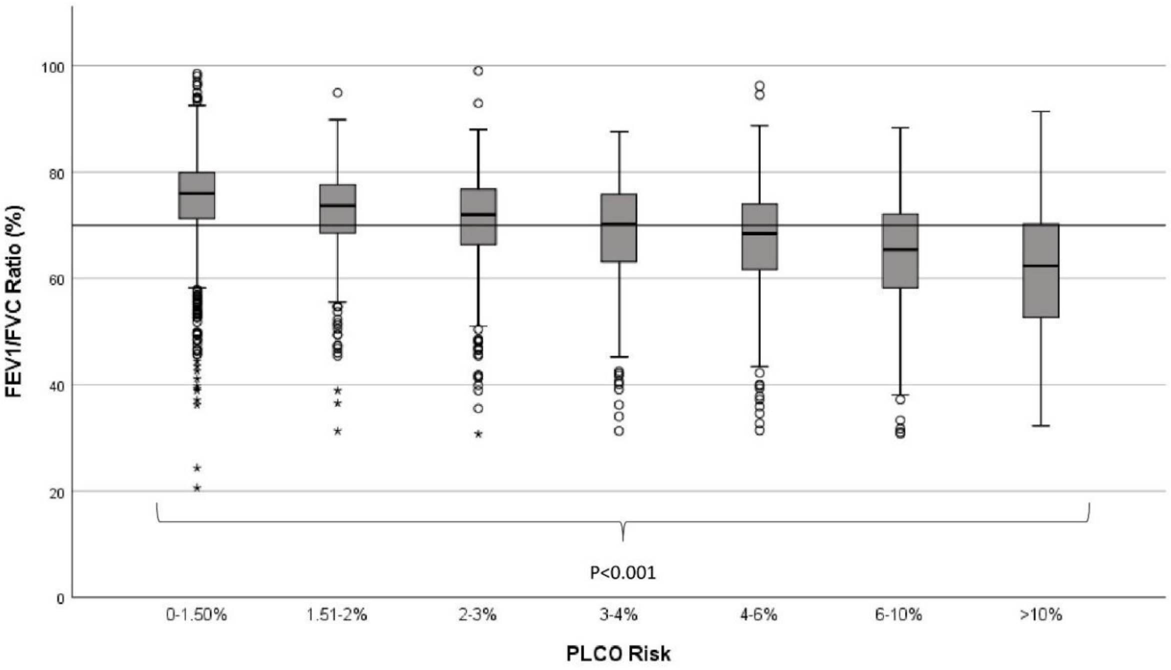


Figure 14. FEV_1/FVC ratio stratified by lung cancer risk group ($PLCO_{M2012}$). Horizontal line represents 70% threshold indicative of airflow obstruction. Reproduced from [318], reuse permitted under CC BY-NC.

Of the 939 T1 attendees who completed a cardiovascular disease (CVD) questionnaire, 22% reported history of CVD (at least one of myocardial infarction, angina, stroke, or transient ischemic attack). Amongst those in the highest lung cancer risk group, 30% reported CVD history, compared to 18% in the $PLCO_{M2012}$ 1.51%-1.99% risk group. The presence of CAC also increased with lung cancer risk; 80% of the highest risk group had CAC compared to 66% in the lowest risk group. Almost 20% of those in the $PLCO_{M2012}$ ≥10% group had severe CAC. Even after adjustment for the multivariable model, CAC was predictive of lung cancer diagnosis with statistical significance (adjOR 2.50, 95%CI 1.11 to 5.64; p=0.028), this association was higher in those with moderate and severe CAC (n=429) ($_{adj}$OR 2.62, 95%CI 1.11- 6.20; p=0.029) compared with those with mild CAC (n=456) ($_{adj}$OR 2.38, 95%CI 0.996-5.67; p=0.051).

3.4 Discussion

In this chapter, I tested the hypothesis that current screening selection strategies are not optimised for use in lung cancer screening programmes catering for socio-economically disadvantaged populations. The main driver of this hypothesis is that screening selection strategies and tools were developed in populations unrepresentative of the highly socio-economically deprived populations most likely to be targeted by community-based lung cancer screening programmes. I used RPM calibration, cancer detection rates, and prevalence of comorbidities in the Manchester LHC pilot as outcomes in assessing this hypothesis.

3.4.1 Calibration

My analysis of the calibration of the $\text{PLCO}_{\text{M2012}}$ RPM, which was used for establishing screening eligibility in the Manchester LHC pilot, found that during only 15 months of screening and associated follow-up, 88% of the predicted 6-year lung cancer incidence rate was reached. Using hypothetical detection rate projections based on NELSON, there was an estimated 1.6-fold increase in lung cancer incidence over 5.5 years of screening in our population compared with the incidence predicted by $\text{PLCO}_{\text{M2012}}$. Analysis of LLP_{V2} calibration also found likely risk underestimation, although by a smaller magnitude than that of $\text{PLCO}_{\text{M2012}}$. A possible interpretation of these findings is that these RPMs underestimate risk in a socio-economically deprived, high-risk population. This has significant implications, as it could result in high-risk participants being deemed ineligible for screening. The higher risk predictions and lower level of overestimation seen in LLP_{V2} may be partially explained by its development in a deprived Liverpool-based population, more akin to the Manchester LHC pilot population than the North American RCT-based population that the $\text{PLCO}_{\text{M2012}}$ RPM was developed in.

A key limitation of this analysis is that the RPMs predict risk in the absence of LDCT screening, whereas we observed detection in the presence of screening. Screening increases incidence over a short-term period by moving lung cancer diagnoses earlier in time, leading to inflated detection rates and potential biases relating to overdiagnosis and lead-time, the magnitude of which are debated in lung cancer screening (see section 1.6.5). Other limitations include that dropout and participation rates, and mortality from other causes, were not considered when calculating projected cancer diagnoses. It may be that when these factors are taken into account, detection rates drop significantly in later LHC screening rounds, correcting for the initial miscalibration.

However, I present further confirmatory evidence of $\text{PLCO}_{\text{M2012}}$ underestimation in this cohort, derived from 5-year follow-up of a subset of the LHC pilot cohort, in section 4.3.4. Other studies have also demonstrated $\text{PLCO}_{\text{M2012}}$ underestimation in certain high-risk subgroups; for example, a US based study found that the RPM underestimated risk by a factor of two to three in Hispanic smokers [319]. Whilst a UK based study compared several RPMs (including $\text{PLCO}_{\text{M2012}}$ and LLP_{V2})

and found overestimation of risk for all models, its analysis took place in biobank settings with participants from high socio-economic backgrounds; the study emphasised that in a less affluent population more representative of high-risk lung cancer screening attendees, the extent of risk overestimation would likely diminish [225]. It is interesting to note that in that study, LLP_{V2} had the highest overestimation of risk whilst $PLCO_{M2012}$ had only slight overestimation. This may be congruent with our results, which showed smaller underestimation of risk in LLP_{V2} compared to $PLCO_{M2012}$. Another UK study applied $PLCO_{M2012}$ and LLP_{V2} to routinely collected primary care data (thereby limiting bias from screening-driven increased incidence) and similarly found that the RPMs overestimated risk for those at lowest risk (comparable to a biobank setting), and underestimated risk for those at highest risk (more comparable to an LHC setting) [320].

If these RPMs are confirmed to underestimate risk in high-risk populations such as that of the Manchester LHC pilot, this would highlight the need for a recalibration of the models to facilitate optimal performance and ensure screening does not exacerbate health inequalities [229,230]. Formal calibration of risk models would require at least 100 events, which is beyond the scope of this study [321]. However, the TLHC rollout and other large screening implementation studies provide ample opportunity for such recalibrations. For example, LLP_{V3} was successfully recalibrated using countrywide health data [224], and demonstrated much improved calibration over LLP_{V2} in comparative studies [225]. However, LLP_{V2} remains the TLHC recommended RPM to date. The prospective YLST will also provide an opportunity to examine RPM calibration in more detail [254].

The way in which RPMs are applied in screening populations with differing demographic factors to the development population is also a crucial factor for consideration. For example, $PLCO_{M2012}$ was originally built using North American educational and ethnic classifications which are not easily transferrable to a UK setting. Consequently, this risk factor is often omitted when the RPM is used in a non-native setting, potentially lowering the selection performance. Some research has been undertaken in the Manchester LHC pilot dataset to examine the impact of using a $PLCO_{M2012noRace}$ model; the study found similar overall performance when race was included or excluded (there were no significant differences in scan numbers and no lung cancers would have been missed), although risk scores were disproportionally impacted in non-white participants [322]. Further research is required, as the majority of the LHC pilot participants were white, precluding detailed analysis. The method by which educational attainment metrics are converted between North American and UK populations is also an important factor to consider, as this is a key surrogate marker for socioeconomic deprivation, which may not be appropriately captured by the RPM when applied in the UK population. Further research, and model recalibration in the setting of intended use, is required as screening programmes develop further.

In summation, the seeming miscalibration of both RPMs used in the Manchester LHC pilot supports the hypothesis that screening selection strategies perform sub-optimally in socio-economically disadvantaged populations. More research is required to assess and recalibrate RPMs in actual screening implementation programmes.

3.4.2 Detection

The second outcome measure I analysed was the number of participants selected for screening and associated cancer detection. There was significant variability between the selection methods. For example, one in four of those with screen detected lung cancer in our cohort would not have qualified for screening in UKLS, which used the LLP_{v2} model at a threshold ≥5%. The lower LLP_{v2} threshold would have classified a similar number of people as screening eligible as $PLCO_{M2012}$, and among the individuals who were screened, it selected 93.5% of those with cancer. It is not possible to establish whether this detection discrepancy would have been compensated for by additional cancers found in the 272 participants only eligible by LLP_{V2}. If eligibility was established by both $PLCO_{M2012}$ and LLP_{v2}, the number screened would have increased by almost 20%. The retrospective nature of our analysis means that we are unable to determine how many of those 20% would have been diagnosed with lung cancer, and therefore, the potential benefit or detriment of concurrently using these two different criteria. As an aside, Previous studies have shown that screening performance may be superior in women [197]; particularly notable, the NELSON RCT found a significantly higher LDCT derived mortality reduction in women compared to men, although the small size of the female subgroup in NELSON limits the certainty of this conclusion [194]. Further research is required to investigate whether this discrepancy indicates that women are having their risk underestimated by the $PLCO_{M2012}$ RPM.

The relatively small difference in performance between $PLCO_{M2012}$ ≥1.51% and LLP_{v2} ≥2.5% criteria is consistent with the UK-based comparative study which demonstrated modest differences in discriminatory ability between the RPMs tested [225]. Other studies report that $PLCO_{M2012}$ outperforms LLP_{V2} on discrimination and calibration [319]. Prospective studies are required to elucidate this further.

RPMs may be augmented with the inclusion of additional clinical variables. We investigated a number of objective measures of ‘smoking damage’ (FEV_1/FVC ratio, CAC and presence of emphysema on LDCT), not included in the $PLCO_{M2012}$ model, to see if they were independently predictive of lung cancer after controlling for $PLCO_{M2012}$ risk score. The presence and degree of CAC remained an independent predictor of lung cancer. If this association is confirmed in other datasets the value of CAC to inform screening strategies, after the baseline round, should be evaluated further. The ratio of FEV_1/FVC showed borderline statistical significance to predict lung cancer after adjusting for $PLCO_{M2012}$ score. Previous studies have shown that the inclusion of lung function test results in lung cancer risk prediction significantly improved its precision [323,324]. As spirometry tests are already conducted as part of the LHC programme (and the results are immediately available), lung function measures could be factored into lung cancer risk prediction and screening selection with ease. A study in the NLST and MILD datasets found that CT derived measures of CVD and COPD may provide small improvements to lung cancer risk prediction in some cases, although validation results were inconsistent [325]. Further research in additional large and diverse cohorts is warranted to confirm whether these factors can improve risk prediction.

NLST eligibility criteria would have missed almost 20% of the screen-detected cancers in our cohort, but would also have screened fewer individuals, resulting in only a marginally lower cancer detection rate than $\mathrm{PLCO}_{\mathrm{M2012}}$. In contrast, interim analysis from the International Lung Screening Trial (ILST) found that when setting the $\mathrm{PLCO}_{\mathrm{M2012}}$ threshold to select the same number of individuals for screening as USPSTF_{2013}, $\mathrm{PLCO}_{\mathrm{M2012}}$ cancer detection was significantly higher (94.7% of total cancers vs. 78.9%; $p=0.0001$) [326]. Other studies have shown RPMs to be superior to NLST criteria for sensitivity and specificity [327].

The high-risk nature of our cohort was also evident when we limited analysis to only those eligible for screening based on NLST criteria, as the lung cancer detection rate was more than double that seen in NLST. Further analysis revealed that the generalised eligibility criteria employed in the NLST was an important factor in the lower detection rate, with the older, more comorbid, and more socio-economically deprived profile of the LHC pilot cohort likely responsible for some of the residual difference. This emphasises the impact RPMs can have on detection rates, as well as the potential added benefit that may be derived by targeting screening to socio-economically disadvantaged populations.

A significant limitation of this analysis is that it was not possible to determine the outcomes of LHC pilot attendees who were ineligible for screening using the $\mathrm{PLCO}_{\mathrm{M2012}}$ threshold but were eligible using $\mathrm{LLP}_{\mathrm{v2}}$ or NLST criteria. Further work is needed to prospectively evaluate the performance of selection methods in high-risk screening populations. The randomised controlled YLST (see section 1.7.1.2) is prospectively evaluating $\mathrm{PLCO}_{\mathrm{M2012}}$ (≥1.51%), $\mathrm{LLP}_{\mathrm{v2}}$ (≥5%) and NLST criteria (extended to age 80) in the context of a TLHC programme, and will provide important information for the optimal approach for screening selection in a UK screening programme [254].

In summation, it seems that using RPMs rather than generalised eligibility criteria for screening selection results in increased efficiency and detection rates, although this improvement may be somewhat attenuated in very high-risk populations. Further research is required to prospectively evaluate which RPM and screening threshold results in the best performance, as well as whether including additional clinical variables in RPMs could improve risk prediction.

3.4.3 Comorbidity

Finally, we analysed the comorbidity profile of the screening cohort. Since the majority of lung cancers diagnosed through screening (for both $PLCO_{M2012}$ and LLP_{v2} risk models) were found in individuals occupying the higher risk groups, cancer detection rate can be increased dramatically by increasing the screening threshold. For example, setting the threshold to $PLCO_{M2012}$ ≥2% (as in the PanCan study) reduces those eligible for screening by 20%, but only 6.5% fewer cancers are detected. Although such a modification could improve screening efficiency, it would fail to take into account the potential reduction in screening efficacy through the selection of a more comorbid population. Similarly, analysis of ILST showed that whilst $PLCO_{M2012}$ improved screening efficiency over generalised eligibility criteria, it also selected an older and more comorbid screening cohort [326].

Our data underlines this by showing how increased lung cancer risk is associated with increased co-morbidity, such as established cardiovascular disease, the presence and severity of airflow obstruction as well as respiratory symptoms. Previous studies have highlighted that individuals with significant comorbidities may derive less benefit from screening, and selecting them may negatively impact the cost-effectiveness of the screening programme. For example, in colorectal cancer screening, individuals with diabetes at age 50 derive approximately 1.8 times fewer 'life years saved per-person' than their counterparts without diabetes [328]. Breast cancer screening for those aged over 79 may only be only cost-effective for the 25% of individuals with the highest life expectancy (based on chronological age and comorbidities) [329]. In socio-economically deprived areas where lung cancer screening is most warranted, disease burden is higher and baseline life expectancy is lower, increasing the potential impact comorbidity could have on the performance of risk-based screening selection [330].

3.4.4 Conclusion

In conclusion, this study supports the hypothesis that current screening selection strategies have some sub-optimal characteristics when applied to a highly socio-economically disadvantaged screening population. In addition to miscalibration and uncertainties surrounding RPM and threshold selection, a key issue is the high levels of comorbidity in the selected screening cohort, which may limit the benefit from screening derived by participants. One novel approach to addressing this challenge is selecting participants based on expected benefit from screening, rather than calculated risk. I investigated this approach in the next chapter.

Chapter Four - 'Life-Gained' Screening Selection: Study of Manchester LHC Pilot

4.1 Introduction

Lung cancer is responsible for a large proportion of cancer mortality worldwide [1]. Screening high-risk populations for lung cancer has been demonstrated to significantly reduce mortality [193,194]. Many lung cancer screening implementation studies and programmes use RPMs to calculate individual lung cancer risk and determine screening eligibility [223,248]. In the previous chapter, I demonstrated how this approach can result in a screening cohort with a high-risk profile, leading to favourable detection rates, but preferentially selects older participants with a higher burden of comorbidity who may be less likely to benefit from having lung cancer detected early. This is congruent with other studies that demonstrate an association between lung cancer risk and comorbidity [231,232]. If confirmed, this may indicate an important drawback to a risk-based approach to screening selection, which may disproportionately affect socio-economically disadvantaged populations who are often at higher risk of comorbidity [331].

Selecting screening participants based on predicted benefit from screening, rather than predicted lung cancer risk, has been proposed as a solution to the issue of comorbidity. One benefit-based tool is the Life Years Gained from Screening-CT (LYFS-CT) [233], a model that calculates an individual's projected days-of-life gained from undergoing screening. It does this by calculating an individual's expected life expectancy in the absence of LDCT screening using a validated mortality model and adjusting life expectancy based on the mortality benefit from screening, with the difference between the two 'life expectancies' constituting the 'life-gained' from screening (detailed description of model in section 1.6.6.3). It is important to note that lung cancer risk is a significant driver of estimated life-gained from screening, as the higher an individual's risk of lung cancer, the higher their chance of gaining benefit from having cancer detected early through screening. Even if an individual has a very favourable comorbidity profile, and therefore would benefit from having cancer detected early, this benefit will only be realised if there is a high enough likelihood that the early detection takes place. Modelling studies have indicated that LYFS-CT may maximise life-years gained from screening when compared to RPMs [233] and it has been recommended for use by The American College of Chest Physicians [235]. However, LYFS-CT has thus far not been tested in a real-world screening programme, an important step in investigating whether it has clinical utility.

This study aimed to compare risk-based and benefit-based selection strategies (see hypotheses and objectives – section 1.8). I did this by retrospectively applying LYFS-CT to the Manchester LHC pilot cohort and analysing the projected outcomes had benefit-based selection been used in this real-world screening setting.

4.2 Methods

4.2.1 Study Cohort and Data Collection

This study examined the Manchester LHC pilot cohort. The design, cohort recruitment protocol, and data collection procedures of the LHC pilot are described in detail in section 2.1. All data were collected prospectively by the clinical staff at the LHC site. In addition to collecting data relating to lung cancer risk and respiratory comorbidities for all participants, data relating to cardiovascular comorbidities including history of myocardial infarction, stroke, angina or transient ischaemic attack was collected in second round attendees only. For the purposes of this study, the mean value for each of these variables was imputed for the remaining cohort where data were not collected.

4.2.2 Analysis Approach

As this is the first time LYFS-CT has been applied to a real-world screening population, first, I provide a basic description of the distribution of life-gained scores and association with demographic factors in the cohort. Following this, I performed an in-depth comparison of the screening performance of risk-based vs. benefit-based selection ($\text{PLCO}_{\text{M2012}}$ vs. LYFS-CT).

Two approaches were taken in analysis. First, to examine the impact a life-gained selection approach would have on screening selection and subsequent cancer detection rate in the LHC pilot, I used the equivalent LYFS-CT threshold to select the same number of individuals eligible for screening as were by $\text{PLCO}_{\text{M2012}}$ ≥1.51%. I examined the differences in number of cancers detected and demographic and clinical characteristics of those who would have been selected by $\text{PLCO}_{\text{M2012}}$ vs. those who would have only been selected by LYFS-CT. I compare 5-year mortality rates between the two subgroups, follow-up comorbidity data were not collected for this analysis.

Second, because comprehensive endpoints were collected only on screened individuals ($\text{PLCO}_{\text{M2012}}$ ≥1.51%), I stratified screened individuals into ‘high risk’ (top 75% of $\text{PLCO}_{\text{M2012}}$ scores) and ‘low risk’ (bottom 25% of $\text{PLCO}_{\text{M2012}}$ scores) groups, and ‘high benefit’ (top 75% of LYFS-CT scores) and ‘low benefit’ (bottom 25% of LYFS-CT scores) groups. These groups were cross-tabulated, resulting in four subgroups: A) high risk/high benefit, B) high risk/low benefit, C) low risk/high benefit and D) low risk/low benefit. I compared demographic characteristics and comorbidities of groups B and C (those ‘selected’ by $\text{PLCO}_{\text{M2012}}$ only vs. those ‘selected’ by LYFS-CT only) and followed up these two subgroups five years post-LHC to see if there was a divergence in mortality and/or development of further comorbidities. Finally, I examined model calibration by comparing the number of deaths and lung cancers in this subgroup after 5-years of follow-up with those predicted by $\text{PLCO}_{\text{M2012}}$ and LYFS-CT in that time period.

4.2.3 Life-Gained Calculations in LHC Pilot Cohort

The protocol and tools used for calculating life-gained scores is described in section 2.3. When LYFS-CT scores were calculated for the Manchester LHC pilot cohort, ethnicity was coded as 'white' for all participants to avoid issues of incorrect risk adjustment due to differing ethnic characteristics between North America and the UK [322], congruent with the protocol in the LHC pilot itself. Self-reported COPD/emphysema/bronchitis was used for the 'lung disease' variable. Missing variables were imputed to the population mean of the LHC pilot participants where data were collected. Coronary heart disease (CHD) was imputed according to angina and heart attack, using the National Health Interview Survey (NHIS) years 2013-2015 population mean as a reference (chance of CHD, rounded to nearest 25%: neither angina/heart attack = 0%; only angina = 50%; only heart attack = 50%; both angina & heart attack = 75%). Atrial fibrillation was used for the 'other heart condition' variable. Where data were not collected in the LHC pilot, variables were coded as not present (liver disease, special equipment and multiple relatives with lung cancer).

4.2.4 Manchester LHC Pilot Follow-up

The procedure for the collection of LHC pilot follow-up data is described in section 2.2. Outcomes were determined by individual primary care medical case record review. In addition to gathering data for those with lung cancer detected in the LHC pilot, data were also gathered for those who were selected by one selection method but not the other (either $\text{PLCO}_{\text{M2012}}$ or LYFS-CT) in either of the two analyses performed. Comorbidities with data available at both the LHC pilot and follow-up were categorised as 'severe' (diabetes mellitus, heart attack, CHD, stroke, previous cancer and kidney disease) or 'other' (COPD, pneumonia, AF, angina and hypertension) according to their published relative risks for mortality [233]. Other chronic or potentially life-shortening conditions with data only available at follow-up were labelled as 'serious'. Minor comorbidities with data only available at follow-up were not analysed.

4.2.5 Model Calibration Calculations

To assess LYFS-CT calibration in this cohort, I used the mortality model component within LYFS-CT to calculate the probability of each individual surviving until the end of the follow-up period based on their age and comorbidities at the LHC. Within the subset of the cohort with follow-up mortality data available, we summed all individuals' risk of dying in the 5-year follow-up period resulting in the total number of predicted deaths in that group. We then compared this with the number of observed deaths in the subgroup. To assess $\text{PLCO}_{\text{M2012}}$ calibration, we summed the $\text{PLCO}_{\text{M2012}}$ risk scores (as proportions not percentages) within the subgroup to result in an expected number of lung cancer diagnoses in 6-years, comparing the total with the observed number of lung cancer diagnoses.

4.2.6 Analysis

Statistical analysis was carried out using RStudio (v.3.6.3) and IBM SPSS Statistics (v.25). Comparison of medians for non-parametrically distributed variables was conducted by Mann Whitney U test. T-test or ANOVA was used for comparison of means. Differences in comorbidities were analysed by comparison of means, despite being non-normally distributed, to allow for more granular comparison of divergence. Differences in frequencies was tested by Chi-squared.

4.3 Results

4.3.1 Overview and Comparison of LYFS-CT and PLCO$_{M2012}$

The distribution of all LYFS-CT scores in the 2541 participants of the LHC pilot are presented in Figure 15. Median estimated days of life gained by three rounds of LDCT screening (LYFS-CT) was 14.8 (range: 0-114.5). As this was the first application of LYFS-CT to a real-world screening programme, there are no previous studies to compare this distribution to or to apply theoretical screening thresholds from. It is important to note that the 'days of life gained' metric 'spreads' the expected screening benefit across the population. For example, an individual with 2% chance of developing lung cancer based on an RPM calculation, and 20 days of life-expectancy gained from undergoing screening based on LYFS-CT, has a 98% chance of gaining no days of life (if lung cancer is not detected) and a 2% chance of gaining 1000 days of life (if lung cancer is detected).

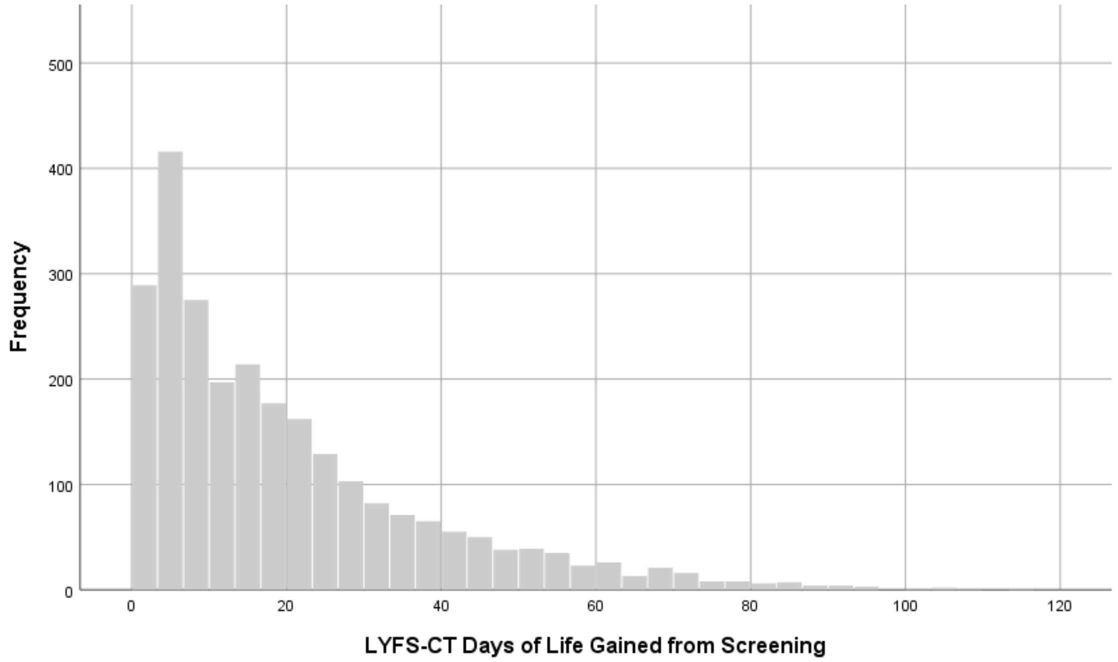


Figure 15. Distribution of LYFS-CT scores for all attendees of the Manchester LHC pilot. LYFS-CT calculates estimated days of life gained from lung cancer screening based on life expectancy, lung cancer risk, and screening mortality benefit.

Women had lower LYFS-CT scores than men, with borderline statistical significance (median: 14.2 vs. 15.2 days; p=0.08). Those in the younger half of the cohort (<64 years) had significantly lower LYFS-CT scores (median: 13.8 vs. 16.2; p<0.001). Age, and other factors associated with lung cancer risk (such as smoking and family history of lung cancer), are significant drivers of estimated life-gained from screening (see Table 4 for comparison of variables in PLCO$_{M2012}$ and LYFS-CT), as benefit is only conferred if there is an adequate chance of lung cancer being diagnosed. Consequently, PLCO$_{M2012}$ score and many of the risk variables that contribute to this show a strong association with increasing LYFS-CT score (Table 5). There is a strong positive correlation between the two scores (Pearson Correlation: 0.89; p<0.001) (Figure 16).

Table 4. Comparison of the variables included in PLCO$_{M2012}$ and LYFS-CT. Below delineation = variables unique to each model.

PLCO$_{M2012}$	**LYFS-CT**
Age	Age
Smoking duration	Smoking Duration
Family history of lung cancer	Family history of lung cancer
COPD diagnosis	Lung Disease
Cigs per day	Cigs per day
Smoking quit time	Smoking quit time
Ethnicity	Ethnicity
Education	Education
BMI	BMI
Previous cancer diagnosis	Sex
Smoking status (current vs. former)	Pack Years
	Hypertension
	Coronary Heart Disease
	Angina
	Heart Attack
	Other Heart Disease
	Stroke
	Diabetes
	Chronic Bronchitis in Past Year
	Kidney Failure
	Liver Disease
	Requires Special Equipment
	Year of Assessment

Table 5. Demographic and clinical characteristics of the complete Manchester LHC Pilot, stratified by LYFS-CT quintile.

	All	**LYFS-CT Quintile**					P-value
		1	2	3	4	5	
Number of Subjects	2541	508	508	508	509	508	-
Median LYFS-CT Score - days of life gained from screening ± IQR (Min-Max)	14.8 ±22.3 (0-114.5)	3.2 ±1.7 (0-4.9)	7.3 ±2.8 (4.9-10.5)	14.8 ±4.1 (10.5-19.1)	24.5 ±6.6 (19.1-33)	46.9 ±20.5 (33-114.5)	<0.001
Median PLCO$_{M2012}$ Score ± IQR (Min-Max)	1.7 ±3.4 (0-33)	0.2 ±0.4 (0-1.4)	0.8 ±0.6 (0-3.2)	1.7 ±0.9 (0-6.9)	3.3 ±1.7 (0-10)	7.3 ±5.7 (0.1-33)	<0.001
% Screened (n)	55.5 (1409)	0	11.2 (57)	71.3 (362)	97 (492)	98 (498)	<0.001
Mean Age ± SD	64.1 ±4.9	62.6 ±5.2	64.7 ±5.9	63.7 ±5.6	63.8 ±5.4	65.9 ±4.9	<0.001
% Female (n)	51 (1296)	53.3 (271)	52.3 (266)	48.2 (245)	53.2 (271)	47.8 (243)	0.19
% Lung Disease (n)	22.2 (565)	5.5 (28)	12.2 (62)	16.5 (84)	26.9 (137)	50 (254)	<0.001
% Family History LC (n)	21.7 (551)	11.6 (59)	18.5 (94)	20.1 (102)	21.6 (110)	36.6 (186)	<0.001
% Asbestos Exposure (n)	24.1 (612)	19.5 (99)	25.6 (130)	24.6 (125)	24.6 (125)	26.2 (133)	0.04
% Previous Cancer (n)	11.5 (291)	12.2 (62)	11.4 (58)	11.6 (59)	10.6 (54)	11.4 (58)	0.96
Median BMI ±IQR	28.5 ±6.9	29.6 ±7.7	28.9 ±6.7	29.6 ±7.1	27.9 ±6.3	26.3 ±6.4	<0.001
Mean Cigs per Day ±SD	20.4 ±13	14.2 ±9.1	17.1 ±11.8	22.5 ±16.1	23.2 ±13.1	25.2 ±10.4	<0.001
Median Years Smoked ±IQR	38 ±22	14 ±14	29 ±15	39 ±11	44 ±8	50 ±7	<0.001
Median Years Quit ±IQR	7 ±23	32 ±17	17 ±22	7 ±16	0 ±6	0 ±0	<0.001
% Prior Pneumonia (n)	14.2 (360)	9.6 (49)	14 (71)	13.4 (68)	14 (71)	19.9 (101)	<0.001
Median IMD Rank ±IQR	2873 ±5873	3126 ±8753	3070 ±8324	2871 ±5319	2872 ±3893	2693 ±3853	<0.001
Median FEV$_1$/FVC Ratio ±IQR	72.7 ±11.8	76.8 ±8.3	73.7 ±9.3	73.5 ±9.2	69.8 ±12.2	65.3 ±14.5	<0.001
Mean Hypertension*	0.5 ±0.3	0.49 ±0	0.5 ±0.35	0.54 ±0.35	0.47 ±0.4	0.46 ±0.38	<0.001

±SD							
Mean Angina* ±SD	0.12 ±0.2	0.12 ±0	0.12 ±0.1	0.13 ±0.24	0.11 ±0.25	0.11 ±0.23	0.42
Mean Heart Attack* ±SD	0.09 ±0.18	0.1 ±0	0.1 ±0.11	0.1 ±0.22	0.09 ±0.22	0.09 ±0.22	0.3
Mean Stroke* ±SD	0.02 ±0.09	0.02 ±0	0.02 ±0.04	0.02 ±0.1	0.02 ±0.1	0.02 ±0.1	0.92
Mean Diabetes* ±SD	0.19 ±0.24	0.19 ±0	0.21 ±0.14	0.24 ±0.32	0.19 ±0.32	0.15 ±0.25	<0.001
Mean Severe Comorbidities ±SD	0.66 ±0.63	0.66 ±0.33	0.69 ±0.44	0.74 ±0.74	0.62 ±0.76	0.6 ±0.73	0.002
Mean Other Comorbidities ±SD	1.05 ±0.75	0.84 ±0.4	0.97 ±0.56	1.05 ±0.79	1.06 ±0.87	1.34 ±0.92	<0.001
Mean Total Comorbidities ±SD	1.7 ±1.1	1.5 ±0.52	1.66 ±0.78	1.79 ±1.3	1.68 ±1.38	1.93 ±1.4	<0.001

*Means used for comorbidities as data were imputed for attendees of the first screening round where data was not collected (see section 4.2.3).

However, as $PLCO_{M2012}$ risk score increases to very high levels, life-gained estimations plateau. To illustrate, average LYFS-CT score increases to 50 days of life gained in the $PLCO_{M2012}$ 0-10% range, but only increases a further 25 days of life gained between $PLCO_{M2012}$ 10% and 32%. This is mediated by the increased burden of comorbidity and lower life expectancy present in those with the highest lung cancer risk scores (Figure 16). This can also be seen when examining LYFS-CT by quintile, with variables such as diabetes, stroke and total number of severe comorbidities being higher in the lower LYFS-CT subgroups (Table 5). In line with these observations, those who underwent screening in the LHC pilot due to having high lung cancer risk scores had higher LYFS-CT scores (median LYFS-CT: 25.5 vs. 5.3; p<0.001) (Figure 17). However, there was significant overlap in LYFS-CT scores between the upper end of the non-screened group and the lower end of the screened group, as well as a minority of individuals with significantly higher LYFS-CT scores who were not screened. This indicates that whilst we would expect significant congruence between screening selection by each method, there would be some divergence. I examine this divergence in more detail in the sections that follow.

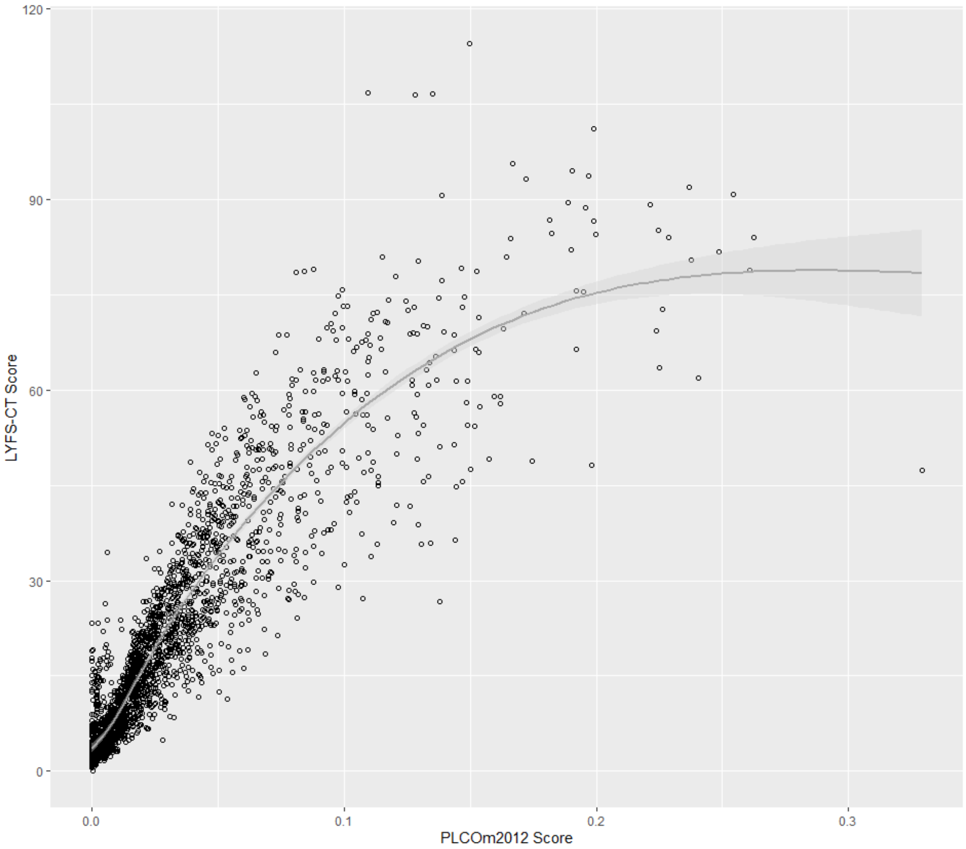


Figure 16. PLCO$_{M2012}$ scores vs. LYFS-CT scores for complete LHC pilot cohort. PLCO$_{M2012}$ scores presented in proportion form as this is how the *lcmodels* package generates the scores. These were converted to percentages in the descriptions to retain consistency with other chapters.

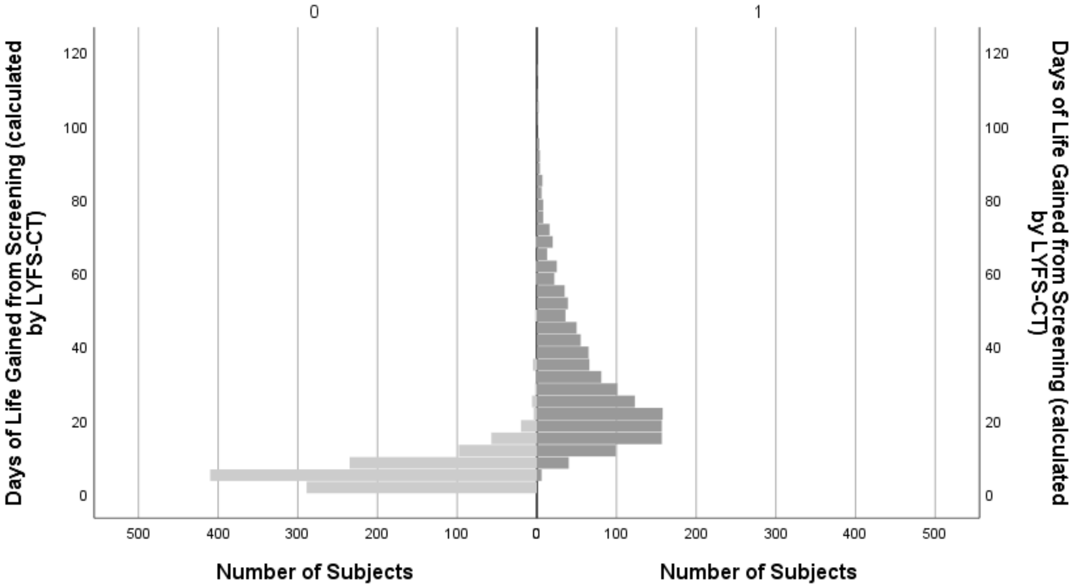


Figure 17. LYFS-CT score distribution stratified by screening status in the LHC pilot. 1 = had LDCT scan in LHC pilot. 0 = did not have LDCT scan in LHC pilot. LYFS-CT calculates estimated days of life gained from lung cancer screening based on life expectancy, lung cancer risk, and screening mortality benefit.

4.3.2 Equivalent Eligibility Thresholds

In this analysis, I simulated a LYFS-CT threshold to select the same number of participants for screening as were by $PLCO_{M2012}$ ≥1.51% in the LHC pilot. Recalculation of $PLCO_{M2012}$ scores in this study found one participant who was incorrectly classified as ineligible in the LHC pilot; we classified this individual by their recalculated risk score in this chapter but retain the 'real-world' calculation in Chapter Three as that was a retrospective study of actual screening performance in the LHC pilot. Thus, eligibility criteria of $PLCO_{M2012}$ ≥1.51% resulted in the selection of 1430 participants for screening (56.3% of the 2541 attendees). A LYFS-CT threshold of ≥12.2 days of life-gained from screening would have resulted in the same number of participants being selected. In the total cohort, 1322 participants (52%) would have been selected by both criteria, 1003 (40%) would have been selected by neither, and 216 (8.6%) would have been selected by either LYFS-CT or $PLCO_{M2012}$.

A representation of a portion of the cohort, with screening thresholds by both models, is presented in Figure 18. Of those selected by $PLCO_{M2012}$, 92.4% would have been eligible for screening by LYFS-CT ≥12.2. Of the 62 lung cancers diagnosed in the LHC pilot, 61 were in the 'Selected by Both' segment, and 1 was in the 'Selected by $PLCO_{M2012}$ Only' subgroup. Those to the left of the $PLCO_{M2012}$ threshold did not receive LDCT screening in the LHC and we are therefore unable to ascertain if they had lung cancer at the LHC. However, in the 5-years post-LHC, there were no unscreened lung cancer diagnoses amongst the 108 subjects selected by LYFS-CT alone, indicating that it is unlikely any additional lung cancers would have been detected had this subset been eligible for screening. In total, more than 98% of cancers detected in the pilot, and all cancers with ≥5-year survival, would have been detected had LYFS-CT ≥12.2 been used as the screening threshold.

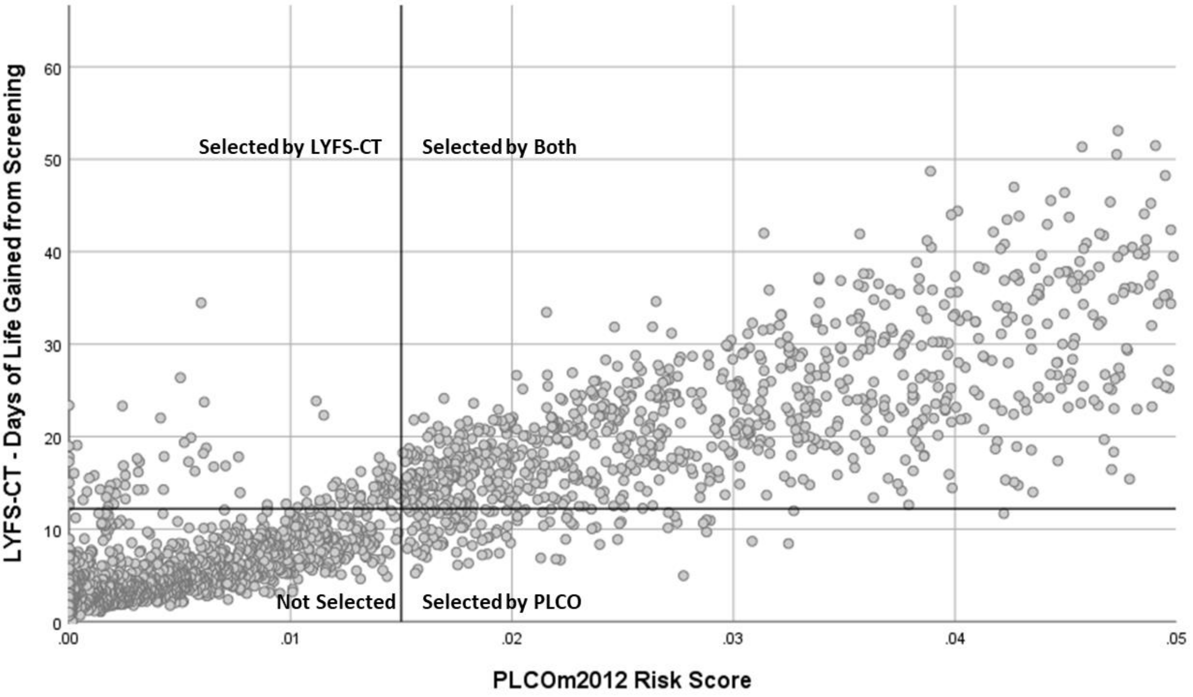


Figure 18. Distribution of LHC pilot participants by LYFS-CT and PLCO$_{M2012}$ scores. Vertical line = PLCO$_{M2012}$ 1.51% threshold. Horizontal Line = LYFS-CT 12.2 days of life gained threshold. Only participants with PLCO$_{M2012}$ score ≤5% displayed in figure to highlight differing selections.

The subgroup who would have been eligible solely by LYFS-CT criteria (n=108) had a median PLCO$_{M2012}$ risk score of 1%, compared to 1.8% in those only selected by PLCO$_{M2012}$ criteria ($p<0.001$). Conversely, those only selected by LYFS-CT (n=108) are projected to have an average of 14.6 days of added life-expectancy from screening compared to 10.3 days in those selected by PLCO$_{M2012}$ ($p<0.001$) (Table 6). There was no significant difference in 5-year mortality between those selected by only one criterion (PLCO$_{M2012}$-only vs. LYFS-CT-only: 6.5% vs. 8.3%; p=0.6), although it should be noted that this study was not powered to observe differences in mortality outcomes. However, there were differences in the comorbidity profiles between the two groups at the LHC pilot (Table 6). Those selected by LYFS-CT criteria only were significantly younger (median age: 63 vs. 69; $p<0.001$) and healthier; they were less overweight (median BMI: 28.3 vs. 30; p=0.02), had fewer comorbidities (median 1.2 vs. 2; p=0.001), and fewer severe comorbidities (median 0.54 vs. 1, $p<0.001$) than those only eligible by risk-based selection (medians could be non-integer values as the dataset included imputed variables). LYFS-CT selected participants were also less deprived (median IMD rank: 3785 vs. 2866; p=0.04) than those selected by PLCO$_{M2012}$; the two median ranks fell within the second most deprived and most deprived IMD deciles respectively.

Table 6. Demographics and calculated risk and benefit scores for the LHC pilot cohort, using the highest 1430 scores for PLCO$_{M2012}$ and LYFS-CT, amongst the 2541 participants of the LHC. Test for median included non-integer imputed values and could therefore be non-integers. Both median and mean values and P-values provided for comorbidities as both are considered in analysis.

		Not selected by PLCO$_{M2012}$ or LYFS-CT	**Selected by both PLCO$_{M2012}$ and LYFS-CT**	**Selected by LYFS-CT only**	**Selected by PLCO$_{M2012}$ only**	**P-value LYFS-CT only vs. PLCO$_{M2012}$ only**
Number of Subjects (%)		1003 (39.5)	1322 (52)	108 (4.3)	108 (4.3)	-
Median Age ±IQR		63 ±9	65 ±9	63 ±9	69 ±8	<0.001
Median BMI ±IQR		29.3 ±7.2	27.7 ±6.6	28.3 ±5.8	30 ±7.5	0.02
Median Cigs per Day ±IQR		15 ±10	20 ±12	10 ±24	20 ±5	0.01
Median Years Smoked ±IQR		20 ±18	45 ±10	39 ±15	36 ±8.4	0.13
Median Years Quit ±IQR		26 ±23	0 ±6	0 ±12.5	14 ±15	<0.001
% Female (n)		52.4 (526)	50 (661)	42.6 (46)	58.3 (63)	0.02
% Exposed to Asbestos (n)		22.2 (223)	25.2 (333)	23.1 (25)	28.7 (31)	0.35
% Previous Cancer Diagnosis (n)		9.5 (95)	11.4 (151)	5.6 (6)	36.1 (39)	<0.001
% COPD (n)		8.28 (83)	34 (449)	11.1 (12)	20.4 (22)	0.06
% Prior Pneumonia (n)		11.5 (115)	16.4 (217)	9.3 (10)	16.7 (18)	0.1
% Family History of LC (n)		14.1 (141)	27.2 (360)	17.6 (19)	28.7 (31)	0.05
Median IMD Rank ±IQR		3109 ±8463	2866 ±3615	3785 ±6387	2866 ±4195	0.04
Severe Comorbidities	Mean ±SD	0.63 ±0.3	0.65 ±0.79	0.59 ±0.23	1.1 ±0.76	<0.001
	Median ±IQR	0.53 ±0	0.54 ±1	0.54 ±0	1 ±1	<0.001
Other Comorbidities	Mean ±SD	0.88 ±0.45	1.18 ±0.9	0.89 ±0.4	1.2 ±0.98	0.003
	Median ±IQR	0.68 ±0	1 ±1	0.69 ±0	1 ±1.3	0.007
Total Comorbidities	Mean ±SD	1.5 ±0.53	1.83 ±1.3	1.5 ±0.46	2.3 ±1.5	<0.001
	Median ±IQR	1.2 ±1	1.2 ±1.2	1.2 ±0.75	2 ±2	0.001
PLCO$_{M2012}$ 6-year LC risk: Median ±IQR (Min-Max)		0.46 ±0.7 (0-0.15)	3.9 ±4 (1.51-31.4)	1 ±1 (0-1.5)	1.8 ±0.6 (1.53-5.34)	<0.001

LYFS-CT days of life-expectancy gained from screening: Median ±IQR (Min-Max)	4.8 ± 4.2 (0-12.2)	27.3 ±22.5 (12.2-114.5)	14.6 ±3.2 (12.27-34.5)	10.3 ±2.9 (4.96-12.15)	<0.001
LYFS-CT Years of Life Expectancy Median ±IQR	24.5 ±7.49	18.06 ±7.63	21.29 ±6.6	18.57 ±5.82	<0.001

4.3.3 Model Performance within Screened Participants

LDCT results, and therefore like-for-like LHC outcome data, were only available for the 1409 participants who had LDCT scans within the LHC pilot due to having risk scores of $\text{PLCO}_{\text{M2012}}$ ≥1.51%. In order to examine how the selection criteria would have performed in a comprehensive screening cohort, I created simulated screening thresholds within this subgroup, comparing the participants with the highest $\text{PLCO}_{\text{M2012}}$ scores (top 75%, threshold: ≥2.27%) vs. those with the highest LYFS-CT scores (top 75%, threshold: ≥17.7 days) (Table 7). These thresholds were selected in order to provide large enough segments in each subgroup for analysis, without being too high as to be completely unrepresentative of a real-world risk threshold in a screening programme. In the screened subgroup, 955 (68%) were selected by both criteria, 249 (18%) by neither, 103 (7.3%) by $\text{PLCO}_{\text{M2012}}$-only and 102 (7.2%) by LYFS-CT-only. A comparison of demographic and LHC outcome variables for all four groups are presented in Table 7.

Participants selected by LYFS-CT-only had a median predicted screening benefit of 20.1 days of life-gained and a predicted lung cancer risk of 1.96%, compared to 15.2 days of life-gained and 2.75% risk in those selected by $\text{PLCO}_{\text{M2012}}$-only (p<0.001). Those selected by LYFS-CT-only were significantly younger (median: 58 vs. 68 years; p<0.001), less overweight (median BMI: 28 vs. 30; p<0.001), had fewer co-morbidities (median = 1 vs. 2; p<0.001) and fewer severe co-morbidities (median = 0 vs. 1; p<0.001) than those selected by $\text{PLCO}_{\text{M2012}}$-only. Consequently, those selected by LYFS-CT-only are predicted to have greater life expectancy (median: 24.2 vs. 17.3; p<0.001) and greater gain in life-expectancy following lung cancer detection (median: 2.4 vs. 1.9 years; p<0.001) and subsequent prevention of death (median: 22.7 vs. 16.1 years; p<0.001) (Table 7). Those selected by $\text{PLCO}_{\text{M2012}}$-only were significantly more likely to have CAC detected on their LDCT scans than those selected by LYFS-CT-only (81.8% vs. 57.4%; p<0.001), providing additional clinical confirmation of the self-reported increased CVD risk in the $\text{PLCO}_{\text{M2012}}$ selected subgroup.

I followed-up the two subgroups selected by one criterion but not the other five years post-LHC, in order to examine whether the differences in comorbidity increased, persisted, or diminished with time. This facilitates analysis of whether the difference in participant selection based on a cross-sectional assessment of clinical information at the LHC reflects genuine differences in long-term health trajectories and outcomes. Five-years post-LHC, the prevalence of severe co-morbidities remained higher in those selected by $\text{PLCO}_{\text{M2012}}$-only compared to LYFS-CT-only, with the mean difference of severe co-morbidities between the two groups increasing from 1.1 to 1.2 (mean severe comorbidities: 1.6 vs. 0.4; p<0.001), differences between the medians remain consistent (1; p<0.001). The difference in ‘other’ (less severe) comorbidities lessened over time from 0.6 to 0.4 but remained significant (mean other comorbidities: 1.5 vs. 1.1; p=0.005) (median: 1 vs. 1; <0.001) (Figure 19). There was no significant difference in new ‘serious’ co-morbidity diagnoses. There were no significant differences in the number of screen-detected lung cancers (2 vs. 1) or deaths 5-years post-LHC (6 vs. 6; p=0.99).

Table 7. Demographics and follow-up outcomes of LHC pilot screened participants, stratified by those selected by risk vs. life-gained selection in 1409 scanned participants.

		Not selected by risk or life-gained	**Selected by risk and life-gained**	**Selected by risk but not life-gained**	**Selected by life-gained but not risk**	**P-value for differences in selection**
Number of Subjects (%)		249 (17.7)	955 (67.8)	103 (7.3)	102 (7.2)	-
At LHC pilot:						
Median Age ±IQR		64 ±9	65 ±8	68 ±6	58 ±7	<0.001
Median BMI ±IQR		30.1 ±6.85	27 ±6.7	30.1 ±7.3	28 ±6.35	<0.001
Median Cigs per Day ±IQR		20 ±10	20 ±10	20 ±10	20 ±10	0.12
Median Years Smoked ±IQR		38 ±11	47 ±8	39 ±8	41.5 ±7	0.001
Median Years Quit ±IQR		9 ±17	0 ±3	11 ±12	0 ±1	<0.001
% Female (n)		49.8 (124)	50.8 (485)	51.5 (53)	47.5 (48)	0.53
% Exposed to Asbestos (n)		22.9 (73)	24.3 (232)	20.4 (21)	29.4 (30)	0.14
% Prior Cancer (n)		11.2 (28)	12.1 (116)	38.8 (40)	0	<0.001
% COPD (n)		31.7 (40)	40.1 (383)	25.2 (26)	13.7 (14)	0.04
% Prior Pneumonia (n)		14.1 (35)	17.7 (169)	13.6 (14)	10.7 (11)	0.54
% Family History of LC (n)		20.9 (52)	30.9 (295)	28.2 (29)	8.8 (9)	<0.001
% Hypertension		48%	55%	56%	38%	0.79
% Coronary Heart Disease		22%	22%	35%	9%	<0.001
% Angina		11%	11%	19%	5%	<0.001
% Prior Heart Attack		10%	9%	19%	4%	<0.001
% Prior Stroke		3%	1%	3%	1%	0.03
% Diabetes		18%	25%	35%	7%	<0.001
Severe Comorbidities †	Mean ± SD	0.7 ± 0.7	0.7 ± 0.8	1.3 ± 1	0.2 ± 0.3	<0.001
	Median ± IQR			1 ± 1.5	0 ± 0.5	<0.001
Other Comorbidities †	Mean ± SD	-	-	1.3 ± 0.8	0.7 ± 0.7	<0.001
	Median ± IQR			1 ± 1.3	0.7 ± 1	<0.001
Total Comorbidities †	Mean ± SD	1.7 ± 1.3	1.9 ± 1.5	2.6 ± 1.7	0.9 ± 0.9	<0.001
	Median ± IQR			2 ± 1	1 ± 1.2	<0.001
$\text{PLCO}_{\text{M2012}}$ 6-year LC risk: Median ± IQR (Min-Max)		1.8 ± 0.3 (1.51-2.27)	4.96 ± 4.39 (2.27-32.9)	2.75 ± 0.84 (2.27-6)	1.96 ± 0.34 (1.51-2.25)	<0.001
LYFS-CT days of life-expectancy gained from screening: Median ± IQR (Min-Max)		13.7 ± 4.4 (5.2-17.7)	34.3 ± 22.5 (17.7-114.5)	15.2 ± 3.8 (4.96-17.64)	20.1 ± 2.9 (17.7-33.4)	<0.001
LYFS-CT Years of Life Expectancy Median ± IQR		21.4 ± 6.1	16.9 ± 6.7	17.2 ± 6.4	24.2 ± 5.6	<0.001
% Scan-Detected Emphysema (n)		45.3 (112/247)	70.2 (658/937)	57.6 (57/99)	52.5 (53/101)	0.47

% Scan-Detected Coronary Artery Calcification (n)		64 (158/247)	78.2 (733/937)	81.8 (81/99)	57.4 (58/101)	<0.001
Number of Lung Cancers Detected		4	55	2	1	-
At 5-year follow-up:						
Number of Lung Cancers with Deaths 5-years post detection		2	22	1	0	-
% Dead ~5 Years Post-Scan ¶		-	-	5.8 (6)	5.9 (6)	0.99
Severe Comorbidities †	Mean ± SD	-	-	1.6 ± 1.2	0.4 ± 0.6	<0.001
	Median ± IQR	-	-	1.2 ± 1.2	0.2 ± 0.7	<0.001
Other Comorbidities †	Mean ± SD	-	-	1.5 ± 1.1	1.1 ± 0.9	0.005
	Median ± IQR	-	-	1 ± 1.3	1 ± 1.3	<0.001
Mean Total Comorbidities ± SD †	Mean ± SD	-	-	3.1 ± 1.9	1.5 ± 1.2	<0.001
	Median ± IQR	-	-	2.4 ± 2.2	1.2 ± 1.3	<0.001
Mean Serious Comorbidities ± SD *	Mean ± SD	-	-	0.15 ± 0.4	0.11 ± 0.35	0.44
	Median ± IQR	-	-	0 ± 0	0 ± 0	0.3

†Severe Comorbidities: diabetes, heart attack, CHD, stroke, previous cancer, kidney disease. Other Comorbidities: COPD, pneumonia, atrial fibrillation, hypertension.

¶ Causes of Death: Selected by risk but not life-gained – Lung Cancer (1), Heart Failure (2), Mouth Cancer (1), Unknown (1). Selected by life-gained but not risk – Pneumonia (1), Hepatocellular Carcinoma (1), Myelodysplastic Syndrome (1), Unknown (3).

* Serious Comorbidities: Thoracic aortic aneurysm, dementia, heart failure, valvular heart disease, liver cirrhosis, abdominal aortic aneurysm, idiopathic pulmonary fibrosis, severe frailty.

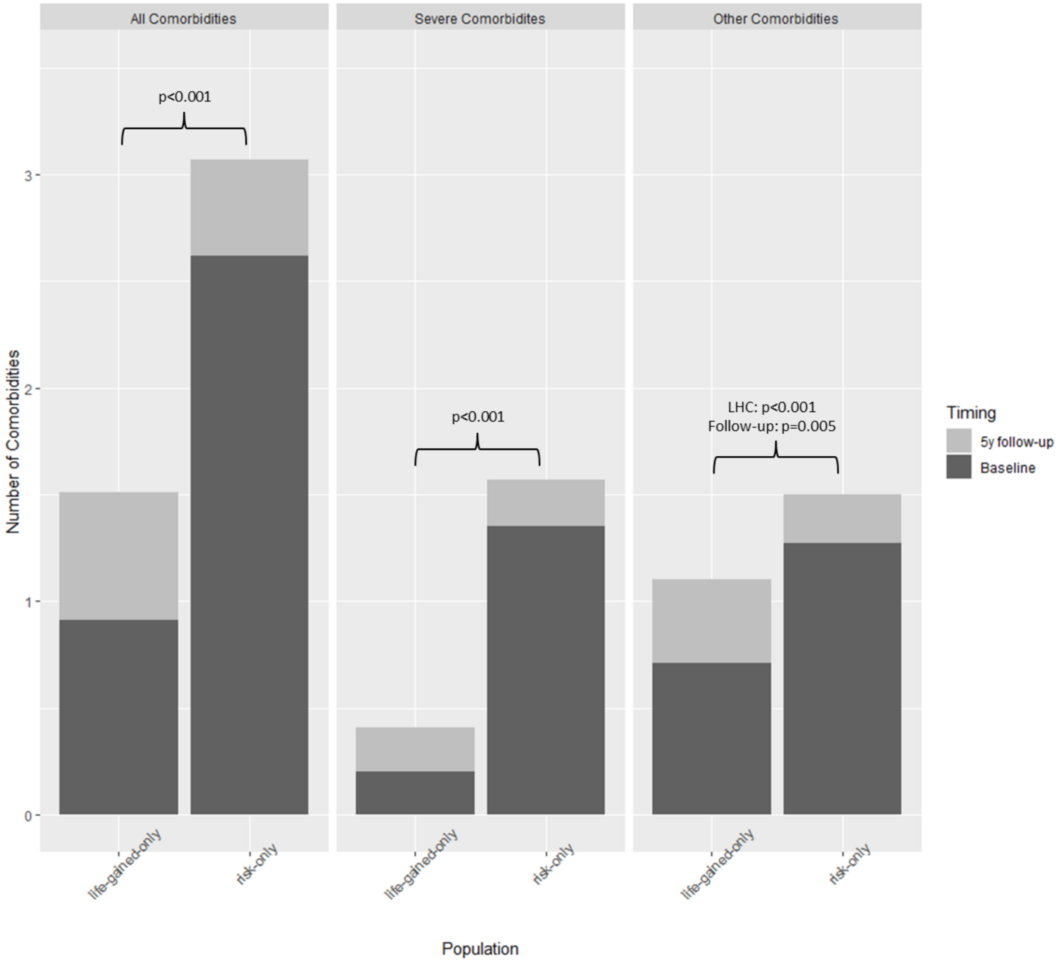


Figure 19. Mean number of comorbidities among screened individuals in the LHC pilot, selected (in top 75% of scores) by life-gained-benefit but not risk vs. selected by risk but not life-gained.

In total, between the two analyses detailed above, 396 LHC pilot participants were followed-up. Those who died were significantly older (mean age: 67 vs. 64; p=0.03), had a higher burden of severe comorbidity at the LHC (mean: 1.08 vs. 0.76; p=0.05), more newly diagnosed severe comorbidities at follow-up (individuals with new severe comorbidity: 44% vs. 19%; p=0.06) (comorbidity classifications presented in Appendix 1), and had a lower probability of survival to 5-years according to the mortality model embedded in the LYFS-CT calculator (0.89 vs. 0.92; p=0.006) (Table 8).

Table 8. Outcomes of LHC pilot participants followed-up at 5-years post-LHC.

		All	**Dead**	**Alive**	**P-value**
Total Participants (%)		396	24 (6.1)	372 (93.9)	-
Mean Age at LHC ±SD		64.2 ±6	66.9 ±5.8	64.1 ±6	0.03
Mean PLCO$_{M2012}$ score ±SD		1.8 ±0.9	1.7 ±0.8	1.9 ±1	0.33
Mean chance of survival to 5-years in LYFS-CT mortality model ±SD		0.91 ±0.05	0.89 ±0.07	0.92 ±0.05	0.006
Mean Number of Comorbidities at LHC ±SD	Severe	0.78 ±0.78	1.08 ±1.03	0.76 ±0.76	0.05
	Other	0.99 ±0.8	1.2 ±0.64	0.98 ±0.81	0.12
	All	1.77 ±1.4	2.3 ±1.5	1.7 ±1.4	0.07
% Subjects with ≥1 new comorbidity at follow-up *	Severe	20 (34/168)	44.4 (4/9)	18.9 (30/159)	0.06
	Other	27.4 (46/168)	22.2 (2/9)	27.7 (44/159)	0.72
	Serious	11.9 (20/168)	22.2 (2/9)	11.3 (18/159)	0.33

* Denominators differ as comorbidity follow-up was not performed on individuals only included in the ‘threshold’ analysis presented in section 4.3.2.

For completeness, I also collected follow-up data for all those with screen-detected lung cancer in the LHC pilot (even if they did not fall into the analysis subgroups detailed above). Of the 62 participants, 60% (n=37) were alive. Those who died had significantly more advanced disease at the time of diagnosis (stage III/IV: 40% vs. 11%; p<0.001) and were less likely to have been treated with surgery (36% vs. 70.3%; p=0.001). Three quarters of the deaths were caused by lung cancer diagnosed during the LHC. Those who died also had higher PLCO$_{M2012}$ scores (mean: 9.6% vs. 7%; p=0.08) and LCDRAT scores (6.4% vs. 4.2%; p=0.04), indicating increased estimated lung cancer risk and increased chance of lung cancer death (Table 9) (LCDRAT is the lung cancer mortality model used in the generation of LYFS-CT scores). The borderline statistical significance of the PLCO$_{M2012}$ association indicated the need for larger studies to confirm this link.

Table 9. Outcomes for those with screen-detected lung cancer in the LHC pilot.

		All	**Dead**	**Alive**	**P-value**
Total Participants (%) *		62	25 (40)	37 (60)	-
Mean Age at Diagnosis ±SD		67.1 ±5	66.6 ±5.6	67.5 ±4.7	0.5
Mean $PLCO_{M2012}$ score at Diagnosis ±SD		8 ±5.6	9.6 ±6.6	7 ±4.7	0.08
Mean risk of lung cancer death in absence of screening (LCDRAT) score at Diagnosis ±SD		5.1 ±4.1	6.4 ±6.6	4.2 ±3.2	0.04
% Stage at Diagnosis (n)	Stage I	66 (41)	44 (11)	81.1 (30)	0.001
	Stage II	11.3 (7)	16 (4)	8.1 (3)	
	Stage III	9.7 (6)	12 (3)	8.1 (3)	
	Stage IV	12.9 (8)	28 (7)	2.7 (1)	
% Treatment – Surgery (n)		56.5 (35)	36 (9)	70.3 (26)	0.008

* Cause of death: lung cancer (n=19), 2nd primary lung cancer (n=1), COVID-19 (n=1), oesophageal cancer (n=1), heart disease (n=1), complications from surgery for lung cancer (n=1), not known (n=1).

4.3.4 LYFS-CT and $PLCO_{M2012}$ Calibration

The mortality model in LYFS-CT predicted 18.8 deaths (9.1%) at 5-years post-LHC in the absence of screening in the subset of LHC attendees with follow-up data in our study, 57% higher than the 12 deaths (5.8%) observed. $PLCO_{M2012}$ predicted the occurrence of 5.7 lung cancers in 6-years in this subgroup; at 5-year follow-up, six lung cancers had been diagnosed. Assuming a further 1.2 lung cancers would be diagnosed in the sixth year, it is likely that there would be 25% more lung cancers diagnosed in this period than predicted by the model. Extrapolated to the complete $PLCO_{M2012}$ eligible screening cohort, this would result in 91 lung cancers diagnosed after 6-years, considerably higher than the 72 predicted by the RPM. Extreme caution must be employed when interpreting this extrapolation, considering the very small case sample size and selection bias in this very specific subgroup; nonetheless, this projection falls between the actual and adjusted NELSON incidence projections reported in section 3.3.3 and Figure 11, adding tentative confirmatory evidence to the previous analysis.

4.4 Discussion

This retrospective study examined the impact of using a benefit-based screening selection method in a real-world lung cancer screening programme. The results indicate that had life-gained-based selection (using the LYFS-CT model) been used at an equivalent threshold to select the same number of screening-eligible participants as $PLCO_{M2012}$ ≥1.51%, the screening threshold would have been ≥12.2 days of life-gained from screening; 92% of the individuals selected by the RPM would still have received screening using this criterion. Crucially, almost all cancers detected in the $PLCO_{M2012}$-selected screening cohort would have been detected had a screening cohort of the same size been selected by LYFS-CT. Attendees who would have only been selected by LYFS-CT using these eligibility thresholds would have been significantly younger and less comorbid. Whilst a previous study has modelled the equivalent LYFS-CT threshold to match USPSTF eligibility criteria (≥16.2 days of life-gained), to my knowledge, this is the first calculation of a LYFS-CT threshold to match a clinically employed RPM threshold in a real-world screening programme.

Whilst the LHC was not large enough to examine differences in subsequent mortality between subgroups, when comparing the top 75% of screened individuals (for whom we had comprehensive LHC and LDCT data) by life-gained vs. risk-based selection, those selected by life-gained-only were younger, more likely to be a current smoker, and had substantially fewer co-morbidities than those selected by risk-only. This is congruent with the findings of the original study in which LYFS-CT was developed [233]. However, unlike in the original study, in our 'equivalent threshold' analysis, women were more likely to be deemed eligible for screening by risk-based selection than life-gained-based selection and had lower LYFS-CT scores overall. Previous studies have shown that screening performance may be superior in women [197]; particularly notable, the NELSON RCT found a significantly higher LDCT derived mortality reduction in women compared to men, although the small size of the female subgroup in NELSON limits the certainty of this conclusion [194]. Further research is required to investigate the impact of varying screening selection methods on the male/female split of the eligible cohort, and any differences in overall life-gained by screening between the sexes.

After five years of follow-up in the subset who would only have been selected for screening by one model and not the other, the increased burden of co-morbidity present in the 'risk-only' subgroup persisted. Additionally, those who died in the follow-up period had a higher burden of comorbidity. Whilst this may indicate that those deemed 'most eligible' for screening by risk-based selection may derive limited benefit from the intervention, it is important to note that the small size of the study and the follow-up subset significantly limits the conclusions we can draw. Large, prospective studies with extended and comprehensive follow-up periods are needed to confirm whether these tentative indications of the potential benefit of using a life-gained-based selection strategy would actually improve the benefit-to-harm trade-offs in lung cancer screening selection.

A further limitation of this analysis is that individuals older than 75 were ineligible for the LHC, and only people with a $PLCO_{M2012}$ score ≥1.51% received LDCT screening. In programmes with more lenient age and risk restrictions, larger differences in comorbidities and outcomes between risk- and benefit-based selection might be expected, as risk increases with age whilst life-gained from screening may not. In the original LYFS-CT study, the most significant divergence between life-gained and risk-based selection only occurred in participants older than 75 [233]. A further limitation is that the comparison of screened individuals did not include $PLCO_{M2012}$-eligible participants who did not attend screening or participants who would have been screen-eligible had LYFS-CT been used alongside $PLCO_{M2012}$ to determine eligibility in the LHC pilot. Additionally, the LYFS-CT scores in our cohort were calculated retrospectively with existing datasets and therefore relied on several imputations and assumptions which may not be clinically accurate. Follow-up data were not collected for the full LHC pilot cohort, nor was the follow-up timepoint fixed in relation to each subject's LHC, further limiting the conclusions I was able to draw from these analyses.

Calibration analyses indicated that the mortality model in LYFS-CT may overestimate mortality risks and the $PLCO_{M2012}$ RPM may underestimate lung cancer risk in the population served by the LHC pilot. This is consistent with my analysis in section 3.3.3, which based solely on the lung cancer detected within the LHC pilot, also showed likely risk underestimation by $PLCO_{M2012}$. Miscalibration of both models is likely caused by their development in populations unlikely to be representative of highly deprived UK cohorts. However, the size of the dataset used for calibration analyses in this study means that there was significant margin for error and extrapolating the outcomes to the full cohort assumes that this subset is representative of the rest of the attendees. Further research is required to investigate whether this miscalibration is present in larger datasets, what the exact nature of any miscalibration is, and whether it has any impact on who is selected for screening. Large studies with extended follow-up are required to recalibrate models in specific populations.

Another potential benefit of LYFS-CT was demonstrated by a recent study showing that augmenting $USPSTF_{2020}$ criteria with LYFS-CT almost totally eliminates racial and ethnic disparities in screening eligibility derived from applying generalised eligibility criteria to a US population [332]. RPMs have also been shown to perform somewhat differently in ethnic subgroups [322]. This study was unable to examine this question, due to the majority white ethnicity of the screening cohort. Further research is required to examine whether this is a significant issue for RPMs and whether LYFS-CT could ameliorate this.

A challenge associated with using LYFS-CT is the ethical implications of restricting screening from those with shorter life expectancies. Whilst clinical practice guidelines regularly recommend factoring life expectancy into screening decisions, some studies show that this may not be well accepted by participants. However, effective shared decision making and communication were shown to assist participants in this regard [333,334]. Risk communication in the context of pre-screening counselling is discussed in detail in Chapter Six; communicating benefit-based metrics to screening participants may add additional complications and requires further research. Even if not ultimately used for screening selection, incorporating estimates of life-gained benefits from screening may be useful for general shared decision-making among individuals in which screening may be preference-sensitive [259]. Considerable further research is required to investigate how LYFS-CT can be integrated successfully into pre-screening counselling and shared decision-making tools. An additional challenge for LYFS-CT implementation is the current lack of an accessible calculator that can be used easily by clinical staff in a screening programme setting. Whilst there are Excel and R versions of the model freely available [335], it would be beneficial to integrate these into a user interface similar to the $\text{PLCO}_{\text{M2012}}$ software used in the TLHC programme. The additional time required to gather the necessary clinical information to calculate LYFS-CT scores is also an important factor to consider in implementation.

In conclusion, this study demonstrates that life-gained-based selection may have the potential to be a useful alternative to risk-based selections for screening programmes. A life-gained approach could maintain high detection rates whilst selecting fewer individuals less likely to benefit from screening, thereby addressing the comorbidity issue which is one of the factors driving the hypothesis that screening selection performs sub-optimally in socio-economically deprived populations. However, due to the limitations of this study, significant further research is required to confirm if LYFS-CT-based selection confers greater gains in realized life-expectancy over risk-based selections. To do this, studies with complete data for LYFS-CT calculation and longer and comprehensive follow-up data, as well as prospective studies, are needed. Further research is also required to ensure the model is well calibrated for the population it is being used in, as well as to establish effective LYFS-CT screening thresholds. Implementation studies are needed to investigate practical aspects of integrating LYFS-CT calculation into screening programmes, and to develop benefit-based communication and counselling for screening participants.

Chapter Five - Genetic Risk Factors for Lung Cancer: Case-Control Study

5.1 Introduction

Lung cancer is the most common global cancer and a leading cause of cancer death [1]. LDCT screening significantly reduces mortality through early diagnosis [193,194] but relies on careful targeting to prevent participants from being exposed to unnecessary harm as well as to ensure overall cost-effectiveness of the programme [215,216]. RPMs are used to calculate individual lung cancer risk scores for screening attendees based on risk factor exposure, allowing for the determination of risk-based screening eligibility [218]. Lung cancer risk is mediated by several risk factors, most notably age and smoking history, but also an array of other factors (see section 1.5), including family history of lung cancer [110,111].

Both environmental and genetic risk factors contribute to the association between family history of lung cancer and increased lung cancer risk [113]. In the past decade, several large GWAS have implicated dozens of SNPs in this association [139,140]. A PRS is a tool that integrates multiple low penetrance SNPs into a tool that can be used to predict an individual's genetic risk of developing a particular condition. Whist several attempts have been made to use SNPs to create a PRS for lung cancer [242,243], thus far, no RPM used clinically for lung cancer risk prediction and screening selection includes a measure of genetic risk in its risk calculation (see sections 1.5.8.2 and 1.6.6.4). Research is needed to investigate whether these PRS tools improve risk prediction and screening selection in high-risk populations likely to be targeted for lung cancer screening.

This study aimed to validate several previously published lung cancer PRS tools in a Manchester-based case-control cohort. The case samples were sourced from patients undergoing surgery for histologically confirmed NSCLC. The control samples were sourced from high-risk attendees of the Manchester LHC pilot, a community-based lung cancer screening programme [248]. I tested the PRS tools for overall efficacy in the cohort and performed subgroup analyses to ascertain PRS effectiveness in specific demographic segments. I explored novel suggestive signals in the genomic data by building a novel PRS which if validated in external cohorts, could provide additional SNPs for inclusion in future PRS tools. Previous studies have often developed and validated PRSs in biobank or RCT settings, which may not be wholly representative of real-world screening population settings [318,336]. A unique aspect of this study is the high-risk nature of the control cohort, having been sourced from the Manchester LHC pilot, thereby being highly representative of community-based lung cancer screening participants.

If a PRS was found to be an effective risk prediction tool, this could be used to improve RPM discrimination, thereby facilitating the targeting of screening to those most at risk, improving programme efficiency, and reducing exposure to screening related harms in those at lower risk of lung cancer. Moreover, genetic risk factors (unlike spirometry and CVD indicators which may also improve risk prediction and are examined in Chapter Three) are not directly implicated in increased risk from other comorbidities. This could allow for the selection of participants at high risk of lung cancer without concomitantly selecting a cohort with high disease burden and shorter life expectancies [233].

5.2 Dataset Processing and Methods Development

5.2.1 Cohort Demographic Data Analysis

An extensive description of cohort recruitment, sample collection, and data gathering can be found in sections 2.1 and 2.4. Comprehensive and detailed demographic and clinical information was available for the control group, as it was sourced from the Manchester LHC pilot. Information was less comprehensive for the case cohort, which was sourced from Manchester Cancer Research Centre (MCRC) Biobank. Consequently, for analysis, cases who had smoked within the previous 12 months were classed as 'current' (even if listed as former). Pack-years for cases were calculated when information was provided in the sample notes; when a range was given for cigarettes-per-day or smoking duration, the mid-point was used. If an explicit pack-year figure was provided, this was used preferentially (even if the calculated figure was different). If no pack-year information was available, imputed scores for current and ex-smokers were used within the cases. Where available, cancer stage was ascertained from the TNM coding in the sample notes, using the International Association for the Study of Lung Cancer (IASLC) TNM classification of malignant tumours 7th edition [337]. FEV_1/FVC ratios above 1 were assumed to be erroneous and were recalculated. Missing BMI and FEV_1/FVC data were imputed according to the mean of the remaining cases or controls as appropriate. Age, calculated pack-years, and smoking status were used to estimate NLST eligibility in the case cohort (age 55-74, ≥30 pack years, current smoker within 15 years [193]).

Demographic differences between the case and control groups, as well as between the recruited control cohort and the remaining eligible control population (T1 or 3-month post-T1 surveillance participants who were screen-negative for lung cancer), were calculated and tested for significance using an Independent-Samples T Test for scale variables, Chi-square test for dichotomous variables (with multiple z-tests and Bonferroni correction for variables with multiple categories), and Mann-Whitney U-Test for medians.

5.2.2 DNA Processing and Quality Control

Detailed protocols for DNA extraction and quality control are described in section 2.6. A schematic of sample processing for the whole cohort is presented in Figure 20. A total of 701 case and 706 control blood samples were provided. Of these, 98% (n=1374) had DNA successfully extracted and were progressed to genotyping; the vast majority of these (n=1370) were successfully genotyped (see section 2.7) and formed the pre-QC case-control dataset.

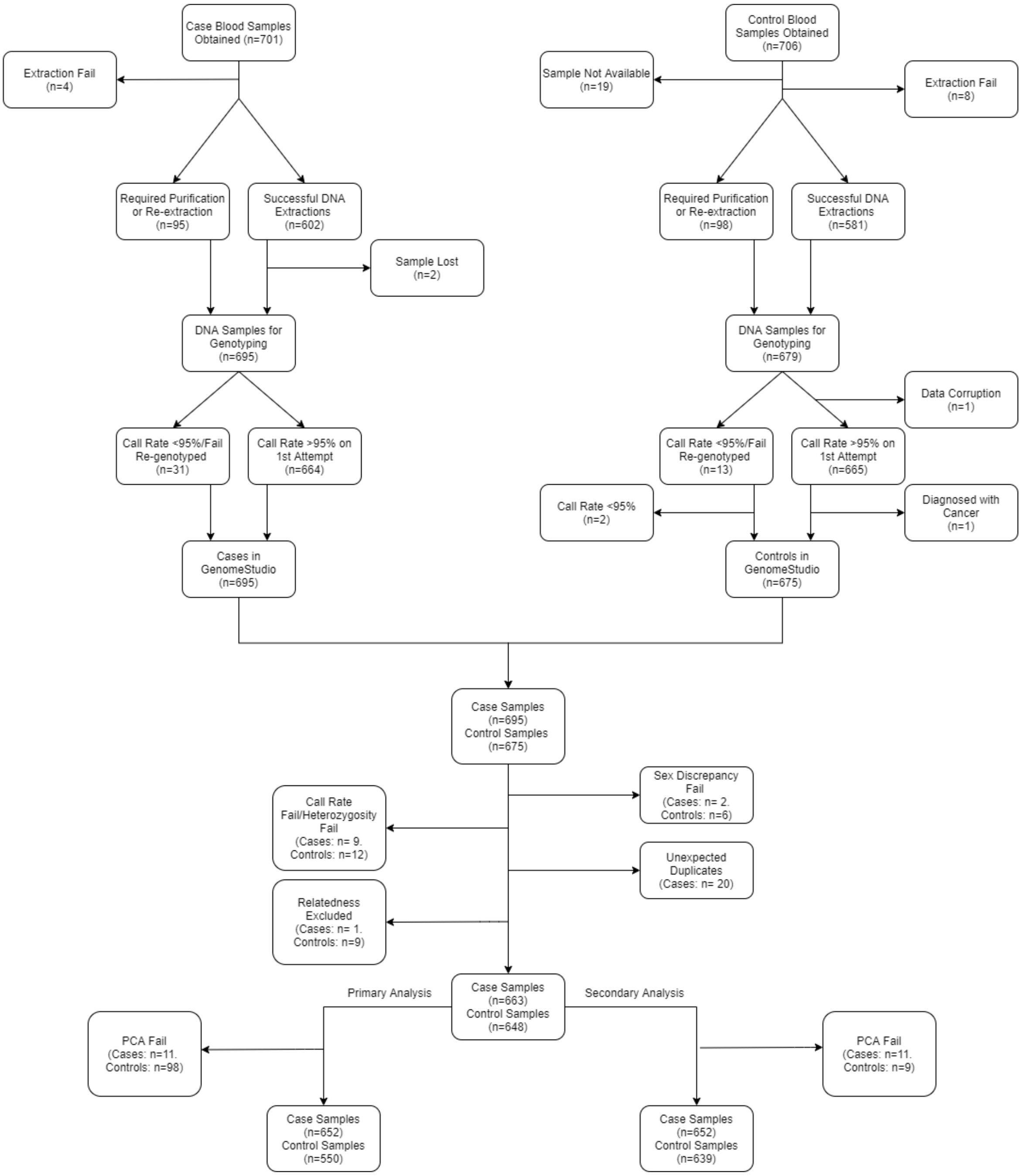

Case Blood Samples Obtained (n=701)
Extraction Fail (n=4)
Required Purification or Re-extraction (n=95)
Successful DNA Extractions (n=602)
Sample Lost (n=2)
DNA Samples for Genotyping (n=695)
Call Rate <95%/Fail Re-genotyped (n=31)
Call Rate >95% on 1st Attempt (n=664)
Cases in GenomeStudio (n=695)
Control Blood Samples Obtained (n=706)
Sample Not Available (n=19)
Extraction Fail (n=8)
Required Purification or Re-extraction (n=98)
Successful DNA Extractions (n=581)
DNA Samples for Genotyping (n=679)
Data Corruption (n=1)
Call Rate <95%/Fail Re-genotyped (n=13)
Call Rate >95% on 1st Attempt (n=665)
Call Rate <95% (n=2)
Diagnosed with Cancer (n=1)
Controls in GenomeStudio (n=675)
Case Samples (n=695) Control Samples (n=675)
Sex Discrepancy Fail (Cases: n= 2. Controls: n=6)
Call Rate Fail/Heterozygosity Fail (Cases: n= 9. Controls: n=12)
Unexpected Duplicates (Cases: n= 20)
Relatedness Excluded (Cases: n= 1. Controls: n=9)
Primary Analysis
Case Samples (n=663) Control Samples (n=648)
Secondary Analysis
PCA Fail (Cases: n=11. Controls: n=98)
PCA Fail (Cases: n=11. Controls: n=9)
Case Samples (n=652) Control Samples (n=550)
Case Samples (n=652) Control Samples (n=639)

Figure 20. Sample processing and quality control flowchart for case-control study.

5.2.3 Genotype Data Quality Control

Detailed procedures for processing and quality control of the genotyping data are described in section 2.8; a schematic of data processing is presented in Figure 20. Of the 1370 samples that entered the quality control workflow, 8 failed the sex discrepancy check, 21 failed due to call rate or heterozygosity (Figure 21), 10 failed due to relatedness, and 20 were unexpected duplicates. All cross-plate positive controls were correctly identified as expected duplicates, with the best performing sample retained for analysis. This resulted in a dataset of 1311 individuals (cases=663; controls=648), 96% of the pre-QC case-control dataset. This dataset progressed to the PCA QC step.

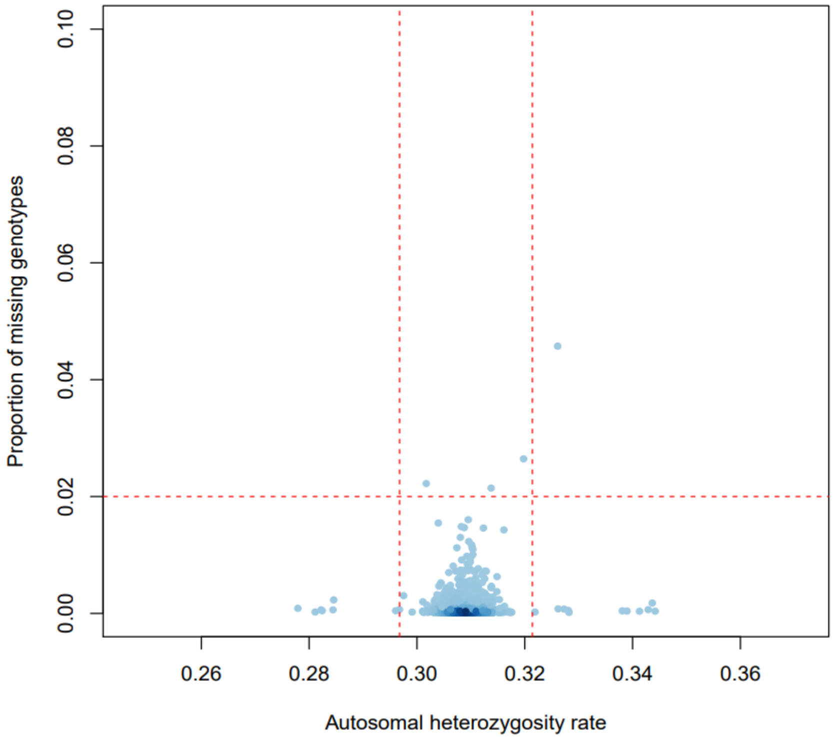


Figure 21. Heterozygosity and call rate quality control plot. Red dashed lines indicate the thresholds for sample exclusion.

PCA identified several individuals not of the predominant Caucasian genetic ancestry group, all of which were excluded (Figure 22). PCA also revealed a subgroup of 91 controls, all originating from the same genotyping plate, that formed a distinct genetic cluster (Figure 26). Troubleshooting by Illumina confirmed processing issues and the presence of artefacts on all four genotyping chips from that run. As time and budget did not allow for re-genotyping, these samples were excluded from the primary analysis to prevent this unusual genetic variation affecting the robustness of the analysis. I repeated the PCA to ensure that all outliers had been appropriately excluded. The lambda value in the Aberrant package was set at 30 (based on a subjective assessment of sample spread), trimming a further 18 outlying samples from the cohort (Figure 23). This resulted in a final study cohort of 1202 samples (Controls: n=550; Cases n=652), 88% of the pre-QC case-control dataset and 85% of the total blood samples that entered the study. The primary analysis cohort had a clear and robust overlapping profile of genetic variation (Figure 24). In the scree plot, only one point clearly deviated indicating that genetic variance could be adequately controlled for with inclusion of one PC as a covariate during association analysis (Figure 25); three PCs were used as covariates to provide additional margin for error.

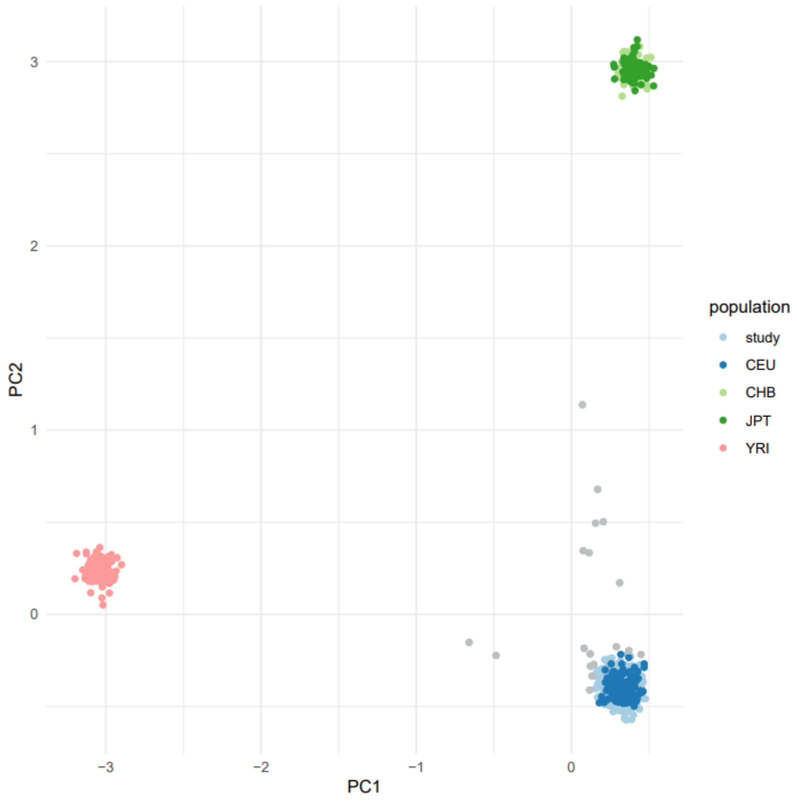


Figure 22. Genetic ancestry of the Manchester cohort dataset vs. the HM3 reference panel. Grey = excluded due to deviation from CEU cluster. CEU: Utah residents with Northern and Western European ancestry from the CEPH collection. CHB: Han Chinese in Beijing, China. JPT: Japanese in Tokyo, Japan. YRI: Yoruba in Ibadan, Nigeria.

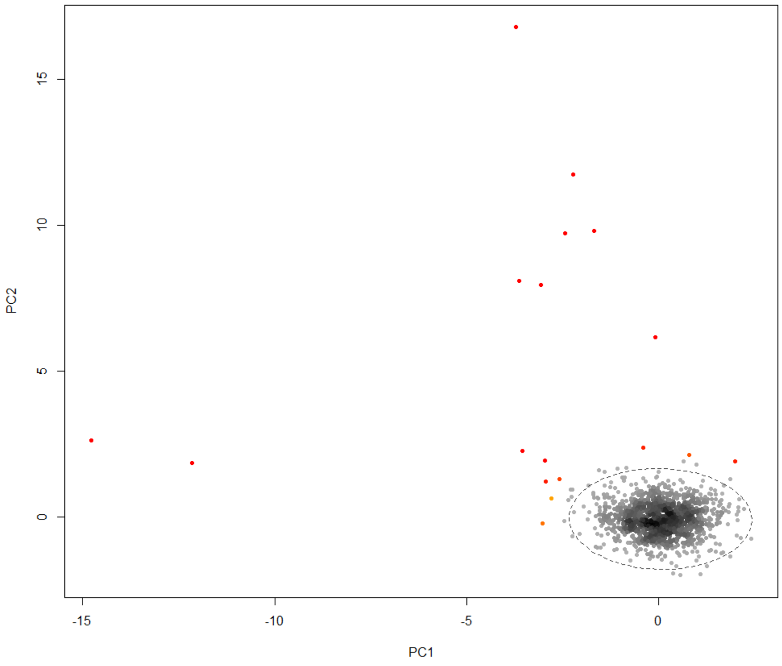


Figure 23. Aberrant plot excluding 18 outliers from the primary analysis cohort. Dashed line = 99% confidence interval of inferred distribution of 'normal' samples. Normal individuals are coloured from black to grey, with darker colouring indicating sample density. Outliers are coloured from orange to red, with redder colours indicating higher posterior probability of being an outlier.

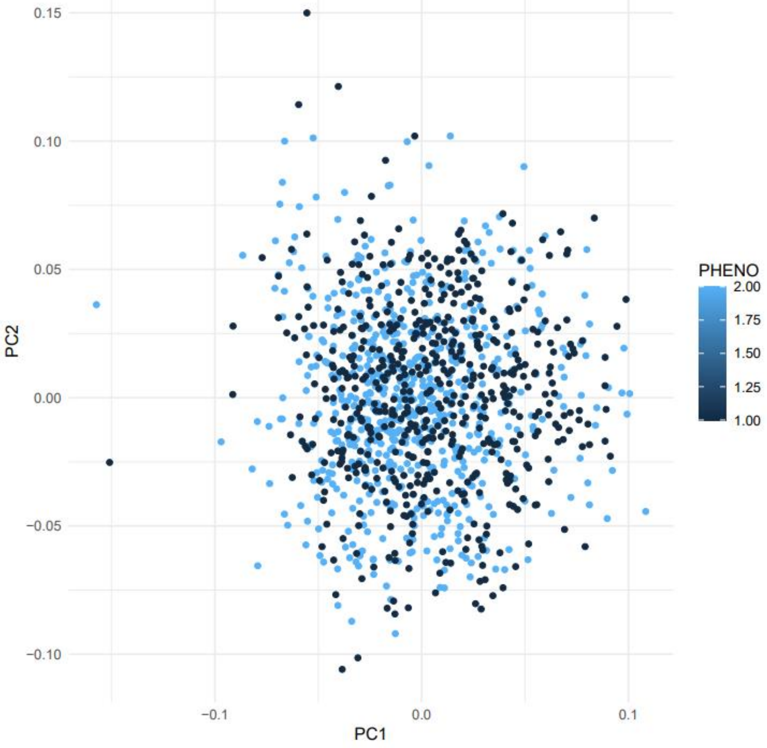


Figure 24. PC1 and PC2 scatterplot for the final primary analysis cohort. Light blue = case; dark blue = control.

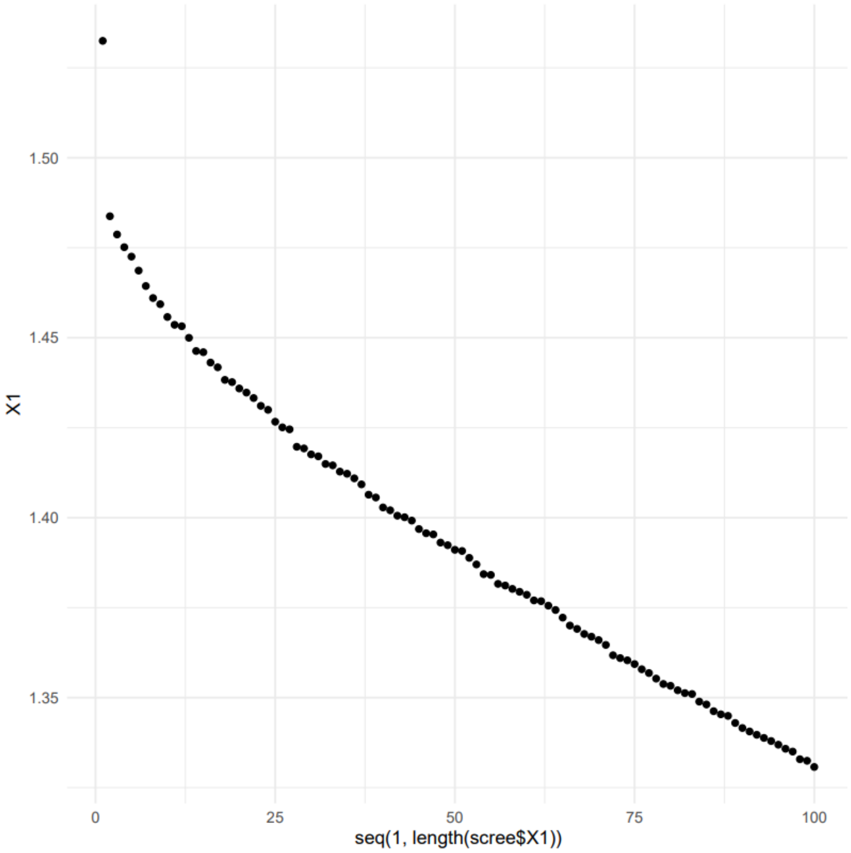


Figure 25. Scree plot for primary analysis. Dots represent eigenvalues plotted against PCs.

I also performed a secondary analysis for comparison purposes, retaining the outlying cluster of control samples (Figure 26). In order to ensure that the additional genetic variance present was accounted for in the association analysis, I restricted the dataset to controls-only and plotted each of the first ten principal components on individual scatterplots (Figure 27). The plots indicated that the aberrant variation was restricted to PC1 and PC2, congruent with the two clearly deviating points on the secondary analysis scree plot (Figure 28). Three principal components were used as covariates during association analysis. The secondary analysis cohort contained 1291 samples, 652 cases and 639 controls, 94% of the pre-QC case-control dataset and 92% of the total blood samples that entered the study.

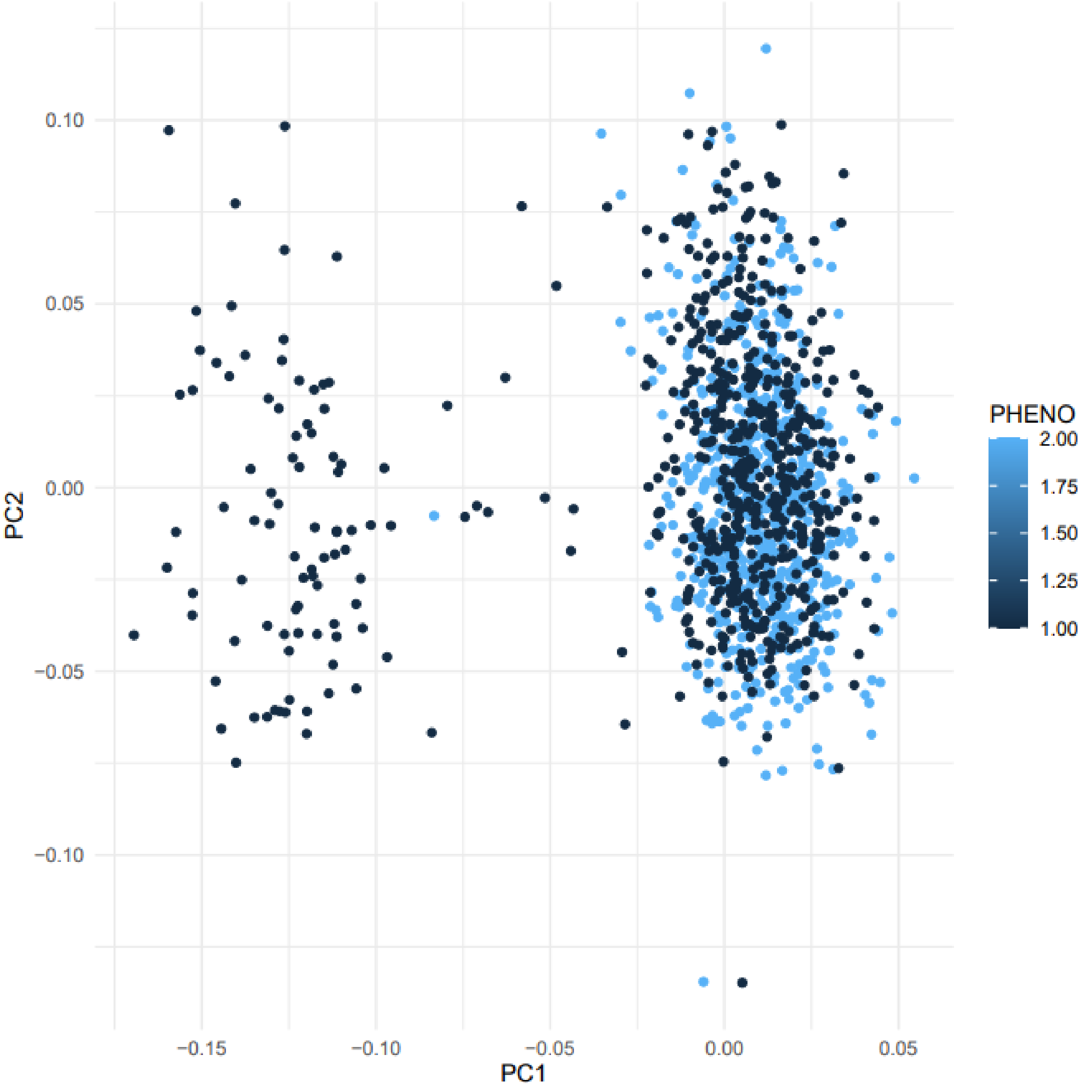


Figure 26. PC1 and PC2 scatterplot for the secondary analysis cohort. Light blue = case; dark blue = control. Outcropping of controls with low PC1 values indicates unexpected abberant genetic variability. These samples were removed from primary analysis but retained in secondary.

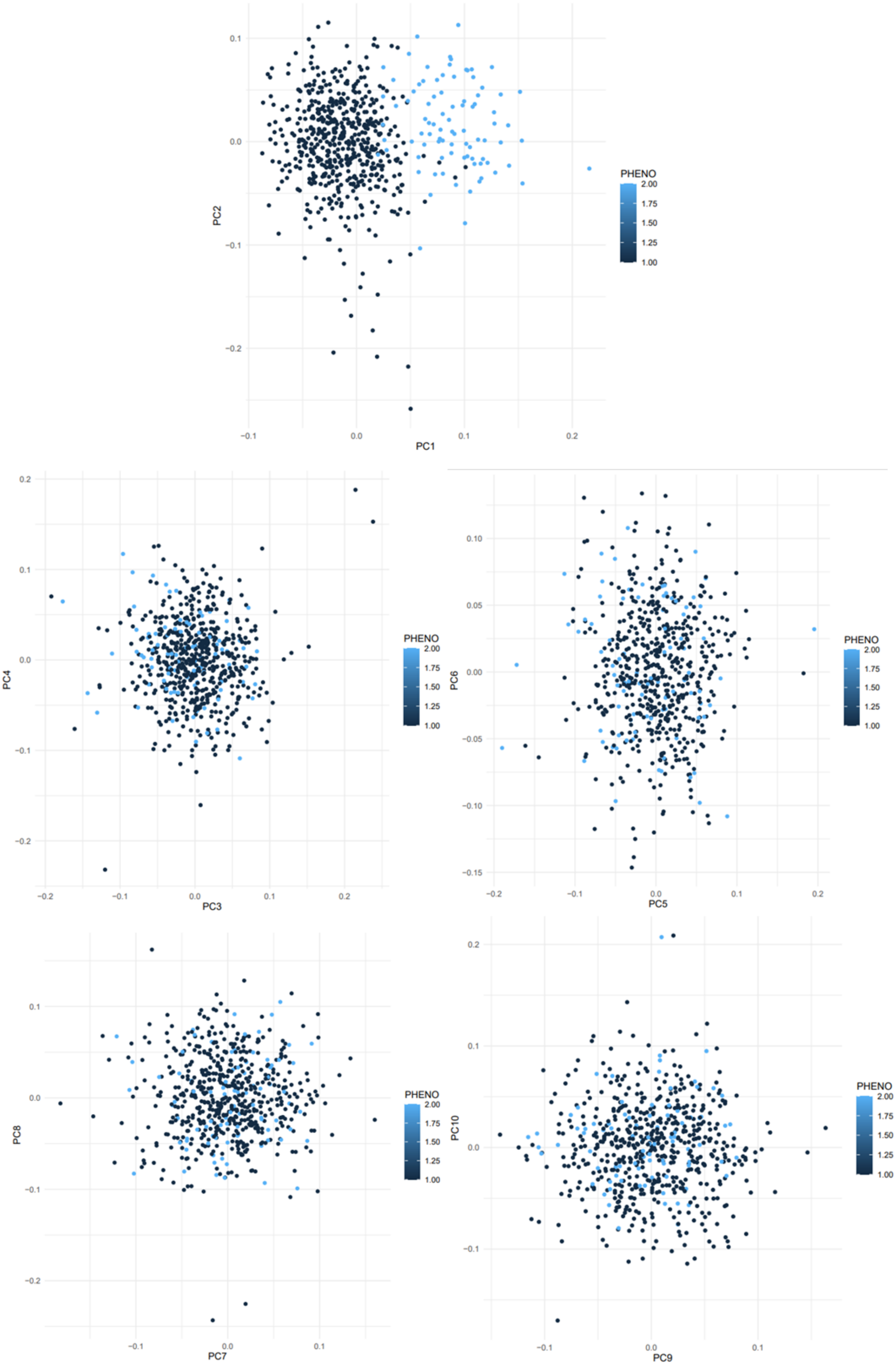


Figure 27. Scatterplots for first ten principal components in the control cohort of the secondary analysis dataset. Light blue = aberrant plate; dark blue = remaining controls.

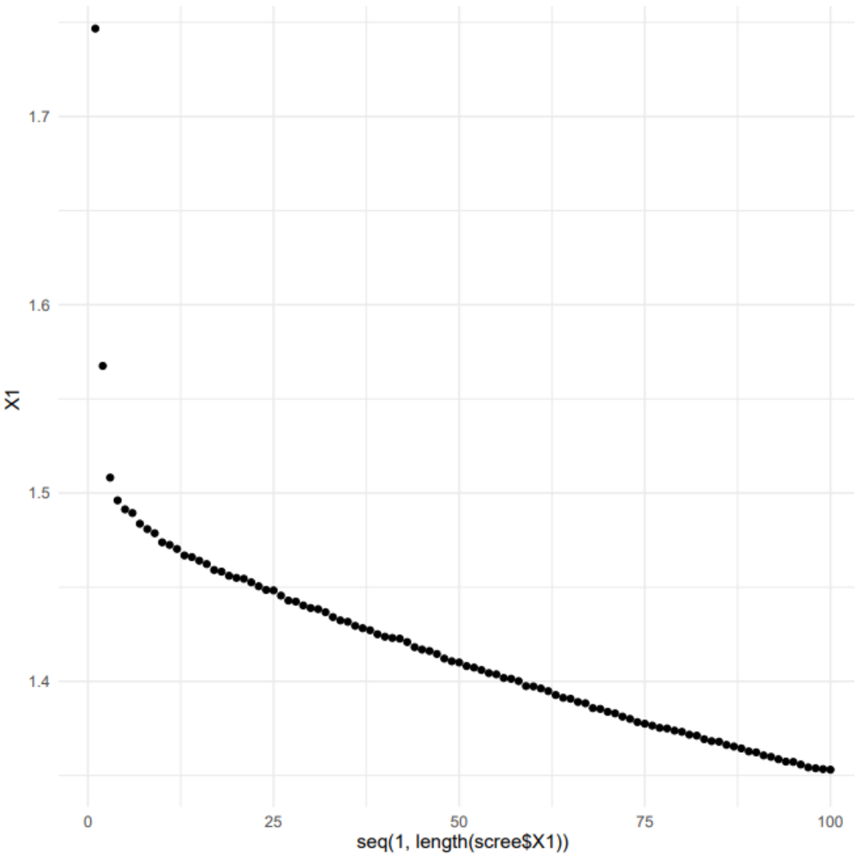


Figure 28. Scree plot for secondary analysis.

5.2.4 Genomic Imputation

Due to previous quality control steps, the genomic imputation was of high quality, with 100% reference overlap, no allelic mismatch, monomorphic sites, SNPs with low call rate, duplicated SNPs or invalid alleles, and no requirement for allele switching or strand flipping (indicating mismatch with reference genome). The overall r^2 value for allele-frequency correlation between the reference panel and uploaded samples was 0.995 (Figure 29).

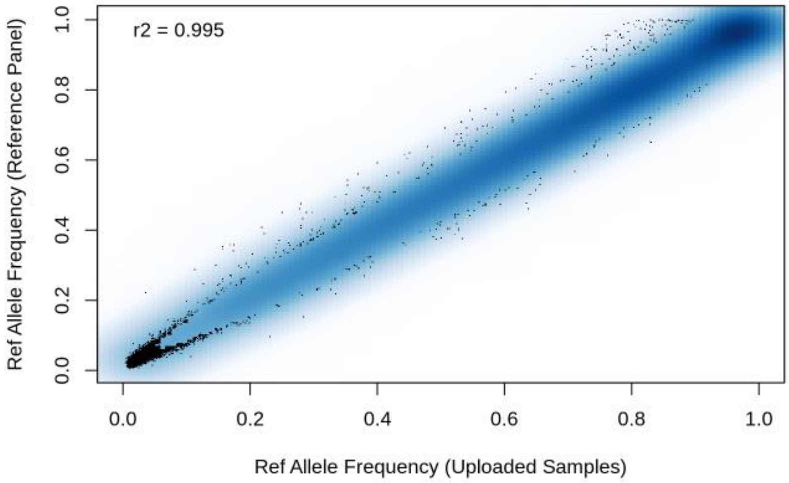


Figure 29. Correlation of densities of allele frequencies between the study samples and imputation reference panel.

5.3 Results

5.3.1 Cohort Demographics

5.3.1.1 Control Cohort

A total of 706 participants of the Manchester LHC pilot T1 screening round provided blood samples, 60.2% of the total eligible cohort (T1 negative and T1 + 3-month surveillance negative participants). The participating control group was highly representative of the total screen-negative participants; the only variable with a statistically significant difference (although not clinically significant) was mean BMI (participating controls = 28.8 vs. remaining controls = 28.1; p=0.03) (Table 10). The participating control group was equally split between males and females and current and former smokers. Mean age was 64.6 years, median pack-year history was 45, 67.4% had no educational qualification and 30.6% had a self-reported COPD diagnosis. Median 6-year lung cancer risk score ($PLCO_{M2012}$) was 3.5% (Table 10).

Table 10. Comparison of control samples included in the study vs. remaining eligible control population.

		Remaining T1 Negative Population	Participating Controls	P-value
Total		467	706	-
% Female (n)		47.5 (245)	49.9 (354)	0.44
Mean Age ±SD		64.7 ±5.4	64.6 ±5.3	0.74
% Current (n)		49.9 (233)	50.4 (356)	0.86
Median Pack Years (±IQR)		44 (±24)	45 (±23)	0.72
Educational Attainment	% Less than GCSE/O Level (n)	70.4 (329)	67.4 (476)	0.55
	% GCSE/O Level (n)	19.2 (85)	18 (127)	
	% A Level (n)	2.6 (12)	4.2 (30)	
	% Some University/College (n)	6 (28)	6.8 (46)	
	% University Degree (n)	2.1 (10)	2.3 (16)	
	% Postgrad (n)	0.6 (3)	1.3 (9)	
Mean BMI ±SD		28.1 ±5.6	28.8 ±5.3	0.03
Median FEV_1/FVC Ratio ±IQR		69.3 ±13.2	70.4 ±12.6	0.06
% COPD (n)		33.4 (156)	30.6 (216)	0.31
Median PLCO Score ±IQR		3.6 ±3.8	3.5 ±3.7	0.77

After sample processing and QC, 550 control samples (77.9%) were eligible for inclusion in the primary analysis and 639 (90.5%) were eligible for inclusion in the secondary analysis (see section 5.2.3 for an explanation of these two analyses) (Figure 20); the only variable with a statistically significant difference between the final control dataset and the samples which were not successfully processed was median FEV_1/FVC ratio (70.2 vs. 71.6; p=0.05) (Table 11).

Table 11. Demographic and clinical characteristics of controls and comparison final primary control cohort vs. failed samples.

		Failed Control Samples	Final Primary Control Dataset	P-value
Total		156	550	-
% Female (n)		48.7 (76)	50.5 (278)	0.68
Mean Age ±SD		65.1 ±5.2	64.5 ±5.3	0.2
% Current Smoker (n)		48.7 (76)	51.3 (282)	0.57
Median Pack Years ±IQR		45 ±28	45 ±21	0.72
Educational Attainment	% Less than GCSE/O Level (n)	71.2 (111)	66.4 (365)	0.4
	% GCSE/O Level (n)	16.7 (26)	18.4 (101)	
	% A Level (n)	3.2 (5)	4.5 (25)	
	% Some University/College (n)	3.8 (6)	7.6 (42)	
	% University Degree (n)	3.2 (5)	2 (11)	
	% Postgrad (n)	1.9 (3)	1.1 (6)	
Mean BMI ±SD		28.6 ±5.3	28.7 ±5.2	0.58
Median FEV_1/FVC Ratio ±IQR		71.6 ±13	70.2 ±13	0.05
% COPD (n)		30.8 (48)	30.5 (168)	0.96
Median PLCO Score ±IQR		3.86 ±4.3	3.37 ±3.5	0.06

5.3.1.2 *Case Cohort*

The MCRC Biobank provided 701 case samples; 55.1% were female, 59% were former smokers, and median age was 69. Pathological subtypes included: adenocarcinomas (64%), squamous cell carcinomas (34%), and large cell carcinomas (2%). Of the 580 (83%) with stage information available, 80% were early stage (stage I = 376; stage II = 86), 17% (n=100) were stage III, and 3% (n=18) were stage IV (Table 12). After sample processing and quality control, 652 case samples (93%) were eligible for inclusion in both analyses (Figure 20). There were no significant differences in the demographic or clinical variables between the cases in the final dataset and those that were not successfully processed (Table 12). Of the 652 cases, 6 (0.92%) required BMI imputation, 88 (13.5%) required FEV_1/FVC imputation, and 206 (31.5%) required pack year imputation (of these, 111 former smokers required imputation and 95 current smokers).

Table 12. Demographic and clinical characteristics of cases and comparison final case cohort vs. failed samples.

		Failed Case Samples	Final Case Dataset	P-value
Total		49	652	-
% Female (n)		53.1 (26)	55.2 (360)	0.77
Median Age ±IQR		70.6 ±11.3	68.9 ±8.8	0.81
% Current Smokers (n)		34.7 (17)	41.9 (273)	0.33
Median Pack Years ±IQR		40 ±45	44 ±25	0.5
Median BMI ±IQR		27.1 ±4.9	26 ±6.7	0.5
Median FEV_1/FVC Ratio ±IQR		65.4 ±20	67 ±16	0.85
% Cancer Stage (n)	I	51 (25)	54 (351)	0.6
	II	8 (4)	13 (82)	
	III	12 (6)	14 (94)	
	IV	6 (3)	2 (15)	
	Unknown	22 (11)	16 (110)	0.3
Cancer Histology	Adenocarcinoma	76 (37)	63 (408)	0.07
	Squamous Cell Carcinoma	24.5 (12)	34.8 (227)	0.14
	Large Cell	0	2.1 (14)	-
	Other	0	0.5 (3)	-

5.3.1.3 Comparison of Case and Control Cohorts

We compared demographic characteristics between the case and control samples included in the analyses. For the primary analysis, the case cohort had a significantly higher proportion of former smokers (58.1% vs. 48.8%; p=0.001), higher average age (median: 69 vs. 65 years; p<0.001), and lower BMI (median 26 vs. 28.6; p<0.001). Clinical information was not sufficient to calculate $PLCO_{M2012}$ scores for cases. A similar proportion of both cases and controls would have been eligible for screening according to NLST eligibility criteria (76% vs. 76.2%; p=0.93).

Table 13. Comparison of demographic variables between post-QC cases and controls for primary analysis.

		Controls	Cases	P-value
Total (%)		550 (45.8)	652 (54.2)	-
% Female (n)		50.5 (278)	55.2 (360)	0.11
Median Age ± IQR		65 ±9	68.9 ±8.8	<0.001
% Current Smokers (n)		51.2 (282)	41.9 (274)	0.001
Median Pack Years ± IQR	Direct	45 ± 21	44 ± 25	0.051
	Imputed*	45 ± 21	40 ± 15	<0.001
Median BMI ± IQR		28.6 ±6.8	26 ±6.7	<0.001
Median FEV_1/FVC Ratio ± IQR	Direct	70.2 ± 13	67 ± 16	<0.001
	Imputed*	70.2 ± 13	65.5 ± 13	<0.001
% NLST eligible (n)		76 (418)	76.2 (497)	0.93

*Imputed values do not match direct values as imputations were calculated separately in smokers and former smokers within the case cohort.

5.3.2 Validation of Previously Published Polygenic Risk Scores – Primary Analysis

5.3.2.1 *Base clinical model*

In order to assess whether genetic risk factors improved risk prediction over the performance of standard clinical variables, I constructed a multivariable model using the clinical factors available in both the case and control datasets ('base clinical model'). This model had an AUC of 0.723 (0.695-0.751) (Figure 30), a Brier score of 0.21 (a score of 0 indicates perfect probabilistic accuracy, a score of 1 indicates perfect probabilistic inaccuracy) and a Nagelkerke R^2 of 0.2 (indicating the estimated proportion of variance explained by the clinical model). As would be expected, higher age, lower BMI, and lower FEV_1/FVC ratio were significantly associated with increased likelihood of lung cancer (see section 1.5). Female sex was also associated with increased lung cancer risk, although only with borderline statistical significance (Table 14). Whilst not statistically significant, it should be noted that there was an inverse relationship between smoking status and lung cancer risk (Table 14), with current smoker status being more prevalent in the controls than the cases in the Manchester cohort (Table 13). This highlights the uniquely high-risk nature of the control cohort, as it is well established that being a current smoker increases lung cancer risk (see section 1.5.1). When all clinical factors were considered independently, age had the highest discriminatory ability with an AUC of 0.682 (0.653-0.712), and sex had the lowest with an AUC of 0.523 (0.495-0.552) (Figure 30).

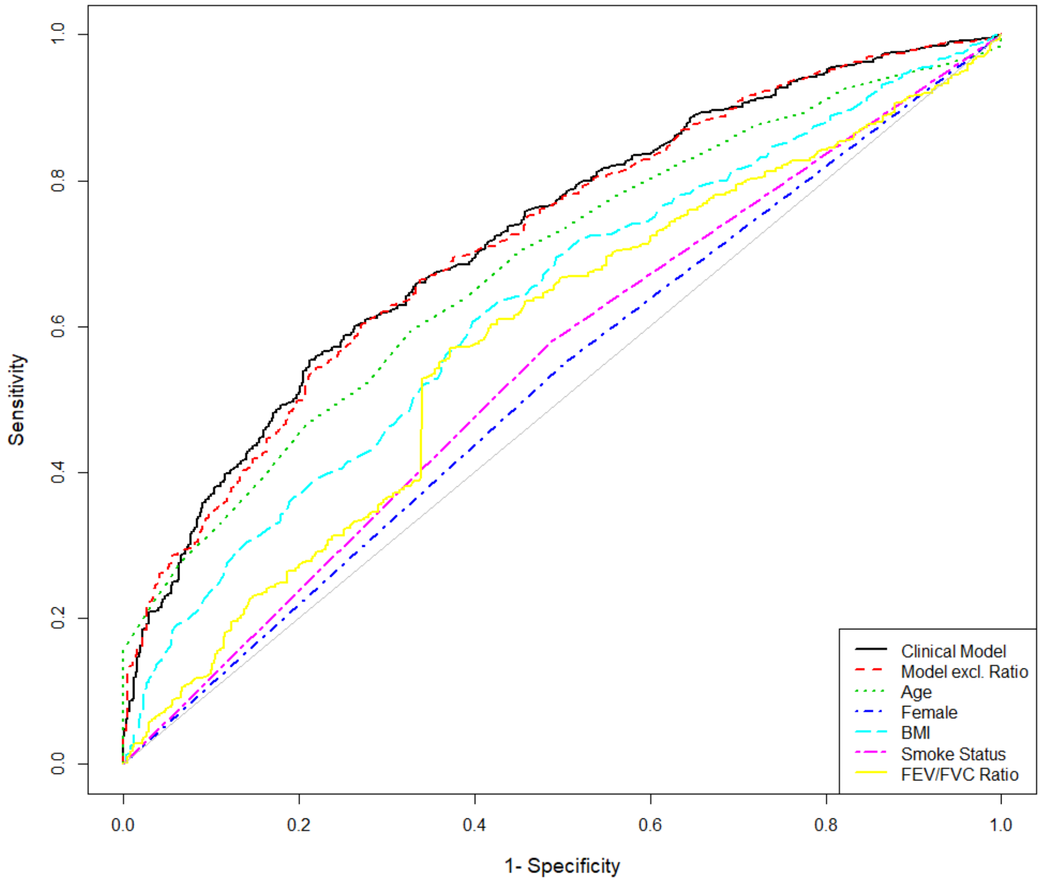


Figure 30. ROC curve for overall base clinical model, clinical model excl. FEV_1/FVC ratio, and individual clinical factors, in primary analysis cohort.

Table 14. Associations between clinical factors and lung cancer from logistic regression in primary analysis cohort.

Factor	**OR – likelihood to be case (95%CI)**	**P-value**	**AUC (95%CI)**
Older Age	1.12 (1.09-1.14)	<0.001	0.682 (0.653-0.712)
Female Sex	1.27 (1-1.63)	0.055	0.523 (0.495-0.552)
Current Smoker Status	0.79 (0.61-1.02)	0.07	0.546 (0.518-0.574)
Higher BMI	0.91 (0.89-0.94)	<0.001	0.629 (0.598-0.66)
Higher FEV_1/FVC Ratio	0.99 (0.97-1)	0.02	0.581 (0.548-0.613)

5.3.2.2 PRS Selection

A survey of the literature revealed eight PRSs for lung cancer which were developed based on large GWAS datasets and externally validated (Table 15). These were: the 109 SNP PRS published by Graff *et. al.* [338], the 35 SNP and expanded 128 SNP PRSs published by Hung *et. al.* [244], the 14 SNP and expanded 19 SNP PRSs published by Fritsche *et. al.* [339], the 19 SNP PRS published by Dai *et. al.* [242], the 6 SNP PRS published by Shi *et. al.* [340], and the 19 SNP PRS published by Jia *et. al.* [341]. It is interesting to note that several of these studies used the same cohort (UK Biobank) for validation. We also tested the 20 SNP *Young* PRS [236] which was published in 2009 (preceding many of the large GWAS studies in the field). A list of all published SNPs absent in our dataset, the proxy SNPs I substituted into the relevant PRSs, and associated linage disequilibrium values, is presented in Table 16; many proxy SNPs had R^2 scores of 1 (indicating perfect predictiveness of the alleles in the SNP of interest), with the majority of the remaining being above 0.8.

Table 15. Summary of PRSs validated in the Manchester cohort, data sourced from GRS Catalog.

First Author (year published) [paper reference]	PRS Development Population	PRS Development Number of Individuals (n=cases)	PRS Validation Population	PRS Validation Number of Individuals (n=cases)
Young (2009) [236]	European (Manual recruitment of SNPs from small COPD and LC studies)	-	European (Clinic recruitment)	930 (446)
Dai (2019) [242]	Chinese (New GWAS) + Chinese & European (meta-analysis)	New GWAS: 19,546 (9298) Meta-analysis: 54,475 (27,120)	Chinese (Prospective – China Kadoorie Biobank)	95,793 (1316)
Shi (2019) [340]	European (5 x GWAS)	258,478	European (Cancer Genome Atlas + eMerge)	14,335 (908)
Graff (2020) [243] [338]	>70% European, pre-June 2018 (3 x GWAS)	183,537	European (UK BioBank)	413,753 (1541)
Jia (2020) [341]	European (5 x GWAS)	293,065	European (UK BioBank)	400,812 (1508)
Fritsche (2020) [339]	European (11x GWAS + meta-analysis)	428,696	European (UK Biobank + Michigan Genomics Initiative)	446,955 (3106)
Hung (2021) [244]	European (8 x GWAS + Training set in 32,341 ILCCO)	310,646	European (UK BioBank)	335,931 (1786)

Table 16. Proxy SNPs used in lieu of published SNPs absent in this dataset. Magnitude of linkage disequilibrium between published and proxy SNPs represented by R^2 and **D'. D'** is a basic measure of the difference between observed and expected frequencies of a haplotype. R^2 also takes allele frequency into account, expressing the correlation between a pair of loci.

Published SNP	Proxy SNP	R^2	D'
rs6920364	rs427824	1	1
rs114544105	rs9274623	0.86	0.96
rs17879961	rs186184919	1	1
rs114928225	rs74787667	0.86	1
rs185666783	rs2318540	1	1
rs12722051	rs2760995	0.77	0.88
rs2518717	rs1985742	0.99	1
rs189146505	rs138833245	0.83	0.97
rs28624856	rs2046144	0.82	0.96
rs9926896	rs183161830	1	1
rs67210567	rs1801272	0.97	1
rs13036436	rs6011779	0.92	0.99
rs71603396	rs35622894	0.77	1
rs13156167	rs13167280	0.61	0.84
rs6912292	rs1535275	1	1
rs182364552	rs2233986	0.99	1
rs116651383	rs3132514	0.76	1
rs141707415	rs2524119	0.88	0.99
rs115494074	rs2507997	0.97	0.99
rs139850307	rs2523589	0.65	0.84
rs139089278	rs2844518	1	1
rs9270868	rs2097432	0.78	0.99
rs190788477	rs9271611	0.74	0.98
rs115566240	rs35656734	0.83	0.95
rs9272307	rs9272306	1	1
rs9273429	rs1049053	0.64	0.92
rs36061084	rs2004038	0.96	1
rs11375254	rs12696594	0.89	0.98
rs2517873	rs376316	0.88	1
rs5879422	rs9374663	1	1
rs35201538	rs10758201	0.97	1
rs200595745	rs11079710	0.87	1
rs1799732	rs11214613	1	1

5.3.2.3 Overall PRS performance

Performance metrics of all the PRSs validated in the Manchester cohort are reported in Table 17. Of the nine published PRS tools validated, eight provided some level of improved discrimination over the clinical model (Figure 31) and had a clear divergence in score distribution between cases and controls (Figure 32). The exception was the PRS published by *Young* et. al., which was developed in 2009 before many of the large lung cancer GWAS studies were performed. It had an independent AUC of 0.5 and did not show improved AUC or net reclassification index (NRI) over the base clinical model; there were several base/strand mismatches between the published SNPs and those observed in this study. More than half of the SNPs in this PRS had a different effect direction in the Manchester cohort to that published in the original study (Appendix 2). Consequently, it was omitted from further analysis.

Of the remaining PRS tools, the 19 SNP *Jia* PRS performed most successfully, with an independent AUC of 0.588. It added 0.015 AUC to the base clinical model (p<0.0001), increasing the overall AUC from 0.723 to 0.738 (0.71-0.766) (Figure 33). More than 67% of the top quintile of PRS scores were cases, compared to 40% of the bottom quintile (p<0.0001) (Table 18). This PRS had the highest categorical NRI, at 0.11 (p<0.0001). Only 16% of SNPs in this PRS displayed a different effect direction to those published, the lowest proportion amongst all PRSs tested. Five SNPs (26%), all in the correct effect direction, reached a p-value threshold of <0.05 in our association analysis (Table 17) (Appendix 2).

The 19 SNP *Fritsche* PRS and the 35 SNP *Hung* PRS also performed relatively well. Both added approximately 0.01 AUC to the base clinical model, increasing overall AUC from 0.723 to 0.733 and 0.734 respectively (p<0.0001). *Hung-35* had an independent AUC of 0.575, slightly higher than *Fritsche-19* at 0.569 (Figure 33); its NRI was also higher and more significant (*Hung-35:* 0.07; p=0.006. vs. *Fritsche-19:* 0.04; p=0.08). 65% of the top quintile of *Fritsche-19* scores and 61% of the top quintile of *Hung-35* scores were cases, compared to 44% and 49% of the bottom quintiles (p=0.002 and p=0.0006) (Table 18). The two iterations of the *Hung* PRS were the only ones among all those validated to have a consistent linear increase of case-proportion across all genomic risk quintiles (Table 18). This may indicate that they add predictive value across the whole cohort, as opposed to only discriminating between those at the highest and lowest genetic risk.

Almost 23% of the SNPs in the *Hung-35* PRS and 21% of the *Fritsche-19* SNPs had a different effect direction in the Manchester cohort compared to the published data. Four SNPs (21%) of the *Fritsche-19* PRS, all in the correct effect direction, were statistically significant in my dataset (p<0.05). Twelve SNPs are shared between all three of the best performing PRSs, with *Hung-35, Fritsche-19* and *Jia* containing only twenty, five and three unique SNPs respectively (Table 17) (Appendix 2).

Considering its small size, the 6 SNP PRS published by *Shi* also performed relatively well. It had an independent AUC of 0.56, adding 0.009 to the base clinical model, increasing overall AUC from 0.723 to 0.732 (p<0.0001). Five of the six (83%) SNPs in this PRS reached statistical significance (p<0.05) in my dataset. However, one of these significant SNPs, rs6495309, was in the incorrect effect direction in my results (published risk allele: ‘T’ OR: 1.3; In my data: ‘T’ OR: 0.69), indicating that C should be considered the risk allele. Other studies confirm that C is the risk allele for this

SNP [342,343]. It is unclear whether in the *Shi* study this is a typographical error, a methodological error, or whether they actually observed a higher frequency of T in the cases. Consequently, I retested this PRS, substituting in C as the risk allele, but maintaining the published allelic weighting. The updated PRS had an independent AUC of 0.569 (95% CI 0.536-0.601), 0.009 higher than the original *Shi* PRS. It added 0.012 of AUC to the base clinical model, increasing overall AUC from 0.723 to 0.735 (95% CI 0.707-0.762; p<0.0001), roughly comparable to *Fritsche-19* and *Hung-35*, despite only including approximately one-third and one-fifth of the total number of SNPs as those two PRSs respectively.

The two PRSs with the largest number of SNPs, *Graff* and *Hung-128*, both displayed modest discrimination, with independent AUCs of 0.553 and 0.562 respectively. Each PRSs added 0.005 and 0.007 AUC (p=0.002; p=0.0001) to the base clinical model respectively, increasing the combined AUC to from 0.723 to 0.728 and 0.73 (Figure 33). Almost 30% of the SNPs in *Graff* and 35% of SNPs in *Hung-128* had reverse effect directions in the Manchester cohort compared to the published data (Table 17) (Appendix 2). Of the functional PRSs, *Dai* performed least effectively, although considering it was developed in a Chinese population and this cohort is European (and it contained a considerable proportion of unique SNPs), it is notable that it still improved AUC above the base clinical model with statistical significance.

Table 17. Performance metrics for the nine PRS validations in the Manchester case-control cohort, as well as a combined PRS of most frequently included SNPs.

First Author (year published) [paper reference]	Number of SNPs in PRS (SNPs available in the Manchester dataset)	Published raw AUC for PRS (95%CI)	AUC in Manchester Cohort (95%CI)	AUC Clinical Model + PRS (95% CI)	Additional AUC over clinical model (p-value – clinical model vs. clinical model + PRS)	Net Reclassification Index – categorical quartiles (p-value)	% SNPs with correct direction of effect (n)
Young (2009) [236]	20 (20)	0.68	0.5 (0.467-0.533)	0.723 (0.695-0.751)	0 (0.88)	-0.003 (0.3)	45 (9)
Dai (2019) [242]	19 (19)	Unknown	0.552 (0.519-0.585)	0.726 (0.698-0.754)	0.003 (0.02)	-0.005 (0.79)	68.4 (13)
Shi (2019) [340]	6 (6)	Unknown	0.56 (0.528-0.593)	0.732 (0.704-0.76)	0.009 (<0.0001)	0.056 (0.02)	83.3 (5)
Graff (2020) [243]	109 (103)	Unknown	0.553 (0.52-0.585)	0.728 (0.7-0.756)	0.005 (0.002)	0.032 (0.14)	71.8 (74)
Jia (2020) [341]	19 (19)	0.591 (0.576-0.606)	0.588 (0.556-0.62)	0.738 (0.71-0.766)	0.015 (<0.0001)	0.11 (0.0001)	84.2 (16)
Fritsche (2020) [339]	14 (14)	0.529 (0.503-0.558)	0.562 (0.529-0.594)	0.731 (0.703-0.759)	0.008 (<0.0001)	0.04 (0.08)	78.6 (11)
	19 (19)	0.552 (0.534-0.569)	0.569 (0.537-0.602)	0.733 (0.705-0.761)	0.01 (<0.0001)	0.04 (0.08)	78.9 (15)
Hung (2021) [244]	35 (35)	Unknown	0.575 (0.542-0.607)	0.734 (0.706-0.762)	0.011 (<0.0001)	0.07 (0.006)	77.1 (27)
	128 (127)	Unknown	0.562 (0.53-0.595)	0.73 (0.702-0.758)	0.007 (0.0001)	0.05 (0.04)	65.5 (83)
Overlapping SNPs in >3 PRSs	11	-	0.57 (0.538-0.602)	0.73 (0.702-0.758)	0.007 (<0.0001)	0.05 (0.02)	-

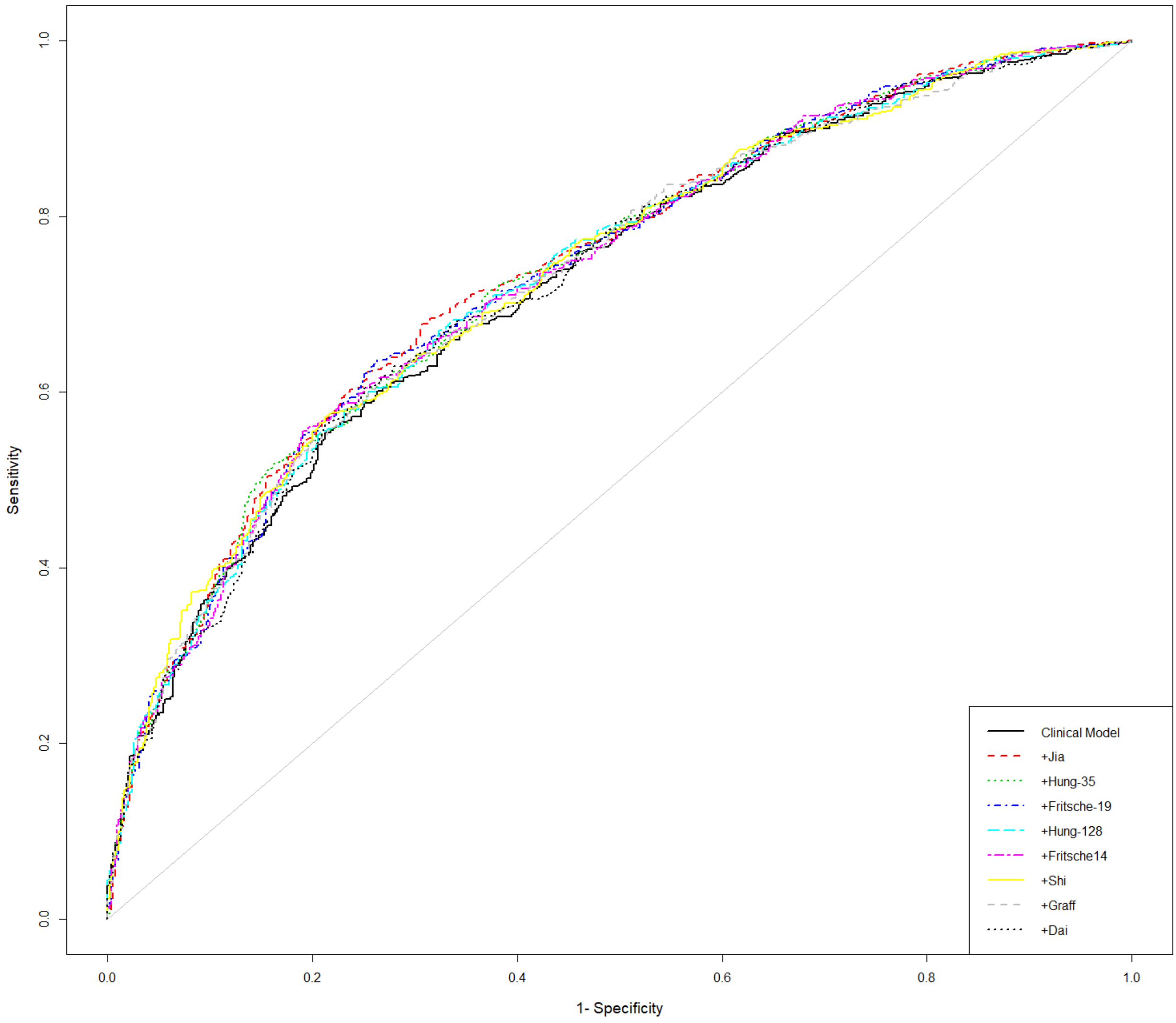


Figure 31. ROC curves for all PRSs validated, showing the range of AUC increases conferred by addition of a PRS (Young PRS not plotted as it does not add AUC).

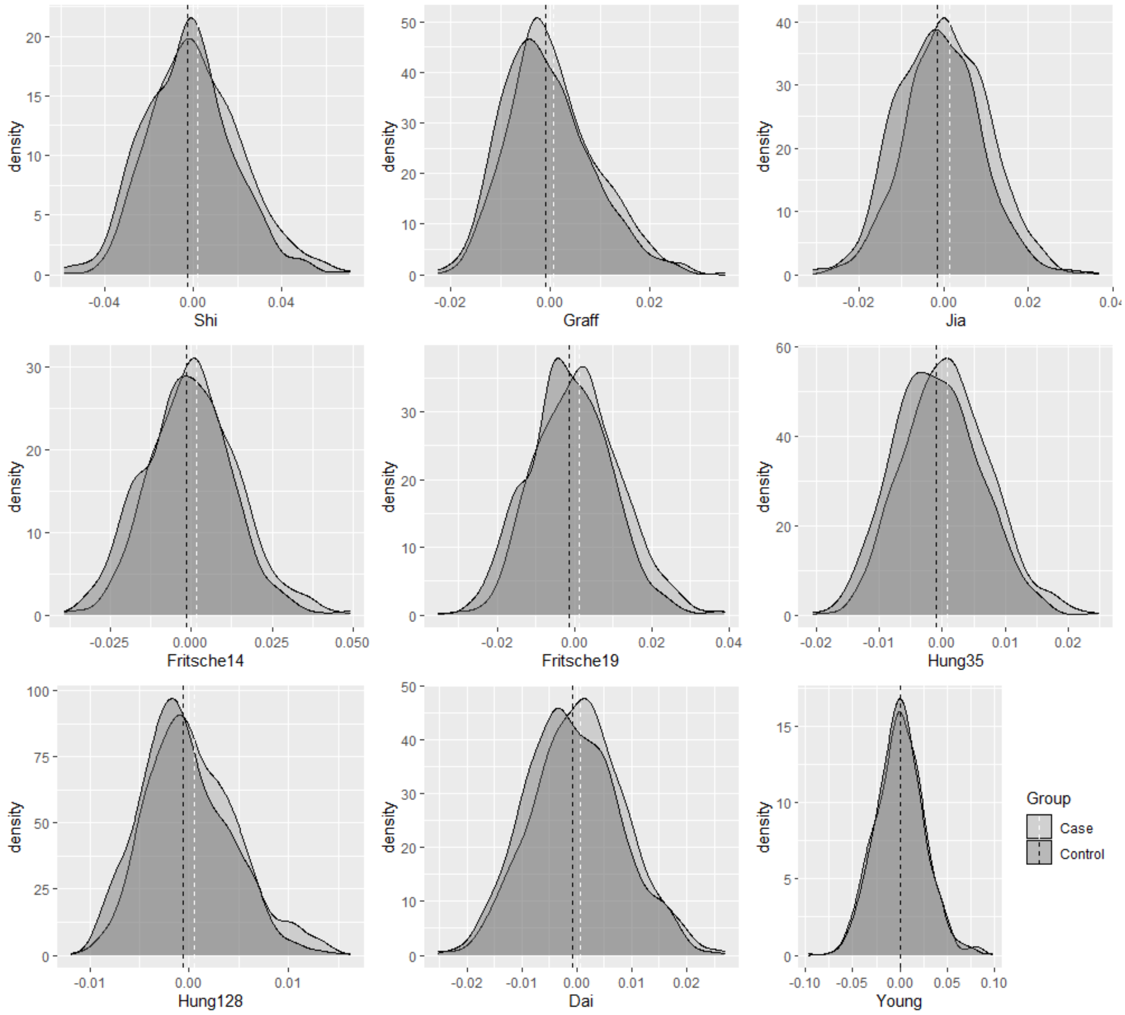
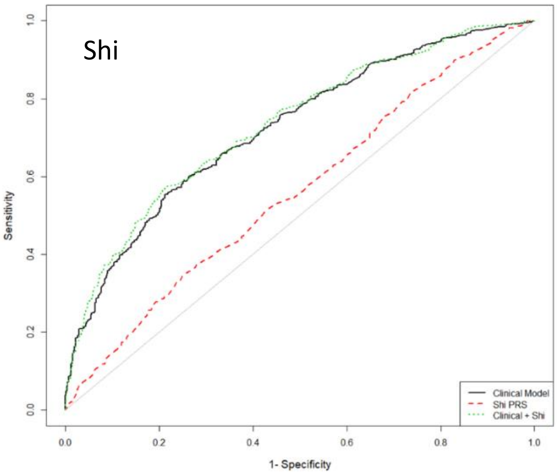
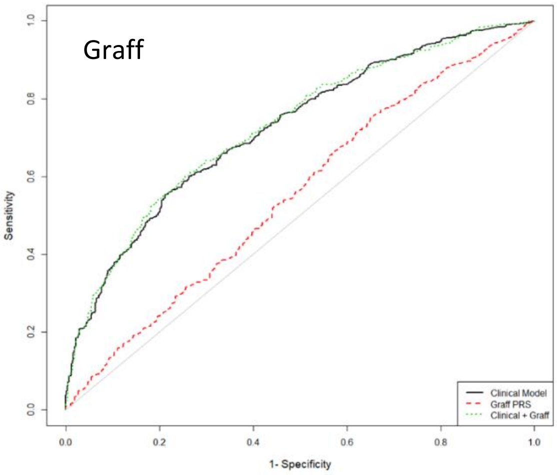


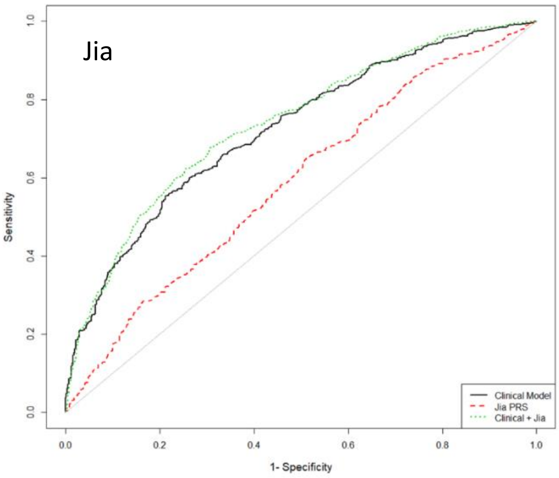
Figure 32. Density plots for each of the PRS tools validated, stratified by case/control grouping. Mean score for overall cohort is normalised to 0. Dashed lines = mean scores for case or control subgroups.

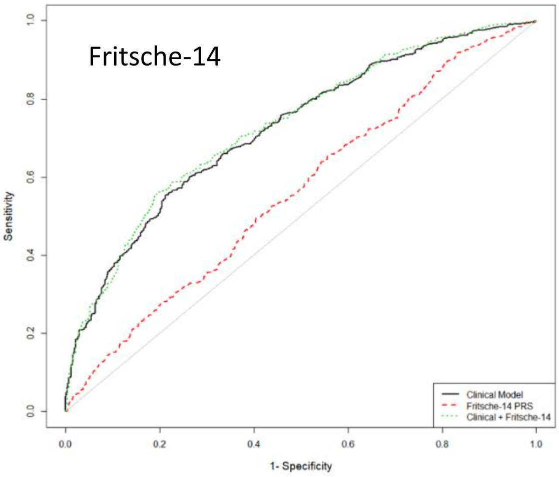
Table 18. Case proportion across the PRS quintiles in the case-control cohort.

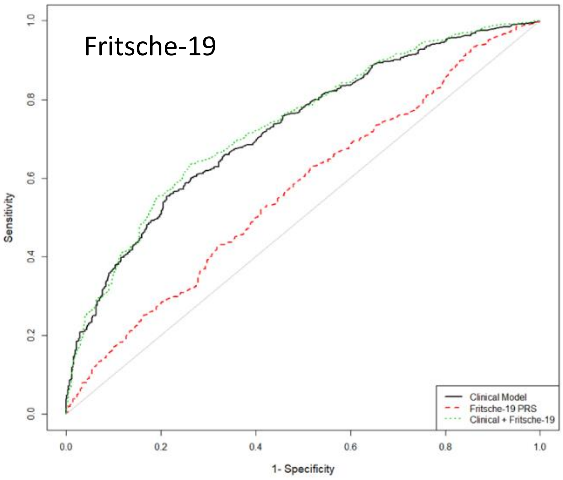
PRS		**Cases as % of PRS Quintile**					**P value – Chi-Square**	**P value – Linear-by-Linear**
		1 – lowest PRS	**2**	**3**	**4**	**5 – highest PRS**		
Young		56	51	54	55	55	0.85	1
Dai		46	52	57	56	61	0.02	0.001
Shi		45	57	53	55	62	0.006	0.002
Graff		45	54	58	56	59	0.02	0.004
Jia		40	52	58	53	67	<0.0001	<0.0001
Fritsche	-14	46	51	57	53	63	0.003	0.0003
	-19	49	48	55	55	64	0.002	0.0001
Hung	-35	44	50	56	60	61	0.0006	<0.0001
	-128	47	50	54	58	61	0.02	0.001

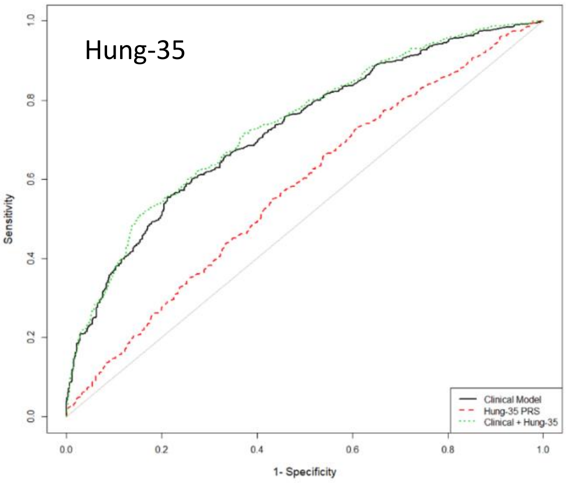

Shi
Sensitivity
1- Specificity
0.0
0.2
0.4
0.6
0.8
1.0
Clinical Model
Shi PRS
Clinical + Shi


Graff
Sensitivity
1- Specificity
0.0
0.2
0.4
0.6
0.8
1.0
Clinical Model
Graff PRS
Clinical + Graff


Jia
Sensitivity
1- Specificity
0.0
0.2
0.4
0.6
0.8
1.0
Clinical Model
Jia PRS
Clinical + Jia


Fritsche-14
Sensitivity
1- Specificity
0.0
0.2
0.4
0.6
0.8
1.0
Clinical Model
Fritsche-14 PRS
Clinical + Fritsche-14


Fritsche-19
Sensitivity
1- Specificity
0.0
0.2
0.4
0.6
0.8
1.0
Clinical Model
Fritsche-19 PRS
Clinical + Fritsche-19


Hung-35
Sensitivity
1- Specificity
0.0
0.2
0.4
0.6
0.8
1.0
Clinical Model
Hung-35 PRS
Clinical + Hung-35

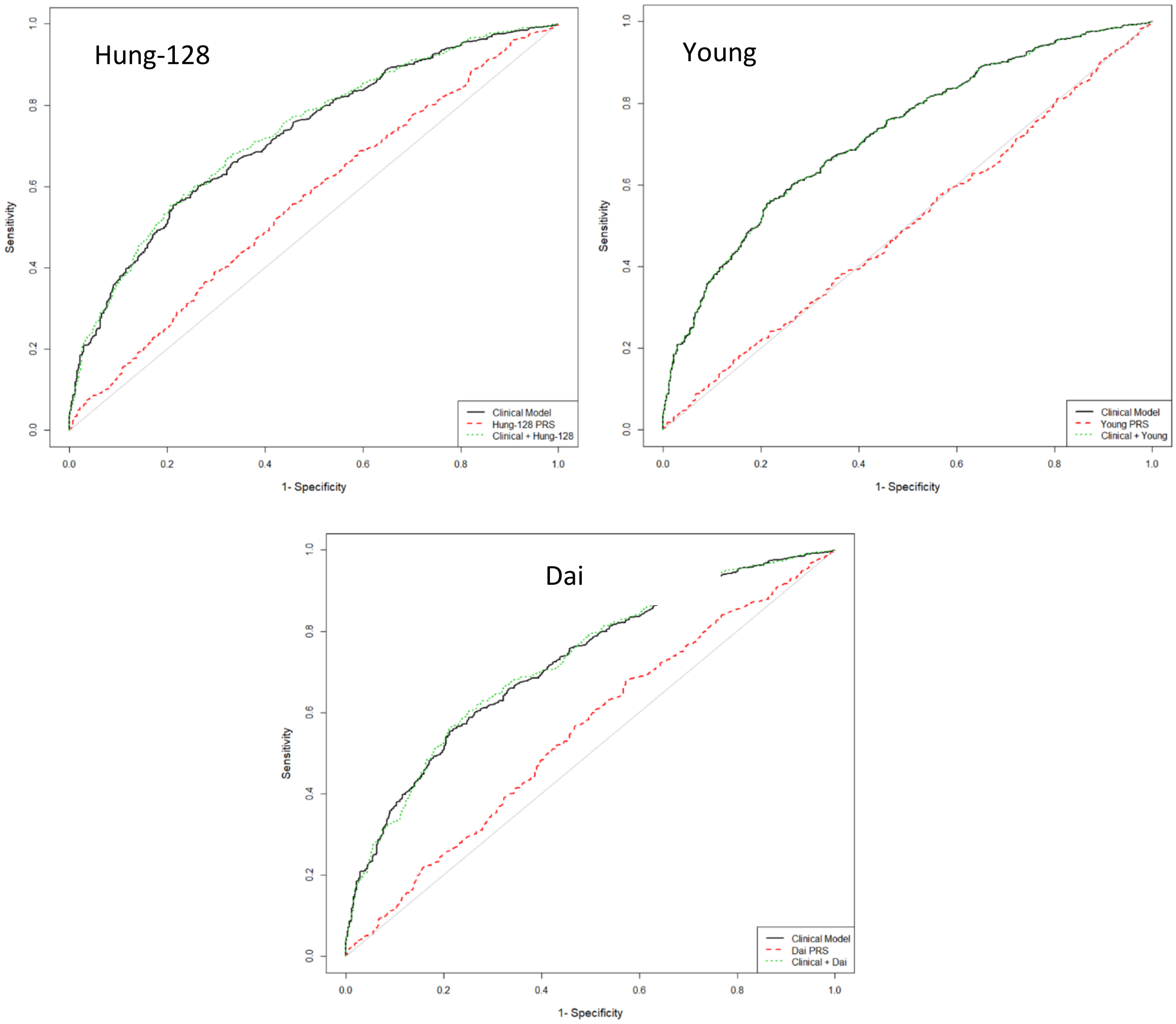


Figure 33. ROC curves for all PRSs, displayed independently and added to base clinical model.

5.3.2.4 Subgroup analysis

I performed subgroup analyses using the *Jia* PRS as it had the best overall performance (Table 19). PRS inclusion resulted in superior improvement in AUC when compared to the base clinical model in the following subgroups: age below the median, women, former smokers, BMI above the median, NLST-eligible, and adenocarcinoma-only cases. In most instances, the AUC improvement derived from the PRS was associated with poorer discrimination in the base clinical model, indicating that PRS utility is strongly linked to the dynamics of existing risk prediction strategies. The only exception was in the BMI subgroups, each of which had the identical performing base clinical models, resulting in the selection with the higher independent PRS AUC (BMI above median) having the larger PRS-derived AUC improvement (0.019 vs. 0.016).

The smallest PRS derived AUC improvement was observed in the NLST-ineligible subgroup (+0.003; p=0.05), which had the highest base clinical model performance (AUC 0.754); independent PRS AUC was 0.567. Whilst the independent PRS AUC was slightly lower in the ‘other histology’ subgroup (0.564), AUC improvement was larger (+0.007; p=0.007) due to a poorer performing base clinical model in that subgroup (0.75). These two subgroups had the highest AUCs when the base clinical model and PRS were combined (0.757).

The largest PRS derived AUC improvements were observed in the age below median and adenocarcinoma-only subgroups, both with an AUC increase of +0.022 (p<0.0001) above the base clinical model. Of those two subgroups, the adenocarcinoma-only selection had the higher overall AUC when the base model and PRS were combined (0.739 vs. 0.709). The age below median subgroup had the lowest combined AUC of all the selections (0.709).

Table 19. Subgroup analyses for *Jia* PRS validation.

Subgroup		% of total cohort (n)	AUC			
			PRS (95% CI)	Base Clinical Model (95% CI)	Base Model + PRS (95% CI)	AUC added by PRS (p-value)
Age	<Median	52.8 (635)	0.601 (0.554-0.648)	0.687 (0.643-0.731)	0.709 (0.666-0.752)	0.022 (<0.0001)
	>Median	47.2 (567)	0.583 (0.536-0.631)	0.71 (0.67-0.751)	0.731 (0.691-0.771)	0.021 (0.0003)
Sex	Female	53.1 (638)	0.609 (0.565-0.652)	0.717 (0.677-0.756)	0.737 (0.698-0.755)	0.02 (<0.0001)
	Male	47 (564)	0.567 (0.52-0.614)	0.726 (0.685-0.767)	0.738 (0.698-0.778)	0.012 (0.001)
Smoke Status	Former	53.7 (646)	0.594 (0.55-0.639)	0.703 (0.663-0.743)	0.722 (0.683-0.761)	0.019 (<0.0001)
	Current	46.3 (556)	0.577 (0.53-0.625)	0.738 (0.697-0.779)	0.749 (0.709-0.789)	0.011 (0.0007)
BMI	<Median	50.1 (602)	0.582 (0.536-0.629)	0.71 (0.669-0.752)	0.726 (0.685-0.767)	0.016 (0.0003)
	>Median	49.9 (600)	0.592 (0.546-0.637)	0.71 (0.669-0.752)	0.729 (0.688-0.769)	0.019 (<0.0001)
NLST	Eligible	76.1 (915)	0.595 (0.558-0.631)	0.73 (0.698-0.763)	0.746 (0.714-0.777)	0.016 (<0.0001)
	Ineligible	23.9 (287)	0.567 (0.501-0.633)	0.754 (0.697-0.811)	0.757 (0.701-0.814)	0.003 (0.048)
Case Histology	Adeno + Controls	79.6 (957)	0.602 (0.566-0.638)	0.717 (0.684-0.749)	0.739 (0.707-0.77)	0.022 (<0.0001)
	Other + Controls	66.1 (795)	0.564 (0.521-0.607)	0.75 (0.713-0.788)	0.757 (0.72-0.794)	0.007 (0.007)

Underline = superior performance seen in that subgroup.

5.3.2.5 PRS Scores and other Clinical Variables in the Control Cohort

Amongst the control cohort (for which additional clinical data were available), there was no significant association between any of the PRS scores and likelihood of being above the $PLCO_{M2012}$ or LLP_{V2} median, and no overall correlation between the RPM scores and any of the PRS scores. There was also no significant association between any PRS score and family history of lung cancer, previous COPD diagnosis, or LDCT detected CAC.

There was an association between scan-detected emphysema and higher *Hung-128* and *Graff* scores (*Hung-128*: p=0.045. *Graff*: p=0.02); 56.7% of the bottom quintile of *Hung-128* scores had emphysema, compared to 69.1% of the top quintile (p=0.078). In *Graff* these figures were 59% and 66% respectively (p=0.12).

Strikingly, in binary logistic regression analysis, previous cancer diagnosis was associated with higher *Graff* (p=0.006)*, Hung-35* (p=0.012) and *Hung-128* (p=0.023) scores with statistical significance, and *Shi* (p=0.077), *Fritsche-14* (p=0.09) and *Fritsche-19* (p=0.08) scores with borderline statistical significance. In total, there were 70 participants with previous cancer diagnoses, including 40% who reported either previous breast cancer (n=14) or skin cancer (n=15). These and the remaining cancer types are presented in Figure 34. Distribution of participants by previous cancer diagnosis status across the PRS quintiles is presented in Table 20. Discrimination analysis with previous cancer diagnosis used as the outcome resulted in AUCs of: *Hung-35:* 0.6 (0.529-0.67), *Graff:* 0.589 (0.517-0.661), *Hung-128:* 0.565 (0.491-0.639), *Fritsche-14:* 0.576 (0.5-0.65), *Fritsche-19:* 0.574 (0.504-0.643) and *Shi:* 0.551 (0.481-0.621).

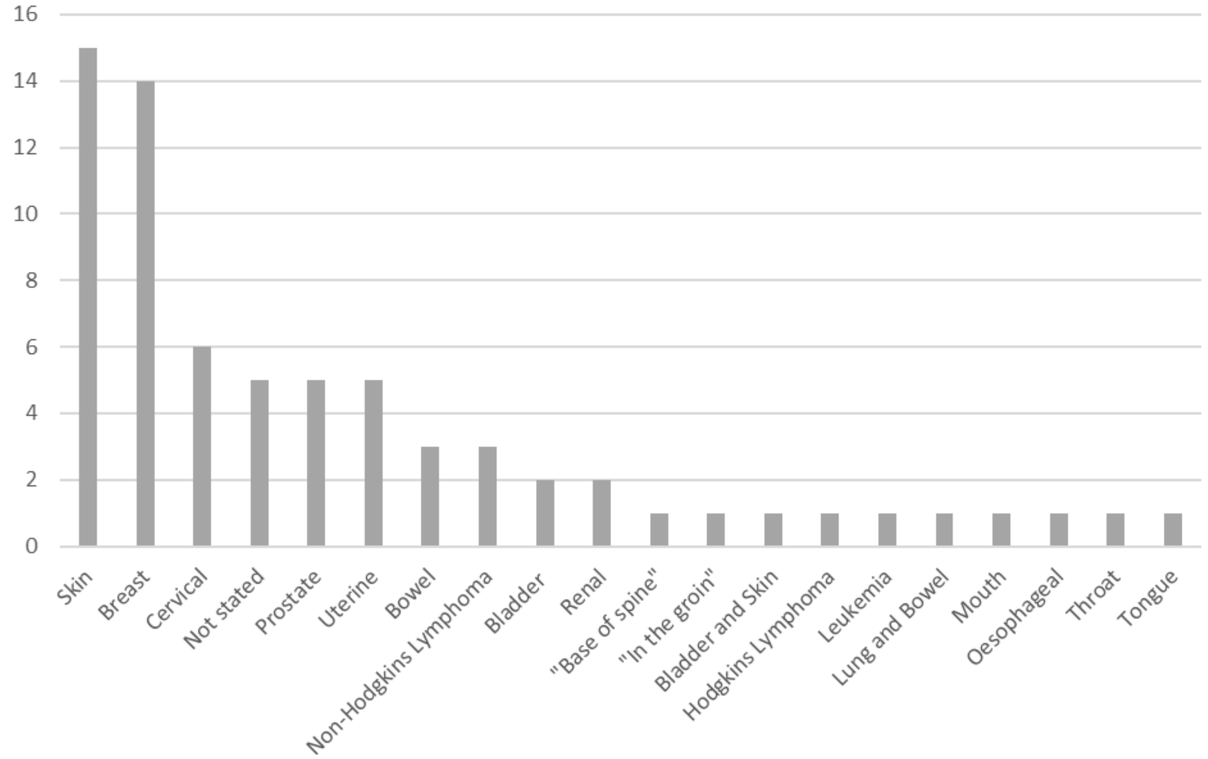


Figure 34. Reported previous cancer types in the 70 LHC pilot participants comprising the primary analysis dataset.

Table 20. Distribution of participants with previous cancer diagnosis across PRS quintiles.

PRS	**% (n) of PRS Quintile with Previous Cancer Diagnosis**					**P-value (linear-by-linear)**
	1 - lowest	2	3	4	5 - highest	
Graff	9.8 (13)	9.1 (10)	13.7 (14)	14.3 (15)	18 (18)	0.033
Hung-35	9.6 (13)	8.4 (10)	12.1 (13)	15.6 (15)	20.4 (19)	0.006
Hung-128	9.4 (12)	12.6 (15)	13.6 (15)	11 (11)	18.1 (17)	0.13
Fritsche-14	10 (13)	9.3 (11)	13.7 (14)	13.5 (15)	19.1 (17)	0.03
Fritsche-19	8.1 (10)	11.2 (14)	15.6 (17)	14 (15)	16.5 (14)	0.05
Jia	11 (16)	12.2 (14)	13.1 (13)	13.3 (15)	15.2 (12)	0.37
Shi	8.2 (11)	16.5 (17)	11.5 (13)	14.9 (17)	14 (12)	0.24
Dai	8.4 (11)	13.2 (15)	16.3 (17)	13.1 (14)	13.8 (13)	0.23
Young	15.2 (16)	11.1 (13)	18.2 (20)	10.1 (11)	9.2 (10)	0.2 (inverse direction)

5.3.2.6 Combined PRS

The SNP that appeared most frequently in the published PRSs was rs4236709 on chromosome 8; it was the only SNP to feature in five PRSs (*Jia*, *Hung-35, Fritsche-14, Graff and Dai*) (Appendix 2). In the Manchester dataset, the G effect allele had an allele frequency of 0.23, matching the MAF reported in three of the original studies. Whilst there was no statistically significant association for this SNP in this study (p=0.89), the raw OR was in the correct direction (A non-effect allele: 0.98. 95%CI 0.81-1.19). However, when adjusted for covariates this reversed (A non-effect allele: 1.01). The independent AUC, calculated using the mean of the reported natural logarithms of the ORs (lnORs) in the studies as the allelic weight, was 0.505 (0.477-0.533).

Eleven SNPs featured in at least four of the PRSs (Table 21). When these were synthesised into a single PRS, it resulted in an independent AUC of 0.57 and a combined base-clinical model and PRS AUC of 0.73, similar to *Hung-128* and *Fritsche-14*; it did not perform as well as *Jia, Hung-35* or *Fritsche-19* (Table 17).

Table 21. Combined PRS, SNPs shared by at least four of the validated PRSs.

SNP	Effect Allele	Effect Magnitude (lnOR)
rs11780471	G	0.141138
rs13080835	G	0.093074
rs4236709	G	0.107861
rs55781567	G	0.261372
rs56113850	C	0.124476
rs66759488	A	0.067721
rs71658797	A	0.129398
rs7705526	A	0.143531
rs77468143	T	0.105544
rs7953330	G	0.107096
rs885518	G	0.122077

5.3.3 Novel SNPs in the Manchester Cohort

5.3.3.1 Associated SNPs and Genes

After running a GWAS style association analysis, the expected vs. observed P-values in the primary analysis dataset showed a clear divergence from the null hypothesis for lower P-values, indicating a higher frequency of more significant hits than would be expected by chance (Figure 35). There was no early-separation of expected and observed frequencies at lower P-values, indicating that population stratification was appropriately controlled for at earlier QC steps [344]. The peaks on the Manhattan plot confirmed that there was correlation of genetic variants with the outcome of interest at several loci. Whilst no SNP reached standard genome-wide significance, considering the small size of this study, the clearly defined peaks visible are highly indicative of a locus of interest (Figure 36).

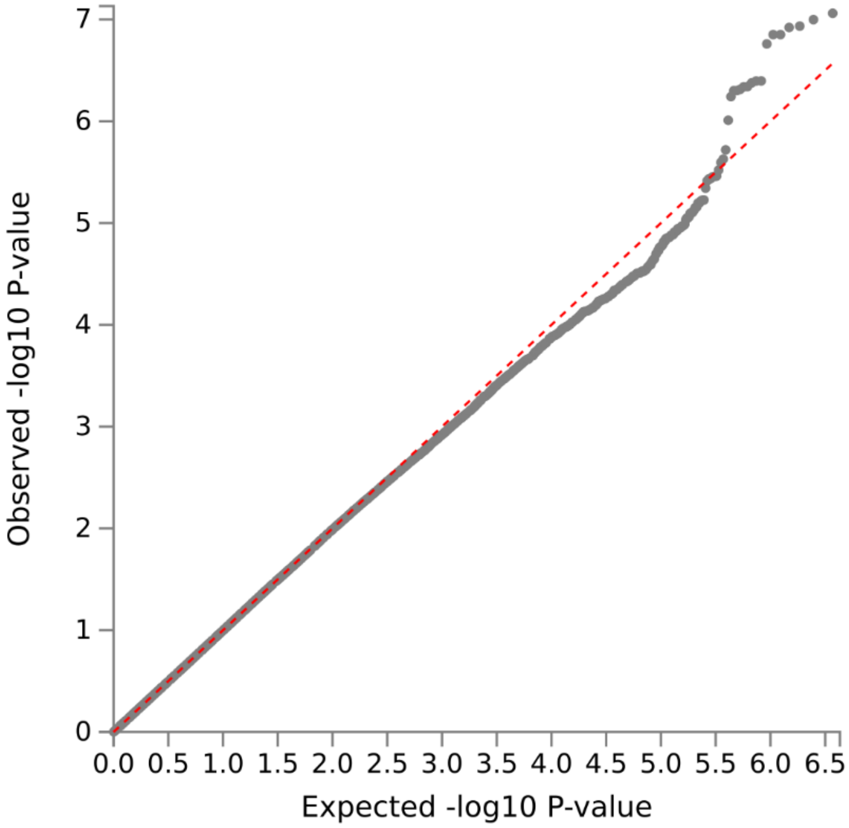


Figure 35. Functional Mapping and Annotation of Genome-Wide Association Studies (FUMA) generated quantile-quantile (Q-Q) plot for Primary Analysis.

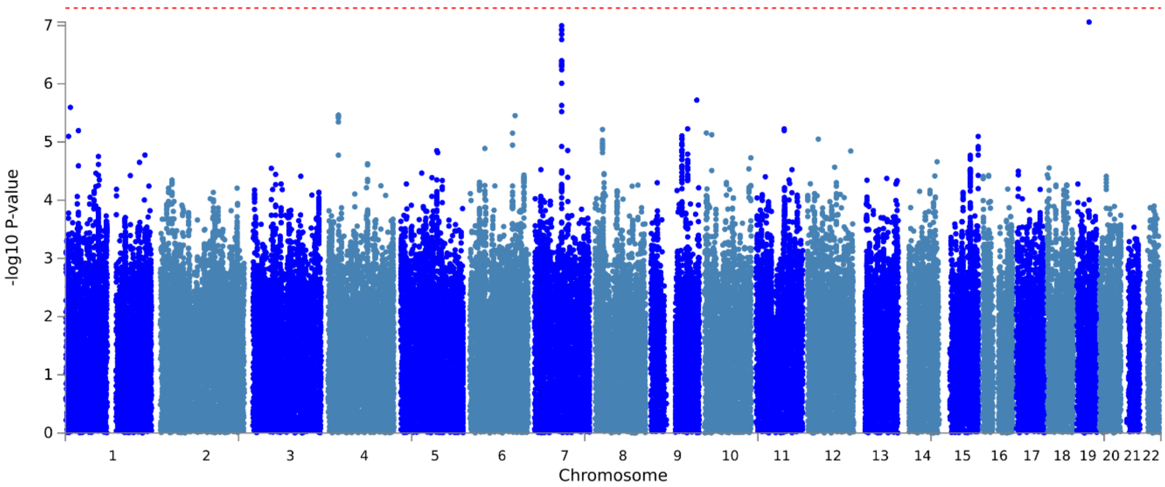


Figure 36. Functional Mapping and Annotation of Genome-Wide Association Studies (FUMA) generated Manhattan plot of Primary Analysis. Dashed line = threshold for genome-wide significance.

FUMA analysis resulted in the identification of 17 genomic risk loci, comprising of 206 candidate SNPs (complete list available in Appendix 3). Ten of the identified loci had multiple SNPs present, whilst seven were lone-SNPs (Table 22). In the Manhattan plot, there was a particularly pronounced peak at chromosome 7, with the lead SNP (rs17389497) reaching a P-value of 1.0048e-07 (Figure 36).

Seven of the seventeen lead SNPs were located on intronic regions of genes, all these genes are protein coding. Of the ten loci with more than a single significant SNP, five were mapped to specific genes. The two mapped genes containing SNPs at the highest significance level were *MAGI2* on chromosome 7 (lead SNP: rs17389497) and *DAPK1* on chromosome 9 (lead SNP: rs4878090). Plots of all the SNPs significantly associated with lung cancer in the Manchester cohort in these genomic regions are presented in Figure 37 and Figure 38. SNPs from these genes are not included in any of the published PRSs validated in section 5.3.2.

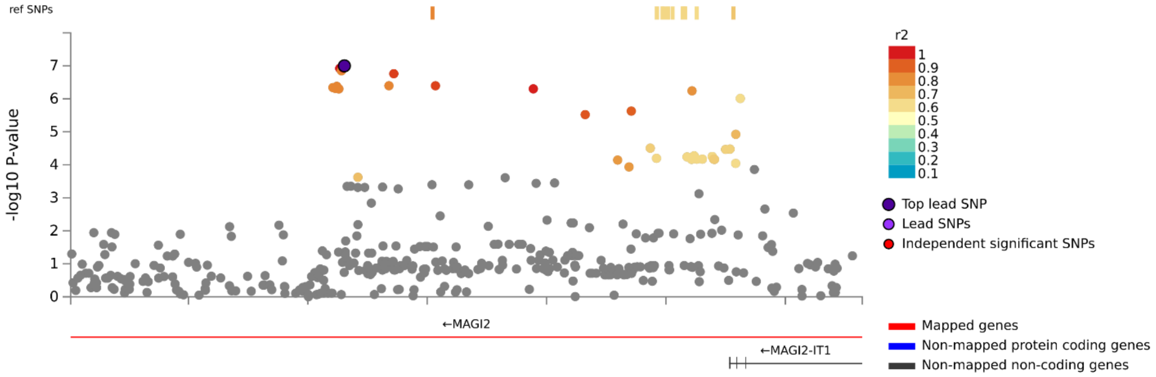


Figure 37. Regional plot for Manchester GWAS SNPs in the *MAGI2* gene. Displays the genomic environment of the SNPs in chromosome 7 that were significantly associated with lung cancer in my association analysis. Purple dot = top lead SNP. Red/orange/yellow dots = independent significant SNPs, linkage disequilibrium with top lead SNP. Red line = mapped gene.

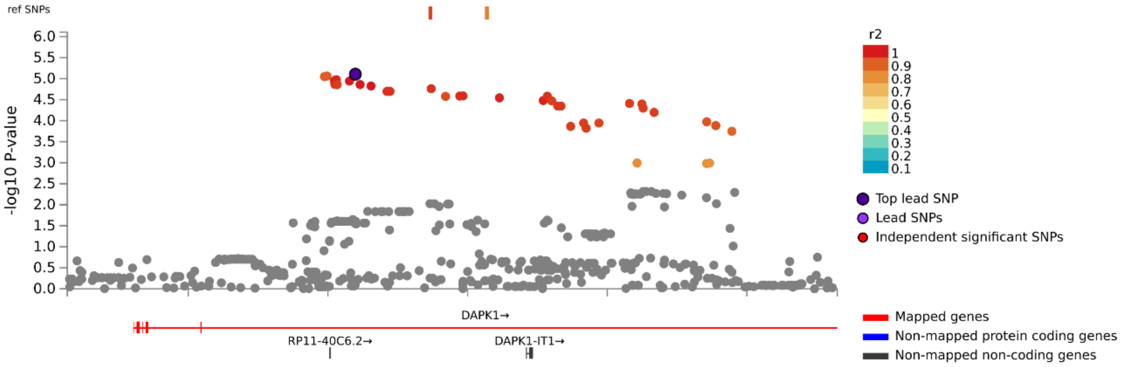


Figure 38. Regional plot for Manchester GWAS SNPs in the *DAPK1* gene. Displays the genomic environment of the SNPs in chromosome 9 that were significantly associated with lung cancer in my association analysis. Purple dot = top lead SNP. Red/orange/yellow dots = independent significant SNPs, linkage disequilibrium with top lead SNP. Red line = mapped gene.

Table 22. SNPs displaying high levels of significance from GWAS analysis of the primary analysis dataset. Shaded = lone SNP at that locus.

Chromosome	**Position**	**P-value**	**Locus Start**	**Locus End**	**Unique Candidate SNPs in Locus**	**Independent Significant SNPs in Locus**	**Lead SNP ID**	**Lead SNP rsID**	**Mapped Genes**	**Gene Type** [345]	**Nearest Gene (Distance)**	**SNP Position**
7	78466130	1.00E-07	78464211	78532462	53	1	7:78466130:C:G	rs17389497	*MAGI2*	Protein Coding	-	Intronic
4	30168938	3.46E-06	30168938	30233080	51	1	4:30168938:G:T	rs76640173	-	-	*RP11-174E22.2* (158999)	Intergenic
9	90143928	7.83E-06	90139505	90197840	40	1	9:90143928:A:G	rs4878090	*DAPK1*	Protein Coding	-	Intronic
1	37521740	6.38E-06	37521551	37534673	19	1	1:37521740:A:G	rs6676142	-	-	*GRIK3* (22009)	Intergenic
8	21502545	6.08E-06	21500728	21513913	12	1	8:21502545:G:T	rs12549783	-	-	*GFRA2* (45369)	Intergenic
15	101658292	8.02E-06	101658292	101667226	10	1	15:101658292:C:T	rs12898233	-	-	*RP11-424I19.1* (540)	Downstream
9	106313968	5.94E-06	106275350	106429018	6	1	9:106313968:A:G	rs79091275	-	-	*RP11-436F21.1* (121793)	Intergenic
11	81372090	5.96E-06	81264068	81372090	3	1	11:81372090:C:T	rs80240928	-	-	*RP11-664H7.2* (98897)	Intergenic
10	21078477	7.50E-06	21056587	21078477	3	1	10:21078477:A:G	rs12244585	*NEBL*	Protein Coding	-	Intronic
6	124413195	7.04E-06	124413195	124438622	2	1	6:124413195:A:G	rs17629528	*NKAIN2*	Protein Coding	-	Intronic
19	35610725	8.71E-08	35610725	35610725	1	1	19:35610725:C:G	rs4806091	*LGI4*	Protein Coding	-	Intronic
9	133023677	1.91E-06	133023677	133023677	1	1	9:133023677:A:G	rs79366318	-	-	*HMCN2* (23204)	Intergenic
1	14233836	2.54E-06	14233836	14233836	1	1	1:14233836:A:T	rs144716332	-	-	*PRDM2* (82261)	Intergenic

6	131852291	3.52E-06	131852291	131852291	1	1	6:131852291:C:T	rs12204890	-	-	*ARG1* (41992)	Intergenic
10	5551233	6.99E-06	5551233	5551233	1	1	10:5551233:G:T	rs35859955	*CALM5*	Protein Coding	-	Intronic
1	8985036	8.00E-06	8985036	8985036	1	1	1:8985036:C:T	rs10864370	-	-	*CALML3-AS1* (4973)	Intergenic
12	31469921	8.91E-06	31469921	31469921	1	1	12:31469921:A:G	rs67228087	*FAM60A*	Protein Coding	-	Intronic

5.3.3.2 Polygenic Risk Score Development

I generated polygenic risk scores for all participants in the Manchester case-control cohort using the 52 SNPs in the top p-value bracket ($p<0.00005$) of the PRSice output (Table 23). The PRS at this P-value threshold had a model fit of 0.47 (Figure 39). When all individuals in the cohort were scored and discrimination analysis performed, this PRS had an independent AUC of 0.843 (95%CI 0.822-0.865) and 0.927 (95%CI 0.912-0.941) when combined with the base clinical model (Figure 40). As the PRS was applied in the same dataset it was developed in, these AUC results are subject to a high degree of inflation [346]. It should be noted that lead SNP designation in particular loci differ between these results and those seen in the FUMA analysis (Table 22); this is likely a result of differing SNP pruning methodologies.

Table 23. Fifty-two SNP PRS developed in the Manchester cohort.

rsID	Chr	Allele A	Allele B	MAF	OR B vs. A	OR CI lower	OR CI higher	P	Beta
rs9791886	7	G	A	0.33026	0.672483	0.566915	0.797709	1.20E-07	-0.55148
rs79366318	9	G	A	0.04277	2.46274	1.57469	3.8516	1.91E-06	1.0697
rs76640173	4	G	T	0.020136	0.348977	0.18737	0.64997	3.46E-06	-1.70726
rs12204890	6	T	C	0.055656	0.504759	0.352791	0.72219	3.52E-06	-0.96511
rs79091275	9	A	G	0.097866	1.51502	1.14704	2.00105	5.94E-06	0.720836
rs80240928	11	C	T	0.035885	0.427131	0.271795	0.671244	5.96E-06	-1.17317
rs12549783	8	G	T	0.478879	1.37607	1.17137	1.61654	6.08E-06	0.438173
rs6676142	1	A	G	0.471167	1.38642	1.18004	1.62889	6.38E-06	0.412207
rs35859955	10	T	G	0.205618	0.683634	0.560565	0.833721	6.99E-06	-0.50859
rs17629528	6	G	A	0.034446	0.401602	0.252119	0.639716	7.04E-06	-1.16411
rs12244585	10	G	A	0.155607	0.652192	0.522534	0.814022	7.50E-06	-0.56167
rs4878090	9	A	G	0.140939	0.662386	0.525861	0.834356	7.83E-06	-0.60434
rs10864370	1	C	T	0.258812	0.700706	0.583469	0.841501	8.00E-06	-0.46962
rs12898233	15	C	T	0.048388	1.89751	1.27185	2.83096	8.02E-06	1.03855
rs67228087	12	G	A	0.058747	0.518916	0.366351	0.735018	8.91E-06	-0.8586
rs58482754	6	T	G	0.180407	1.41761	1.14628	1.75318	1.29E-05	0.563692
rs11766078	7	T	C	0.101924	0.67888	0.520664	0.885173	1.40E-05	-0.68143
rs76919286	5	T	C	0.040131	0.451113	0.29497	0.689912	1.41E-05	-0.98126
rs77599664	5	G	A	0.036255	0.472425	0.303151	0.736217	1.53E-05	-1.21989
rs10990760	9	A	C	0.167878	1.43524	1.15302	1.78655	1.61E-05	0.522484
rs61824298	1	A	G	0.053963	0.554755	0.386945	0.79534	1.68E-05	-1.00558
rs2869032	15	C	T	0.199779	1.33933	1.09622	1.63635	1.70E-05	0.497026
rs9419171	10	T	C	0.298442	1.50669	1.26417	1.79574	1.87E-05	0.432071
rs117161983	14	G	A	0.026222	0.381141	0.222783	0.652064	2.19E-05	-1.30755
rs11120170	1	G	A	0.406878	1.26872	1.07684	1.49478	2.24E-05	0.455825
rs35933466	4	G	A	0.056602	0.592662	0.417492	0.841328	2.36E-05	-0.88557
rs11165236	1	G	A	0.222006	1.42614	1.17207	1.73527	2.43E-05	0.489492

rs117355684	12	A	G	0.03879	0.547217	0.358651	0.834924	2.71E-05	-1.04518
rs12232670	18	A	G	0.261702	1.39282	1.15775	1.67561	2.79E-05	0.440224
rs7638857	3	A	C	0.254167	0.703026	0.584773	0.845193	2.84E-05	-0.44752
rs11599670	10	A	G	0.057325	2.15418	1.47733	3.14114	2.90E-05	0.836599
rs34498099	7	G	T	0.020364	0.326291	0.17431	0.610785	2.99E-05	-1.43081
rs55716073	17	G	A	0.354226	0.764132	0.646163	0.90364	3.18E-05	-0.44852
rs173780	5	G	A	0.150087	0.656735	0.524459	0.822373	3.42E-05	-0.56896
rs72712347	1	C	A	0.041044	2.01287	1.29908	3.11886	3.44E-05	1.06042
rs62501800	8	C	T	0.094506	1.78971	1.34094	2.38865	3.51E-05	0.647157
rs62246869	3	T	C	0.072782	1.76695	1.27605	2.44671	3.60E-05	0.743982
rs2621166	18	T	C	0.31425	1.47733	1.24001	1.76007	3.67E-05	0.407044
rs6940739	6	A	G	0.175384	1.52187	1.22606	1.88907	3.69E-05	0.50219
rs117542875	16	G	T	0.022752	0.435589	0.247942	0.765252	3.76E-05	-1.43804
rs17601891	1	A	G	0.036144	2.35798	1.45743	3.81496	3.78E-05	1.06945
rs80319952	3	C	T	0.084289	0.610578	0.456632	0.816425	3.88E-05	-0.70922
rs8046839	16	C	T	0.196946	0.752877	0.615567	0.920815	3.90E-05	-0.5282
rs118124361	11	G	A	0.109925	1.57807	1.21043	2.05736	3.96E-05	0.637515
rs62076880	18	T	C	0.04385	0.469631	0.313262	0.704055	4.14E-05	-0.9643
rs75794645	13	A	G	0.020776	0.331831	0.178761	0.615975	4.22E-05	-1.44574
rs945879	1	C	T	0.185552	0.684317	0.55673	0.841144	4.47E-05	-0.47149
rs11678507	2	C	T	0.462036	0.705195	0.600088	0.828711	4.50E-05	-0.37127
rs111748305	5	A	C	0.044475	0.54598	0.367573	0.810979	4.59E-05	-0.92324
rs2477469	13	G	A	0.301962	0.703396	0.58939	0.839454	4.67E-05	-0.40971
rs422331	6	C	T	0.400596	1.36755	1.15982	1.61249	4.91E-05	0.371941
rs74738049	12	A	C	0.060951	0.549926	0.391295	0.772867	4.99E-05	-0.77818

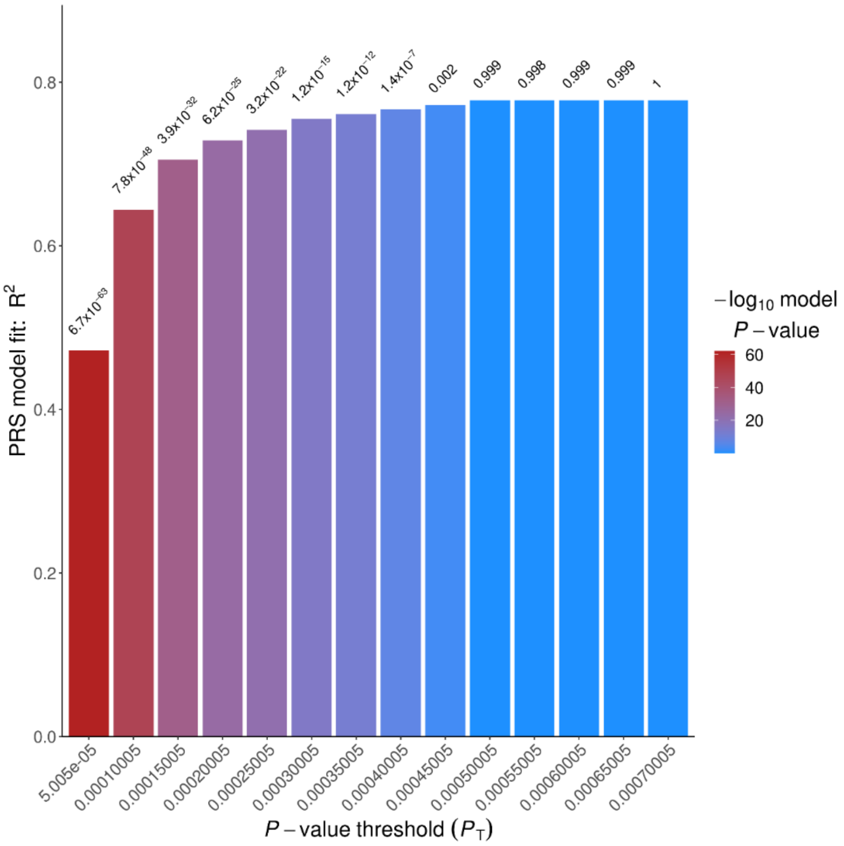


Figure 39. Chart showing the model fit of various PRSice-2 generated PRSs, with increasingly lenient P-value thresholds (and more SNPs included).

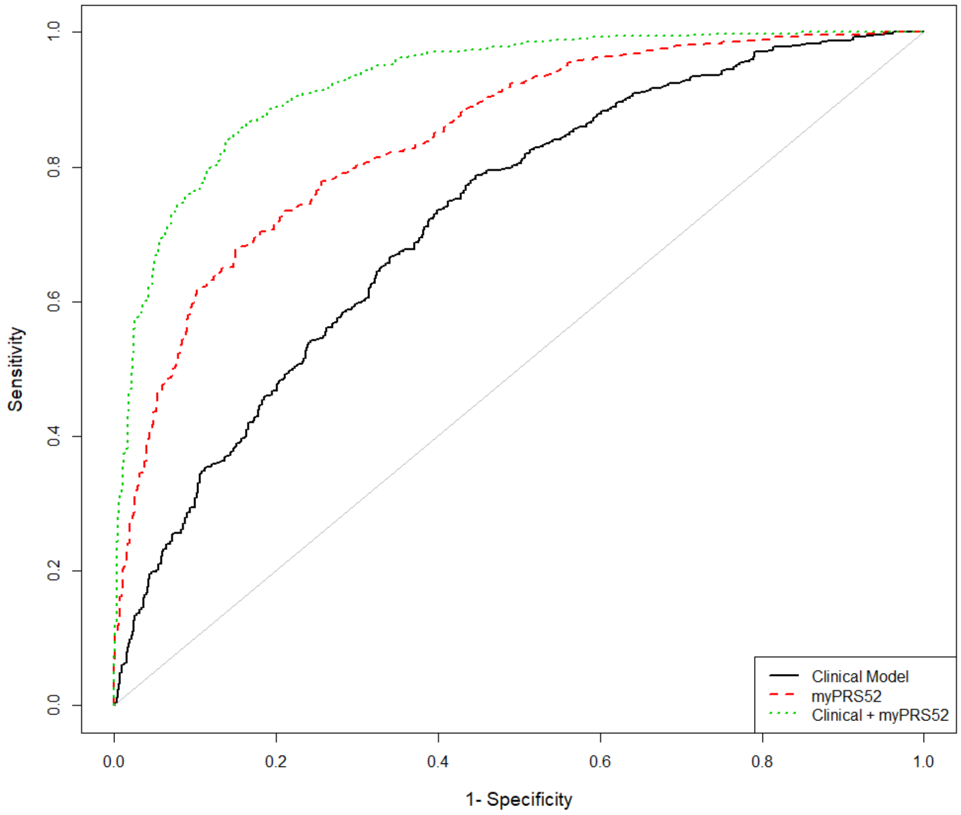


Figure 40. ROC curves for PRS52 in the Manchester case-control cohort.

5.3.4 Secondary Analysis

The secondary analysis cohort had a larger control sample set than the primary analysis cohort; it retained a significant number of control samples that had aberrant PCA output (see section 5.2.3). These were excluded from the primary analysis to ensure robustness; here, I reran the analysis with these samples included to test whether it produces different results. The total dataset consisted of 639 control samples and 652 case samples (Figure 20). The demographics of the 91 controls that were only included in the secondary analysis dataset were broadly similar to those of the controls included in both the primary and secondary analysis datasets, except for mean age and median PLCO$_{M2012}$ scores which were significantly higher in the secondary-only cohort (mean age: 65.6 vs. 64.5; p=0.05. median PLCO$_{M2012}$: 4.2 vs. 3.4; p=0.02) (Table 24). The case cohort was the same in both primary and secondary analyses. When comparing case and control demographics in the secondary analysis dataset, most comparisons remained the same as in the primary analysis. The difference in median pack years for non-imputed values became statistically significant (45 vs. 44; p=0.04), although the actual difference was minor (Table 25).

Table 24. Comparison of secondary analysis-only controls and primary & secondary analysis controls.

		Controls in Both Primary and Secondary Analysis	Controls in Secondary Analysis Only	P-value
Total		548	91	-
% Female (n)		50.5 (277)	48.4 (44)	0.7
Mean Age ±SD		64.5 ±5.3	65.6 ±4.9	0.05
% Current Smoker (n)		51 (280)	43.4 (40)	0.21
Median Pack Years ±IQR		45 ± 21	45 ± 28	0.6
Educational Attainment	% Less than GCSE/O Level (n)	66.6 (365)	71.4 (65)	0.55
	% GCSE/O Level (n)	18.1 (99)	13.2 (12)	
	% A Level (n)	4.6 (25)	5.5 (5)	
	% Some University/College (n)	7.7 (42)	4.4 (4)	
	% University Degree (n)	2 (11)	3.3 (3)	
	% Postgrad (n)	1.1 (6)	2.2 (2)	
Mean BMI ±SD		28.9 ±5.2	28.2 ±5	0.32
Median FEV$_1$/FVC Ratio ±IQR		70.2 ±13	70.2 ±12	0.65
% COPD (n)		30.7 (168)	35.2 (32)	0.39
Median PLCO Score ±IQR		3.4 ±3.4	4.2 ±4.7	0.02

Table 25. Comparison of case and control demographics in secondary analysis.

		Controls	Cases	P-value
Total (%)		639 (49.5)	652 (50.5)	-
% Female (n)		50.2 (321)	55.2 (360)	0.07
Median Age ± IQR		65 ±9	69 ±9	<0.001
% Current Smokers (n)		50.1 (320)	41.9 (273)	0.004
Median Pack Years ± IQR	Direct	45 ±23	44 ±25	0.04
	Imputed	45 ±23	40 ±15	<0.001
Median BMI ± IQR		28.3 ±6.8	26 ±6.7	<0.001
Median FEV_1/FVC Ratio ± IQR	Direct	70.2 ± 13	67 ± 16	<0.001
	Imputed	70.2 ±12.5	65.5 ±13	<0.001
% NLST eligible (n)		75.6 (483)	76.3 (498)	0.74

The AUC for the base clinical model was 0.716 (0.6881-0.7433), slightly lower than in the primary analysis. All raw AUCs associated with the functional PRS validations were either the same or slightly lower than those seen in the primary analysis. There were no significant differences in p-value for additional AUC over the clinical model between the primary and secondary analysis (Table 26). Due to the lack of significant differences, I did not perform detailed subgroup analyses for the secondary analysis and assumed that the primary analysis cohort was as representative of the wider population as the secondary.

For PRS generation using PRSice, the equivalent p-value threshold which resulted in a 52 SNP PRS in the primary analysis resulted in a 57 SNP PRS in the secondary analysis. This PRS had an AUC of 0.835 (0.814-0.857) and 0.91 (0.895-0.925) when combined with the base clinical model, both slightly lower than the equivalent metrics from the primary analysis.

Table 26. Comparison of PRS performance in primary and secondary analyses.

PRS	Primary AUC (95%CI)	Secondary AUC (95%CI)	Primary AUC Clinical Model + PRS (95% CI)	Secondary AUC Clinical Model + PRS (95% CI)	Primary Additional AUC over clinical model (p-value – clinical model vs. clinical model + PRS)	Secondary Additional AUC over clinical model (p-value – clinical model vs. clinical model + PRS)
Dai	0.552 (0.519-0.585)	0.55 (0.519-0.581)	0.726 (0.698-0.754)	0.72 (0.692-0.747)	0.003 (0.02)	0.004 (0.01)
Shi	0.56 (0.528-0.593)	0.56 (0.524-0.587)	0.732 (0.704-0.76)	0.724 (0.697-0.751)	0.009 (<0.0001)	0.008 (<0.0001)
Graff	0.553 (0.52-0.585)	0.55 (0.519-0.748)	0.728 (0.7-0.756)	0.721 (0.694-0.748)	0.005 (0.002)	0.005 (0.002)
Jia	0.588 (0.556-0.62)	0.586 (0.555-0.617)	0.738 (0.71-0.766)	0.731 (0.704-0.758)	0.015 (<0.0001)	0.015 (<0.0001)
Fritsche-14	0.562 (0.529-0.594)	0.562 (0.531-0.593)	0.731 (0.703-0.759)	0.725 (0.698-0.752)	0.008 (<0.0001)	0.009 (<0.0001)
Fritsche-19	0.569 (0.537-0.602)	0.569 (0.538-0.6)	0.733 (0.705-0.761)	0.726 (0.699-0.754)	0.01 (<0.0001)	0.01 (<0.0001)
Hung-35	0.575 (0.542-0.607)	0.573 (0.542-0.604)	0.734 (0.706-0.762)	0.727 (0.7-0.754)	0.011 (<0.0001)	0.01 (<0.0001)
Hung-128	0.562 (0.53-0.595)	0.557 (0.526-0.589)	0.73 (0.702-0.758)	0.723 (0.696-0.75)	0.007 (0.0001)	0.007 (0.0003)

5.4 Discussion

5.4.1 Validation of Published PRSs

In this chapter, I tested the hypothesis that assessing the genetic profile of ever-smokers improves the prediction of lung cancer risk. To do this, I validated several previously published polygenic risk score tools for lung cancer prediction in a Manchester based case-control cohort. Most of these tools were developed and validated in large RCT or biobank-based datasets; consequently, validation of their efficacy in a high-risk cohort, representative of a screening population, is an important step in assessing the potential benefit of using genetic risk factors in real-world lung cancer screening selection.

My results showed that eight of the nine PRSs tested were predictive of lung cancer risk in the Manchester cohort. When applied in conjunction with a base model comprised of several non-genetic risk factors, discrimination was significantly improved by the addition of each of these eight PRSs, albeit by varying magnitudes (AUC added to base clinical model ranged from 0.003 to 0.015 depending on the PRS). The only PRS that did not improve discrimination was the *Young* PRS, which was developed over a decade ago, demonstrating the importance of building PRS tools based on large GWAS datasets rather than self-selected candidate genes. Of the 251 unique SNPs tested across all the PRSs, 37 (15%) reached nominal statistical significance ($p<0.05$) and 170 (68%) were in the correct effect direction in the Manchester dataset. Fundamentally, these results demonstrate that the inclusion of robust measures of genetic risk could improve lung cancer risk prediction.

The best performing PRS (*Jia*) had an independent AUC of 0.59 and added approximately 0.015 AUC to the base clinical model. The *Fritsche-19* and *Hung-35* PRSs also added more than 0.01 AUC to the base clinical model. None of these PRSs had been previously validated in a cohort including individuals recruited from an actual lung cancer screening programme. Whilst the AUCs of these PRSs may appear relatively modest, the benefits of slight improvement in model discrimination and risk prediction may aggregate when utilised in risk stratification of a large population in the context of a screening programme [347]. For example, studies have demonstrated that even with limited AUC improvement, the *Jia* and *Hung* PRSs could be used to effectively modulate the screening commencement age for individual smokers in a population [244,348]. By highlighting segments of the screening cohort at even higher disease risk due to their genetic profile, PRSs could also be used to target the provision of chemoprevention drugs, assist in the triage of indeterminate screening results, or regulate the frequency of screening rounds [349].

Alternative approaches to AUC in assessing the clinical utility of PRSs have been suggested; net reclassification index (NRI) is one example [312]. Using this metric, adding the *Jia* PRS to the base clinical model would have resulted in cases being 11% more likely to move up a risk quartile than down. This was notably higher than the NRI observed in other PRSs tested, even if the magnitude of difference in AUC did not appear as stark. This finding is consistent with a previous study that demonstrated the *Graff* PRS's ability to successfully reclassify non-cases into lower risk categories, even though AUC appeared modest [243]. This demonstrates how PRSs could be used to direct screening away from those who may not benefit, limiting exposure to harms. It should be noted that there has been significant criticism published regarding the statistical robustness of NRI, and results should be interpreted with caution [350,351]. However, when considered in combination with AUC and observations of the proportion of case individuals in each of the PRS quintiles, it is clear that these PRSs have potential clinical utility.

In the Manchester cohort, the *Jia* PRS displayed superior discrimination among women, former smokers, and adenocarcinoma-only subgroups. Many SNPs are associated with specific lung cancer histological subtypes, although several significant SNPs are shared [12,352]. PRS construction must ensure that there is a sufficiently diverse array of SNPs on the panel to predict several types of lung cancer. Evidence for sex-stratification in PRS performance in several disease areas has been published previously [353,354]. There is evidence of increased genetic lung cancer risk in women compared to men [93,94,355]; further research is required to assess this discrepancy and ensure PRS tools perform adequately in all segments of the screening cohort.

I found a significant association between at least four of the PRS tools and previous cancer diagnosis. Sample size was not large enough to allow for stratification by individual cancer type. The *Graff* PRS, which had the second largest cross-cancer association after *Hung-35* in our study, has previously been shown to demonstrate pleiotropy, having the highest cross-cancer association out of all of the cancer PRSs tested in two large biobank cohorts [338]. The inclusions of SNPs from *TERT-CLPTM1L* and *HLA* in the PRS is likely to have contributed to this association, as both are well established general cancer susceptibility loci [356–358]. To my knowledge, the *Hung-35* PRS (which had the most significant association with previous cancer diagnosis in our study) has not had its cross-cancer predictiveness tested previously. SNP pleiotropy may be useful in developing a cross-cancer PRS for risk prediction [338]. If such a PRS were employed in individuals attending lung cancer screening, this may facilitate the integration of additional diagnostic and preventative services into the screening programme, increasing benefit to participants and overall programme cost effectiveness. On the other hand, it may also lead to the selection of screening participants at higher risk of other cancers which may reduce life expectancy and benefit to be gained from lung cancer screening. It should be noted that both $PLCO_{M2012}$ and LLP_{V2} already include previous cancer diagnosis as a factor for lung cancer risk prediction. Several theories have been advanced to explain the increased risk of lung cancer after another cancer diagnosis (see section 1.5.6). Further research is required to examine the proportion of predictiveness conferred by a PRS that is already being provided by previous cancer diagnosis and other clinical factors.

It is important to note that a statistically significant improvement in risk prediction does not always indicate a clinically significant improvement. It may not be cost effective to include an auxiliary biomarker to risk prediction tools (particularly one as costly as genetic testing) when the additional predictiveness it confers is marginal compared to traditional risk factors such as age and smoking history. Previous studies have demonstrated that RPMs for lung cancer perform extremely successfully without the inclusion of genetic factors [225,243]. Unless a very specific target population which will particularly benefit from SNP testing is identified, an additional tranche of SNPs are discovered significantly improving PRS predictiveness, or genetic testing reduces in price considerably, it may not be cost effective to pursue PRS integration into RPMs. Formal cost effectiveness analysis is required once PRS tools are fully developed, validated and optimised.

A study published in February 2022 claimed to be the most extensive and fully adjusted prospective study of lung cancer incidence to date [27]. It developed a 33 SNP PRS for lung cancer and applied it prospectively in a UK Biobank cohort of close to 346,000 participants, followed-up for an average of seven years resulting in 1687 cases of incident lung cancer. It reported that high genetic risk and smoking status were both independently and robustly associated with lung cancer risk. In never-smokers, there was no difference in lung cancer risk based on genetic risk factors. This is unlike other studies which have shown risk stratification in never-smokers driven by genetics [242,243], although it should be noted that the prospective nature of this study limited its statistical power and the authors posit that an increase in incident lung cancer cases might result in a signal emerging in never-smokers [27]. Unfortunately, the study was published too late for the PRS to be validated here; however, 26 of the 33 SNPs in the PRS were present in PRSs I validated, so it can be assumed that this PRS would perform similarly to those in this study.

5.4.2 Novel PRS Development

In addition to validating previously published lung cancer PRSs, I developed a novel 52 SNP PRS in the Manchester dataset. Whilst internal cross-validation or external validation of the PRS was not performed (significantly limiting the robustness of any conclusions drawn), two particularly promising genomic loci had a high level of significance after association analysis.

One cluster of associated SNPs was located on the *MAGI2* gene [359]. This gene is primarily expressed in the brain. Its protein product is involved in anchoring, and providing a scaffolding for, cellular signalling proteins and maintaining the structure of neuronal synaptic junctions. It is implicated in several neurodegenerative diseases [360]. The gene is also expressed at lower levels in non-brain tissues. Previous studies have demonstrated that MAGI2 regulates the activity of PTEN, an important tumour suppressor [361]. There is evidence of *MAGI2* mutations in various cancer types [362], and it has been suggested that *MAGI2* itself might act as an independent tumour suppressor [360]. Some studies have proposed MAGI2 as a biomarker that could be used to predict prostate cancer aggressiveness and recurrence, informing treatment and surveillance decisions [363–366].

In recent years, studies have reported an association between the *MAGI2-AS3* long non-coding RNA (lncRNA) and lung cancer [367]. Under-expression of this molecule has been observed in NSCLC tissues and cell lines [368], and is associated with poor survival in patients [369]. Upregulation of *MAGI2-AS3* decreases NSCLC cell viability and invasiveness. The molecule may have potential as a prognostic biomarker or as a target for cancer therapy [367]. Interestingly, a SNP near *MAGI2* (rs2714700; risk allele=C) has been implicated in nicotine dependence in the UK Biobank cohort; the SNP was independent of other genetic loci implicated in smoking behaviour [370], such as the well characterised *CHRNA5* gene which is strongly linked to lung cancer risk (see section 1.5.8.2). In our dataset, this SNP was significantly associated with lung cancer (risk allele: C. $_{adj}$OR: 1.2. p=0.04), even after adjustment for smoking status. However, there was no significant linkage disequilibrium between this SNP and the lead *MAGI2* SNPs from the FUMA or PRSice analyses. *MAGI2* has not been previously implicated in lung cancer risk prediction, and none of the PRSs tested in the validation portion of this study included SNPs from this gene. If further validation studies with larger cohorts and additional smoking history-based adjustments confirm this association, it may be an excellent candidate for inclusion in future lung cancer PRS tools.

The second locus of interest was in the *DAPK1* gene on chromosome 9. This gene is involved in the modulation of cell apoptosis and autophagy, and functions as a tumour suppressor. DAPK1 under-expression has been reported in several cancer types [371–374], and may be an independent prognostic biomarker in clear cell renal cancer [372]. One study showed that upregulating DAPK1 resulted in increased natural killer cell activity and reduced tumour immune evasion in gastric cancer cells [371]. There is some evidence linking *DAPK1* function and lung cancer. For example, one study of 135 patients with stage I NSCLC found that 44% had hypermethylation of the gene (repressing expression); those with the altered gene expression had significantly poorer survival (p=0.007). This was the only independent predictor of disease-specific survival among all parameters tested [375].

To my knowledge, there are no published studies linking SNPs on *DAPK1* with lung cancer risk. None of the chromosome 9 SNPs included in the PRSs validated in this study were located in or near *DAPK1*, nor were any of them in linkage disequilibrium with the lead *DAPK1* SNP identified by FUMA. If validated in external datasets, this locus may offer additional SNPs for inclusion in PRSs for lung cancer.

5.4.3 Strengths and Limitations

A significant limitation of this study is that clinical and demographic data for the case cohort were too limited to facilitate the calculation of $PLCO_{M2012}$ risk scores. Lung cancer screening implementation in England relies on individual RPM scores and risk thresholds for screening selection (see Chapter Four). Had $PLCO_{M2012}$ scores been available for the whole cohort, it would have been possible evaluate whether the PRSs validated would have augmented risk prediction over and above the actual methods being used for selection in screening programmes. Some of the variables included in RPMs such as family history and tobacco smoke exposure might already be accounting for a portion of the risk impact conferred by genetic variants. Considering genetic risk factors in combination with demographic and lifestyle risk factors and testing them in actual screening populations (such as has been done with the BOADICEA and Tyrer-Cuzick breast cancer RPMs [376,377]) ensures that personal risk is not overestimated, and that the genetic component of the RPM has independent utility in a screening selection context. The demographic variables that were available facilitated the construction of a base clinical model that was used as a substitute for an RPM in our analysis. Age was the main contributing factor to discrimination, which would be expected from RPM based risk prediction. However, the AUC derived from this model was lower than those of RPMs observed in comparative studies [225]; better performing non-genetic RPMs may reduce the added benefit of including genetic risk factors [243].

Additional demographic variables would also have allowed for more granular adjustment of SNP effect sizes and significance in GWAS analysis and novel PRS development. This would improve our understanding of the contribution the SNPs are making to risk prediction compared to other clinical variables which may interact with the genetic factors. Smoking-based factors such as pack years, years quit, and age started, for which data were not complete in the case cohort are particularly important to adjust for, as SNPs associated with lung cancer risk often mediate their impact by influencing smoking behaviour [378].

The case cohort in this study was derived from a biobank rather than a screening programme setting, and the cancers were likely to have been diagnosed in clinic rather than through screening. Whilst the 80% early-stage distribution of the cancers in the case cohort are a fair representation of the expected stage distribution in UK-based screening programmes [204], there may be differences in the clinical and genetic characteristics of these cases compared to screen-detected cases. It should be noted that as the cases in this study were sourced from the MCRC Biobank, we would not expect to see the ‘healthy volunteer bias’ often reported as a limitation in studies which validate PRSs in the UK Biobank cohort [27]. Although we do not have individual-level deprivation data for the cases, the MCRC Biobank collects samples from patients located in the same socio-economically disadvantaged areas of Manchester from which the control cohort was sourced. This, and the generally well-matched nature of the case and control cohorts, are significant strengths of this study. Going forward, studies (ideally prospective) in which both the case and control cohorts are derived from the same screening programme, with uniform demographic data available, will provide more robust evidence as to the impact of PRS inclusion on risk prediction.

This study was restricted to individuals of European descent, a common feature of many PRS studies for lung cancer. The *Dai* PRS, developed and validated in Chinese datasets, is the only published and validated lung cancer PRS tool tailored for use in a non-European population. In the Manchester cohort, this PRS performed least successfully; it can be assumed that applying a European-developed PRS in a Chinese population would have similar results, as many SNPs associated with lung cancer risk are specific to a particular ethnicity. This highlights the importance of the development and validation of PRS tools in a wide variety of populations, ensuring that the use of PRS in screening selection does not exacerbate health inequalities, with White-European patients benefiting from disease risk prediction of a superior accuracy than participants of other ethnicities [379,380]. The lack of non-European GWAS and Biobank datasets is a significant challenge in the development of PRSs for diverse populations [349,381].

Despite these limitations, this study contributes important evidence supporting the hypothesis that lung cancer risk prediction in ever-smokers can be improved by considering genetic risk factors. I demonstrated that PRS tools predominantly developed in RCT or biobank populations functioned successfully in a case-control cohort highly representative of a target population for lung cancer screening. Consequently, there is potential for a PRS to be integrated into RPMs used for lung cancer risk prediction, improving model discrimination and thereby refining screening selection in community-based screening programmes. The exceptionally high-risk nature of the control cohort in this study adds robustness to this conclusion; this cohort would not be expected to be protected from lung cancer due to limited exposure to other risk factors. Similarly, novel SNPs that emerged in this cohort, whilst requiring external validation, are derived from a cohort highly representative of Manchester-based screening attendees and could potentially be included in future PRS development.

A unique added benefit of using genetic factors in lung cancer risk prediction is that they are often independent of other risk factor-associated comorbidities such as COPD and CVD. The factors that contribute most to risk prediction in RPMs are age and smoking history. A limitation of using such an approach for screening selection is that the criteria favour older and more comorbid attendees for screening, who are at the greatest risk of lung cancer but may have limited benefit to be gained from screening. This phenomenon was observed in the Manchester LHC pilot, with multiple measures of respiratory and cardiovascular comorbidity associated with calculated lung cancer risk and screen-detected lung cancer (see section 3.3.5). This chapter adds additional evidence to this observation, with the case cohort being significantly older and having lower FEV_1/FEV ratio (indicative of COPD) than the control cohort. One approach to mitigating this limitation is to employ benefit-based selection, such as LYFS-CT (see Chapter Four), which considers life expectancy when determining screening eligibility. An alternative approach may be to integrate additional non-comorbidity linked factors, such as polygenic risk, which stay constant through an individual's life and are not necessarily linked to increased risk of other diseases [382]. These approaches are not mutually exclusive, as LYFS-CT relies on lung cancer risk prediction in its determination of the estimated benefit an individual will receive from screening [228]; improved risk prediction therefore enhances both benefit-based and risk-based screening selection.

5.4.4 Implementation and Next Steps

With further research and validation, it is not inconceivable that a PRS for lung cancer could be deemed viable for inclusion in pre-screening risk calculation. At that point, practical considerations relating to clinical implementation must be considered. Several biomarker studies embedded within lung cancer screening trials and programmes have provided evidence for the acceptability of blood collection from participants within screening settings, and that a pipeline for blood storage and transportation, as well as the subsequent extraction of DNA and genotyping, is feasible [383–385]. Whilst this provides good proof of concept, establishing a participant's PRS after their initial contact with the screening service means that it could not be used to inform screening eligibility at their initial assessment. The PRS could still be used to inform decisions regarding screening interval, or to exclude low-risk individuals from further scans, but this limits the potential utility and effectiveness of the PRS. Asking participants to attend a separate clinic some time prior to screening for blood extraction would solve this issue but may reduce uptake and compliance amongst the target population. The genetic testing of saliva, rather than blood, could be an effective solution to this implementation challenge, as mailed collection kits can be returned by the participants for genotyping prior to any in-person contact with the screening service. Saliva collection has been shown to be acceptable to participants and a viable source of DNA for genotyping in several screening studies [386–388].

Appropriate genetic counselling infrastructure must be implemented for PRS to become a routine tool for screening selection. An individual's understanding of their lung cancer risk, how genetic factors influence their risk, and what impact this knowledge has on health behaviour and anxiety, are all important psychological considerations, some of which I investigate in Chapter Six. Research examining patient interest in PRS testing in other disease areas has revealed broadly positive attitudes [389,390]; patients also seem to receive their genetic risk score without significant distress or anxiety and are able to recall the information accurately [391,392]. Development of tools for counselling patients in polygenic risk is ongoing [393]. This research will need to be replicated in lung cancer screening populations, particularly considering that those at high risk of lung cancer often live in deprived areas, have low educational attainment and may have limited health literacy [248].

In conclusion, in this chapter I demonstrated that genetic risk factors could improve lung cancer risk prediction in ever-smokers and may assist with screening selection. Further studies refining the PRS tools, validating novel SNPs, integrating PRSs into RPMs, and testing them prospectively in real-world screening settings are necessary to further advance this research area. Ultimately, the routine adoption of a PRS tool within a lung screening programme will depend on its clinical impact and cost effectiveness. As it stands, the small (but significant) improvement in AUC conferred by inclusion of a PRS in an RPM is unlikely to position wide-scale genetic testing in the context of lung cancer screening to be cost-effective, especially considering how successful lung cancer risk prediction using standard risk factors is. On the other hand, an effective PRS might reduce the total number of people eligible for screening or reduce the frequency of screening, increasing programme efficiency. It might also favour the selection of those who have a lower smoking exposure and therefore a lower burden of comorbidity who have ‘more to gain’ from screening. To reduce the cost of the test, the PRS could be targeted to those close to the risk threshold (above and below) rather than being used more broadly. Ongoing research may also identify specific subgroups that will gain particular benefit from PRS testing. For example, whilst not currently recommended [394], if lung cancer screening for never-smokers becomes clinically viable in the coming decades, genetic risk prediction will take on a much more significant role. This is because never-smokers are not exposed to the primary risk factor for lung cancer, adding more predictive weight to secondary factors such as family history and genetic variants. It is also expected that genetic testing will continue to reduce in price and it may become a regularly used clinical tool for a range of conditions, which may further reduce the overall cost of applying it in lung cancer screening selection [395]. Once a well optimised PRS tool, applicable to a range of populations and a variety of lung cancer types, becomes available, formal cost effectiveness analysis will be required to determine the best approach to application, as has been performed in other disease areas [396–398].

Chapter Six - Lung Cancer Risk Perception: Study of Manchester LHC Programme

6.1 Introduction

Lung cancer is the leading cause of cancer death globally [1]. Providing LDCT screening to high-risk individuals has been shown to reduce lung cancer mortality significantly. NLST and NELSON, the two RCTs that demonstrated this mortality benefit, used generalised eligibility criteria to determine screening eligibility [193,194]. Targeting screening based on individually calculated lung cancer risk is an important method of improving programme efficiency and reducing harm to participants [218]; implementation studies and programmes (including NHS England's TLHC programme [48]) often use RPMs to individually calculate participants' lung cancer risk scores and determine screening eligibility based on a risk threshold. Consequently, screening programme attendees (many of whom are from socio-economically deprived areas) are having their individual lung cancer risk score calculated, communicated, and used to determine screening eligibility, in real-world clinical settings.

The number of participants who will have their risk calculated is expected to increase dramatically in the coming years. Despite this, there is no recommended risk communication procedure included in the TLHC protocol [48]. Furthermore, studies investigating participants' perceptions of lung cancer risk, knowledge of the disease, and the psychological impact of concerns regarding lung cancer, are very limited (see section 1.7.2). There is an urgent need for further research in this area. Understanding participant risk perception is an important step in ensuring risk communication effectively facilitates shared decision making [256]. Having an awareness of participants' disease knowledge and worry would enable pre-screening counselling and screening related messaging to be targeted and delivered appropriately. There may also be opportunities to leverage risk perception and risk communication as methods to mediate health behaviours such as screening adherence and smoking cessation [274,275].

This study aimed to examine lung cancer risk perception, disease knowledge and lung cancer-specific worry in attendees of the Manchester LHC programme, a community-based lung cancer screening programme (see objectives - section 1.8). I examined associations between these measures, as well as between responses and a range of key demographic factors. To my knowledge, this is the first study to examine the relationship between these outcomes and the individually calculated actual lung cancer risk scores of the participants in such a setting.

6.2 Questionnaire Development and Methods

6.2.1 Measures

I developed the study questionnaire based on previous studies of a similar nature in other disease areas. The questionnaire evaluated three broad areas: lung cancer risk perception, disease knowledge and lung cancer-specific worry (full questionnaire available in Appendix 3 - 8.4). Each area was assessed with multiple individual questions:

6.2.1.1 Risk Perception

Three questions assessed personal lung cancer risk perception. To assess perceived absolute risk, participants were asked: *"In an imaginary room filled with 100 people exactly like you, how many do you think will get lung cancer in the next 6 years?"* The participant responded by completing the phrase: *"___out of 100"*. This question is based on similar questions found in previous risk perception studies [265]. The question was adapted to use plain language, as well as frequencies rather than percentages, as these techniques have been shown to aid in patient understanding of risk [399]. Six years was chosen as the risk timeframe to facilitate comparison with participants' actual $\text{PLCO}_{\text{M2012}}$ lung cancer risk score.

Perceived comparative risk was assessed by asking participants to estimate their risk of developing lung cancer in comparison to other people their age: *"What do you think your chances of developing lung cancer are compared to other people your age in the UK population?"* Participants responded on a 5-point Likert scale ranging from *'much less likely'* to *'much more likely'*. Participants were also asked to compare their lung cancer risk with other smokers (for smokers) or other former smokers (for former smokers) on the 5-point Likert scale. These questions are adapted from previously published lung and breast cancer risk perception studies [400–402].

6.2.1.2 Disease Knowledge

Three questions assessed general knowledge of lung cancer. The first related to disease incidence in the UK population: *"What do you think the risk of developing lung cancer is for **any person** in the general population of the UK?"* The participant was asked to select one of six options presented in both odds and percentage formats. This is a truncated version of a question from a previously published study [265]. Participants were asked to identify factors which increase an individual's risk of developing lung cancer by selecting *'Yes'*, *'No'* or *'Not Sure'* for each of 12 risk factors: age, air pollution, asbestos exposure, asthma, being overweight, changes or mutations in your genes, chronic obstructive pulmonary disease (COPD), family history of lung cancer, poor diet, radon exposure, smoking and second-hand smoke. Finally, participants were asked: "*Think about all the people who are diagnosed with lung cancer in the UK in one year. What percentage* *of these patients do you think will survive for… a. 1 year. b. 5 years. C. 10 years ...following their diagnosis?"* Participants provided an estimate for each of these timeframes.

6.2.1.3 Disease Specific Worry

Participants were asked to provide both the frequency of any lung cancer-specific worry they had experienced in the previous two weeks, and the impact of any lung cancer-specific worry on their mood. Answers were on a 4-point scale ranging from *'not at all' to 'nearly every day'* for frequency, and *'not at all'* to *'a lot'* for impact. These questions were adapted from the Lerman Cancer Worry Score instrument, with the timescale in the frequency question added to correspond to the PHQ-4 instrument [403]. The final question was the PHQ-4 measure of anxiety and depression, a validated brief four-part survey for detection of depressive and anxiety disorders [404,405]. PHQ-4 scores are rated as normal (0-2), mild (3-5), moderate (6-8) and severe (9-12). A total score of ≥3 for the first two questions suggests anxiety; a total score of ≥3 for the last two questions suggests depression. I used PHQ-4 results as an additional covariate when analysing responses but did not employ it as a primary outcome variable.

6.2.2 Patient and Public Involvement and Engagement

Prior to commencement of the study, the questionnaire was reviewed by a specialist cancer patient and public involvement and engagement (PPIE) panel convened by the Manchester Biomedical Research Centre (BRC). All participants (n=3) would have been eligible for invitation to the Manchester Lung Health Check Programme (and therefore, eligible for this study) based on age and smoking history. One had family history of cancer. Participants completed the survey and provided oral feedback as to the ease of comprehension of the questions, ease of providing answers and general readability. Participants' feedback was positive and indicated that the questionnaire was generally understandable and did not result in undue distress. Several amendments were made to the questionnaire to address points raised by the panel (Table 27).

Table 27. Changes made to study survey after PPIE consultation.

Question	Original	Updated	Comments
1	What do you think the risk of developing lung cancer is **for any person** in the general population (UK)?	What do you think the risk of developing lung cancer is **for any person** in the general population of the UK? *Please circle one answer.*	Readability improvements and answer prompt.
1	Eleven answer options ranging from 'Inevitable' to 'Very Unlikely'	Six answer options ranging from '1 chance in 2 - 50%' to '1 chance in 100 - 1%'.	Participants felt overwhelmed by number of options.
2	What do you think your risk of developing lung cancer is over the next 6-years? (Complete a or b) a. 1 in ____ OR b. ____ % **Alternative Wording:** In an imaginary room filled with 100 people just like you, how many will get lung cancer in the next 6 years? ___out of 100	In an imaginary room filled with 100 people exactly like you, how many do you think will get lung cancer in the next 6 years? *Please fill in the blank.* _____out of 100	Participants preferred the 'alternative wording'. "Just like you" adapted to "Exactly like you" for improved scientific accuracy. Answer prompt added.
4 + 5	-	*If you are an **ex-smoker**, please **skip** question 4* *If you are a **current smoker**, please **skip** question 5:*	Additional instructional prompts provided. Shaded boxes added to indicate optional questions.
6	Have you ever been told a score indicating your personal risk of lung cancer?	Have you ever been told a score or percentage indicating your personal risk of lung cancer? *Please circle one answer.*	Readability improvements and answer prompt.
7	Lots of small changes in your genes - **Yes/No/Not Sure** A few major changes in your genes - **Yes/No/Not Sure**	Changes or mutations in your genes - **Yes/No/Not Sure**	Original deemed too complex for audience.
8	What percentage of people diagnosed with lung cancer survive… a. 1 year? ______% b. 5 years? _____% c. 10 years? _____%	Think about **all** the people who are diagnosed with lung cancer in the UK in one year. a) What percentage do you think will survive for 1 year after diagnosis? _____% b) What percentage do you think will survive for 5 years after diagnosis? _____% c) What percentage do you think will survive for 10 years after diagnosis? _____%	Readability improvements
End	-	**If anything in this questionnaire has caused you concern, please discuss it with the nurse during your Lung Health Check.**	Added to conclusion of questionnaire.

6.2.3 Analysis

This study took place in the Manchester LHC programme, an expansion of the Manchester LHC pilot. A detailed description of screening programme structure and study recruitment is presented in section 2.12. Due to time constraints endemic to operating this study within the context of a clinical service, not all participants completed the whole questionnaire. Therefore, the denominators vary for different analyses in the study (Table 28). Questionnaires were included in analysis if the participant completed at least one question comprehensively. The subsections of the risk factor question were treated as distinct questions for this purpose, with any remaining blank responses classified as 'blank' for basic descriptive analysis, or 'don't know' for association analysis. If a participant completed the comparative risk question for both current smokers and former smokers, both responses were discounted. Incomplete and non-valid responses to the three-part survival question (e.g. survival estimates increasing with time rather than decreasing) were excluded as it indicated a misunderstanding of the question. I describe the responses of the complete cohort, and then stratify them by screening eligibility.

Statistical analysis was carried out using IBM SPSS Statistics V.25. Comparison of medians was conducted by Mann-Whitney U test. Comparison of means was conducted by T test. P values for trend were calculated by Pearson's Chi-Squared test or fitting linear models, using a single independent variable with 1 degree of freedom. Correlations between two scale variables were measured with Pearson's Correlation or Spearman's Rho.

Associations between demographic variables and questionnaire responses were assessed using binary logistic regression models, with univariable and multivariable analyses performed in each instance. In order to investigate associations between the outcome of interest and overall lung cancer risk score, as well as individual demographic factors that contribute to the risk score, I built two models for the logistic regression analyses. Model A included $\text{PLCO}_{\text{M2012}}$ score, sex, PHQ-4 score, and either lung cancer worry frequency or perceived comparative risk (depending on the dependant variable selected) as covariates. Model B substituted $\text{PLCO}_{\text{M2012}}$ score for individual demographic variables: age, family history of lung cancer, pack-years, smoking status, BMI, previous cancer diagnosis, and educational attainment, as well as sex, PHQ-4 score and either lung cancer worry frequency or perceived comparative risk, as in Model A.

6.3 Results

6.3.1 Study Cohort Characteristics

A total of 371 first-round participants of the Manchester LHC programme were recruited for this study. Of those, 243 (65.5%) consented to their responses being linked with the clinical data collected during the LHC ('identified'). There was minimal difference between the questionnaire responses provided by identified and non-identified individuals (Table 28); the only difference in responses that approached statistical significance were indicators of depression (14% vs. 25%; p=0.03) and PHQ-4 score (mean: 2.5 vs. 2.9; p=0.06) which were lower in the identified subset, although mean PHQ-4 scores in both groups remained within the 'normal' category. Demographic characteristics of the identified subset are presented in Table 29. Median age was 68 (IQR ±11), 56% were female, 60% had no educational qualifications, 96% were white (when ethnicity was provided), 20% were current smokers and the median pack-year history was 23 (IQR ± 31). Median index of multiple deprivation (IMD) was 5631 (IQR ±10679); more than 53% of participants fell in the two most deprived IMD deciles.

Table 28. Study questionnaire responses for all respondents and stratified by consent to linked clinical data.

Measures			**All**	**Consent for data linkage**		**P value**
				Yes	**No**	
Number (%)			371	243 (65.5)	128 (34.5)	-
Risk Perception						
Perceived Absolute Risk Score	Median ± IQR		20 ± 22	20 ± 20	15 ± 24	0.67
	Total Respondents		368	242	126	-
Perceived Comparative Risk. % (n)	Lower		21 (76)	21 (51)	20 (25)	0.9
	The Same		49 (180)	49 (118)	49 (62)	
	Higher		30 (111)	30 (71)	32 (40)	
	Total Respondents		367	240	127	-
Population Risk Estimate % (n)	Underestimate		19 (70)	19 (46)	19 (24)	0.47
	Correct		11 (39)	9 (22)	13 (17)	
	Overestimate		70 (259)	72 (172)	68 (87)	
	Total Respondents		368	240	128	-
Exposure to Risk Score Previously	% Yes (n)		4 (12)	5 (10)	2 (2)	0.19
	Total Respondents		304	200	104	-
Disease Knowledge						
Mean number of risk factors correctly identified (/11)			6 ± 2	6 ± 2	6 ± 2	0.45
Mean Estimated Survival ±SD	1-year		61 ± 2	59 ± 2	65 ± 3	0.1
	5-years		39 ± 22	37 ± 21	22 ± 3	0.03
	10-years		22 ± 20	20 ± 19	22 ± 3	0.08
	Total Valid Respondents		185	123	62	-
Lung Cancer Worry						
Lung Cancer Specific Worry	Frequency % (n)	Any	18 (60)	17 (38)	20 (22)	0.58
		None	82 (276)	83 (185)	81 (91)	0.5
		Several days	12 (41)	12 (27)	12 (14)	
		More than half the days	3 (11)	3 (6)	4 (5)	
		Nearly every day	2 (8)	2 (5)	3 (3)	
		Total Respondents	336	223	113	-
	Impact % (n)	Any	38 (129)	37 (82)	41 (47)	0.4
		None	62 (209)	63 (142)	59 (67)	0.3
		A small amount	29 (99)	29 (65)	30 (34)	
		Quite a lot	7 (24)	6 (13)	10 (11)	
		A lot	2 (6)	2 (4)	2 (2)	
		Total Respondents	338	224	114	-
Mental Health Score						
PHQ-4	Mean Score (±SD)		2.5 ± 3.3	2.2 ± 3.1	2.9 ± 3.6	0.06
	Overall Classification % (n)	Normal	67 (215)	70 (150)	60 (65)	0.03
		Mild	16 (50)	16 (34)	15 (16)	
		Moderate	9 (29)	7 (15)	13 (14)	
		Severe	9 (29)	7 (16)	12 (13)	
	Total Respondents		323	215	108	-
	Depression Indicated % (n)		18 (58)	14 (31)	25 (27)	0.02
	Anxiety Indicated % (n)		22 (70)	19 (41)	27 (29)	0.11

Table 29. Demographic characteristics for the identified subset of the study participants.

Variable			Value
Total			243
% Female (n)			55.6 (135)
Median Age (±IQR)			68 (±11)
Median BMI (±IQR)			28.6 (±7.7)
% Current Smoker (n)			19.8 (48)
Median Pack Years (± IQR)			23 (±31)
% Family History Lung Cancer (n)			16.9 (41)
% Previous Cancer Diagnosis (n)			15.2 (37)
Index of Multiple Deprivation (IMD)	Median Rank (±IQR)		5631 (±10679)
	% Decile (n)	1	40.3 (98)
		2	13.2 (32)
		3	6.2 (15)
		4	27.2 (66)
		5	4.1 (10)
		6	5.3 (13
		7	3.7 (9)
% Educational Attainment (n)	Left school age <16 with no qualification		51 (124)
	Left school with no qualification		9.1 (22)
	CSEs or equivalent		4.1 (10)
	O-levels		4.5 (11)
	A-levels or equivalent		1.2 (3)
	Some college		25.1 (61)
	College graduate		3.7 (9)
	Postgraduate		0.8 (2)
	Prefer not to say		0.4 (1)
% Ethnicity (n)	White - British		61.7 (150)
	White - Irish		2.5 (6)
	White - Other		0.4 (1)
	Asian – Pakistani		0.8 (2)
	Black - Caribbean		0.4 (1)
	Black - Other		0.4 (1)
	Other		0.4 (1)
	Not Known		20.6 (50)
	Not Stated		12.3 (30)
% Asbestos Exposure (n)			15.2 (37)
PLCO$_{M2012}$ Lung Cancer Risk – % 6-year	Median (±IQR)		1 (±2.5)
	Mean (±SD)		2.34 (±3.68)

6.3.2 Questionnaire Responses

6.3.2.1 Risk Perception

The median perceived absolute 6-year risk of lung cancer was 20% (IQR ±22). When participants were asked about their comparative risk of developing lung cancer compared to others their age, approximately half classified themselves at *'the same'* level of risk, 21% considered themselves at 'lower' comparative risk (*'less likely'* or '*much less likely')* and 30% classified themselves as at 'higher' comparative risk (*'more likely'* or *'much more likely')* (Table 28). Respondents with above average perceived absolute risk (≥20%) were more likely to have 'higher' comparative risk compared to those with below average perceived absolute risk (34% vs. 26%; p=0.04). Median perceived absolute risk increased progressively across the three perceived comparative risk groups; those in the 'lower' comparative risk group had a median perceived absolute risk of 14%, those in 'the same' comparative group had a median of 20%, and those in the 'higher' group had a median of 23% (p=0.02). Only 4% reported having received their lung cancer risk score previously.

6.3.2.2 Lung Cancer-Specific Worry

The majority (82%) of the study cohort reported no lung cancer-specific worry in the previous two weeks, 12% reported worry on *'several days',* 3% on *'more than half the days'* and 2% *'nearly every day'.* Participants were also asked if lung cancer-specific worry had any impact on their mood: 62% reported no impact, 29% reported *'a small amount'*, 7% *'quite a lot'* and 2% *'a lot'* of impact (Table 28). There was a significant difference in perceived risk between those who reported worry and those who did not. Amongst those who reported worry, median perceived absolute risk was 25%, compared to 20% in those with no worry (p=0.008). Perceived comparative risk was also higher, with 48.3% of those who reported worry in the 'higher' comparative risk group, compared to 24.7% of those who did not report worry (p<0.001). Only 7% of those with worry were in the 'lower' comparative risk group, compared to 23.6% of those without worry (p<0.001).

6.3.2.3 Lung Cancer Knowledge

Only 11% of respondents identified the correct lifetime risk of developing lung cancer for anyone in the UK (7% risk - an average of UK male and female lifetime lung cancer risk [406]). More than 70% of respondents overestimated population risk, with almost half selecting the highest two options (25% or 50% risk) (Table 28) (Figure 41). Those with below average perceived absolute risk were more likely to choose the correct population risk estimate (16% vs. 6%; p=0.002) and less likely to overestimate (55% vs. 84%; <0.001).

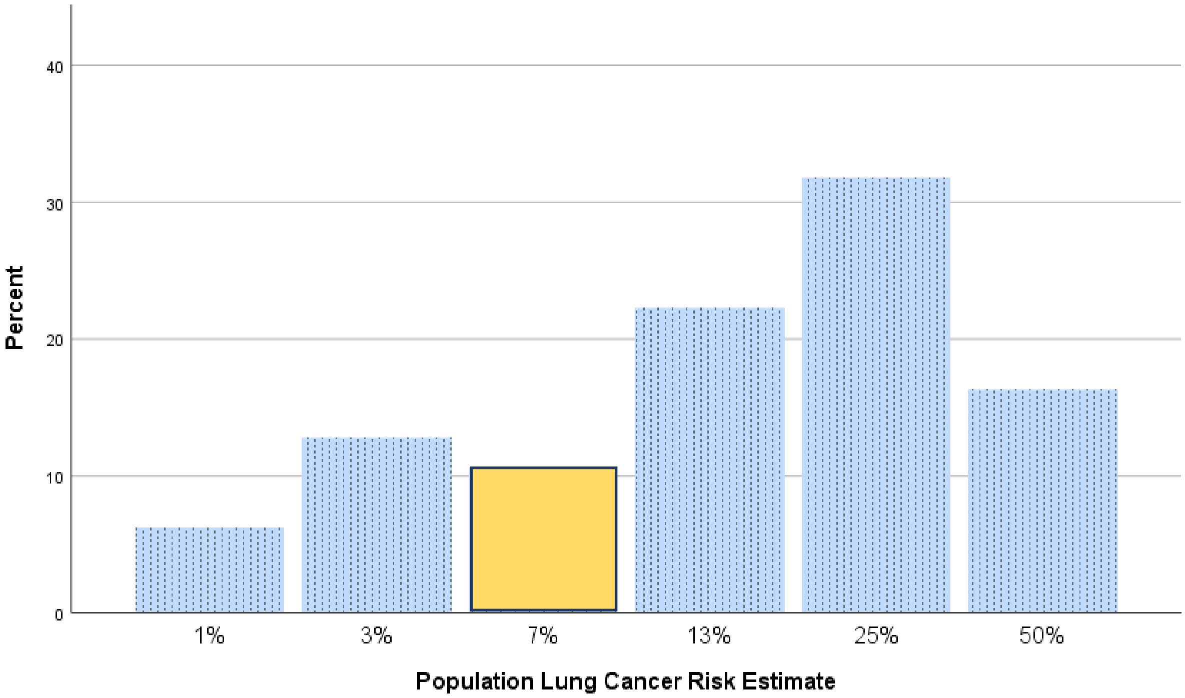


Figure 41. Study participant estimates of UK lung cancer incidence. Shaded bar = correct response.

Estimated survival after a lung cancer diagnosis was also widely overestimated; the reported mean one-, five- and ten-year survival were 61% ±24, 39% ±21, and 22% ±22 respectively. This compares to actual UK survival of 41%, 16% and 9.5% [13] (Figure 42).

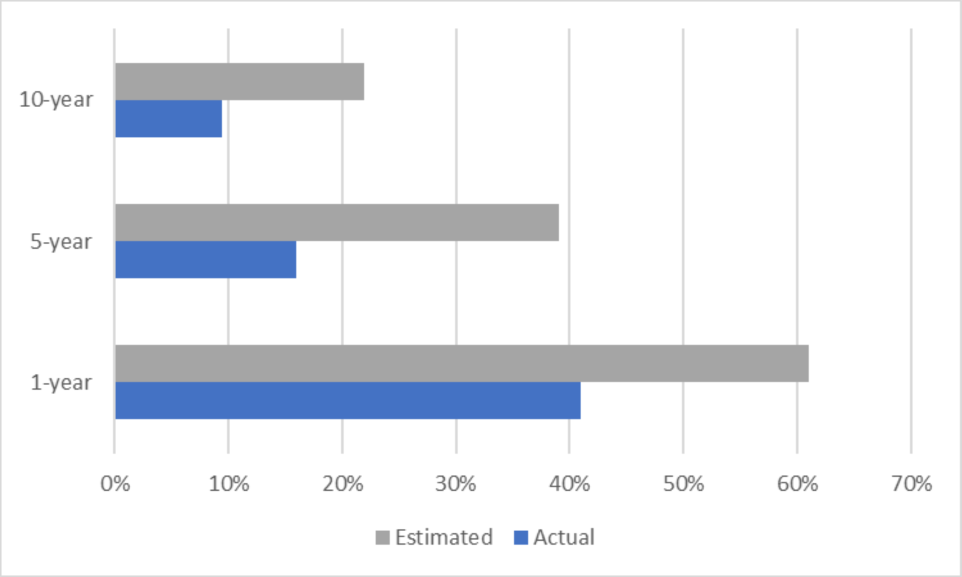


Figure 42. Study participant mean survival estimates vs. actual mean survival.

Participants correctly identified an average of 6.4 lung cancer risk factors out of eleven. Smoking, asbestos exposure, passive smoking and pollution were the risk factors classified correctly most frequently (smoking: 93%; asbestos: 92%; passive smoking 84%; pollution 76%) (Figure 43).

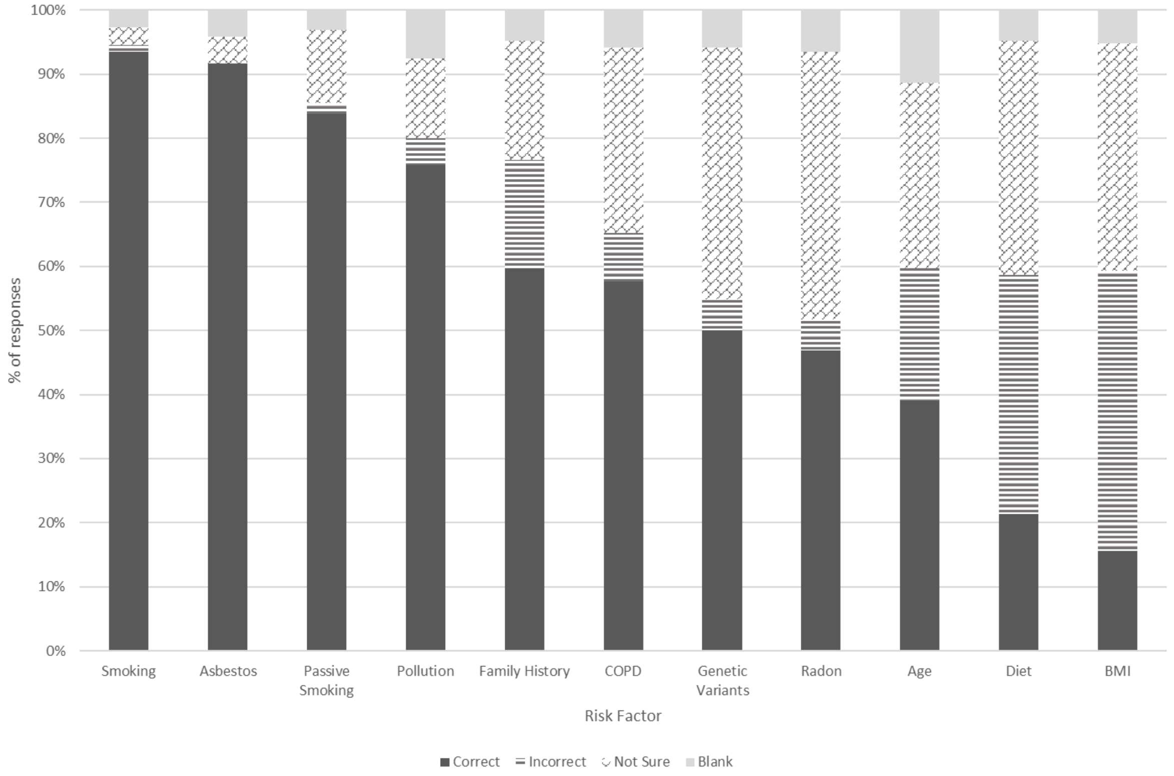


Figure 43. Lung cancer risk factor identification in study participants. Reused from [407], permitted under CC BY 4.0.

Of the analysed risk factor identification responses (inclusion criteria in section 6.2.3), participants with a family history of lung cancer were significantly more likely to correctly identify it as a risk factor compared to those without (78% vs. 57%; p=0.02). Former smokers were significantly more likely than current smokers to identify passive smoking (89% vs. 76%; p=0.04) and air pollution (79% vs. 59%; p=0.01) as lung cancer risk factors. Men were significantly more likely than women to correctly identify radon as a risk factor (59% vs. 39%; p=0.009). Having an educational qualification was significantly associated with correctly identifying smoking (100% vs. 92%; p=0.01) and passive smoking (97% vs. 80%; p<0.001) as risk factors. Age, BMI, IMD decile, previous cancer diagnosis, and asbestos exposure were not associated with any of the risk factor identifications.

6.3.2.4 PHQ-4 Mental Health Score

Two thirds of PHQ-4 scores were classified as normal (score 0-2), 16% as mild (score 3-5) and 18% as moderate or severe (score 6-12). When split into depression and anxiety indicators, 18% had scores suggestive of depression and 22% of anxiety (Table 28). Those with higher PHQ-4 scores had higher perceived comparative risk (proportion of each PHQ-4 category in *‘higher’* comparative group: 25% → 31% → 50%; p=0.001) and more lung cancer worry by both frequency (9% → 37% → 35%; p<0.001) and impact (27% → 55% → 66%; p<0.001). Those in the highest PHQ-4 subgroup had higher perceived absolute risk than the other two groups (median: 25% vs. 20%; p=0.06).

6.3.3 Questionnaire Responses Stratified by Risk Score and Screening Eligibility

There was no correlation between individual perceived absolute risk and individual calculated risk scores (as calculated by $PLCO_{M2012}$) in the identified cohort (p=0.4). Overall, perceived absolute risk was approximately twenty-times higher than actual risk score (median: 1% vs. 20%; p<0.001) (Figure 44). Almost 95% (n=227) of respondents perceived their absolute risk to be above the screening eligibility threshold eligible ($PLCO_{M2012}$ score ≥1.51%), dramatically higher than the 42% (n=101) who were actually eligible. Median calculated risk score increased in line with perceived comparative risk. Those with lowest perceived comparative risk ('much less likely') had a median calculated risk score of 0.3%, increasing to 0.35%, 1.1% and 1.6% in the 'less likely', 'the same' and 'more likely' group respectively, with the highest median calculated risk score of 2.9% seen in the 'much more likely' group (p=0.004) (Figure 45). Those who were eligible for screening were approximately twice as likely to have *'higher'* comparative risk than those who were ineligible (41% vs. 21%; p<0.0001) (Table 30).

Those eligible for screening were more likely to report lung cancer-specific worry (any worry: 27% vs. 10%; p=0.001) and it was more likely to impact their mood (any impact: 51% vs. 27%; p=0.0003). More than two-thirds of those who reported worry were eligible for screening, compared to just over a third of those who did not report worry (p=0.01); median calculated risk score was 2.6% in those with lung cancer worry, compared to 1% in those without (p=0.008). The screening-eligible subgroup also scored higher on the PHQ-4 test for mental health (mean 2.75 vs. 1.82; p=0.03) and were twice as likely to have a score indicative of depression (20% vs. 10%; p=0.05) (Table 30).

6.3.4 Screening Adherence

Whilst our study was not powered to examine screening adherence as a primary outcome, we performed exploratory analysis to investigate whether any of the questionnaire responses were associated with participants returning for the second round of screening. In total, 101 participants of the identified subset were eligible for screening and had a baseline CT scan (T0). Of these, 87 returned 12 months later for the second-round scan (T0+12), equivalent to a screening adherence of 86%. There was no statistically significant difference in calculated risk, perceived risk, lung cancer worry or mental health between those who attended T0+12 and those who did not, although the small number (n=14) of T0+12 non-attendees precludes detailed analysis.

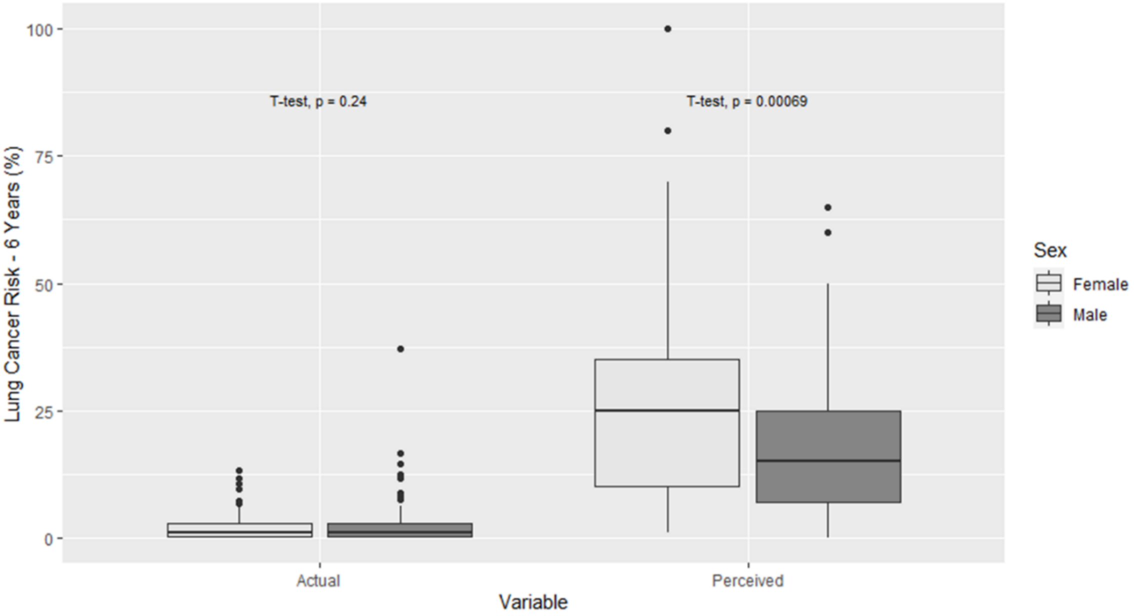


Figure 44. Actual and perceived absolute lung cancer risk score stratified by sex. Reused from [407], permitted under CC BY 4.0

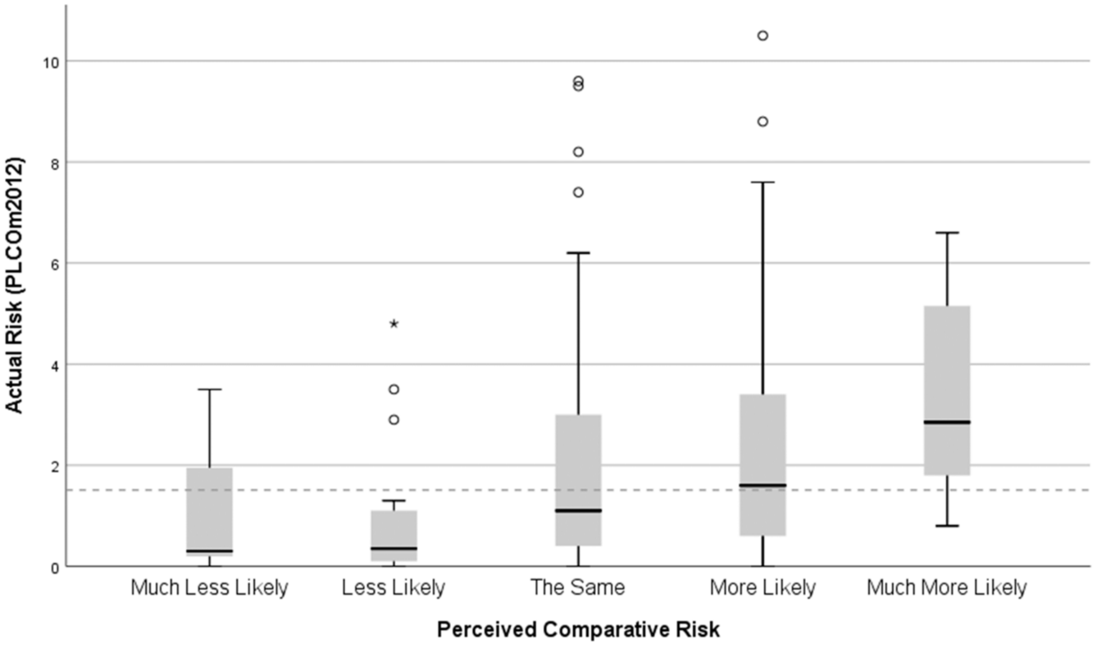


Figure 45. Perceived comparative risk vs. actual calculated risk ($PLCO_{m2012}$). Reused from [407], permitted under CC BY-4.0

Table 30. Questionnaire responses for identified subset and stratified by screening eligibility, based on PLCO$_{M2012}$ calculated risk score.

Demographic Variable			All Identified	Eligibility for screening		P value
				Ineligible (PLCO$_{M2012}$<1.51%)	**Eligible** (PLCO$_{M2012}$≥1.51%)	
Number			243	142	101	-
Risk Perception						
Perceived Absolute Risk Score Median ± IQR			20 ± 20	20 ± 16	20 ± 20	0.43
Perceived Comparative Risk % (n)		Lower	21 (51)	31 (43)	8 (8)	<0.0001
		The Same	49 (118)	48 (68)	51 (50)	
		Higher	30 (71)	21 (30)	41 (41)	
Population Risk Estimate % (n)		Underestimate	19 (46)	18 (26)	20 (20)	0.63
		Correct	9 (22)	11 (15)	7 (7)	
		Overestimate	72 (172)	71 (100)	73 (72)	
% Exposure to Risk Score Previously (n)			5 (10)	3 (4)	7.6 (6)	0.17
Disease Knowledge						
Mean number risk factors correctly identified (11)			6 ± 2	6 ± 2	6 ± 2	0.8
Mean Estimated Survival (±SD)		1-year	59 ± 23	59 ± 22	59 ± 24	0.83
		5-years	37 ± 21	37.7 ± 21	34 ± 22	0.41
		10-years	20 ± 19	20 ± 20	18 ± 19	0.56
Lung Cancer Worry						
Lung Cancer Specific Worry	Frequency % (n)	Any	17 (38)	10 (13)	27 (25)	0.001
		None	83 (185)	90 (118)	73 (67)	0.001
		Several days	12 (27)	8 (10)	19 (17)	
		More than half the days	3 (6)	2 (2)	4 (4)	
		Nearly every day	2 (5)	1 (1)	4 (4)	
	Impact % (n)	Any	37 (82)	27 (35)	51 (47)	0.0003
		None	63 (142)	73 (96)	50 (46)	0.001
		A small amount	29 (65)	22 (29)	39 (36)	
		Quite a lot	6 (13)	2 (3)	11 (10)	
		A lot	2 (4)	2 (3)	1 (1)	
Mental Health Score						
PHQ-4	Mean Score (±SD)		2.21 ± 3.1	1.82 ± 2.8	2.75 ± 3.5	0.03
	Overall Classification	Normal	70 (150)	74 (92)	64 (58)	0.06
		Mild	16 (34)	15 (19)	17 (15)	
		Moderate	7 (15)	5 (6)	10 (9)	
		Severe	7 (16)	6 (7)	10 (9)	
	% Depression Indicated (n)		14 (31)	10 (13)	20 (18)	0.05
	% Anxiety Indicated (n)		19 (41)	17 (21)	22 (20)	0.33

6.3.5 Demographic Factors Associated with Risk Perception

To identify demographic factors associated with risk perception, I constructed two multivariable logistic regression models, described in detail in section 6.2.3. Model A used $PLCO_{M2012}$ risk score as a covariate (Table 31), whilst model B substituted risk score for individual demographic factors (Table 32).

6.3.5.1 Perceived Absolute Risk

Female sex was the only demographic factor associated with above-average perceived absolute risk in both models (A: $_{adj}$OR 2.4, 95%CI 1.3-4.3; p=0.004) (B: $_{adj}$OR 2.1, 95%CI 1.13-3.8; p=0.02) (Table 31). This is consistent with the significantly higher perceived absolute risk amongst women when compared to men (median: 25% vs. 15%; p=0.001). It is important to highlight that there was no difference in calculated $PLCO_{M2012}$ risk score according to sex (Figure 44).

6.3.5.2 Perceived Comparative Risk

The factors significantly associated with *'higher'* perceived comparative risk were PHQ-4 score in model A (PHQ-4 $_{adj}$OR 1.1, 95%CI 1-1.2; p=0.02) (Table 31), and family history of lung cancer in model B ($_{adj}$OR 4.03, 95%CI 1.74-9.3; p=0.001) (Table 32). Mean PHQ-4 score was 3.58 (±3.9) in those with *'higher´* perceived comparative risk, significantly higher than the mean of 1.96 (±2.9) in the rest of the cohort (p<0.001). Almost half of respondents with a family history of lung cancer placed themselves in the *'more likely´* and *'much more likely'* perceived comparative risk categories, compared to just a quarter of those without family history (p=0.002).

Whilst not significantly predictive in the multivariable models, 46% of current smokers placed themselves in the *'higher'* group, compared to 26% of former smokers (p=0.008). Even when current smokers and former smokers were asked to compare their risk to others within the same smoking category (thereby accounting for smoking history in the comparative risk estimates), former smokers were still more likely to estimate themselves to be in the *'lower'* group (38% lower vs. 18% higher), whilst current smokers tended towards the *'higher'* groups (32% higher vs. 11% lower).

The factors significantly associated with *'lower'* perceived comparative risk in model A were male sex and lower $PLCO_{M2012}$ calculated risk score (Female: $_{adj}$OR 0.48, 95%CI 0.24-0.98; p=0.04. $PLCO_{M2012}$ score: $_{adj}$OR 0.85, 95%CI 0.71-1.01; p=0.06) (Table 31). In model B, significantly associated factors were no family history of lung cancer and a lower pack-year history (Family History: $_{adj}$OR 0.21, 95%CI 0.05-0.95; p=0.04. Pack Years: $_{adj}$OR 0.96, 95%CI 0.93-0.98; p=0.001) (Table 32).

6.3.6 Demographic Factors Associated with Lung Cancer Worry

I used the multivariable models to test for factors associated with lung cancer worry. Respondents were categorised into those who reported any frequency of lung cancer worry vs. those who reported no worry, and those who reported any lung cancer-worry mediated impact on mood vs. those who reported no impact. In model A, $PLCO_{M2012}$ score, female sex, and PHQ-4 score were all significantly associated with the presence of worry and impact on mood (Table 31). In model B, alongside female sex and PHQ-4 score which were associated with both presence of worry and impact on mood, lack of educational qualification was significantly associated with presence of worry ($_{adj}$OR 4.3, 95%CI 1.56-11.7; p=0.005) and increasing pack years was a significantly associated with impact on mood ($_{adj}$OR 1.02, 95%CI 1-1.03; p=0.03) (Table 32).

6.3.7 Demographic Factors Associated with Mental Health

Whilst PHQ-4 data were collected principally for use as a covariate and was not a primary outcome of this study, I ran an exploratory analysis to identify demographic factors associated with poorer mental health metrics.

Having a PHQ-4 score indicative of anxiety was significantly associated with female sex in both model A ($_{adj}$OR 2.69, 95%CI 1.14-6.34; p=0.02) and model B ($_{adj}$OR 2.87, 95%CI 1.18-7; p=0.02) (Table 31 & Table 32). Indeed, significantly more women had a PHQ-4 score consistent with anxiety than men (26% vs. 10%; p=0.002). In model A, high perceived comparative risk was also significantly associated with anxiety ($_{adj}$OR 2.24, 95%CI 1.05-4.85; p=0.04). In model B, presence of lung cancer worry and high perceived comparative risk were associated with anxiety, but with borderline statistical significance (Table 32).

In model A, having a PHQ-4 score indicative of depression was significantly associated with high perceived comparative risk ($_{adj}$OR 2.46, 95%CI 1.08-5.6; p=0.03); lung cancer worry was associated with borderline significance (Table 31). In model B, younger age and lack of educational qualifications were significantly associated with depression (age: $_{adj}$OR 0.9, 95%CI 0.84-0.97; p=0.008. Education: $_{adj}$OR 0.37, 95%CI 0.14-0.1; p=0.05) (Table 32).

Table 31. Binary logistic regression modelling (model 'A') for factors associated with risk perception, lung cancer worry, depression and anxiety (PHQ-4).

			Covariates									
			$PLCO_{m2012}$ Score		**Female Sex**		**PHQ-4 Score**		**Lung Cancer Worry Frequency – Any**		**High Perceived Comparative Risk**	
		N in analysis	OR (95% CI)	P	OR (95% CI)	P	OR (95% CI)	P	OR (95% CI)	P	OR (95% CI)	P
Dependant Variable	**Perceived Absolute Risk ≥ Median**	213	1.09 (0.99-1.21)	0.07	2.36 (1.31-4.26)	0.004	1.04 (0.95-1.15)	0.4	0.92 (0.4-2.1)	0.84	-	-
	High Perceived Comparative Risk	212	1.08 (0.99-1.19)	0.09	1.41 (0.72-2.74)	0.3	1.12 (1.02-1.24)	0.02	1.94 (0.85-4.45)	0.12	-	-
	Low Perceived Comparative Risk	212	0.85 (0.72-1.01)	0.06	0.48 (0.24-0.98)	0.04	0.96 (0.84-1.09)	0.51	0.55 (0.15-2.04)	0.37	-	-
	Lung Cancer Worry Frequency - Any	212	1.15 (1.03-1.29)	0.01	5.39 (1.87-15.6)	0.002	1.16 (1.03-1.3)	0.01	-	-	1.9 (0.83-4.4)	0.13
	Lung Cancer Worry Impact Mood - Any	213	1.13 (1.02-1.24)	0.02	2.45 (1.29-4.64)	0.006	1.2 (1.08-1.32)	0.001	-	-	0.94 (0.48-1.86)	0.86
	PHQ-4 Anxiety	212	1.03 (0.95-1.13)	0.48	2.69 (1.14-6.34)	0.02	-	-	2.04 (0.84-4.97)	0.12	2.24 (1.05-4.85)	0.04
	PHQ-4 Depression	212	1.07 (0.97-1.18)	0.16	1.01 (0.42-2.43)	0.98	-	-	2.52 (0.96-6.6)	0.06	2.46 (1.08-5.6)	0.03

Table 32. Binary logistic regression modelling (model 'B') for factors associated with risk perception, lung cancer worry, depression and anxiety (PHQ-4).

			Dependant Variable						
			Perceived Absolute Risk ≥ Median	High Perceived Comparative Risk	Low Perceived Comparative Risk	Lung Cancer Worry Frequency - Any	Lung Cancer Worry Impact Mood - Any	PHQ-4 Anxiety	PHQ-4 Depression
		N in analysis	204	203	203	203	204	204	206
Covariates	**Female**	**OR (95% CI)**	2.07 (1.13-3.8)	1.25 (0.61-2.58)	0.54 (0.24-1.19)	4.8 (1.7-13.6)	2.36 (1.2-4.6)	2.87 (1.18-7)	0.97 (0.37-2.5)
		P	0.02	0.54	0.13	0.003	0.01	0.02	0.97
	LC Worry – Freq. Any	**OR (95% CI)**	0.98 (0.42-2.3)	1.81 (0.73-4.5)	0.64 (0.16-2.59)	-	-	2.35 (0.93-5.95)	2.23 (0.79-6.3)
		P	0.96	0.2	0.54	-	-	0.07	0.13
	PHQ-4	**OR (95% CI)**	1.08 (0.97-1.19)	1.09 (0.98-1.2)	0.95 (0.81-1.1)	1.17 (1.04-1.32)	1.2 (1.07-1.33)	-	-
		P	0.18	0.11	0.5	0.01	0.001	-	-
	High Perceived Comparative Risk	**OR (95% CI)**	-	-	-	1.91 (0.78-4.7)	0.81 (0.39-1.69)	2.15 (0.95-4.84)	1.99 (0.8-4.95)
		P	-	-	-	0.16	0.57	0.07	0.14
	Age	**OR (95% CI)**	1.03 (0.98-1.08)	0.96 (0.91-1.01)	1.05 (0.98-1.11)	0.97 (0.9-1.05)	0.98 (0.94-1.04)	1 (0.94-1.06)	0.9 (0.84-0.97)
		P	0.22	0.17	0.17	0.46	0.52	0.88	0.008
	Family History LC	**OR (95% CI)**	1.2 (0.54-2.65)	4.03 (1.74-9.3)	0.21 (0.05-0.95)	1.38 (0.48-3.9)	1.97 (0.85-4.58)	1.2 (0.48-3.02)	1.3 (0.44-3.9)
		P	0.67	0.001	0.04	0.55	0.12	0.7	0.63
	Pack Years	**OR (95% CI)**	1 (0.99-1.02)	1.01 (1-1.03)	0.96 (0.93-0.98)	1.02 (1-1.04)	1.02 (1-1.03)	1 (0.98-1.02)	1.01 (1-1.03)
		P	0.94	0.09	0.001	0.13	0.03	0.86	0.29
	Current Smoking	**OR (95% CI)**	1.6 (0.69-3.73)	2.06 (0.85-5.02)	0.85 (0.2-3.5)	1.09 (0.38-3.14)	1.14 (0.48-2.68)	1.23 (0.45-3.38)	2.14 (0.71-6.4)
		P	0.28	0.11	0.82	0.87	0.78	0.69	0.18
	BMI	**OR (95% CI)**	1.01 (0.96-1.06)	1.02 (0.97-1.08)	1.03 (0.97-1.1)	0.97 (0.9-1.04)	0.99 (0.93-1.04)	1.01 (0.94-1.07)	1.07 (1-1.14)
		P	0.76	0.43	0.36	0.43	0.59	0.88	0.06
	Previous Cancer	**OR (95% CI)**	0.92 (0.41-2.1)	0.97 (0.36-2.66)	0.95 (0.34-2.66)	0.51 (0.11-2.4)	1.4 (0.57-3.4)	1.15 (0.38-3.5)	1.81 (0.5-6.6)
		P	0.84	0.96	0.92	0.39	0.47	0.8	0.45
	Educational Qualification	**OR (95% CI)**	0.89 (0.48-1.64)	0.99 (0.48-2.02)	1.45 (0.63-3.3)	0.24 (0.09-0.64)	1.22 (0.62-2.39)	1.42 (0.62-3.24)	0.37 (0.14-0.1)
		P	0.7	0.97	0.38	0.005	0.38	0.4	0.05

6.4 Discussion

This cross-sectional study examined lung cancer risk perception, disease knowledge and lung cancer-specific worry in ever-smoking attendees of a community-based lung cancer screening service located in highly socio-economically deprived areas. To my knowledge, it is the first to compare perceived lung cancer risk with actual calculated risk in this setting. Whilst it would have been unrealistic to expect highly accurate perceived absolute risk on an individual level, it is notable that as a cohort, the average perceived absolute risk was 20-times higher than actual risk score calculated by the $\text{PLCO}_{\text{M2012}}$ RPM. In addition, we observed no correlation between perceived absolute risk and actual risk score. Conversely, there was a robust association between increasing perceived comparative risk and increasing calculated risk; family history of lung cancer was a significant factor driving this association. Perceived absolute and comparative risk, as well as levels of lung cancer-specific worry and general anxiety, were significantly higher in women than in men, despite no increased actual lung cancer risk in this cohort. These data provide first insights into how community lung cancer screening attendees perceive their cancer risk, with important implications for pre-screening counselling, informed participation, and public health messaging.

Shared decision making has been identified as an important aspect of lung cancer screening implementation [211,260]. Understanding how the participant perceives their lung cancer risk and being able to effectively communicate risk in order to correct inaccurate perceptions are important foundational steps in facilitating shared decision making; a patient's healthcare decisions need to be based on an accurate awareness of the potential risks and benefits associated with the disease and intervention they are considering [256]. Previous studies have shown that inflated and inaccurate estimations of absolute risk are common, both in the context of perceived personal and population risk of disease [265–267,273]. In our study, participants who overestimated personal absolute lung cancer risk were more likely to overestimate population absolute lung cancer risk, suggesting that limited health literacy or understanding of absolute risk may be driving both phenomena. In order to address these misperceptions, some have emphasised the importance of clearly communicating absolute risk scores to screening participants in order to improve absolute risk perception accuracy and facilitate shared decision making [266,273]. However, others have questioned the benefits of such a strategy, both in the context of helping participants make decisions about screening, as well as influencing subsequent health behaviours [267,273,408–413].

Employing comparative risk terminology during risk counselling may be a better approach. Participants in our study displayed a fairly accurate perception of their risk in comparative terms, with exposure to lung cancer risk factors, particularly family history of the disease, being associated with higher perceived comparative risk. A recently published risk perception study in the Pan-Canadian Early Detection of Lung Cancer (PanCan) study cohort also highlighted family history as the most significant risk factor associated with increased perceived comparative risk. Like our study, it demonstrated a direct association between comparative risk perception and actual lung cancer risk, albeit not in a community-based lung cancer screening programme located in highly socio-economically deprived areas [274]. Other studies have also implicated family history of lung cancer as an important driver of perceived comparative risk [273,401], alongside other risk factors such as smoking status [270,279,402,414]. It is important to note that relying solely on comparative risk metrics could inappropriately bias a patient's decisions [399]. Employing a variety of formats, language, and decision aids may be important when communicating risk [415].

Our findings indicate that certain subgroups of the study cohort were more likely to display adverse psychological indicators. For example, women had increased perceived lung cancer risk, lung cancer-specific worry (both frequency and impact), and general anxiety. This is congruent with other studies which report increased perceived disease risk and anxiety amongst women [414,416–418]. We also found that screening-eligible (high risk) participants had significantly higher perceived comparative risk compared to screening-ineligible (low risk) participants. They were also almost three-times as likely to report lung cancer-specific worry, twice as likely to report lung cancer-specific worry impacting their mood and twice as likely to have a PHQ-4 score indicative of depression. These results may reveal segments of screening attendees who may benefit from additional pre-screening counselling to ensure they are adequately supported through the risk calculation and screening process. It also begs the question as to whether LHC attendance would increase or decrease worry and anxiety in each of the subgroups; further research is required to address this.

In addition to the association between actual risk and perceived comparative risk, our results show that measures of poor mental health were associated with higher perceived comparative risk, but lung cancer-specific worry was not. The nature of the relationship between mental health, worry, risk perception and actual risk is complex and requires further research to establish the direction of causality [267]. Whilst previous studies have demonstrated a link between lung cancer worry and risk factors such as smoke exposure and family history, in-line with our results [269–271,419], the PanCan risk perception study reported no association between actual risk and lung cancer worry. It also showed no link between worry and perceived risk, in common with our study [274]. Our study highlights the importance of including a measure of mental health as a covariate in lung cancer risk perception studies, as it may be a confounding variable when examining links between the other psychological measures.

An additional benefit of communicating risk in comparative terms may be that it facilitates a direct interaction with existing risk perception in the participant, thereby providing an opportunity to influence health behaviours linked to risk perception [420–422]. There is some evidence that lung cancer risk perception is correlated with intention to quit smoking and success in quitting [274,423]. A large prospective study in the SUMMIT cohort (an ongoing study taking place in London, trialling LDCT and blood-based cancer screening) found that increased risk perception is associated with screening uptake, although several other psychological constructs such as perceived lung cancer controllability, survival, willingness to be treated, and perceived benefit of screening were also implicated and may be more significant in high-risk populations [414]. Other studies have shown that interventions aimed at increasing uptake by utilising risk perception may not be effective [408,409], and psychological interventions to increase screening uptake must take a complex, multilevel approach rather than a one-off individual communication targeting a single psychological construct [414]. Further research is required to elucidate the most effective interventions for improving the uptake of positive health behaviours. Considering that more than 95% of this study cohort had never received a personalized lung cancer risk score before, the LHC programme could be an ideal setting for both investigating and implementing such strategies.

The lung cancer risk factors identified most accurately by our study participants were smoking, passive smoking, asbestos exposure, and exposure to pollution. A previous UK study also found these factors to be the most well-known [424], possibly highlighting the success of public health messaging in recent decades. In contrast, some of the factors included in RPMs used to establish screening eligibility, such as age and COPD, had much lower recognition. Genetic factors, which are emerging as a potentially important additional risk factor for screening selection (see Chapter Five), were only successfully identified as a risk factor by 50% of the study cohort. Considerable additional research will be required to ensure genetic risk information is communicated ethically and successfully before it can be practically implemented into real-world screening programmes; similar research has taken place in breast cancer screening [391,393,425] (see section 5.4.4). Participants with family history of lung cancer were significantly more likely to correctly identify it as a risk factor. Increasing awareness of familial lung cancer risk has been suggested as an important method of promoting protective behaviours [426]. Conversely, current smokers were less likely to identify passive smoking or pollution as risk factors. This is congruent with a large US study that reported significantly higher perceived harm from second-hand smoke among non-smokers and non-combustible users when compared to smokers [427]. Another study found that perceived risk of passive smoking was positively associated with intention to quit and successful smoking cessation [428]. Further research is required to establish if there is a causal link between this awareness and smoking cessation, and whether this could be a useful approach in public health messaging [429].

The cohort significantly overestimated lung cancer survival rates, a phenomenon previously described [430,431]. Raising awareness of lesser-known risk factors and true survival rates may have an impact on health behaviours; previous studies have shown that accurate risk factor and survival rate awareness (as opposed to an overly fatalistic attitude [414,432]) can affect risk perception and may improve screening compliance and other protective behaviours [431,433,434].

With its cross-sectional design, our study was not able to investigate the impact of LHC attendance on long-term risk perception accuracy, mental health, worry, or health-related behaviours such as smoking cessation. Further research at multiple time-points is required to investigate the impact LHC attendance has on these factors. We were also unable to examine risk perception in those who declined invitation to screening, and thereby investigate factors that influence screening uptake, an important area for future research; additional psychological constructs such as perceived disease controllability, stigma, and perceived effectiveness of risk-reducing behaviours are emerging as important mediators of health behaviour alongside risk perception [414] and should also be examined in future research. We used a single RPM ($\mathrm{PLCO}_{\mathrm{M2012}}$) to calculate actual risk score in our study, alternative RPMs may provide different risk scores for individuals; research into RPM performance in the LHC setting is ongoing (see Chapter Three). Benefit-based screening selection is emerging as a viable and potentially superior alternative to risk-based selection (see Chapter Four); this study did not examine perceived benefits or participant understanding of risk vs. benefit considerations, an important area of future research if benefit-based selection implementation continues. The single-centre, majority-White British profile of our study may prevent extrapolation of our findings to other populations; further research in a variety of diverse screening programmes is required. Whilst there were no major differences in questionnaire responses between identified and non-identified participants, we were not able to analyse differences in demographics between these groups (or indeed among those who attended the LHC but declined to participate in this study) and cannot exclude the possibility of selection bias.

Despite these limitations, our study provides important information about risk perception in attendees of a ‘real-world’, targeted screening programme located in communities with high levels of socio-economic deprivation. Furthermore, participants were surveyed at the closest possible timepoint prior to risk calculation and counselling. In other study settings (such as screening trials), participation bias and healthy volunteer effect may result in a study cohort less representative of actual screening attendees [318,336]. These factors contribute to the novelty of our findings and position our results and their associated implications for risk communication and pre-screening counselling to be directly applicable to the target population.

Chapter Seven – Discussion and Future Work

7.1 Project Overview

In March 2022, the UK National Screening Committee (UK NSC) (the advisory body which recommends screening policy to the UK government), published an interim report regarding targeted LDCT screening for lung cancer [253]. In this report, the UK NSC proposed that screening high-risk individuals for lung cancer should be recommended nationwide. The TLHC strategy for screening implementation, including the use of multivariable risk models for screening selection, was deemed “feasible, practical and effective”. It also found that LDCT screening for lung cancer is likely to be cost-effective for the NHS. One of the areas the report highlighted as an area requiring additional consideration was: “which multivariable risk assessment tool or combination of tools should be used to maximise efficiency within the screening programme.” This objective highlights the pressing need for further research into how risk prediction tools can be operated optimally in a real-world screening setting, aligning perfectly with one of the hypotheses of this thesis, that screening selection strategies perform sub-optimally in the socio-economically disadvantaged populations most likely to be targeted for lung cancer screening.

In Chapter Three, I presented evidence that supports this hypothesis by demonstrating that risk prediction models used in a community based LHC pilot may be poorly calibrated, had significant performance variation based on RPM and threshold selected, and favoured older, more comorbid participants for screening. It is important to note that despite these concerns, my analyses were congruent with a large volume of previous research demonstrating that risk-based selection does result in an effective, successful, and efficient screening programme. I also demonstrated that risk-based screening selection resulted in a highly comorbid screening cohort, highlighting what may be a strategic oversight in the UK NSC recommendation to “maximise efficiency within the screening programme”. High efficiency is indeed crucial in ensuring the screening programme is cost-effective and provides benefit to the population. However, if efficiency is narrowly defined as maximal detection rates from screening, aiming to maximise screening efficiency means optimising risk prediction as much as possible and directing screening away from those at lower risk, concurrently increasing the age and comorbidity burden in the screening cohort. This would result in very high cancer detection rates but could result in increased overdiagnosis and limited actual life expectancy benefit for the screening attendees.

In Chapter Four, I assessed one approach to addressing this concern: calculating individual life-gained from screening scores (using LYFS-CT) and employing these scores in benefit-based screening selection. I demonstrated that risk-based and benefit-based screening selection would have resulted in broadly similar screening cohorts in the Manchester LHC pilot, but participants favoured by benefit-based selection were younger and had a lower burden of comorbidity. This divergence in comorbidity persisted after five years of follow-up, which may indicate that this subgroup would be expected to have a longer life-expectancy and more to gain from screening. Unfortunately, the size of this study, and more specifically the size of the follow-up subset and length of follow-up period, did not allow for conclusions regarding differences in mortality or cancer detection rates between the risk-selected or benefit-selected cohorts. However, to my knowledge, this was the first time LYFS-CT has been applied to real-world screening participants, and my results demonstrated that this framework for screening selection certainly has the potential to address the issue of the high burden of comorbidity in the RPM-selected screening cohort. Developing a benefit-based approach to screening selection may still adhere to the UK NSC recommendation to “maximise efficiency within the screening programme” if ‘efficiency’ is defined more holistically, focussing on maximising the life-years gained from screening for participants per scan, rather than simply increasing the lung cancer detection rate.

In Chapter Five, I presented evidence that supported the hypothesis that considering genetic risk factors in ever-smokers has the potential to improve risk prediction for lung cancer. I did this by validating several previously published polygenic risk score tools in a case-control cohort highly representative of the target population for community-based lung cancer screening. Most of the PRSs improved the discrimination of a base clinical model with statistical significance, indicating that they could improve risk prediction when integrated into an RPM. These PRSs were developed in large RCT or biobank populations, so validation in this cohort is an important novel finding providing additional evidence regarding their clinical utility. I also identified novel SNPs on two genes of interest which were not included in the PRSs previously published. If validated in external study populations, SNPs from these loci may be candidates for inclusion in future PRSs. Including genetic factors in risk prediction is a method of generally optimising the risk prediction framework (as the inclusion of any additional biomarker would) and may also assist in selecting high-risk screening participants without simultaneously selecting a very comorbid cohort. This is an important added benefit and ties two of the primary aims of this thesis together.

In the final section of this thesis, I examined how socio-economically disadvantaged participants of a community-based lung cancer screening programme perceive personal and population risk of lung cancer, what their knowledge of the disease is, and how much lung cancer-specific worry they experience. My findings have implications for pre-screening risk communication, an area with very limited previous research in the lung cancer screening field. The results may also be useful for ensuring worried, anxious, or fatalist participants receive the appropriate supportive counselling whilst attending screening. A further potential implication is developing tailored behavioural medicine strategies which use risk perception and communication as tools to influence positive health behaviours such as smoking cessation and screening adherence. My study was unable to investigate such interventions, but it does provide an important foundational snapshot of participant attitudes which should be built upon in future research. The ever-expanding implementation of lung cancer screening provides ample opportunity for this research to be developed further.

7.2 Future Work

The results presented in this thesis highlight the urgent need for the development of large prospective studies embedded in community-based lung cancer screening programmes. Such studies would facilitate prospective screening selection using multiple RPMs, allowing for like-for-like performance comparisons between the various methods of selection. Full cohort extended follow-up would allow for model recalibration, ensuring RPMs are functioning adequately in the target screening population. The YLST is currently seeking to prospectively address this question but will not be assessing $\text{PLCO}_{\text{M2012}}$ ≥1.51% vs. LLP_{V2} ≥2.5%, the two criteria which were most closely matched in the results presented here [254]. Other RPMs which have gained prominence since the YLST protocol was published [225] should also be tested in this setting, necessitating the advancement of additional prospective studies to ensure optimal risk prediction is achieved. With the ongoing rollout of the TLHC (including in Manchester) resulting in tens of thousands of participants passing though screening programmes, as well as the potential final recommendation of a national targeted lung cancer screening programme by the UK NSC at the end of 2022, there will be many opportunities to perform further extensive prospective studies assessing risk-based selection using several RPMs at varying selection thresholds. This would facilitate model recalibration and allow for a fully informed assessment of which selection method performs best in the socio-economically deprived populations most at risk of lung cancer.

Prospective studies of this nature are also the key to establishing whether benefit-based selection should be considered alongside, or as an alternative to, risk-based selection. Whilst my results provide an indication of the potential for this selection method, considerable additional research is required to investigate if it provides an actual mortality benefit. To my knowledge, there are no current plans to consider LYFS-CT alongside RPMs in TLHC-based prospective studies and the UK NSC interim report does not mention benefit-based selection [253]. Testing LYFS-CT in screening programmes will require careful pre-emptive planning, as some of the variables needed to calculate LYFS-CT are not routinely collected during an LHC. Even if prospective studies in which LYFS-CT is used for screening selection are not deemed feasible, it would be advantageous to collect the necessary data during LHCs to facilitate large retrospective studies. I believe this should be a priority in lung cancer screening research moving forward.

It is important to note that whilst the results presented in this thesis indicate that using a PRS would result in a statistically significant improvement to lung cancer risk prediction, the clinical significance of the relatively modest effects derived from genetic risk factors remains unclear. Existing risk prediction strategies harnessing standard clinical risk factors function very successfully in lung cancer screening selection [225]. In order to be cost effective, it is likely that adding a PRS will either require the identification of specific subgroups of the population who would be particularly likely to benefit from incrementally improved risk prediction, a drastic reduction in the cost or increase in the ubiquity of genetic testing, or the identification of a significant additional tranche of SNPs responsible for lung cancer heritability. The ongoing development of GWAS cohorts of ever increasing sample sizes (Million Veterans Project, UK Biobank, 23andme, and others) is an important step in both identifying these additional SNPs, as well as precisely defining the effect size of each variant to improve predictive ability [435]. Methodological advances such as advanced genomic imputation [435], machine learning [436], and functional studies [437] will also assist in improving PRS tools. It is absolutely crucial that GWAS and associated studies take place in datasets representing populations of diverse and non-European ethnicities to ensure PRS tools are applicable to non-European screening attendees and health disparities are not exacerbated. This represents a significant gap in the field until this point [438].

In addition to continued study to further develop PRSs, considerable research is required to validate these tools in real-world screening programmes. Priority should be given to conducting validation studies in which both the cases and controls are derived from the same screening cohort, as well as comparing PRS performance with the actual RPM used for screening selection (neither of which I was able to do in this study). Further research is also required to ensure PRSs predict risk for all lung cancer subtypes, and for all subgroups of the screening cohort.

The strategies for improving risk prediction and screening selection analysed in this thesis are not the only ones available. We did not analyse every RPM published (some of which have demonstrated superior performance to those tested in this thesis) [225], nor did we consider RPMs that integrate LDCT results into ongoing risk prediction and determination of screening schedule [439,440]. Other biomarkers such as nasal swab RNA sequencing [441], circulating tumour DNA [442], and epigenetic risk factors [443] may also have the potential to augment risk prediction. Currently, research into these biomarkers is relatively siloed, although there are some attempts to integrate genetic and epigenetic risk factors into a combined risk calculator in some disease areas [444]. Plans have been developed to undertake a study at Bristol University examining the epigenetic profile of the samples from my case-control cohort; it is hoped that the data generated from my project and this study will be able to be integrated to provide further insight into the determinants of lung cancer risk.

Finally, I think the prospect of utilising risk perception and pre-screening counselling to influence participants' beliefs and health behaviours is particularly exciting. There is extremely mixed evidence as to the effectiveness of such interventions, but systematic studies examining a variety of communication techniques and tools in a structured community lung cancer screening programme are rare. The TLHC rollout provides an ideal setting for such studies. Studies examining the impact of simple LHC attendance (without any additional behavioural intervention) on risk perception, worry, mental health, and health behaviours are also extremely important.

In conclusion, this thesis has successfully addressed its overall aim: to improve lung cancer screening selection in socio-economically disadvantaged communities. I used data and samples sourced directly from individuals in these communities to identify several areas of concern that may inhibit optimal screening selection in the target population for lung cancer screening. I considered two key approaches to addressing these areas of concern, and demonstrated that with further research, they may be important strategies in further improving screening selection. I also examined how real-world screening participants perceive lung cancer risk and the implications this may have for pre-screening counselling, an under-researched and important area of screening implementation directly relevant to risk-based screening selection; this aspect of my thesis highlights the real patients who stand to benefit from lung cancer screening. My thesis addresses two of the four key focus areas for successful lung cancer screening implementation identified by Van Der Aalst *et.al.* [245]. With the ongoing expansion of lung cancer screening in the UK, and the likely development of a national screening programme, I hope that the research contained in this thesis contributes to the accumulating evidence that lung cancer screening has significant potential to reduce mortality and save lives, particularly in socioeconomically deprived communities. However, this thesis also highlights the many questions still unanswered with regards to the optimal way to select individuals for lung cancer screening to maximise benefit. Further research, funding, and organisation is absolutely crucial to continue to address these issues and give lung cancer screening the best chance at being efficacious and cost-effective. A unified approach to digitizing, amalgamating, and analysing data from all LHCs across the country (and perhaps internationally), in a way similar to that seen when the COVID-19 vaccine was rolled out to the population [445], would facilitate a wealth of diverse and well powered prospective and retrospective studies, aiming to answer some of the questions still remaining in this field.

References

1 Sung H, Ferlay J, Siegel RL, *et al.* Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* Published Online First: 2021. doi:10.3322/caac.21660

2 Wong MCS, Lao XQ, Ho KF, *et al.* Incidence and mortality of lung cancer: global trends and association with socioeconomic status. *Sci Reports 2017 71* 2017;**7**:1–9. doi:10.1038/s41598-017-14513-7

3 Lung cancer incidence statistics | Cancer Research UK. https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/lung-cancer/incidence#heading-Two (accessed 9 Feb 2022).

4 Fidler-Benaoudia MM, Torre LA, Bray F, *et al.* Lung cancer incidence in young women vs. young men: A systematic analysis in 40 countries. *Int J cancer* 2020;**147**:811–9. doi:10.1002/IJC.32809

5 Lung cancer statistics | Cancer Research UK. https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/lung-cancer (accessed 29 Apr 2021).

6 Lung cancer mortality statistics | Cancer Research UK. https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/lung-cancer/mortality#heading-Two (accessed 10 Feb 2022).

7 Smittenaar CR, Petersen KA, Stewart K, *et al.* Cancer incidence and mortality projections in the UK until 2035. *Br J Cancer* 2016;**115**:1147. doi:10.1038/BJC.2016.304

8 Hanahan D, Weinberg RA. The Hallmarks of Cancer. *Cell* 2000;**100**:57–70. doi:10.1016/S0092-8674(00)81683-9

9 Hanahan D, Weinberg RA. Hallmarks of cancer: The next generation. *Cell* 2011;**144**:646–74. doi:10.1016/J.CELL.2011.02.013/ATTACHMENT/3F528E16-8B3C-4D8D-8DE5-43E0C98D8475/MMC1.PDF

10 Gridelli C, Rossi A, Carbone DP, *et al.* Non-small-cell lung cancer. *Nat Rev Dis Prim* 2015;**1**:15009. doi:10.1038/nrdp.2015.9

11 Sher T, Dy GK, Adjei AA. Small Cell Lung Cancer. *Mayo Clin Proc* 2008;**83**:355–67. doi:10.4065/83.3.355

12 O'Brien TD, Jia P, Aldrich MC, *et al.* Lung cancer: One disease or many? *Hum Hered* 2018;**83**:65. doi:10.1159/000488942

13 Lung cancer survival statistics | Cancer Research UK. https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/lung-cancer/survival#collapseTwo (accessed 2 Jun 2020).

14 Quaresma M, Coleman MP, Rachet B. 40-year trends in an index of survival for all cancers combined and survival adjusted for age and sex for each cancer in England and Wales, 1971–2011: a population-based study. *Lancet* 2015;**385**:1206–18. doi:10.1016/S0140-6736(14)61396-9

15 Rosen RD, Sapra A. TNM Classification. *StatPearls NCBI* Published Online First: 23 February 2021.https://www.ncbi.nlm.nih.gov/books/NBK553187/ (accessed 15 Feb 2022).

16 Cancer survival in England - adults diagnosed - Office for National Statistics.

https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/datasets/cancersurvivalratescancersurvivalinenglandadultsdiagnosed (accessed 29 Apr 2021).

17 Rami-Porta R, Bolejack V, Crowley J, *et al.* The IASLC Lung Cancer Staging Project: Proposals for the Revisions of the T Descriptors in the Forthcoming Eighth Edition of the TNM Classification for Lung Cancer. *J Thorac Oncol* 2015;**10**:990–1003. doi:10.1097/JTO.0000000000000559

18 Lung cancer overview - NICE Pathways. https://pathways.nice.org.uk/pathways/lung-cancer#content=view-node%3Anodes-symptoms-and-signs-indicating-urgent-chest-x-ray-and-urgent-and-immediate-referral (accessed 17 Oct 2018).

19 O’dowd E, Mackenzie J, Balata H. Lung cancer for the non-respiratory physician. *Clin Med (Northfield Il)* 2021;**21**:e578–83. doi:10.7861/CLINMED.2021-0657

20 Lemjabbar-Alaoui H, Hassan OUI, Yang YW, *et al.* Lung cancer: biology and treatment options. *Biochim Biophys Acta* 2015;**1856**:189. doi:10.1016/J.BBCAN.2015.08.002

21 Blake-Cerda M, Lozano-Ruíz F, Maldonado-Magos F, *et al.* Consolidative stereotactic ablative radiotherapy (SABR) to intrapulmonary lesions is associated with prolonged progression-free survival and overall survival in oligometastatic NSCLC patients: A prospective phase 2 study. *Lung Cancer* 2021;**152**:119–26. doi:10.1016/J.LUNGCAN.2020.12.029

22 Soldà F, Lodge M, Ashley S, *et al.* Stereotactic radiotherapy (SABR) for the treatment of primary non-small cell lung cancer; systematic review and comparison with a surgical cohort. *Radiother Oncol* 2013;**109**:1–7. doi:10.1016/J.RADONC.2013.09.006

23 Recommendations | Lung cancer: diagnosis and management | Guidance | NICE.

24 Brown KF, Rumgay H, Dunlop C, *et al.* The fraction of cancer attributable to modifiable risk factors in England, Wales, Scotland, Northern Ireland, and the United Kingdom in 2015. *Br J Cancer* 2018;**118**:1130–41. doi:10.1038/s41416-018-0029-6

25 Ridge CA, McErlean AM, Ginsberg MS. Epidemiology of lung cancer. *Semin Intervent Radiol* 2013;**30**:93–8. doi:10.1055/s-0033-1342949

26 Parkin DM. Tobacco-attributable cancer burden in the UK in 2010. *Br J Cancer* 2011;**105**:S6–13. doi:10.1038/bjc.2011.475

27 Zhang P, Chen P-L, Li Z-H, *et al.* Association of smoking and polygenic risk with the incidence of lung cancer: a prospective cohort study. *Br J Cancer 2022* 2022;:1–10. doi:10.1038/s41416-022-01736-3

28 Proctor RN. The history of the discovery of the cigarette–lung cancer link: evidentiary traditions, corporate denial, global toll. *Tob Control* 2012;**21**:87–91. doi:10.1136/TOBACCOCONTROL-2011-050338

29 Doll R, Bradfordhill A. The Mortality of Doctors in Relation to Their Smoking Habits. *Br Med J* 1954;**1**:1451. doi:10.1136/BMJ.1.4877.1451

30 Doll R, Peto R, Boreham J, *et al.* Mortality from cancer in relation to smoking: 50 years observations on British doctors. *Br J Cancer* 2005;**92**:426–9. doi:10.1038/sj.bjc.6602359

31 Singhavi H, Ahluwalia JS, Stepanov I, *et al.* Tobacco carcinogen research to aid understanding of cancer risk and influence policy. *Laryngoscope Investig Otolaryngol* 2018;**3**:372. doi:10.1002/LIO2.204

32 Cancer - How Tobacco Smoke Causes Disease: The Biology and Behavioral Basis for

Smoking-Attributable Disease - NCBI Bookshelf. https://www.ncbi.nlm.nih.gov/books/NBK53010/ (accessed 7 May 2021).

33 Westcott PMK, To MD. The genetics and biology of KRAS in lung cancer. Chin. J. Cancer. 2013;**32**:63–70. doi:10.5732/cjc.012.10098

34 Kempf E, Rousseau B, Besse B, *et al.* KRAS oncogene in lung cancer: focus on molecularly driven clinical trials. *Eur Respir Rev* 2016;**25**:71–6. doi:10.1183/16000617.0071-2015

35 Lubin JH, Caporaso NE. Cigarette smoking and lung cancer: modeling total exposure and intensity. *Cancer Epidemiol Biomarkers Prev* 2006;**15**:517–23. doi:10.1158/1055-9965.EPI-05-0863

36 Schivo M, Avdalovic M V., Murin S. Non-cigarette tobacco and the lung. *Clin Rev Allergy Immunol* 2014;**46**:34–53. doi:10.1007/S12016-013-8372-0

37 Waziry R, Jawad M, Ballout RA, *et al.* The effects of waterpipe tobacco smoking on health outcomes: an updated systematic review and meta-analysis. *Int J Epidemiol* 2017;**46**:32–43. doi:10.1093/IJE/DYW021

38 Henley SJ, Thun MJ, Chao A, *et al.* Association between exclusive pipe smoking and mortality from cancer and other diseases. *J Natl Cancer Inst* 2004;**96**:853–61. doi:10.1093/JNCI/DJH144

39 (US) O on S and H. *The Health Consequences of Involuntary Exposure to Tobacco Smoke*. Centers for Disease Control and Prevention (US) 2006. http://www.ncbi.nlm.nih.gov/pubmed/20669524 (accessed 19 Oct 2018).

40 Hackshaw AK, Law MR, Wald NJ. The accumulated evidence on lung cancer and environmental tobacco smoke. *BMJ* 1997;**315**:980–8.http://www.ncbi.nlm.nih.gov/pubmed/9365295 (accessed 19 Oct 2018).

41 Peres J. No clear link between passive smoking and lung cancer. J. Natl. Cancer Inst. 2013;**105**:1844–6. doi:10.1093/jnci/djt365

42 Kim CH, Mu L, Wu M, *et al.* Evaluating the evidence for the relationship between passive smoking and lung cancer. Int. J. Cancer. 2014;**135**:2232–3. doi:10.1002/ijc.28860

43 Tindle HA, Stevenson Duncan M, Greevy RA, *et al.* Lifetime Smoking History and Risk of Lung Cancer: Results From the Framingham Heart Study. *JNCI J Natl Cancer Inst* Published Online First: 16 May 2018. doi:10.1093/jnci/djy041

44 Pirie K, Peto R, Reeves GK, *et al.* The 21st century hazards of smoking and benefits of stopping: a prospective study of one million women in the UK. *Lancet* 2013;**381**:133–41. doi:10.1016/S0140-6736(12)61720-6

45 White MC, Holman DM, Boehm JE, *et al.* Age and Cancer Risk: A Potentially Modifiable Relationship. *Am J Prev Med* 2014;**46**:S7. doi:10.1016/J.AMEPRE.2013.10.029

46 Anisimov VN. The relationship between aging and carcinogenesis: a critical appraisal. *Crit Rev Oncol Hematol* 2003;**45**:277–304. doi:10.1016/S1040-8428(02)00121-X

47 Krist AH, Davidson KW, Mangione CM, *et al.* Screening for Lung Cancer: US Preventive Services Task Force Recommendation Statement. *JAMA - J Am Med Assoc* 2021;**325**:962–70. doi:10.1001/jama.2021.1117

48 Targeted Screening for Lung Cancer with Low Radiation Dose Computed Tomography Standard Protocol prepared for the Targeted Lung Health Checks Programme. www.england.nhs.uk/cancer (accessed 5 May 2021).

49 Tas F, Ciftci R, Kilic L, *et al.* Age is a prognostic factor affecting survival in lung cancer

patients. *Oncol Lett* 2013;**6**:1507–13. doi:10.3892/ol.2013.1566

50 Hutchings SJ, Rushton L. Occupational cancer in Britain. *Br J Cancer 2012 1071* 2012;**107**:S8–17. doi:10.1038/BJC.2012.113

51 McCormack V, Peto J, Byrnes G, *et al.* Estimating the asbestos-related lung cancer burden from mesothelioma mortality. *Br J Cancer* 2012;**106**:575–84. doi:10.1038/bjc.2011.563

52 Report on Carcinogens, Fifteenth Edition. http://www.cdc.gov/ (accessed 13 Feb 2022).

53 Yang H, Rivera Z, Jube S, *et al.* Programmed necrosis induced by asbestos in human mesothelial cells causes high-mobility group box 1 protein release and resultant inflammation. *Proc Natl Acad Sci U S A* 2010;**107**:12611–6. doi:10.1073/pnas.1006542107

54 Klebe S, Leigh J, Henderson DW, *et al.* Asbestos, Smoking and Lung Cancer: An Update. *Int J Environ Res Public Heal 2020, Vol 17, Page 258* 2019;**17**:258. doi:10.3390/IJERPH17010258

55 Olsson AC, Vermeulen R, Schüz J, *et al.* Exposure-Response Analyses of Asbestos and Lung Cancer Subtypes in a Pooled Analysis of Case-Control Studies. *Epidemiology* 2017;**28**:288–99. doi:10.1097/EDE.0000000000000604

56 Lee PN. Relation between exposure to asbestos and smoking jointly and the risk of lung cancer. *Occup Environ Med* 2001;**58**:145–53.http://www.ncbi.nlm.nih.gov/pubmed/11171926 (accessed 25 Oct 2018).

57 Loomis D, Grosse Y, Lauby-Secretan B, *et al.* The carcinogenicity of outdoor air pollution. *Lancet Oncol* 2013;**14**:1262–3. doi:10.1016/S1470-2045(13)70487-X

58 Yang WS, Zhao H, Wang X, *et al.* An evidence-based assessment for the association between long-term exposure to outdoor air pollution and the risk of lung cancer. *Eur J Cancer Prev* 2016;**25**:163–72. doi:10.1097/CEJ.0000000000000158

59 Cui P, Huang Y, Han J, *et al.* Ambient particulate matter and lung cancer incidence and mortality: a meta-analysis of prospective studies. *Eur J Public Health* 2015;**25**:324–9. doi:10.1093/EURPUB/CKU145

60 Hamra GB, Guha N, Cohen A, *et al.* Outdoor particulate matter exposure and lung cancer: a systematic review and meta-analysis. *Environ Health Perspect* 2014;**122**:906–11. doi:10.1289/EHP/1408092

61 Silverman DT. Diesel Exhaust Causes Lung Cancer – Now What? *Occup Environ Med* 2017;**74**:233. doi:10.1136/OEMED-2016-104197

62 Sato T, Shimosato T, Klinman DM. Silicosis and lung cancer: current perspectives. *Lung Cancer Targets Ther* 2018;**9**:91. doi:10.2147/LCTT.S156376

63 Auvinen A, Pershagen G. Indoor radon and deaths from lung cancer. *BMJ* 2009;**338**:a3128. doi:10.1136/BMJ.A3128

64 Garcia-Rodriguez JA. Radon gas-the hidden killer: What is the role of family doctors? Can. Fam. Physician. 2018.

65 *WHO HANDBOOK ON INDOOR RADON A PUBLIC HEALTH PERSPECTIVE*. 2009. http://apps.who.int/iris/bitstream/handle/10665/44149/9789241547673_eng.pdf;jsessionid=C70F2996715CF9A39A0FA2E2B9D2BC8A?sequence=1 (accessed 23 Oct 2018).

66 Riaz SP, Horton M, Kang J, *et al.* Lung cancer incidence and survival in England: an analysis by socioeconomic deprivation and urbanization. *J Thorac Oncol* 2011;**6**:2005–10. doi:10.1097/JTO.0B013E31822B02DB

67 Jack RH, Gulliford MC, Ferguson J, *et al.* Geographical inequalities in lung cancer

management and survival in South East England: evidence of variation in access to oncology services? *Br J Cancer* 2003;**88**:1025–31. doi:10.1038/SJ.BJC.6600831

68 Rachet B, Ellis L, Maringe C, *et al.* Socioeconomic inequalities in cancer survival in England after the NHS cancer plan. *Br J Cancer* 2010;**103**:446–53. doi:10.1038/SJ.BJC.6605752

69 English indices of deprivation 2019 - GOV.UK. https://www.gov.uk/government/statistics/english-indices-of-deprivation-2019 (accessed 14 Feb 2022).

70 Lung cancer incidence statistics | Cancer Research UK. https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/lung-cancer/incidence#ref-5 (accessed 14 Feb 2022).

71 Mackenbach JP, Huisman M, Andersen O, *et al.* Inequalities in lung cancer mortality by the educational level in 10 European populations. *Eur J Cancer* 2004;**40**:126–35. doi:10.1016/J.EJCA.2003.10.018

72 Hart CL, Hole DJ, Gillis CR, *et al.* Social class differences in lung cancer mortality: risk factor explanations using two Scottish cohort studies. *Int J Epidemiol* 2001;**30**:268–74. doi:10.1093/ije/30.2.268

73 Graham H, Inskip HM, Francis B, *et al.* Pathways of disadvantage and smoking careers: evidence and policy implications. *J Epidemiol Community Health* 2006;**60 Suppl 2**:7–12. doi:10.1136/JECH.2005.045583

74 WHO. equity social determinants and public health programms. *Equity, Soc Determ public Heal Program* Published Online First: 2010. doi:10.1016/j.memsci.2014.01.055

75 Second-hand smoke exposure - Environmental Health Indicators. http://www.ehinz.ac.nz/indicators/indoor-environment/second-hand-smoke-exposure/#Rateshigherinmoredeprivedareas (accessed 31 Oct 2018).

76 Shaw A, Ritchie D, Semple S, *et al.* Reducing Children’s Exposure to Second Hand Smoke in the Home A Literature Review. 2012.

77 Graham H, Der G. Patterns and predictors of smoking cessation among British women. 1999.

78 Li X, Sundquist J, Zöller B, *et al.* Neighborhood deprivation and lung cancer incidence and mortality: A multilevel analysis from Sweden. *J Thorac Oncol* 2015;**10**:256–63. doi:10.1097/JTO.0000000000000417/ATTACHMENT/40AB51B3-6BA7-4B5C-8446-A09E972E4C55/MMC1.DOCX

79 Shohaimi S, Welch A, Bingham S, *et al.* Area deprivation predicts lung function independently of education and social class. *Eur Respir J* 2004;**24**:157–61. doi:10.1183/09031936.04.00088303

80 Adult smoking habits in the UK - Office for National Statistics. https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/healthandlifeexpectancies/bulletins/adultsmokinghabitsingreatbritain/2019 (accessed 14 Feb 2022).

81 Risch HA, Howe GR, Jain M, *et al.* Are Female Smokers at Higher Risk for Lung Cancer than Male Smokers?A Case-Control Analysis by Histologic Type. *Am J Epidemiol* 1993;**138**:281–93. doi:10.1093/OXFORDJOURNALS.AJE.A116857

82 Gasperino J. Gender is a risk factor for lung cancer. *Med Hypotheses* 2011;**76**:328–31. doi:10.1016/J.MEHY.2010.10.030

83 Cerny D, Cerny T, Ess S, *et al.* Lung cancer in the Canton of St. Gallen, Eastern Switzerland:

sex-associated differences in smoking habits, disease presentation and survival. *Onkologie* 2009;**32**:569–73. doi:10.1159/000232581

84 Wakelee HA, Chang ET, Gomez SL, *et al.* Lung cancer incidence in never smokers. *J Clin Oncol* 2007;**25**:472–8. doi:10.1200/JCO.2006.07.2983

85 Jemal A, Miller KD, Ma J, *et al.* Higher Lung Cancer Incidence in Young Women Than Young Men in the United States. *N Engl J Med* 2018;**378**:1999–2009. doi:10.1056/NEJMoa1715907

86 Gasperino J, Rom WN. Gender and Lung Cancer. 2004. doi:10.3816/CLC.2004.n.013

87 Cassidy A, Myles JP, van Tongeren M, *et al.* The LLP risk model: an individual risk prediction model for lung cancer. *Br J Cancer* 2008;**98**:270–6. doi:10.1038/sj.bjc.6604158

88 Stapelfeld C, Dammann C, Maser E. Sex-specificity in lung cancer risk. *Int J Cancer* 2020;**146**:2376–82. doi:10.1002/ijc.32716

89 Tammemägi MC, Katki HA, Hocking WG, *et al.* Selection criteria for lung-cancer screening. *N Engl J Med* 2013;**368**:728–36. doi:10.1056/NEJMoa1211776

90 Olak J, Colson Y. Gender differences in lung cancer: Have we really come a long way, baby? *J Thorac Cardiovasc Surg* 2004;**128**:346–51. doi:10.1016/J.JTCVS.2004.05.025

91 Kiyohara C, Ohno Y. Sex differences in lung cancer susceptibility: A review. *Gend Med* 2010;**7**:381–401. doi:10.1016/J.GENM.2010.10.002

92 Hsu L-H, Chu N-M, Kao S-H. Estrogen, Estrogen Receptor and Lung Cancer. *Int J Mol Sci* 2017;**18**. doi:10.3390/ijms18081713

93 Dogan S, Shen R, Ang DC, *et al.* Molecular epidemiology of EGFR and KRAS mutations in 3,026 lung adenocarcinomas: higher susceptibility of women to smoking-related KRAS-mutant cancers. *Clin Cancer Res* 2012;**18**:6169–77. doi:10.1158/1078-0432.CCR-11-3265

94 Kure EH, Ryberg D, Hewer A, *et al.* p53 mutations in lung tumours: relationship to gender and lung DNA adduct levels. *Carcinogenesis* 1996;**17**:2201–5. doi:10.1093/CARCIN/17.10.2201

95 Global Strategy For The Diagnosis, Management, And Prevention of Chronic Obstructive Pulomonary Disease (2018 Report). *Glob Initiat Chronic Obstr Lung Dis*https://goldcopd.org/wp-content/uploads/2017/11/GOLD-2018-v6.0-FINAL-revised-20-Nov_WMS.pdf

96 Decramer M, Janssens W, Miravitlles M. Chronic obstructive pulmonary disease. *Lancet* 2012;**379**:1341–51. doi:10.1016/S0140-6736(11)60968-9

97 Azad N, Rojanasakul Y, Vallyathan V. Inflammation and Lung Cancer: Roles of Reactive Oxygen/Nitrogen Species. *J Toxicol Environ Health* 2008;**11**:1–15. doi:10.1080/10937400701436460

98 Brenner DR, McLaughlin JR, Hung RJ. Previous lung diseases and lung cancer risk: a systematic review and meta-analysis. *PLoS One* 2011;**6**:e17479. doi:10.1371/journal.pone.0017479

99 Denholm R, Schüz J, Straif K, *et al.* Is Previous Respiratory Disease a Risk Factor for Lung Cancer? *Am J Respir Crit Care Med* 2014;**190**:549–59. doi:10.1164/rccm.201402-0338OC

100 Brenner DR, Boffetta P, Duell EJ, *et al.* Previous lung diseases and lung cancer risk: a pooled analysis from the International Lung Cancer Consortium. *Am J Epidemiol* 2012;**176**:573–85. doi:10.1093/aje/kws151

101 Denholm R, Schüz J, Straif K, *et al.* Is previous respiratory disease a risk factor for lung cancer? *Am J Respir Crit Care Med* 2014;**190**:549–59. doi:10.1164/rccm.201402-0338OC

102 Koshiol J, Rotunno M, Consonni D, *et al.* Lower risk of lung cancer after multiple pneumonia diagnoses. *Cancer Epidemiol Biomarkers Prev* 2010;**19**:716–21. doi:10.1158/1055-9965.EPI-09-0873

103 Liang H-Y, Li X-L, Yu X-S, *et al.* Facts and fiction of the relationship between preexisting tuberculosis and lung cancer risk: A systematic review. *Int J Cancer* 2009;**125**:2936–44. doi:10.1002/ijc.24636

104 Dracham CB, Shankar A, Madan R. Radiation induced secondary malignancies: A review article. Radiat. Oncol. J. 2018. doi:10.3857/roj.2018.00290

105 Rheingold SR, Neugut AI, Meadows AT. Secondary Cancers: Incidence, Risk Factors, and Management. Published Online First: 2003.https://www.ncbi.nlm.nih.gov/books/NBK12712/ (accessed 15 Feb 2022).

106 Geurts TW, Nederlof PM, Van Den Brekel MWM, *et al.* Pulmonary Squamous Cell Carcinoma following Head and Neck Squamous Cell Carcinoma: Metastasis or Second Primary? *Clin Cancer Res* 2005;**11**:6608–14. doi:10.1158/1078-0432.CCR-05-0257

107 Pagedar NA, Jayawardena A, Charlton ME, *et al.* Second primary lung cancer after head and neck cancer: implications for screening computed tomography. *Ann Otol Rhinol Laryngol* 2015;**124**:765. doi:10.1177/0003489415582259

108 Sanikini H, Yuan JM, Butler LM, *et al.* Body mass index and lung cancer risk: A pooled analysis based on nested case-control studies from four cohort studies. *BMC Cancer* 2018;**18**:1–10. doi:10.1186/s12885-018-4124-0

109 Zhu H, Zhang S. Body mass index and lung cancer risk in never smokers: a meta-analysis. *BMC Cancer* 2018;**18**:635. doi:10.1186/s12885-018-4543-y

110 TOKUHATA GK, LILIENFELD AM. Familial aggregation of lung cancer among hospital patients. *Public Heal reports (Washington, DC 1896)* 1963;**78**:277–83.http://www.ncbi.nlm.nih.gov/pubmed/13985328 (accessed 12 Nov 2018).

111 Matakidou A, Eisen T, Houlston RS. Systematic review of the relationship between family history and lung cancer risk. *Br J Cancer* 2005;**93**:825–33. doi:10.1038/sj.bjc.6602769

112 Coté ML, Liu M, Bonassi S, *et al.* Increased risk of lung cancer in individuals with a family history of the disease: a pooled analysis from the International Lung Cancer Consortium. *Eur J Cancer* 2012;**48**:1957–68. doi:10.1016/j.ejca.2012.01.038

113 Jonsson S, Thorsteinsdottir U, Gudbjartsson DF, *et al.* Familial Risk of Lung Carcinoma in the Icelandic Population. *JAMA* 2004;**292**:2977. doi:10.1001/jama.292.24.2977

114 Wu AH, Fontham ET, Reynolds P, *et al.* Family history of cancer and risk of lung cancer among lifetime nonsmoking women in the United States. *Am J Epidemiol* 1996;**143**:535–42.http://www.ncbi.nlm.nih.gov/pubmed/8610670 (accessed 11 Nov 2018).

115 Schwartz AG, Yang P, Swanson GM. Familial risk of lung cancer among nonsmokers and their relatives. *Am J Epidemiol* 1996;**144**:554–62.http://www.ncbi.nlm.nih.gov/pubmed/8797515 (accessed 29 Nov 2018).

116 Mucci LA, Hjelmborg JB, Harris JR, *et al.* Familial Risk and Heritability of Cancer Among Twins in Nordic Countries. *JAMA* 2016;**315**:68–76. doi:10.1001/jama.2015.17703

117 Torkamani A, Wineinger NE, Topol EJ. The personal and clinical utility of polygenic risk scores. Nat. Rev. Genet. 2018;**19**:581–90. doi:10.1038/s41576-018-0018-x

118 Li X, Hemminki K. Inherited predisposition to early onset lung cancer according to histological type. *Int J Cancer* 2004;**112**:451–7. doi:10.1002/ijc.20436

119 Li X, Hemminki K. Familial multiple primary lung cancers: a population-based analysis from Sweden. *Lung Cancer* 2005;**47**:301–7. doi:10.1016/J.LUNGCAN.2004.07.048

120 Gaughan EM, Cryer SK, Yeap BY, *et al.* Family history of lung cancer in never smokers with non-small-cell lung cancer and its association with tumors harboring EGFR mutations. *Lung Cancer* 2013;**79**:193–7. doi:10.1016/j.lungcan.2012.12.002

121 Schneider K, Zelley K, Nichols KE, *et al. Li-Fraumeni Syndrome. 1999 Jan 19 [updated 2013 Apr 11]*. University of Washington, Seattle; 1993-2019 2019. https://www.ncbi.nlm.nih.gov/books/NBK1311/ (accessed 13 Aug 2020).

122 Li FP, Fraumeni JF. Soft-tissue sarcomas, breast cancer, and other neoplasms. A familial syndrome? *Ann Intern Med* Published Online First: 1969. doi:10.7326/0003-4819-71-4-747

123 Lynch HT, McComb RD, Osborn NK, *et al.* Predominance of brain tumors in an extended Li-Fraumeni (SBLA) kindred, including a case of Sturge-Weber syndrome. *Cancer* Published Online First: 2000. doi:10.1002/(SICI)1097-0142(20000115)88:2<433::AID-CNCR26>3.0.CO;2-0

124 Yamamoto H, Yatabe Y, Toyooka S. Inherited lung cancer syndromes targeting never smokers. Transl. Lung Cancer Res. 2018. doi:10.21037/tlcr.2018.06.01

125 Ohtsuka K, Ohnishi H, Kurai D, *et al.* Familial lung adenocarcinoma caused by the EGFR V843I germ-line mutation. *J Clin Oncol* Published Online First: 2011. doi:10.1200/JCO.2010.31.4492

126 Yu HA, Arcila ME, Fleischut MH, *et al.* Germline EGFR T790M mutation found in multiple members of a familial cohort. *J Thorac Oncol* 2014;**9**:554–8. doi:10.1097/JTO.0000000000000052

127 Suda K, Onozato R, Yatabe Y, *et al.* EGFR T790M mutation: A double role in lung cancer cell survival? *J Thorac Oncol* 2009;**4**:1–4. doi:10.1097/JTO.0b013e3181913c9f

128 Bailey-Wilson JE, Sellers TA, Elston RC, *et al.* Evidence for a major gene effect in early-onset lung cancer. *J La State Med Soc* 1993;**145**:157–62.

129 Yang P, Schwartz AG, McAllister AE, *et al.* Lung cancer risk in families of nonsmoking probands: Heterogeneity by age at diagnosis. *Genet Epidemiol* 1999;**17**:253–73. doi:10.1002/(SICI)1098-2272(199911)17:4<253::AID-GEPI2>3.0.CO;2-K

130 Wu PF, Lee CH, Wang MJ, *et al.* Cancer aggregation and complex segregation analysis of families with female non-smoking lung cancer probands in Taiwan. *Eur J Cancer* 2004;**40**:260–6. doi:10.1016/j.ejca.2003.08.021

131 Amos CI, Pinney SM, Li Y, *et al.* A susceptibility locus on chromosome 6q greatly increases lung cancer risk among light and never smokers. *Cancer Res* 2010;**70**:2359–67. doi:10.1158/0008-5472.CAN-09-3096

132 You M, Wang D, Liu P, *et al.* Fine mapping of chromosome 6q23-25 region in familial lung cancer families reveals RGS17 as a likely candidate gene. *Clin Cancer Res* 2009;**15**:2666–74. doi:10.1158/1078-0432.CCR-08-2335

133 Michael AJ, Lu Y, Liu Y, *et al.* RGS17, an overexpressed gene in human lung and prostate cancer, induces tumor cell proliferation through the cyclic AMP-PKA-CREB pathway. *Cancer Res* Published Online First: 2009. doi:10.1158/0008-5472.CAN-08-3495

134 Chen HY, Yu SL, Ho BC, *et al.* R331W missense mutation of oncogene YAP1 is a germline

risk allele for lung adenocarcinoma with medical actionability. *J Clin Oncol* Published Online First: 2015. doi:10.1200/JCO.2014.59.3590

135 Xiong D, Wang Y, Kupert E, *et al.* A recurrent mutation in PARK2 is associated with familial lung cancer. *Am J Hum Genet* 2015;**96**:301–8. doi:10.1016/j.ajhg.2014.12.016

136 6-162206852-G-A | gnomAD. https://gnomad.broadinstitute.org/variant/6-162206852-G-A?dataset=gnomad_r2_1 (accessed 23 Jul 2020).

137 Bossé Y, Amos CI. A Decade of GWAS Results in Lung Cancer. *Cancer Epidemiol Biomarkers Prev* 2018;**27**:363–79. doi:10.1158/1055-9965.EPI-16-0794

138 Ioannidis JPA, Boffetta P, Little J, *et al.* Assessment of cumulative evidence on genetic associations: Interim guidelines. *Int J Epidemiol* 2008;**37**:120–32. doi:10.1093/ije/dym159

139 Wang J, Liu Q, Yuan S, *et al.* Genetic predisposition to lung cancer: Comprehensive literature integration, meta-analysis, and multiple evidence assessment of candidate-gene association studies. *Sci Rep* 2017;**7**. doi:10.1038/s41598-017-07737-0

140 Liu C, Cui H, Gu D, *et al.* Genetic polymorphisms and lung cancer risk: Evidence from meta-analyses and genome-wide association studies. Lung Cancer. 2017. doi:10.1016/j.lungcan.2017.08.026

141 Wacholder S, Chanock S, Garcia-Closas M, *et al.* Assessing the probability that a positive report is false: An approach for molecular epidemiology studies. *J Natl Cancer Inst* 2004;**96**:434–42. doi:10.1093/jnci/djh075

142 McKay JD, Hung RJ, Han Y, *et al.* Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. *Nat Genet* Published Online First: 2017. doi:10.1038/ng.3892

143 Improgo MRD, Scofield MD, Tapper AR, *et al.* From smoking to lung cancer: the CHRNA5/A3/B4 connection. *Oncogene* 2010;**29**:4874–84. doi:10.1038/onc.2010.256

144 Saccone SF, Hinrichs AL, Saccone NL, *et al.* Cholinergic nicotinic receptor genes implicated in a nicotine dependence association study targeting 348 candidate genes with 3713 SNPs. *Hum Mol Genet* 2007;**16**:36–49. doi:10.1093/hmg/ddl438

145 Chen L-S, Baker T, Hung RJ, *et al.* Genetic Risk Can Be Decreased: Quitting Smoking Decreases and Delays Lung Cancer for Smokers With High and Low CHRNA5 Risk Genotypes - A Meta-Analysis. *EBioMedicine* 2016;**11**:219–26. doi:10.1016/j.ebiom.2016.08.012

146 Hung RJ, McKay JD, Gaborieau V, *et al.* A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. *Nature* 2008;**452**:633–7. doi:10.1038/nature06885

147 Amos CI, Wu X, Broderick P, *et al.* Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. *Nat Genet* 2008;**40**:616–22. doi:10.1038/ng.109

148 Liu P, Vikis HG, Wang D, *et al.* Familial aggregation of common sequence variants on 15q24-25.1 in lung cancer. *J Natl Cancer Inst* 2008;**100**:1326–30. doi:10.1093/jnci/djn268

149 Zhao D, Yang C, Zhou X, *et al.* Association between CLPTM1L polymorphisms (rs402710 and rs401681) and lung cancer susceptibility: evidence from 27 case–control studies. *Mol Genet Genomics* 2014;**289**:1001–12. doi:10.1007/s00438-014-0868-7

150 Yang Y-C, Fu W-P, Zhang J, *et al.* rs401681 and rs402710 confer lung cancer susceptibility by regulating TERT expression instead of CLPTM1L in East Asian populations.

Carcinogenesis 2018;**39**:1216–21. doi:10.1093/carcin/bgy084

151 Marchesi M, Andersson E, Villabona L, *et al.* HLA-dependent tumour development: A role for tumour associate macrophages? *J Transl Med* 2013;**11**. doi:10.1186/1479-5876-11-247

152 Chatterjee N, Shi J, García-Closas M. Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nat Rev Genet* 2016;**17**:392–406. doi:10.1038/nrg.2016.27

153 Pashayan N, Duffy SW, Neal DE, *et al.* Implications of polygenic risk-stratified screening for prostate cancer on overdiagnosis. *Genet Med* 2015;**17**:789–95. doi:10.1038/gim.2014.192

154 Hsu L, Jeon J, Brenner H, *et al.* A model to determine colorectal cancer risk using common genetic susceptibility loci. *Gastroenterology* 2015;**148**:1330-9.e14. doi:10.1053/j.gastro.2015.02.010

155 Natarajan P, Young R, Stitziel NO, *et al.* Polygenic Risk Score Identifies Subgroup With Higher Burden of Atherosclerosis and Greater Relative Benefit From Statin Therapy in the Primary Prevention Setting. *Circulation* 2017;**135**:2091–101. doi:10.1161/CIRCULATIONAHA.116.024436

156 Desikan RS, Fan CC, Wang Y, *et al.* Genetic assessment of age-associated Alzheimer disease risk: Development and validation of a polygenic hazard score. *PLoS Med* 2017;**14**:e1002258. doi:10.1371/journal.pmed.1002258

157 Fahed AC, Wang M, Homburger JR, *et al.* Polygenic background modifies penetrance of monogenic variants conferring risk for coronary artery disease, breast cancer, or colorectal cancer. Published Online First: 2019. doi:10.1101/19013086

158 Mavaddat N, Michailidou K, Dennis J, *et al.* Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. *Am J Hum Genet* 2019;**104**:21–34. doi:10.1016/j.ajhg.2018.11.002

159 NHS screening - NHS. https://www.nhs.uk/conditions/nhs-screening/ (accessed 16 Feb 2022).

160 Wilson J, Jungner G. Principles and practice of screening for disease. *World Heal Organ* 1986;:11–3.https://apps.who.int/iris/handle/10665/37650

161 Andermann A, Blancquaert I, Beauchamp S, *et al.* Revisiting Wilson and Jungner in the genomic age: a review of screening criteria over the past 40 years. *Bull World Health Organ* 2008;**86**:317. doi:10.2471/BLT.07.050112

162 Sturdy S, Miller F, Hogarth S, *et al.* Half a Century of Wilson &amp; Jungner: Reflections on the Governance of Population Screening. *Wellcome Open Res 2020 5158* 2020;**5**:158. doi:10.12688/wellcomeopenres.16057.2

163 Dobrow MJ, Hagens V, Chafe R, *et al.* Consolidated principles for screening based on a systematic review and consensus process. *C Can Med Assoc J* 2018;**190**:E422. doi:10.1503/CMAJ.171154

164 Receiver Operating Characteristic Curve in Diagnostic Test Assessment - ScienceDirect. https://www.sciencedirect.com/science/article/pii/S1556086415306043 (accessed 16 Feb 2022).

165 Pinsky PF. Principles of Cancer Screening. *Surg Clin North Am* 2015;**95**:953. doi:10.1016/J.SUC.2015.05.009

166 Evaluating Screening Programs. https://sphweb.bumc.bu.edu/otlt/mph-modules/ep/ep713_screening/EP713_Screening8.html (accessed 25 Feb 2022).

167 Cancer Prevention and Screening. *Cancer Prev Screen* Published Online First: 28 August 2018. doi:10.1002/9781118990957

168 Heijnsdijk EAM, Csanádi M, Gini A, *et al.* All-cause mortality versus cancer-specific mortality as outcome in cancer screening trials: A review and modeling study. *Cancer Med* 2019;**8**:6127. doi:10.1002/CAM4.2476

169 HEAD TO HEAD.

170 Dobbin KK, Ebell M. Should we expect all-cause mortality reductions in large screening studies? *Br J Gen Pract* 2018;**68**:290. doi:10.3399/BJGP18X696545

171 Spix C, Berthold F, Hero B, *et al.* Correction factors for self-selection when evaluating screening programmes. doi:10.1177/0969141315597959

172 Tanner NT, Dai L, Bade BC, *et al.* Assessing the generalizability of the national lung screening trial: Comparison of patients with stage 1 disease. *Am J Respir Crit Care Med* 2017;**196**:602–8. doi:10.1164/RCCM.201705-0914OC/SUPPL_FILE/DISCLOSURES.PDF

173 Mascalchi M, Sali L. Lung cancer screening with low dose CT and radiation harm—from prediction models to cancer incidence data. *Ann Transl Med* 2017;**5**. doi:10.21037/ATM.2017.06.41

174 Radiation risk with digital mammography in breast screening - GOV.UK. https://www.gov.uk/government/publications/breast-screening-radiation-risk-with-digital-mammography/radiation-risk-with-digital-mammography-in-breast-screening (accessed 17 Feb 2022).

175 Denis B, Gendre I, Sauleau EA, *et al.* Harms of colonoscopy in a colorectal cancer screening programme with faecal occult blood test: a population-based cohort study. *Dig Liver Dis* 2013;**45**:474–80. doi:10.1016/J.DLD.2013.01.006

176 Welch HG, Black WC. Overdiagnosis in Cancer. *JNCI J Natl Cancer Inst* 2010;**102**:605–13. doi:10.1093/jnci/djq099

177 Carter SM, Barratt A. What is overdiagnosis and why should we take it seriously in cancer screening? | PHRP. *https://www.phrp.com.au/* 2017;**27**. doi:10.17061/PHRP2731722

178 Black WC. Overdiagnosis: An Underrecognized Cause of Confusion and Harm in Cancer Screening. *JNCI J Natl Cancer Inst* 2000;**92**:1280–2. doi:10.1093/JNCI/92.16.1280

179 Carter JL, Coletti RJ, Harris RP. Quantifying and monitoring overdiagnosis in cancer screening: a systematic review of methods. *BMJ* 2015;**350**. doi:10.1136/BMJ.G7773

180 Chiolero A, Paccaud F, Aujesky D, *et al.* How to prevent overdiagnosis. *Swiss Med Wkly 2015 3* 2015;**145**. doi:10.4414/SMW.2015.14060

181 Tosteson ANA, Fryback DG, Hammond CS, *et al.* Consequences of false-positive screening mammograms. *JAMA Intern Med* 2014;**174**:954–61. doi:10.1001/jamainternmed.2014.981

182 Rasmussen JF, Siersma V, Malmqvist J, *et al.* Psychosocial consequences of false positives in the Danish Lung Cancer CT Screening Trial: A nested matched cohort study. *BMJ Open* 2020;**10**:34682. doi:10.1136/bmjopen-2019-034682

183 Brodersen J, Siersma VD. Long-term psychosocial consequences of false-positive screening mammography. *Ann Fam Med* 2013;**11**:106–15. doi:10.1370/afm.1466

184 Petticrew MP, Sowden AJ, Lister-Sharp D, *et al.* False-negative results in screening programmes: Systematic review of impact and implications. Health Technol. Assess. (Rockv). 2000.

185 Wu GX, Raz DJ, Brown L, *et al.* Psychological Burden Associated with Lung Cancer Screening: A Systematic Review. *Clin Lung Cancer* 2016;**17**:315. doi:10.1016/J.CLLC.2016.03.007

186 Andersson E, Dai Ydrefelt Y, Johannesson M, *et al.* Surveillance of indeterminate pulmonary nodules detected with CT in a Swedish population-based study (SCAPIS): psychosocial consequences and impact on health-related quality of life—a multicentre prospective cross-sectional study. *BMJ Open* 2021;**11**:e048721. doi:10.1136/BMJOPEN-2021-048721

187 Fontana RS, Sanderson DR, Woolner LB, *et al.* Lung cancer screening: the Mayo program. *J Occup Med* 1986;**28**:746–50. doi:10.1097/00043764-198608000-00038

188 Kubik A, Parkin DM, Khlat M, *et al.* Lack of benefit from semi-annual screening for cancer of the lung: follow-up report of a randomized controlled trial on a population of high-risk males in Czechoslovakia. *Int J cancer* 1990;**45**:26–33. doi:10.1002/IJC.2910450107

189 Oken MM, Hocking WG, Kvale PA, *et al.* Screening by chest radiograph and lung cancer mortality: The Prostate, Lung, Colorectal, and Ovarian (PLCO) randomized trial. *JAMA - J Am Med Assoc* 2011;**306**:1865–73. doi:10.1001/jama.2011.1591

190 Manser R, Irving LB, Stone C, *et al.* Screening for lung cancer. In: Manser R, ed. *Cochrane Database of Systematic Reviews*. Chichester, UK: : John Wiley & Sons, Ltd 2004. CD001991. doi:10.1002/14651858.CD001991.pub2

191 Kaneko M, Eguchi K, Ohmatsu H, *et al.* Peripheral lung cancer: screening and detection with low-dose spiral CT versus radiography. *Radiology* 1996;**201**:798–802. doi:10.1148/radiology.201.3.8939234

192 International Early Lung Cancer Action Program Investigators, Henschke CI, Yankelevitz DF, *et al.* Survival of Patients with Stage I Lung Cancer Detected on CT Screening. *N Engl J Med* 2006;**355**:1763–71. doi:10.1056/NEJMoa060476

193 National Lung Screening Trial Research Team TNLSTR, Aberle DR, Adams AM, *et al.* Reduced lung-cancer mortality with low-dose computed tomographic screening. *N Engl J Med* 2011;**365**:395–409. doi:10.1056/NEJMoa1102873

194 de Koning HJ, van der Aalst CM, de Jong PA, *et al.* Reduced Lung-Cancer Mortality with Volume CT Screening in a Randomized Trial. *N Engl J Med* 2020;**382**:503–13. doi:10.1056/NEJMoa1911793

195 Paci E, Puliti D, Lopes Pegna A, *et al.* Mortality, survival and incidence rates in the ITALUNG randomised lung cancer screening trial. *Thorax* 2017;**72**:825–31. doi:10.1136/thoraxjnl-2016-209825

196 Wille MMW, Dirksen A, Ashraf H, *et al.* Results of the randomized danish lung cancer screening trial with focus on high-risk profiling. *Am J Respir Crit Care Med* 2016;**193**:542–51. doi:10.1164/rccm.201505-1040OC

197 Becker N, Motsch E, Trotter A, *et al.* Lung cancer mortality reduction by LDCT screening—Results from the randomized German LUSI trial. *Int J Cancer* 2020;**146**:1503–13. doi:10.1002/ijc.32486

198 Pastorino U, Silva M, Sestini S, *et al.* Prolonged lung cancer screening reduced 10-year mortality in the MILD trial: new confirmation of lung cancer screening efficacy. *Ann Oncol* Published Online First: 1 April 2019. doi:10.1093/annonc/mdz117

199 Field JK, Vulkan D, Davies MPA, *et al.* Lung cancer mortality reduction by LDCT screening: UKLS randomised trial results and international meta-analysis. *Lancet Reg Heal – Eur*

2021;**0**:100179. doi:10.1016/J.LANEPE.2021.100179

200 Marcus PM, Bergstralh EJ, Zweig MH, *et al.* Extended Lung Cancer Incidence Follow-up in the Mayo Lung Project and Overdiagnosis. *JNCI J Natl Cancer Inst* 2006;**98**:748–56. doi:10.1093/jnci/djj207

201 National Lung Screening Trial Research Team T. Lung Cancer Incidence and Mortality with Extended Follow-up in the National Lung Screening Trial. *J Thorac Oncol* 2019;**14**:1732–42. doi:10.1016/j.jtho.2019.05.044

202 Heleno B, Siersma V, Brodersen J. Estimation of Overdiagnosis of Lung Cancer in Low-Dose Computed Tomography Screening: A Secondary Analysis of the Danish Lung Cancer Screening Trial. *JAMA Intern Med* 2018;**178**:1420. doi:10.1001/JAMAINTERNMED.2018.3056

203 Ebell MH, Lin KW. Accounting for the Harms of Lung Cancer Screening. *JAMA Intern Med* 2018;**178**:1422–3. doi:10.1001/JAMAINTERNMED.2018.3061

204 Balata H, Ruparel M, O’Dowd E, *et al.* Analysis of the baseline performance of five UK lung cancer screening programmes. *Lung Cancer* 2021;**161**:136–40. doi:10.1016/J.LUNGCAN.2021.09.012

205 Li WWL, Lee TW, Yim APC. Quality of life after lung cancer resection. *Thorac Surg Clin* 2004;**14**:353–65. doi:10.1016/S1547-4127(04)00023-4

206 Lee J, Kim Y, Kim HY, *et al.* Feasibility of implementing a national lung cancer screening program: Interim results from the Korean Lung Cancer Screening Project (K-LUCAS). *Transl Lung Cancer Res* 2021;**10**:723–36. doi:10.21037/TLCR-20-700

207 Heerink WJ, de Bock GH, de Jonge GJ, *et al.* Complication rates of CT-guided transthoracic lung biopsy: meta-analysis. *Eur Radiol* 2017;**27**:138–48. doi:10.1007/s00330-016-4357-8

208 Rampinelli C, De Marco P, Origgi D, *et al.* Exposure to low dose computed tomography for lung cancer screening and risk of cancer: Secondary analysis of trial data and risk-benefit analysis. *BMJ* Published Online First: 2017. doi:10.1136/bmj.j347

209 Kummer S, Waller J, Ruparel M, *et al.* Psychological outcomes of low-dose CT lung cancer screening in a multisite demonstration screening pilot: the Lung Screen Uptake Trial (LSUT). *Thorax* 2020;**75**:1065–73. doi:10.1136/THORAXJNL-2020-215054

210 Brain K, Lifford KJ, Carter B, *et al.* Long-term psychosocial outcomes of low-dose CT screening: results of the UK Lung Cancer Screening randomised controlled trial. *Thorax* 2016;**71**:996–1005. doi:10.1136/THORAXJNL-2016-208283

211 Moyer VA. Screening for lung cancer: U.S. preventive services task force recommendation statement. *Ann Intern Med* 2014;**160**:330–8. doi:10.7326/M13-2771

212 Kovalchik SA, Tammemagi M, Berg CD, *et al.* Targeting of low-dose CT screening according to the risk of lung-cancer death. *N Engl J Med* 2013;**369**:245–54. doi:10.1056/NEJMoa1301851

213 Pinsky PF, Zhu CS, Kramer BS. Lung cancer risk by years since quitting in 30+ pack year smokers. *J Med Screen* 2015;**22**:151–7. doi:10.1177/0969141315579119

214 Reitsma M, Kendrick P, Anderson J, *et al.* Reexamining Rates of Decline in Lung Cancer Risk after Smoking Cessation: A meta-analysis. *Ann Am Thorac Soc* 2020;:AnnalsATS.201909-659OC. doi:10.1513/AnnalsATS.201909-659OC

215 Marcus MW, Raji OY, Field JK. Lung cancer screening: identifying the high risk cohort. *J Thorac Dis* 2015;**7**:S156-62. doi:10.3978/j.issn.2072-1439.2015.04.19

216 Hüsing A, Kaaks R. Risk prediction models versus simplified selection criteria to determine eligibility for lung cancer screening: an analysis of German federal-wide survey and incidence data. *Eur J Epidemiol* 2020;**1**:3. doi:10.1007/s10654-020-00657-w

217 Hüsing A, Kaaks R. Risk prediction models versus simplified selection criteria to determine eligibility for lung cancer screening: an analysis of German federal-wide survey and incidence data. *Eur J Epidemiol* 2020;**35**:899–912. doi:10.1007/s10654-020-00657-w

218 Tammemägi MC. Selecting lung cancer screenees using risk prediction models-where do we go from here. *Transl lung cancer Res* 2018;**7**:243–53. doi:10.21037/tlcr.2018.06.03

219 Weber M, Yap S, Goldsbury D, *et al.* Identifying high risk individuals for targeted lung cancer screening: Independent validation of the PLCO $_{m2012}$ risk prediction tool. *Int J Cancer* 2017;**141**:242–53. doi:10.1002/ijc.30673

220 Tammemägi MC, Church TR, Hocking WG, *et al.* Evaluation of the lung cancer risks at which to screen ever- and never-smokers: screening rules applied to the PLCO and NLST cohorts. *PLoS Med* 2014;**11**:e1001764. doi:10.1371/journal.pmed.1001764

221 Cassidy A, Myles JP, van Tongeren M, *et al.* The LLP risk model: an individual risk prediction model for lung cancer. *Br J Cancer* 2008;**98**:270–6. doi:10.1038/sj.bjc.6604158

222 Raji OY, Duffy SW, Agbaje OF, *et al.* Predictive accuracy of the Liverpool Lung Project risk model for stratifying patients for computed tomography screening for lung cancer: a case-control and cohort validation study. *Ann Intern Med* 2012;**157**:242–50. doi:10.7326/0003-4819-157-4-201208210-00004

223 Field JK, Duffy SW, Baldwin DR, *et al.* The UK Lung Cancer Screening Trial: a pilot randomised controlled trial of low-dose computed tomography screening for the early detection of lung cancer. *Health Technol Assess (Rockv)* 2016;**20**:1–146. doi:10.3310/hta20400

224 Field JK, Vulkan D, Davies MPA, *et al.* Liverpool Lung Project lung cancer risk stratification model: Calibration and prospective validation. *Thorax* 2021;**76**:161–8. doi:10.1136/THORAXJNL-2020-215158

225 Robbins HA, Alcala K, Swerdlow AJ, *et al.* Comparative performance of lung cancer risk models to define lung screening eligibility in the United Kingdom. *Br J Cancer 2021 12412* 2021;**124**:2026–34. doi:10.1038/s41416-021-01278-0

226 Bach PB, Kattan MW, Thornquist MD, *et al.* Variations in Lung Cancer Risk Among Smokers. *JNCI J Natl Cancer Inst* 2003;**95**:470–8. doi:10.1093/JNCI/95.6.470

227 Hoggart C, Brennan P, Tjonneland A, *et al.* A risk model for lung cancer incidence. *Cancer Prev Res (Phila)* 2012;**5**:834–46. doi:10.1158/1940-6207.CAPR-11-0237

228 Katki HA, Kovalchik SA, Berg CD, *et al.* Development and validation of risk models to select ever-smokers for CT lung-cancer screening. *JAMA* 2016;**315**:2300. doi:10.1001/JAMA.2016.6255

229 Melzer AC, Wilt TJ. Expanded Access to Lung Cancer Screening—Implementing Wisely to Optimize Health. *JAMA Netw Open* 2021;**4**:e210275–e210275. doi:10.1001/JAMANETWORKOPEN.2021.0275

230 S S, D D, X M, *et al.* Participation in lung cancer screening programs: are there gender and social differences? A systematic review. *Public Health Rev* 2018;**39**. doi:10.1186/S40985-018-0100-0

231 Ten Haaf K, Bastani M, Cao P, *et al.* A Comparative Modeling Analysis of Risk-Based Lung Cancer Screening Strategies. *J Natl Cancer Inst* 2020;**112**:466–79. doi:10.1093/jnci/djz164

232 Tammemägi MC, Ruparel M, Tremblay A, *et al.* USPSTF2013 versus PLCOm2012 lung cancer screening eligibility criteria (International Lung Screening Trial): interim analysis of a prospective cohort study. *Lancet Oncol* 2021;**0**. doi:10.1016/S1470-2045(21)00590-8

233 Cheung LC, Berg CD, Castle PE, *et al.* Life-Gained-Based Versus Risk-Based Selection of Smokers for Lung Cancer Screening. *Ann Intern Med* Published Online First: 22 October 2019. doi:10.7326/M19-1263

234 NHIS - About the National Health Interview Survey. https://www.cdc.gov/nchs/nhis/about_nhis.htm (accessed 21 Feb 2022).

235 Mazzone PJ, Silvestri GA, Souter LH, *et al.* Screening for Lung Cancer: CHEST Guideline and Expert Panel Report – Executive Summary. *Chest* Published Online First: 13 July 2021. doi:10.1016/J.CHEST.2021.07.003

236 Young RP, Hopkins RJ, Hay BA, *et al.* A gene-based risk score for lung cancer susceptibility in smokers and ex-smokers. *Postgrad Med J* 2009;**85**:515–24. doi:10.1136/pgmj.2008.077107

237 Kim TJ, Kim HY, Goo JM, *et al.* Risk prediction model for lung cancer screening. J. Korean Soc. Radiol. 2019;**80**:860–71. doi:10.3348/jksr.2019.80.5.860

238 Spitz MR, Etzel CJ, Dong Q, *et al.* An Expanded Risk Prediction Model for Lung Cancer. *Cancer Prev Res* 2008;**1**:250–4. doi:10.1158/1940-6207.CAPR-08-0060

239 Raji OY, Agbaje OF, Duffy SW, *et al.* Incorporation of a Genetic Factor into an Epidemiologic Model for Prediction of Individual Risk of Lung Cancer: The Liverpool Lung Project. *Cancer Prev Res* 2010;**3**:664–9. doi:10.1158/1940-6207.CAPR-09-0141

240 Marcus MW, Raji OY, Duffy SW, *et al.* Incorporating epistasis interaction of genetic susceptibility single nucleotide polymorphisms in a lung cancer risk prediction model. *Int J Oncol* 2016;**49**:361–70. doi:10.3892/ijo.2016.3499

241 Field JK, Chen Y, Marcus MW, *et al.* The contribution of risk prediction models to early detection of lung cancer. *J Surg Oncol* 2013;**108**:304–11. doi:10.1002/jso.23384

242 Dai J, Lv J, Zhu M, *et al.* Identification of risk loci and a polygenic risk score for lung cancer: a large-scale prospective cohort study in Chinese populations. *Lancet Respir Med* 2019;**7**:881–91. doi:10.1016/S2213-2600(19)30144-4

243 Kachuri L, Graff RE, Smith-Byrne K, *et al.* Pan-cancer analysis demonstrates that integrating polygenic risk scores with modifiable risk factors improves risk prediction. *Nat Commun* 2020;**11**:1–11. doi:10.1038/s41467-020-19600-4

244 Hung RJ, Warkentin MT, Brhane Y, *et al.* Assessing lung cancer absolute risk trajectory based on a polygenic risk model. *Cancer Res* 2021;**81**:1607–15. doi:10.1158/0008-5472.CAN-20-1237

245 Van Der Aalst CM, Ten Haaf K, de Koning HJ. Implementation of lung cancer screening: what are the main issues? *Transl Lung Cancer Res* 2021;**10**:1050. doi:10.21037/TLCR-20-985

246 Aberle DR, Adams AM, Berg CD, *et al.* Baseline characteristics of participants in the randomized national lung screening trial. *J Natl Cancer Inst* 2010;**102**:1771–9. doi:10.1093/JNCI/DJQ434

247 McRonald FE, Yadegarfar G, Baldwin DR, *et al.* The UK Lung Screen (UKLS): Demographic profile of first 88,897 approaches provides recommendations for population screening. *Cancer Prev Res* 2014;**7**:362–71. doi:10.1158/1940-6207.CAPR-13-0206/36159/AM/THE-UK-LUNG-SCREEN-UKLS-DEMOGRAPHIC-PROFILE-OF

248 Crosbie PA, Balata H, Evison M, *et al.* Implementing lung cancer screening: Baseline results from a community-based 'Lung Health Check' pilot in deprived areas of Manchester. Thorax. 2018. doi:10.1136/thoraxjnl-2017-211377

249 QRISK3. https://www.qrisk.org/ (accessed 20 Apr 2022).

250 Crosbie PA, Balata H, Evison M, *et al.* Second round results from the Manchester 'Lung Health Check' community-based targeted lung cancer screening pilot. *Thorax* 2018;:thoraxjnl-2018-212547. doi:10.1136/thoraxjnl-2018-212547

251 Manchester's Lung Health Check Pilot. http://cancervanguard.nhs.uk/wp-content/uploads/2017/04/Macmillan-lung-health-check-pilot-Manchester.pdf (accessed 30 Nov 2021).

252 Lung Health Checks | Lung Cancer | Cancer Research UK. https://www.cancerresearchuk.org/about-cancer/lung-cancer/getting-diagnosed/lung-health-checks (accessed 22 Feb 2022).

253 Lung cancer - UK National Screening Committee (UK NSC) - GOV.UK. https://view-health-screening-recommendations.service.gov.uk/lung-cancer/ (accessed 29 Mar 2022).

254 Crosbie PA, Gabe R, Simmonds I, *et al.* Protocol: Yorkshire Lung Screening Trial (YLST): protocol for a randomised controlled trial to evaluate invitation to community-based low-dose CT screening for lung cancer versus usual care in a targeted population at risk. *BMJ Open* 2020;**10**:e037075. doi:10.1136/BMJOPEN-2020-037075

255 Quaife SL, Ruparel M, Dickson JL, *et al.* Lung screen uptake trial (LSUT): Randomized controlled clinical trial testing targeted invitation materials. *Am J Respir Crit Care Med* 2020;**201**:965–75. doi:10.1164/RCCM.201905-0946OC/SUPPL_FILE/DISCLOSURES.PDF

256 Elwyn G, Frosch D, Thomson R, *et al.* Shared Decision Making: A Model for Clinical Practice. *J Gen Intern Med* 2012;**27**:1361. doi:10.1007/S11606-012-2077-6

257 Childress JF, Childress MD. What does the evolution from informed consent to shared decision making teach us about authority in health care? *AMA J Ethics* 2020;**22**:E423–9. doi:10.1001/AMAJETHICS.2020.423

258 Caverly TJ, Hayward RA. Dealing with the Lack of Time for Detailed Shared Decision-making in Primary Care: Everyday Shared Decision-making. *J Gen Intern Med* 2020;**35**:3045. doi:10.1007/S11606-020-06043-2

259 Caverly TJ, Cao P, Hayward RA, *et al.* Identifying Patients for Whom Lung Cancer Screening Is Preference-Sensitive A Microsimulation Study. doi:10.7326/M17-2561

260 NCA - Screening for Lung Cancer with Low Dose Computed Tomography (LDCT) (CAG-00439N) - Decision Memo. https://www.cms.gov/medicare-coverage-database/view/ncacal-decision-memo.aspx?proposed=N&NCAId=274 (accessed 14 Mar 2022).

261 Peretti-Watel P, Fressard L, Bocquier A, *et al.* Perceptions of cancer risk factors and socioeconomic status. A French study. *Prev Med Reports* 2016;**3**:171–6. doi:10.1016/j.pmedr.2016.01.008

262 Tilburt JC, James KM, Sinicrope PS, *et al.* Factors influencing cancer risk perception in high risk populations: A systematic review. *Hered Cancer Clin Pract* 2011;**9**:2. doi:10.1186/1897-4287-9-2

263 Brawarsky P, Eibensteiner K, Klinger E V., *et al.* Accuracy of self-perceived risk for common conditions. *Cogent Med* 2018;**5**:1463894. doi:10.1080/2331205x.2018.1463894

264 Diaz D, Fix B, Caruso R, *et al.* Worry about Lung Cancer is Related to Numeracy and Risk Perception of Diseases Associated with Smoking. *Am J Heal Educ* 2020;**51**:14–21. doi:10.1080/19325037.2019.1687364

265 Evans D, Burnell L, Hopwood P, *et al.* Perception of risk in women with a family history of breast cancer. *Br J Cancer* 1993;**67**:612–4. doi:10.1038/bjc.1993.112

266 Evans DGR, Donnelly LS, Harkness EF, *et al.* Breast cancer risk feedback to women in the UK NHS breast screening population. *Br J Cancer* 2016;**114**:1045–52. doi:10.1038/bjc.2016.56

267 Sivell S, Elwyn G, Gaff CL, *et al.* How Risk is Perceived, Constructed and Interpreted by Clients in Clinical Genetics, and the Effects on Decision Making: Systematic Review. *J Genet Couns* 2008;**17**:30–63. doi:10.1007/s10897-007-9132-1

268 Brawarsky P, Eibensteiner K, Klinger E V., *et al.* Accuracy of self-perceived risk for common conditions. *Cogent Med* 2018;**5**. doi:10.1080/2331205x.2018.1463894

269 Park ER, Ostroff JS, Rakowski W, *et al.* Risk perceptions among participants undergoing lung cancer screening: baseline results from the National Lung Screening Trial. *Ann Behav Med* 2009;**37**:268–79. doi:10.1007/s12160-009-9112-9

270 Chen L-S, Kaphingst KA, Tseng T-S, *et al.* How are lung cancer risk perceptions and cigarette smoking related?-testing an accuracy hypothesis. *Transl Cancer Res* 2016;**5**:S964–71. doi:10.21037/tcr.2016.10.75

271 Chen LS, Kaphingst KA. Risk perceptions and family history of lung cancer: differences by smoking status. *Public Health Genomics* 2011;**14**:26–34. doi:10.1159/000294151

272 Ziebarth NR. Lung cancer risk perception biases. *Prev Med (Baltim)* 2018;**110**:16–23. doi:10.1016/J.YPMED.2018.01.020

273 Han PKJ, Lary C, Black A, *et al.* Effects of Personalized Risk Information on Patients Referred for Lung Cancer Screening with Low-Dose CT. *Med Decis Mak* 2019;:0272989X1987596. doi:10.1177/0272989X19875966

274 Turner J, Pond GR, Tremblay A, *et al.* Risk Perception Among a Lung Cancer Screening Population. *Chest* Published Online First: 2 March 2021. doi:10.1016/j.chest.2021.02.050

275 Ferrer RA, Klein WMP. Risk perceptions and health behavior. Curr. Opin. Psychol. 2015;**5**:85–9. doi:10.1016/j.copsyc.2015.03.012

276 Cataldo JK. High-risk older smokers' perceptions, attitudes, and beliefs about lung cancer screening. *Cancer Med* 2016;**5**:753–9. doi:10.1002/cam4.617

277 Raz DJ, Wu G, Nelson RA, *et al.* Perceptions and Utilization of Lung Cancer Screening Among Smokers Enrolled in a Tobacco Cessation Program. *Clin Lung Cancer* 2019;**20**:e115–22. doi:10.1016/j.cllc.2018.09.013

278 Lam ACL, Aggarwal R, Cheung S, *et al.* Predictors of participant nonadherence in lung cancer screening programs: a systematic review and meta-analysis. Lung Cancer. 2020;**146**:134–44. doi:10.1016/j.lungcan.2020.05.013

279 Chalian H, Khoshpouri P, Assari S. Demographic, social, and behavioral determinants of lung cancer perceived risk and worries in a national sample of American adults; Does lung cancer risk matter? *Med* 2018;**54**. doi:10.3390/MEDICINA54060097

280 Bell NR, Dickinson JA, Grad R, *et al.* Understanding and communicating risk: Measures of outcome and the magnitude of benefits and harms. *Can Fam Physician* 2018;**64**:181./pmc/articles/PMC5851390/ (accessed 24 Feb 2022).

281 Zipkin DA, Umscheid CA, Keating NL, *et al.* Evidence-based risk communication: a systematic review. *Ann Intern Med* 2014;**161**:270–80. doi:10.7326/M14-0295

282 Green DB, Pua BB, Crawford CB, *et al.* Screening for Lung Cancer: Communicating With Patients. *https://doi.org/102214/AJR1718836* 2017;**210**:497–502. doi:10.2214/AJR.17.18836

283 Should I Screen: Lung cancer screening decision aid. https://shouldiscreen.com/English/home (accessed 24 Feb 2022).

284 Balata H. The Impact of a Community Based Lung Cancer Screening Programme. Thesis, Univ. Manchester. https://www.research.manchester.ac.uk/portal/en/theses/the-impact-of-a-community-based-lung-cancer-screening-programme(8683fbb3-0237-46fe-b59f-55fcca7c091b).html (accessed 11 Oct 2021).

285 Callister MEJ, Baldwin DR, Akram AR, *et al.* British Thoracic Society guidelines for the investigation and management of pulmonary nodules: accredited by NICE. *Thorax* 2015;**70**:ii1–54. doi:10.1136/THORAXJNL-2015-207168

286 FLETCHER, M. C. Standardised questionnaire on respiratory symptoms : a statement prepared and approved by the MRC Committee on the Aetiology of Chronic Bronchitis (MRC breathlessness score). *BMJ* 1960;**2**:1665.https://ci.nii.ac.jp/naid/10025106705 (accessed 11 Oct 2021).

287 ISCED Mappings | UNESCO UIS. http://uis.unesco.org/en/isced-mappings (accessed 27 Feb 2020).

288 R Package for Individual Risks of Lung Cancer and Lung Cancer Death - National Cancer Institute. https://dceg.cancer.gov/tools/risk-assessment/lcmodels (accessed 15 Feb 2021).

289 (EN) - Gentra Puregene Handbook - QIAGEN. https://www.qiagen.com/gb/resources/resourcedetail?id=a9e6a609-4600-4b03-afbd-974318590ce5&lang=en (accessed 25 May 2021).

290 FlexiGene DNA Handbook - (EN) - QIAGEN. https://www.qiagen.com/gb/resources/resourcedetail?id=6b147421-7846-411e-9993-bb01563b807e&lang=en (accessed 25 May 2021).

291 QIAamp DNA Blood Midi/Maxi Handbook - QIAGEN. https://www.qiagen.com/gb/resources/resourcedetail?id=bf32146a-77fd-40c2-8743-c28974f7935b&lang=en (accessed 25 May 2021).

292 Lite N. T123-TECHNICAL BULLETIN Interpretation of Nucleic Acid 260/280 Ratios. www.thermoscientific.com/nanodrop

293 DNA Clean & Concentrator-25. https://www.zymoresearch.com/collections/dna-clean-concentrator-kits-dcc/products/dna-clean-concentrator-25 (accessed 25 May 2021).

294 Qubit^{TM} dsDNA BR Assay Kit. https://www.thermofisher.com/order/catalog/product/Q32850#/Q32850 (accessed 25 May 2021).

295 Amos CI, Dennis J, Wang Z, *et al.* The OncoArray Consortium: A Network for Understanding the Genetic Architecture of Common Cancers. *Cancer Epidemiol Biomarkers Prev* 2017;**26**:126–35. doi:10.1158/1055-9965.EPI-16-0106

296 Amplify DNA Infinium HTS Assay Manual Workflow Checklist.

297 Evaluation of Infinium Genotyping Assay Controls Training Guide. 2012. https://support.illumina.com/content/dam/illumina-support/courses/eval-inf-

controls/story_content/external_files/Infinium_Controls_Training_Guide.pdf (accessed 21 Aug 2019).

298 Chang CC, Chow CC, Tellier LCAM, *et al.* Second-generation PLINK: Rising to the challenge of larger and richer datasets. *Gigascience* 2015;**4**. doi:10.1186/S13742-015-0047-8/2707533

299 Marees AT, Kluiver H de, Stringer S, *et al.* A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *Int J Methods Psychiatr Res* 2018;**27**. doi:10.1002/MPR.1608

300 Illumina. Illumina ® SNP Genotyping TECHNICAL NOTE making sense out of life. Published Online First: 2005.www.illumina.com (accessed 11 Oct 2021).

301 McCarthy Tools. https://www.well.ox.ac.uk/~wrayner/tools/ (accessed 11 Oct 2021).

302 KING: Relationship Inference Software. https://www.kingrelatedness.com/ (accessed 2 Mar 2022).

303 Abraham G, Qiu Y, Inouye M. FlashPCA2: principal component analysis of Biobank-scale genotype datasets. *Bioinformatics* 2017;**33**:2776–8. doi:10.1093/BIOINFORMATICS/BTX299

304 Altshuler DM, Gibbs RA, Peltonen L, *et al.* Integrating common and rare genetic variation in diverse human populations. *Nat 2010 4677311* 2010;**467**:52–8. doi:10.1038/nature09298

305 Bellenguez C, Strange A, Freeman C, *et al.* A robust clustering algorithm for identifying problematic samples in genome-wide association studies. *Bioinformatics* 2012;**28**:134–5. doi:10.1093/BIOINFORMATICS/BTR599

306 Das S, Forer L, Schönherr S, *et al.* Next-generation genotype imputation service and methods. *Nat Genet* 2016;**48**:1284–7. doi:10.1038/NG.3656

307 PGS Catalog - lung cancer [MONDO:0008903] (Polygenic Trait). https://www.pgscatalog.org/trait/MONDO_0008903/ (accessed 11 Oct 2021).

308 Machiela MJ, Chanock SJ. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics* 2015;**31**:3555–7. doi:10.1093/BIOINFORMATICS/BTV402

309 Wickham H. *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York 2016. https://ggplot2.tidyverse.org.

310 Robin X, Turck N, Hainard A, *et al.* pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* 2011;**12**:1–8. doi:10.1186/1471-2105-12-77/TABLES/3

311 Kundu S, Aulchenko YS, Van Duijn CM, *et al.* PredictABEL: an R package for the assessment of risk prediction models. *Eur J Epidemiol* 2011;**26**:261. doi:10.1007/S10654-011-9567-4

312 Pencina MJ, D'Agostino RB, D'Agostino RB, *et al.* Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med* 2008;**27**:157–72. doi:10.1002/SIM.2929

313 Watanabe K, Taskesen E, Van Bochoven A, *et al.* Functional mapping and annotation of genetic associations with FUMA. *Nat Commun 2017 81* 2017;**8**:1–11. doi:10.1038/s41467-017-01261-5

314 Choi SW, O'Reilly PF. PRSice-2: Polygenic Risk Score software for biobank-scale data. *Gigascience* 2019;**8**. doi:10.1093/GIGASCIENCE/GIZ082

315 Gallery - Lung Health Check. https://mft.nhs.uk/lunghealthcheck/about-the-lung-health-

check/gallery/ (accessed 9 Dec 2021).

316 Sigel K, Wisnivesky JP. Comorbidity Profiles of Patients with Lung Cancer: A New Approach to Risk Stratification? *https://doi.org/101513/AnnalsATS201706-442ED* 2017;**14**:1512–3. doi:10.1513/ANNALSATS.201706-442ED

317 Saghir Z, Dirksen A, Ashraf H, *et al.* CT screening for lung cancer brings forward early disease. The randomised Danish Lung Cancer Screening Trial: status after five annual screening rounds with low-dose CT. *Thorax* 2012;**67**:296–301. doi:10.1136/THORAXJNL-2011-200736

318 Lebrett MB, Balata H, Evison M, *et al.* Analysis of lung cancer risk model (PLCO $_{M2012}$ and LLP $_{v2}$) performance in a community-based lung cancer screening programme. *Thorax* 2020;:thoraxjnl-2020-214626. doi:10.1136/thoraxjnl-2020-214626

319 Katki HA, Kovalchik SA, Petito LC, *et al.* Implications of 9 risk prediction models for selecting ever-smokers for CT lung-cancer screening. *Ann Intern Med* 2018;**169**:10. doi:10.7326/M17-2701

320 O'Dowd EL, Haaf K ten, Kaur J, *et al.* Selection of eligible participants for screening for lung cancer using primary care data. *Thorax* 2021;:thoraxjnl-2021-217142. doi:10.1136/THORAXJNL-2021-217142

321 Collins GS, Ogundimu EO, Altman DG. Sample size considerations for the external validation of a multivariable prognostic model: A resampling study. *Stat Med* 2016;**35**:214–26. doi:10.1002/sim.6787

322 Bradley P, Ryan D, Craig C, *et al.* P61.02 The Use of PLCOm2012 vs PLCOm2012noRace Risk Prediction Models in a UK Lung Cancer Screening Programme. *J Thorac Oncol* 2021;**16**:S1172. doi:10.1016/J.JTHO.2021.08.637

323 Tammemagi MC, Lam SC, McWilliams AM, *et al.* Incremental value of pulmonary function and sputum DNA image cytometry in lung cancer risk prediction. *Cancer Prev Res (Phila)* 2011;**4**:552–61. doi:10.1158/1940-6207.CAPR-10-0183

324 Muller DC, Johansson M, Brennan P. JOURNAL OF CLINICAL ONCOLOGY Lung Cancer Risk Prediction Model Incorporating Lung Function: Development and Validation in the UK Biobank Prospective Cohort Study. *J Clin Oncol* 2017;**35**:861–9. doi:10.1200/JCO.2016.69.2467

325 Schreuder A, Jacobs C, Lessmann N, *et al.* Combining pulmonary and cardiac computed tomography biomarkers for disease-specific risk modelling in lung cancer screening. *Eur Respir J* 2021;**58**. doi:10.1183/13993003.03386-2020

326 Tammemägi MC, Ruparel M, Tremblay A, *et al.* USPSTF2013 versus PLCOm2012 lung cancer screening eligibility criteria (International Lung Screening Trial): interim analysis of a prospective cohort study. *Lancet Oncol* 2022;**23**:138–48. doi:10.1016/S1470-2045(21)00590-8/ATTACHMENT/FC1A2043-9EC0-47D8-B708-BF106ED17CA8/MMC1.PDF

327 ten Haaf K, Jeon J, Tammemägi MC, *et al.* Risk prediction models for selection of lung cancer screening candidates: A retrospective validation study. *PLOS Med* 2017;**14**:e1002277. doi:10.1371/JOURNAL.PMED.1002277

328 Dinh TA, Alperin P, Walter LC, *et al.* Impact of comorbidity on colorectal cancer screening cost-effectiveness study in diabetic populations. *J Gen Intern Med* 2012;**27**:730–8. doi:10.1007/s11606-011-1972-6

329 Mandelblatt JS, Schechter CB, Yabroff KR, *et al.* Toward optimal screening strategies for older women. Costs, benefits, and harms of breast cancer screening by age, biology, and

health status. *J Gen Intern Med* 2005;**20**:487–96. doi:10.1111/j.1525-1497.2005.0116.x

330 Health state life expectancies by national deprivation deciles, England - Office for National Statistics. https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/healthinequalities/bulletins/healthstatelifeexpectanciesbyindexofmultipledeprivationimd/2016to2018 (accessed 12 Mar 2022).

331 Charlton J, Rudisill C, Bhattarai N, *et al.* Impact of deprivation on occurrence, outcomes and health care costs of people with multiple morbidity. *J Health Serv Res Policy* 2013;**18**:215. doi:10.1177/1355819613493772

332 Landy R, Young CD, Skarzynski M, *et al.* Using Prediction Models to Reduce Persistent Racial and Ethnic Disparities in the Draft 2020 USPSTF Lung Cancer Screening Guidelines. *J Natl Cancer Inst* 2021;**113**:1590–4. doi:10.1093/JNCI/DJAA211

333 Torke AM, Schwartz PH, Holtz LR, *et al.* Older adults and forgoing cancer screening: 'I think it would be strange'. *JAMA Intern Med* 2013;**173**:526–31. doi:10.1001/JAMAINTERNMED.2013.2903

334 Schoenborn NL, Lee K, Pollack CE, *et al.* Older Adults' Views and Communication Preferences About Cancer Screening Cessation. *JAMA Intern Med* 2017;**177**:1121–8. doi:10.1001/JAMAINTERNMED.2017.1778

335 https://dceg.cancer.gov/tools/risk-assessment/lcrisks.

336 Howard DH, Richards TB, Bach PB, *et al.* Comorbidities, smoking status, and life expectancy among individuals eligible for lung cancer screening. *Cancer* 2015;**121**:4341–7. doi:10.1002/CNCR.29677

337 Goldstraw P, Crawley J, Chansky K. No The IASLC Lung Cancer Staging Project: proposals for the revision of the TNM stage groupings in the forthcoming (seventh) edition of the TNM Classification of malignant tumours. *J Thorac Oncol* 2007;**2**:706–14.

338 Graff RE, Cavazos TB, Thai KK, *et al.* Cross-cancer evaluation of polygenic risk scores for 16 cancer types in two large cohorts. *Nat Commun 2021 121* 2021;**12**:1–9. doi:10.1038/s41467-021-21288-z

339 LG F, S P, LJ B, *et al.* Cancer PRSweb: An Online Repository with Polygenic Risk Scores for Major Cancer Traits and Their Evaluation in Two Independent Biobanks. *Am J Hum Genet* 2020;**107**:815–36. doi:10.1016/J.AJHG.2020.08.025

340 Shi Z, Yu H, Wu Y, *et al.* Systematic evaluation of cancer-specific genetic risk score for 11 types of cancer in The Cancer Genome Atlas and Electronic Medical Records and Genomics cohorts. *Cancer Med* 2019;**8**:3196–205. doi:10.1002/CAM4.2143

341 G J, Y L, W W, *et al.* Evaluating the Utility of Polygenic Risk Scores in Identifying High-Risk Individuals for Eight Common Cancers. *JNCI cancer Spectr* 2020;**4**. doi:10.1093/JNCICS/PKAA021

342 Wang J, Liu Q, Yuan S, *et al.* Genetic predisposition to lung cancer: comprehensive literature integration, meta-analysis, and multiple evidence assessment of candidate-gene association studies. *Sci Rep* 2017;**7**. doi:10.1038/S41598-017-07737-0

343 Yi X, Li W, Wang Y, *et al.* The relationship between CHRNA5/A3/B4 gene cluster polymorphisms and lung cancer risk: An updated meta-analysis and systematic review. *Medicine (Baltimore)* 2021;**100**:e24355. doi:10.1097/MD.0000000000024355

344 Ehret GB. Genome-Wide Association Studies: Contribution of Genomics to Understanding Blood Pressure and Essential Hypertension. *Curr Hypertens Rep* 2010;**12**:17.

doi:10.1007/S11906-009-0086-6

345 Biotypes. http://www.ensembl.org/info/genome/genebuild/biotypes.html (accessed 21 Mar 2022).

346 Choi SW, Mak TSH, O'Reilly PF. A guide to performing Polygenic Risk Score analyses. *Nat Protoc* 2020;**15**:2759. doi:10.1038/S41596-020-0353-1

347 Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation* 2007;**115**:928–35. doi:10.1161/CIRCULATIONAHA.106.672402

348 Rosenberger A, Muttray N, Hung RJ, *et al.* Gene–gene interaction of AhRwith and within the Wntcascade affects susceptibility to lung cancer. *Eur J Med Res* 2022;**27**. doi:10.1186/S40001-022-00638-7

349 Sud A, Turnbull C, Houlston R. Will polygenic risk scores for cancer ever be clinically useful? *NPJ Precis Oncol* 2021;**5**. doi:10.1038/S41698-021-00176-1

350 Kerr KF, Wang Z, Janes H, *et al.* Net Reclassification Indices for Evaluating Risk-Prediction Instruments: A Critical Review. *Epidemiology* 2014;**25**:114. doi:10.1097/EDE.0000000000000018

351 Pepe MS, Fan J, Feng Z, *et al.* The Net Reclassification Index (NRI): A Misleading Measure of Prediction Improvement Even with Independent Test Data Sets. *Stat Biosci* 2015;**7**:282–95. doi:10.1007/S12561-014-9118-0/FULLTEXT.HTML

352 O'Brien TD, Jia P, Caporaso NE, *et al.* Weak sharing of genetic association signals in three lung cancer subtypes: Evidence at the SNP, gene, regulation, and pathway levels. *Genome Med* 2018;**10**:1–14. doi:10.1186/S13073-018-0522-9/FIGURES/5

353 Roberts MR, Sordillo JE, Kraft P, *et al.* Sex-stratified polygenic risk score identifies individuals at increased risk of basal cell carcinoma. *J Invest Dermatol* 2020;**140**:971. doi:10.1016/J.JID.2019.09.020

354 Manikpurage HD, Eslami A, Perrot N, *et al.* Polygenic Risk Score for Coronary Artery Disease Improves the Prediction of Early-Onset Myocardial Infarction and Mortality in Men. *Circ Genomic Precis Med* 2021;**14**:697–708. doi:10.1161/CIRCGEN.121.003452

355 Shi X, Young S, Morahan G. Identification of genetic variants associated with sex-specific lung-cancer risk. *Cancers (Basel)* 2021;**13**. doi:10.3390/CANCERS13246379/S1

356 Karnes JH, Bastarache L, Shaffer CM, *et al.* Phenome-wide scanning identifies multiple diseases and disease severity phenotypes associated with HLA variants. *Sci Transl Med* 2017;**9**. doi:10.1126/SCITRANSLMED.AAI8708

357 Smedby KE, Foo JN, Skibola CF, *et al.* GWAS of follicular lymphoma reveals allelic heterogeneity at 6p21.32 and suggests shared genetic susceptibility with diffuse large B-cell lymphoma. *PLoS Genet* 2011;**7**. doi:10.1371/JOURNAL.PGEN.1001378

358 Bordeira Gaspar T, Sá A, Manuel Lopes J, *et al.* Telomere maintenance mechanisms in cancer. *mdpi.com* doi:10.3390/genes9050241

359 MAGI2 Gene - GeneCards | MAGI2 Protein | MAGI2 Antibody. https://www.genecards.org/cgi-bin/carddisp.pl?gene=MAGI2 (accessed 22 Mar 2022).

360 Hu Y, Li Z, Guo L, *et al.* MAGI-2 Inhibits cell migration and proliferation via PTEN in human hepatocarcinoma cells. *Arch Biochem Biophys* 2007;**467**:1–9. doi:10.1016/J.ABB.2007.07.027

361 Wu X, Hepner K, Castelino-Prabhu S, *et al.* Evidence for regulation of the PTEN tumor suppressor by a membrane-localized multi-PDZ domain containing scaffold protein MAGI-

2. *Proc Natl Acad Sci U S A* 2000;**97**:4233. doi:10.1073/PNAS.97.8.4233

362 Goldstein J, Borowsky AD, Goyal R, *et al.* MAGI-2 in prostate cancer: an immunohistochemical study. *Hum Pathol* 2016;**52**:83–91. doi:10.1016/J.HUMPATH.2016.01.003

363 David SN, Arnold Egloff SA, Goyal R, *et al.* MAGI2 is an independent predictor of biochemical recurrence in prostate cancer. *Prostate* 2018;**78**:616–22. doi:10.1002/PROS.23506

364 Goldstein J, Goyal R, Roland JT, *et al.* MAGI-2 Is a Sensitive and Specific Marker of Prostatic Adenocarcinoma A Comparison With AMACR. *Am J Clin Pathol* 2016;**146**:294–302. doi:10.1093/AJCP/AQW111

365 Cao Z, Ji J, Wang FB, *et al.* MAGI-2 downregulation: a potential predictor of tumor progression and early recurrence in Han Chinese patients with prostate cancer. *Asian J Androl* 2020;**22**:616–22. doi:10.4103/AJA.AJA_142_19

366 Mahdian R, Nodouzi V, Asgari M, *et al.* Expression profile of MAGI2 gene as a novel biomarker in combination with major deregulated genes in prostate cancer. *Mol Biol Rep* 2014;**41**:6125–31. doi:10.1007/S11033-014-3491-0/FIGURES/3

367 Kai-xing L, Cheng C, Rui L, *et al.* Roles of lncRNA MAGI2-AS3 in human cancers. *Biomed Pharmacother* 2021;**141**:111812. doi:10.1016/J.BIOPHA.2021.111812

368 Hao X, Sci KY-ERMP, 2019 undefined. LncRNA MAGI2-AS3 suppresses the proliferation and invasion of non-small cell lung carcinoma through miRNA-23a-3p/PTEN axis. *europeanreview.org*https://www.europeanreview.org/wp/wp-content/uploads/7399-7407.pdf (accessed 22 Mar 2022).

369 Li F, Hu Q, Pang Z, *et al.* LncRNA MAGI2-AS3 Upregulates Cytokine Signaling 1 by Sponging miR-155 in Non-Small Cell Lung Cancer. *Cancer Biother Radiopharm* 2020;**35**:72–6. doi:10.1089/CBR.2019.2898

370 Quach BC, Bray MJ, Gaddis NC, *et al.* Expanding the genetic architecture of nicotine dependence and its shared genetics with multiple traits. *Nat Commun 2020 111* 2020;**11**:1–13. doi:10.1038/s41467-020-19265-z

371 Guo Z, Zhou C, Zhou L, *et al.* Overexpression of DAPK1-mediated inhibition of IKKβ/CSN5/PD-L1 axis enhances natural killer cell killing ability and inhibits tumor immune evasion in gastric cancer. *Cell Immunol* 2022;**372**:104469. doi:10.1016/J.CELLIMM.2021.104469

372 Song Z, Li Z, Han W, *et al.* Low DAPK1 expression correlates with poor prognosis and sunitinib resistance in clear cell renal cell carcinoma. *Aging (Albany NY)* 2020;**13**:1842–58. doi:10.18632/AGING.103638

373 Chen D, Zhou XZ, Lee TH. Death-Associated Protein Kinase 1 as a Promising Drug Target in Cancer and Alzheimer's Disease. *Recent Pat Anticancer Drug Discov* 2019;**14**:144. doi:10.2174/1574892814666181218170257

374 Yuan W, Chen J, Shu Y, *et al.* Correlation of DAPK1 methylation and the risk of gastrointestinal cancer: A systematic review and meta-analysis. *PLoS One* 2017;**12**:e0184959. doi:10.1371/JOURNAL.PONE.0184959

375 Tang X, Khuri FR, Jack Lee J, *et al.* Hypermethylation of the death-associated protein (DAP) kinase promoter and aggressiveness in stage I non-small-cell lung cancer. *J Natl Cancer Inst* 2000;**92**:1511–6. doi:10.1093/JNCI/92.18.1511

376 Lee A, Mavaddat N, Wilcox AN, *et al.* BOADICEA: a comprehensive breast cancer risk

prediction model incorporating genetic and nongenetic risk factors. *Genet Med* 2019;**21**:1708–18. doi:10.1038/s41436-018-0406-9

377 Evans DGR, Harkness EF, Brentnall AR, *et al.* Breast cancer pathology and stage are better predicted by risk stratification models that include mammographic density and common genetic variants. *Breast Cancer Res Treat* 2019;**176**:141–8. doi:10.1007/s10549-019-05210-2

378 Li Y, Xiao X, Han Y, *et al.* Genome-wide interaction study of smoking behavior and non-small cell lungcancer risk in Caucasian population. *Carcinogenesis* 2018;**39**:336. doi:10.1093/CARCIN/BGX113

379 Martin AR, Kanai M, Kamatani Y, *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat Genet* 2019;**51**:584–91. doi:10.1038/s41588-019-0379-x

380 Knoppers BM, Bernier A, Moreno PG, *et al.* Of Screening, Stratification, and Scores. *J Pers Med* 2021;**11**. doi:10.3390/JPM11080736

381 Fritsche LG, Ma Y, Zhang D, *et al.* On cross-ancestry cancer polygenic risk scores. *PLOS Genet* 2021;**17**:e1009670. doi:10.1371/JOURNAL.PGEN.1009670

382 Lewis CM, Vassos E. Polygenic risk scores: From research tools to clinical instruments. *Genome Med* 2020;**12**:1–11. doi:10.1186/S13073-020-00742-5/TABLES/2

383 Crosbie PA, Gabe R, Simmonds I, *et al.* Yorkshire Lung Screening Trial (YLST): protocol for a randomised controlled trial to evaluate invitation to community-based low-dose CT screening for lung cancer versus usual care in a targeted population at risk. *BMJ Open* 2020;**10**:e037075. doi:10.1136/bmjopen-2020-037075

384 Manchester Lung Health Study - Full Text View - ClinicalTrials.gov. https://clinicaltrials.gov/ct2/show/NCT04409444 (accessed 8 Oct 2020).

385 Carozzi FM, Bisanzi S, Falini P, *et al.* Molecular profile in body fluids in subjects enrolled in a randomised trial for lung cancer screening: Perspectives of integrated strategies for early diagnosis. *Lung Cancer* 2010;**68**:216–21. doi:10.1016/j.lungcan.2009.06.015

386 Evans DG, Astley S, Stavrinos P, *et al.* Project 2: assessment of predictive value of new genetic variants. Published Online First: 2016.https://www.ncbi.nlm.nih.gov/books/NBK379496/ (accessed 9 Oct 2020).

387 Abraham JE, Maranian MJ, Spiteri I, *et al.* Saliva samples are a viable alternative to blood samples as a source of DNA for high throughput genotyping. *BMC Med Genomics* 2012;**5**. doi:10.1186/1755-8794-5-19

388 Eeles RA, ni Raghallaigh H. BARCODE 1: A pilot study investigating the use of genetic profiling to identify men in the general population with the highest risk of prostate cancer to invite for targeted screening. *J Clin Oncol* 2020;**38**:1505–1505. doi:10.1200/jco.2020.38.15_suppl.1505

389 Leventhal KG, Tuong W, Peshkin BN, *et al.* ‘Is it really worth it to get tested?’: Primary care patients’ impressions of predictive SNP testing for colon cancer. In: *Journal of Genetic Counseling*. NIH Public Access 2013. 138–51. doi:10.1007/s10897-012-9530-x

390 Hall MJ, Ruth KJ, Chen DYT, *et al.* Interest in genomic SNP testing for prostate cancer risk: A pilot survey. *Hered Cancer Clin Pract* 2015;**13**. doi:10.1186/s13053-015-0032-3

391 Young MA, Forrest LE, Rasmussen VM, *et al.* Making Sense of SNPs: Women’s Understanding and Experiences of Receiving a Personalized Profile of Their Breast Cancer Risks. *J Genet Couns* 2018;**27**:702–8. doi:10.1007/s10897-017-0162-z

392 Nusbaum R, Leventhal KG, Hooker GW, *et al.* Translational genomic research: Protocol development and initial outcomes following SNP testing for colon cancer risk. *Transl Behav Med* 2013;**3**:17–29. doi:10.1007/s13142-012-0149-0

393 Forrest LE, Sawyer SD, Hallowell N, *et al.* High-risk women's risk perception after receiving personalized polygenic breast cancer risk information. *J Community Genet* 2019;**10**:197–206. doi:10.1007/s12687-018-0378-0

394 Kerpel-Fronius A, Tammemägi M, Cavic M, *et al.* Screening for Lung Cancer in Individuals Who Never Smoked: An International Association for the Study of Lung Cancer Early Detection and Screening Committee Report. *J Thorac Oncol* 2022;**17**:56–66. doi:10.1016/J.JTHO.2021.07.031

395 Adeyemo A, Balaconis MK, Darnes DR, *et al.* Responsible use of polygenic risk scores in the clinic: potential benefits, risks and gaps. *Nat Med 2021 2711* 2021;**27**:1876–84. doi:10.1038/s41591-021-01549-6

396 Callender T, Emberton M, Morris S, *et al.* Polygenic risk-tailored screening for prostate cancer: A benefit-harm and cost-effectiveness modelling study. *PLoS Med* 2019;**16**. doi:10.1371/journal.pmed.1002998

397 Pashayan N, Morris S, Gilbert FJ, *et al.* Cost-effectiveness and Benefit-to-Harm Ratio of Risk-Stratified Screening for Breast Cancer A Life-Table Model. *JAMA Oncol* 2018;**4**:1504–10. doi:10.1001/jamaoncol.2018.1901

398 Naber SK, Kundu S, Kuntz KM, *et al.* Cost-effectiveness of risk-stratified colorectal cancer screening based on polygenic risk: Current status and future potential. *JNCI Cancer Spectr* 2020;**4**. doi:10.1093/jncics/pkz086

399 Fagerlin A, Zikmund-Fisher BJ, Ubel PA. Helping patients decide: ten steps to better risk communication. *J Natl Cancer Inst* 2011;**103**:1436–43. doi:10.1093/jnci/djr318

400 The Long Island Breast Cancer Research Project. https://static-content.springer.com/esm/art%3A10.1007%2Fs13187-019-01530-5/MediaObjects/13187_2019_1530_MOESM1_ESM.pdf (accessed 26 May 2019).

401 Chen LS, Kaphingst KA. Risk perceptions and family history of lung cancer: Differences by smoking status. *Public Health Genomics* 2011;**14**:26–34. doi:10.1159/000294151

402 Greillier L, Cortot AB, Viguier J, *et al.* Perception of Lung Cancer Risk: Impact of Smoking Status and Nicotine Dependence. *Curr Oncol Rep* 2018;**20**:1–7. doi:10.1007/s11912-017-0650-1

403 Lerman C, Trock B, Rimer BK, *et al.* Psychological side effects of breast cancer screening. *Health Psychol* 1991.

404 Kroenke K, Spitzer RL, Williams JBW, *et al.* PHQ-4. ;**50**:613–21. doi:10.1176/appi.psy.50.6.613

405 Arroll B, Goodyear-Smith F, Crengle S, *et al.* Validation of PHQ-2 and PHQ-9 to screen for major depression in the primary care population. *Ann Fam Med* 2010;**8**:348–53. doi:10.1370/afm.1139

406 CR S, KA P, K S, *et al.* Cancer incidence and mortality projections in the UK until 2035. *Br J Cancer* 2016;**115**:1147–55. doi:10.1038/BJC.2016.304

407 Lebrett MB, Crosbie EJ, Yorke J, *et al.* Risk perception and disease knowledge in attendees of a community-based lung cancer screening programme. *Lung Cancer* 2022;**168**:1–9. doi:10.1016/J.LUNGCAN.2022.04.003

408 Usher-Smith JA, Silarova B, Sharp SJ, *et al.* Effect of interventions incorporating personalised cancer risk information on intentions and behaviour: A systematic review and meta-analysis of randomised controlled trials. *BMJ Open* 2018;**8**. doi:10.1136/bmjopen-2017-017717

409 Bayne M, Fairey M, Silarova B, *et al.* Effect of interventions including provision of personalised cancer risk information on accuracy of risk perception and psychological responses: A systematic review and meta-analysis. *Patient Educ Couns* 2020;**103**. doi:10.1016/j.pec.2019.08.010

410 Freeman ALJ. How to communicate evidence to patients. Drug Ther. Bull. 2019. doi:10.1136/dtb.2019.000008

411 Sherratt FC, Marcus MW, Robinson J, *et al.* Utilizing Lung Cancer Risk Prediction Models to Promote Smoking Cessation: Two Randomized Controlled Trials. *Am J Heal Promot* 2018;**32**:1196–205. doi:10.1177/0890117116673820

412 Edwards AGK, Naik G, Ahmed H, *et al.* Personalised risk communication for informed decision making about taking screening tests. *Cochrane Database Syst Rev* 2013;**2013**:1–96. doi:10.1002/14651858.CD001865.pub3

413 Klein WMP, Ferrer RA, Kaufman AR. How (or Do) People “Think” About Cancer Risk, and Why That Matters. *JAMA Oncol* 2020;**6**:983–4. doi:10.1001/JAMAONCOL.2020.0170

414 Quaife SL, Waller J, Dickson JL, *et al.* Psychological Targets for Lung Cancer Screening Uptake: A Prospective Longitudinal Cohort Study. *J Thorac Oncol* 2021;**16**:2016–28. doi:10.1016/J.JTHO.2021.07.025/ATTACHMENT/38BB4ED4-CD41-4A0D-A48E-2A1A7C64ADBD/MMC1.DOCX

415 Simpkin AL, Armstrong KA. Communicating Uncertainty: a Narrative Review and Framework for Future Research. J. Gen. Intern. Med. 2019;**34**. doi:10.1007/s11606-019-04860-8

416 Lewis A, Duch R. Gender differences in perceived risk of COVID-19. *Soc Sci Q* Published Online First: 29 October 2021. doi:10.1111/SSQU.13079

417 Brown GD, Largey A, McMullan C. The impact of gender on risk perception: Implications for EU member states’ national risk assessment processes. *Int J Disaster Risk Reduct* 2021;**63**:102452. doi:10.1016/J.IJDRR.2021.102452

418 McLean CP, Asnaani A, Litz BT, *et al.* Gender Differences in Anxiety Disorders: Prevalence, Course of Illness, Comorbidity and Burden of Illness. *J Psychiatr Res* 2011;**45**:1027. doi:10.1016/J.JPSYCHIRES.2011.03.006

419 Zajac LE, Klein WMP, McCaul KD. Absolute and comparative risk perceptions as predictors of cancer worry: Moderating effects of gender and psychological distress. *J Health Commun* 2006;**11**. doi:10.1080/10810730600637301

420 Ferrer RA, Klein WMP. Risk perceptions and health behavior. Curr. Opin. Psychol. 2015;**5**. doi:10.1016/j.copsyc.2015.03.012

421 Brewer NT, Chapman GB, Gibbons FX, *et al.* Meta-analysis of the relationship between risk perception and health behavior: the example of vaccination. *Health Psychol* 2007;**26**:136–45. doi:10.1037/0278-6133.26.2.136

422 Sheeran P, Harris PR, Epton T. Does heightening risk appraisals change people’s intentions and behavior? A meta-analysis of experimental studies. *Psychol Bull* 2014;**140**:511–43. doi:10.1037/A0033065

423 Gibbons FX, McGovern PG, Lando HA. Relapse and risk perception among members of a

smoking cessation clinic. *Health Psychol* 1991;**10**:42–5. doi:10.1037/0278-6133.10.1.42

424 Simon AE, Juszczyk D, Smyth N, *et al.* Knowledge of lung cancer symptoms and risk factors in the UK: Development of a measure and results from a population-based survey. *Thorax* 2012;**67**. doi:10.1136/thoraxjnl-2011-200898

425 de Jonge ETM, Vlasselaer J, Van de Putte G, *et al.* The construct of breast cancer risk perception: need for a better risk communication? *Facts, views Vis ObGyn* 2009;**1**.

426 Chen LS, Kaphingst KA. Risk Perceptions and Family History of Lung Cancer: Differences by Smoking Status. *Public Health Genomics* 2010;**14**:26. doi:10.1159/000294151

427 Kruger J, Patel R, Kegler M, *et al.* Perceptions of harm from secondhand smoke exposure among U.S. adults, 2009–2010. *Tob Induc Dis* 2016;**14**. doi:10.1186/S12971-016-0069-8

428 Madewell ZJ. The belief that secondhand smoke causes serious illness among Chinese smokers: Smoking cessation and intention to quit. *Tob Prev Cessat* 2018;**4**. doi:10.18332/TPC/82813

429 Song A V., Glantz SA, Halpern-Felsher BL. Perceptions of Second-hand Smoke Risks Predict Future Adolescent Smoking Initiation. *J Adolesc Health* 2009;**45**:618. doi:10.1016/J.JADOHEALTH.2009.04.022

430 Ziebarth NR. Lung cancer risk perception biases. *Prev Med (Baltim)* 2018;**110**:16–23. doi:10.1016/j.ypmed.2018.01.020

431 Mazières J, Pujol JL, Kalampalikis N, *et al.* Perception of lung cancer among the general population and comparison with other cancers. *J Thorac Oncol* 2015;**10**:420–5. doi:10.1097/JTO.0000000000000433

432 Quaife SL, Marlow LAV, McEwen A, *et al.* Attitudes towards lung cancer screening in socioeconomically deprived and heavy smoking communities: informing screening communication. *Health Expect* 2017;**20**:563–73. doi:10.1111/HEX.12481

433 Vernon SW, Vogel VG, Halabi S, *et al.* Factors associated with perceived risk of breast cancer among women attending a screening program. *Breast Cancer Res Treat* Published Online First: 1993. doi:10.1007/BF00666426

434 McCaffery K, Wardle J, Waller J. Knowledge, attitudes, and behavioral intentions in relation to the early detection of colorectal cancer in the United Kingdom. *Prev Med (Baltim)* Published Online First: 2003. doi:10.1016/S0091-7435(03)00016-1

435 Loos RJF. 15 years of genome-wide association studies and no signs of slowing down. *Nat Commun 2020 111* 2020;**11**:1–3. doi:10.1038/s41467-020-19653-5

436 Nicholls HL, John CR, Watson DS, *et al.* Reaching the End-Game for GWAS: Machine Learning Approaches for the Prioritization of Complex Disease Loci. *Front Genet* 2020;**11**:350. doi:10.3389/FGENE.2020.00350/BIBTEX

437 Gallagher MD, Chen-Plotkin AS. The Post-GWAS Era: From Association to Function. *Am J Hum Genet* 2018;**102**:717. doi:10.1016/J.AJHG.2018.04.002

438 Genetics for all. *Nat Genet 2019 514* 2019;**51**:579–579. doi:10.1038/s41588-019-0394-y

439 Tammemägi MC, Ten Haaf K, Toumazis I, *et al.* Development and Validation of a Multivariable Lung Cancer Risk Prediction Model That Includes Low-Dose Computed Tomography Screening Results: A Secondary Analysis of Data From the National Lung Screening Trial. *JAMA Netw Open* 2019;**2**:e190204–e190204. doi:10.1001/JAMANETWORKOPEN.2019.0204

440 Robbins H, Berg C, … LC-JJ of the, *et al.* Identification of candidates for longer lung cancer

screening intervals following a negative low-dose computed tomography result. *academic.oup.com*https://academic.oup.com/jnci/article-abstract/111/9/996/5445482 (accessed 30 Mar 2022).

441 Mazzone PJ, Lamb C, Rieger-Christ KM, *et al.* Early candidate nasal swab classifiers developed using machine learning and whole transcriptome sequencing may improve early lung cancer detection. *https://doi.org/101200/JCO20213915_suppl8551* 2021;**39**:8551–8551. doi:10.1200/JCO.2021.39.15_SUPPL.8551

442 Waldeck S, Mitschke J, Wiesemann S, *et al.* Early assessment of circulating tumor DNA after curative-intent resection predicts tumor recurrence in early-stage and locally advanced non-small-cell lung cancer. *Mol Oncol* 2022;**16**:527–37. doi:10.1002/1878-0261.13116

443 Langevin SM, Kratzke RA, Kelsey KT. Epigenetics of Lung Cancer. *Transl Res* 2015;**165**:74. doi:10.1016/J.TRSL.2014.03.001

444 Pashayan N, Reisel D, Widschwendter M. Integration of genetic and epigenetic markers for risk stratification: opportunities and challenges. *Per Med* 2016;**13**:93. doi:10.2217/PME.15.53

445 Rosen B, Waitzberg R, Israeli A. Israel’s rapid rollout of vaccinations for COVID-19. *Isr J Health Policy Res* 2021;**10**:1–14. doi:10.1186/S13584-021-00440-6/FIGURES/1

Appendices

8.1 Appendix 1 – Comorbidity Classification for Life-Gained Study Follow-up

Severe	**Other**	**Serious**	**Minor**
2nd primary lung cancer Acute coronary syndrome Bowel cancer Breast cancer Cerebrovascular disease Chronic kidney disease Coronary artery disease Coronary artery disease Diabetes type II Endometrial cancer Hepatocellular carcinoma Ischaemic heart disease Malignant melanoma Mouth cancer Myelodysplastic syndrome Myocardial infarction NSTEMI Prostate cancer Renal cell cancer Salivary gland cancer Thyroid cancer	Atrial Fibrillation COPD Hypertension Pneumonia	Abdominal aortic aneurysm Alzheimer’s disease Aortic aneurysm Corticobasal degeneration Dementia Emphysema Heart failure Idiopathic pulmonary fibrosis Interstitial lung disease Liver cirrhosis Pneumothorax Pulmonary embolism Severe frailty Small vessel cerebrovascular disease Thoracic aortic aneurysm Valvular heart disease	Acoustic neuroma Alcohol problem drinking Barrett's oesophagus Benign pleural effusion Bronchiectasis Cervical disc prolapse Cholecystitis Cholelithiasis Cognitive impairment Colitis Crohn's disease DVT Deep vein thrombosis Degenerative disc disease Diabetic foot ulcer Diabetic retinopathy Diverticular disease Diverticulitis Epilepsy Fatty liver Fibromyalgia Gallstones Gastric ulcer Gastro-oesophageal reflux disease Glaucoma Gout Graves' disease Hiatus hernia Housebound Hypercholesterolaemia Hyperlipidaemia Hyperthyroidism Iron deficiency anaemia Lower limb DVT Lower limb ischaemia Lumbar spondylosis Obstructive sleep apnoea Oesophagitis Opioid dependant Osteoarthritis Osteopenia Osteoporosis Peripheral vascular disease Polymyalgia rheumatica Primary hyperparathyroidism Psoriasis Pulmonary Embolus Reflux

8.2 Appendix 2 - SNP-level Data for PRSs Validated in Manchester Cohort

	Source Paper												Proxy SNPs (if applicable)					Validation				SNP-level Stats in Our Study													
	RS	Chr	Position	Locus	Gene	Ref Allele	Effect Allele	MAF/*RAF*	OR	95% OR	P	Alternative RS ID	RS	Position	MAF	R2	Correlated Alleles	RS used in validation	Effect Allele used in Validation	Effect Magnitude used in Validation: ln(OR)	SNP also appears in...	A Allele	B Allele	MAF	OR (B vs. A)	OR lower	OR upper	P	PC Adjusted Beta	Minor Allele	Minor Allele Match	Risk Allele Based on Beta	Risk Allele/Effect Direction Match	P<0.05	Notes
Jia	rs71658797	1	77967507	1p31.1			A	0.10	1.14		3.25E-11							rs71658797	A	0.128	Hung-35, Fritsche-14, Graff	T	A	0.132595	1.0729	0.84633	1.36013	0.76778	0.039568	A	Yes	A	Yes	No	
	rs13080835	3	189357199	3q28			G	0.51	1.06		0.00000125							rs13080835	G	0.057	Hung-35, Fritsche-14, Graff	G	T	0.498428	1.00669	0.85743	1.18193	0.98531	-0.00168	G	Yes	G	Yes	No	
	rs7705526	5	1285974	5p15.33			A	0.34	1.12		1.01E-18							rs7705526	A	0.117	Hung-35, Fritsche-14, Graff	C	A	0.344027	1.01008	0.85308	1.19597	0.45172	-0.07169	A	Yes	C	No	No	
	rs2853677	5	1287194	5p15.33			G	0.42	1.12		2.66E-18							rs2853677	G	0.111	Hung-128, Graff	G	A	0.452979	0.90644	0.77141	1.0651	0.61236	-0.04581	G	Yes	G	Yes	No	
	rs465498	5	1325803	5p15.33			A	0.58	1.15		2.68E-32							rs465498	A	0.141		A	G	0.391847	0.83908	0.71196	0.988903	0.04034	-0.18584	G	Yes	A	Yes	Yes	
	rs3115672	6	31727897	6p21.33			T	0.10	1.18		3.47E-18							rs3115672	T	0.166		C	T	0.138501	1.19111	0.94249	1.50531	0.09437	0.216733	T	Yes	T	Yes	No	
	rs6920364	6	167376466	6q27			C	0.46	1.07		1.29E-08		rs427824	167403918	0.5	1	G=G,C=T	rs427824	T	0.068	Hung-35	T	G	0.497524	0.94766	0.80714	1.11263	0.15241	-0.12873	G	NA	T	Yes	No	
	rs4236709	8	32410110	8p12			G	0.22	1.07		0.00000588							rs4236709	G	0.064	Hung-35, Fritsche-14, Graff, Dai	G	A	0.23445	0.98229	0.81274	1.18722	0.88974	0.014861	G	Yes	A	No	No	
	rs11780471	8	27344719	8p21.2			G	0.94	1.15		1.69E-08							rs11780471	G	0.141	Hung-35, Fritsche-19, Graff	G	A	0.0553898	1.06462	0.7489	1.51344	0.36496	0.178825	A	Yes	A	No	No	
	rs885518	9	21830157	9p21.3			G	0.10	1.09		0.00000213							rs885518	G	0.088	Hung-35, Fritsche-19, Graff	A	G	0.100562	1.39647	1.0633	1.83404	0.05472	0.286665	G	Yes	G	Yes	No	
	rs62560775	9	22052068	9p21.3			G	0.10	1.11		0.000000602							rs62560775	G	0.100	Fritsche-19, Graff	A	G	0.100047	1.00097	0.7661	1.30786	0.37769	0.130727	G	Yes	G	Yes	No	
	rs1056562	11	118125625	11q23.3			T	0.48	1.07		1.92E-08							rs1056562	T	0.066	Hung-35, Fritsche-19	T	C	0.475475	0.87993	0.74922	1.03343	0.22531	-0.11177	T	Yes	T	Yes	No	
	rs7953330	12	998819	12p13.33			G	0.69	1.09		6.1E-12							rs7953330	G	0.087	Hung-35, Fritsche-14, Graff	G	C	0.294417	0.73843	0.61933	0.880422	0.00055	-0.34006	C	Yes	G	Yes	Yes	
	rs11571833	13	32972626	13q13.1			T	0.01	1.60		6.12E-16							rs11571833	T	0.472	Hung-35, Fritsche-14	A	T	0.0123943	1.10548	0.5332	2.29197	0.24531	0.476934	T	Yes	T	Yes	No	
	rs66759488	15	47577451	15q21.1			A	0.36	1.07		2.83E-08							rs66759488	A	0.068	Hung-35, Shi, Graff	G	A	0.367899	1.32582	1.12143	1.56747	0.00271	0.274821	A	Yes	A	Yes	Yes	
	rs77468143	15	49376624	15q21.1			T	0.75	1.09		0.000000001							rs77468143	T	0.083	Hung-35, Fritsche-14, Dai	T	G	0.222111	0.82449	0.68005	0.999597	0.06449	-0.20643	G	Yes	T	Yes	No	
	rs55781567	15	78857986	15q25.1			G	0.37	1.30		3.1E-103							rs55781567	G	0.260	Hung-35, Fritsche-14, Graff	C	G	0.371797	1.24259	1.05177	1.46803	0.00106	0.306993	G	Yes	G	Yes	Yes	
	rs8042374	15	78908032	15q25.1			A	0.78	1.26		2.91E-60							rs8042374	A	0.232		A	G	0.211308	0.76644	0.62993	0.93252	0.00036	-0.39073	G	Yes	A	Yes	Yes	
	rs56113850	19	41353107	19q13.2			C	0.56	1.13		5.02E-19							rs56113850	C	0.123	Hung-35, Fritsche-14, Graff	T	C	0.420448	1.03358	0.87854	1.21598	0.35254	0.088223	T	Yes	C	Yes	No	
Shi	rs13314271	3	189357602		TP63		T		1.13									rs13314271	T	0.122		T	C	0.498054	1.00308	0.85436	1.17769	0.95605	-0.00503	T	NA	T	Yes	No	
	rs9258375	6	29752808		IFITM4P,HCG4		G		1.25									rs9258375	G	0.223		A	G	0.129278	1.4	1.09691	1.78683	0.01128	0.340031	G	NA	G	Yes	Yes	
	rs66759488	15	47577451		SQRDL,SEMA6D		A		1.07									rs66759488	A	0.068	Jia, Hung-35, Graff	G	A	0.367899	1.32582	1.12143	1.56747	0.00271	0.274821	A	NA	A	Yes	Yes	
	rs4887053	15	78712699		CRABP1,IREB2		C		1.25									rs4887053	C	0.223		A	C	0.193531	1.35246	1.10428	1.65643	1.88E-05	0.49397	A	NA	C	Yes	Yes	
	rs1051730	15	78894339		CHRNA3		A		1.31									rs1051730	A	0.270		G	A	0.367304	1.27028	1.07462	1.50156	0.0003	0.337659	A	NA	A	Yes	Yes	
	rs6495309	15	78915245		CHRNA3,CHRNB4		T		1.3									rs6495309	T	0.262		C	T	0.200889	0.79524	0.65123	0.971096	0.00088	-0.37233	T	NA	C	No	Yes	Incorrect risk allele reported in Shi
Fritsche (14 & 19)	rs71658797	1	77967507			T	A		1.14		3.00E-11							rs71658797	A	0.128	Jia, Hung-35, Graff	T	A	0.132595	1.0729	0.84633	1.36013	0.76778	0.039568	A	NA	A	Yes	No	
	rs13080835	3	189357199			T	G		1.12		7.00E-12							rs13080835	G	0.111	Jia, Hung-35, Graff	G	T	0.498428	1.00669	0.85743	1.18193	0.98531	-0.00168	G	NA	G	Yes	No	
	rs7705526	5	1285974			C	A		1.25		4.00E-35							rs7705526	A	0.222	Jia, Hung-35, Graff	C	A	0.344027	1.01008	0.85308	1.19597	0.45172	-0.07169	A	NA	C	No	No	
	rs380286	5	1320247			A	G		1.15		2.00E-32							rs380286	G	0.141	Graff	G	A	0.389351	0.83436	0.70782	0.983518	0.0321	-0.19414	A	NA	G	Yes	Yes	
	rs3999544	6	26285867			T	C		1.08		1.00E-10							rs3999544	C	0.076	Graff	C	T	0.47932	0.87522	0.74526	1.02785	0.01706	-0.21434	C	NA	C	Yes	Yes	
	rs3094604	6	31434111			A	G		1.25		4.00E-19							rs3094604	G	0.223		A	G	0.191348	1.19677	0.97457	1.46964	0.12205	0.179965	G	NA	G	Yes	No	
	rs4236709	8	32410110			A	G		1.13		1.00E-10							rs4236709	G	0.124	Jia, Hung-35, Graff, Dai	G	A	0.23445	0.98229	0.81274	1.18722	0.88974	0.014861	G	NA	A	No	No	
	rs11591710	10	105687632			A	C		1.16		6.00E-11							rs11591710	C	0.151	Hung-35, Graff	A	C	0.132841	0.9333	0.73711	1.18172	0.6147	-0.06745	C	NA	A	No	No	
	rs7953330	12	998819			C	G		1.16		7.00E-13							rs7953330	G	0.146	Jia, Hung-35, Graff	G	C	0.294417	0.73843	0.61933	0.880422	0.00055	-0.34006	C	NA	G	Yes	Yes	
	rs11571833	13	32972626			A	T		1.60		6.00E-16							rs11571833	T	0.472	Jia, Hung-35	A	T	0.0123943	1.10548	0.5332	2.29197	0.24531	0.476934	T	NA	T	Yes	No	
	rs77468143	15	49376624			G	T		1.17		2.00E-16							rs77468143	T	0.155	Jia, Hung-35, Dai	T	G	0.222111	0.82449	0.68005	0.999597	0.06449	-0.20643	G	NA	T	Yes	No	
	rs55781567	15	78857986			C	G		1.30		3.00E-103							rs55781567	G	0.260	Jia, Hung-35, Graff	C	G	0.371797	1.24259	1.05177	1.46803	0.00106	0.306993	G	NA	G	Yes	Yes	
	rs28408315	15	79185180			A	G		1.10		4.00E-12							rs28408315	G	0.092		G	A	0.250378	1.03662	0.86122	1.24773	0.74178	-0.03475	A	NA	G	Yes	No	
	rs56113850	19	41353107			T	C		1.13		5.00E-19							rs56113850	C	0.123	Jia, Hung-35, Graff	T	C	0.420448	1.03358	0.87854	1.21598	0.35254	0.088223	T	NA	C	Yes	No	
Fritsche (19)	rs1611182	6	29781049			G	T		1.10									rs1611182	T	0.095		T	G	0.466227	0.9287	0.79067	1.09083	0.46334	-0.06614	T	NA	T	Yes	No	
	rs11780471	8	27344719			A	G		1.15									rs11780471	G	0.141	Jia, Hung-35, Graff	G	A	0.0553898	1.06462	0.7489	1.51344	0.36496	0.178825	A	NA	A	No	No	
	rs885518	9	21830157			A	G		1.17									rs885518	G	0.155	Jia, Hung-35, Graff	A	G	0.100562	1.39647	1.0633	1.83404	0.05472	0.286665	G	NA	G	Yes	No	
	rs62560775	9	22052068			A	G		1.18									rs62560775	G	0.166	Jia, Graff	A	G	0.100047	1.00097	0.7661	1.30786	0.37769	0.130727	G	NA	G	Yes	No	
	rs1056562	11	118125625			C	T		1.11									rs1056562	T	0.102	Jia, Hung-35	T	C	0.475475	0.87993	0.74922	1.03343	0.22531	-0.11177	T	NA	T	Yes	No	
Hung (35 & 128)	rs71658797	1	77967507	p31.1	AK5	T	A	*0.1*	1.14	1.09-1.18	3.25E-11							rs71658797	A	0.131	Jia, Fritsche-14, Graff	T	A	0.132595	1.0729	0.84633	1.36013	0.76778	0.039568	A	Yes	A	Yes	No	
	rs13080835	3	189357199	q28	TP63	G	T	*0.5*	0.94	0.92-0.97	1.25E-06							rs13080835	T	-0.062	Jia, Fritsche-14, Graff	G	T	0.498428	1.00669	0.85743	1.18193	0.98531	-0.00168	G	NA	G	Yes	No	
	rs7705526	5	1285974	p15.33	TERT	C	A	*0.34*	1.12	1.10-1.15	1.01E-18							rs7705526	A	0.113	Jia, Fritsche-14, Graff	C	A	0.344027	1.01008	0.85308	1.19597	0.45172	-0.07169	A	Yes	C	No	No	
	rs112290073	5	1286032	p15.33	TERT	G	A	*0.01*	1.39	1.20-1.61	1.16E-05							rs112290073	A	0.329		G	A	0.0130894	1.21545	0.59464	2.48438	0.87621	-0.06414	A	Yes	G	No	No	
	rs2736098	5	1294086	p15.33	TERT	C	T	*0.28*	1.14	1.10-1.18	4.36E-13							rs2736098	T	0.131		C	T	0.300064	1.0164	0.85311	1.21095	0.87806	0.015062	T	Yes	T	Yes	No	
	rs2853668	5	1300025	p15.33	TERT	G	T	*0.24*	1.08	1.05-1.11	9.36E-09							rs2853668	T	0.077		G	T	0.245629	0.99814	0.8284	1.20266	0.77875	-0.02819	T	Yes	G	No	No	
	rs401681	5	1322087	p15.33	CLPTM1L	C	T	*0.43*	0.87	0.85-0.89	3.25E-30							rs401681	T	-0.139		C	T	0.390184	0.82812	0.70256	0.976107	0.02791	-0.19884	T	Yes	C	Yes	Yes	
	rs466502	5	1325767	p15.33	CLPTM1L	A	G	*0.44*	0.91	0.89-0.93	1.99E-15							rs466502	G	-0.094		A	G	0.39184	0.83904	0.71192	0.988851	0.0403	-0.18587	G	Yes	A	Yes	Yes	
	rs6903823	6	28354519	p22.1	ZKSCAN3	A	G	*0.2*	1.07	1.04-1.10	1.09E-05							rs6903823	G	0.068		A	G	0.251974	1.18799	0.98656	1.43055	0.03763	0.213907	G	Yes	G	Yes	Yes	
	rs116822326	6	31434111	p21.33		A	G	*0.16*	1.15	1.12-1.19	5.29E-19	rs3094604						rs3094604	G	0.140		A	G	0.191348	1.19677	0.97457	1.46964	0.12205	0.179965	G	Yes	G	Yes	No	
	rs2855812	6	31504943	p21.33	MICB	G	T	*0.22*	1.05	1.03-1.08	1.26E-04							rs2855812	T	0.049		G	T	0.278702	1.2391	1.03488	1.48362	0.00204	0.316512	T	Yes	T	Yes	Yes	
	rs805262	6	31628733	p21.33	C6orf47	C	T	*0.47*	1.07	1.04-1.09	1.50E-08							rs805262	T	0.068		C	T	0.495424	0.97987	0.83458	1.15045	0.55906	-0.05293	T	Yes	C	No	No	
	rs6916278	6	31678774	p21.33	F-LY6G6D	G	A	*0.05*	0.9	0.86-0.95	1.15E-04							rs6916278	A	-0.105		G	A	0.0700183	0.85065	0.62178	1.16376	0.56179	-0.10095	A	Yes	G	Yes	No	
	rs3129763	6	32590925	p21.32		G	A	*0.22*	1.12	1.09-1.15	8.48E-16							rs3129763	A	0.113		G	A	0.235371	1.11893	0.92553	1.35274	0.21614	0.131788	A	Yes	A	Yes	No	
	rs114544105	6	32667852	p21.32	HLA-DQB1	G	A	*0.2*	1.06	1.02-1.09	9.14E-04		rs9274623	32635998	0.2	0.8568	G=G,A=T	rs9274623	T	0.058		G	T	0.269127	1.10857	0.92467	1.32904	0.10526	0.163833	T	Yes	T	Yes	No	
	rs6920364	6	167376466	q27		G	C	*0.46*	1.07	1.05-1.10	1.29E-08		rs427824	167403918	0.5	1	G=G,C=T	rs427824	T	0.068	Jia	T	G	0.497524	0.94766	0.80714	1.11263	0.15241	-0.12873	G	Yes	T	Yes	No	
	rs11780471	8	27344719	p21.2	EPHX2	G	A	*0.06*	0.87	0.83-0.91	1.69E-08							rs11780471	A	-0.139	Jia, Fritsche-19, Graff	G	A	0.0553898	1.06462	0.7489	1.51344	0.36496	0.178825	A	Yes	A	No	No	
	rs4236709	8	32410110	p12	NRG1	A	G	*0.22*	1.07	1.04-1.10	5.88E-06							rs4236709	G	0.068	Jia, Fritsche-14, Graff, Dai	G	A	0.23445	0.98229	0.81274	1.18722	0.88974	0.014861	G	Yes	A	No	No	
	rs885518	9	21830157	p21.3	MTAP	A	G	*0.1*	1.09	1.05-1.13	2.13E-0							rs885518	G	0.086	Jia, Fritsche-19, Graff	A	G	0.100562	1.39647	1.0633	1.83404	0.05472	0.286665	G	Yes	G	Yes	No	
	rs2007153	9	136503819	q34.2	DBH	C	T	*0.37*	0.96	0.93-0.98	2.49E-04							rs2007153	T	-0.041		T	C	0.384224	1.17867	0.99952	1.38992	0.09847	0.150052	T	Yes	C	Yes	No	
	rs11591710	10	105687632	q24.33		A	C	*0.14*	1.07	1.04-1.11	3.53E-05							rs11591710	C	0.068	Fritsche-14, Graff	A	C	0.132841	0.9333	0.73711	1.18172	0.6147	-0.06745	C	Yes	A	No	No	
	rs1056562	11	118125625	q23.3	MPZL2	C	T	*0.48*	1.07	1.04-1.09	1.92E-08							rs1056562	T	0.068	Jia, Fritsche-19	T	C	0.475475	0.87993	0.74922	1.03343	0.22531	-0.11177	T	Yes	T	Yes	No	
	rs7953330	12	998819	p13.33	WNK1	G	C	*0.31*	0.92	0.89-0.94	6.10E-12							rs7953330	C	-0.083	Jia, Fritsche-14, Graff	G	C	0.294417	0.73843	0.61933	0.880422	0.00055	-0.34006	C	Yes	G	Yes	Yes	
	rs11571833	13	32972626	q13.1	BRCA2	A	T	*0.01*	1.6	1.43-1.80	6.12E-16							rs11571833	T	0.470	Jia, Fritsche-14	A	T	0.0123943	1.10548	0.5332	2.29197	0.24531	0.476934	T	Yes	T	Yes	No	
	rs689647	15	43762196	q15.3	TP53BP1	C	T	*0.11*	0.93	0.90-0.97	2.11E-04							rs689647	T	-0.073		C	T	0.107737	1.04482	0.80624	1.354	0.88003	-0.02179	T	Yes	C	Yes	No	
	rs66759488	15	47577451	q21.1	SEMA6D	G	A	*0.36*	1.07	1.04-1.10	2.83E-08							rs66759488	A	0.068	Jia, Shi, Graff	G	A	0.367899	1.32582	1.12143	1.56747	0.00271	0.274821	A	Yes	A	Yes	Yes	
	rs77468143	15	49376624	q21.1		T	G	*0.25*	0.92	0.90-0.95	1.00E-09							rs77468143	G	-0.083	Jia, Fritsche-14, Dai	T	G	0.222111	0.82449	0.68005	0.999597	0.06449	-0.20643	G	Yes	T	Yes	No	
	rs3885951	15	78825917	q25.1	HYKK	A	G	*0.11*	1.17	.12-1.23b	4.09E-10							rs3885951	G	0.157	Graff	A	G	0.0944913	1.19085	0.90296	1.57053	0.04471	0.318058	G	Yes	G	Yes	Yes	
	rs55781567	15	78857986	q25.1	CHRNA5	C	G	*0.37*	1.3	1.27-1.33	3.08E-103							rs55781567	G	0.262	Jia, Fritsche-14, Graff	C	G	0.371797	1.24259	1.05177	1.46803	0.00106	0.306993	G	Yes	G	Yes	Yes	
	rs7177699	15	79089734	q25.1	ADAMTS7	T	C	*0.44*	1.13	1.11-1.16	5.98E-26							rs7177699	C	0.122		T	C	0.482525	1.03115	0.87817	1.21078	0.19027	0.118445	C	Yes	C	Yes	No	
	rs62070270	17	29936962	q11.2	EFCAB5	A	G	*0.45*	1.03	1.01-1.06	7.24E-03							rs62070270	G	0.030		A	G	0.421083	0.92289	0.7845	1.08569	0.0758	-0.16818	G	Yes	A	No	No	
	rs1542752	17	72938100	q25.1	OTOP3	C	T	*0.16*	1.04	1.01-1.07	1.24E-02							rs1542752	T	0.039		T	C	0.143511	0.83812	0.66547	1.05555	0.24014	-0.15316	T	Yes	T	Yes	No	
	rs56113850	19	41353107	q13.2	CYP2A6	C	T	*0.44*	0.88	0.86-0.91	5.02E-19							rs56113850	T	-0.128	Jia, Fritsche-14, Graff	T	C	0.420448	1.03358	0.87854	1.21598	0.35254	0.088223	T	Yes	C	Yes	No	
	rs41309931	20	62326579	q13.33	RTEL1	G	T	*0.12*	1.08	1.04-1.12	2.23E-05							rs41309931	T	0.077	Graff	G	T	0.132071	1.10521	0.87119	1.40209	0.59422	0.073295	T	Yes	T	Yes	No	
	rs17879961	22	29121087	q12.1	CHEK2	A	G	*0.01*	0.6	0.52-0.70	1.54E-10		rs186184919	29008888	0.005	1	A=C,G=T	rs186184919	T	-0.511		C	T	0.000767471	0.00876	5.43E-09	14131	0.4486	-1.35003	T	Yes	C	Yes	No	
	rs71641333	1	78743005	p31.1	MGC27382	T	A	*0.06*	1.14	1.09,1.21	4.49E-07							rs71641333	A	0.131		T	A	0.0797967	1.18444	0.87857	1.5968	0.33894	0.165559	A	Yes	A	Yes	No	
	rs78062588	1	154566225	q21.3	ADAR	T	C	*0.06*	0.88	0.84,0.93	4.60E-07							rs78062588	C	-0.128		T	C	0.0703387	0.88603	0.648	1.21149	0.23003	-0.21351	C	Yes	T	Yes	No	
	rs114737056	1	168511081	q24.2	XCL2	G	A	*0.12*	0.91	0.88,0.94	5.80E-07							rs114737056	A	-0.094		G	A	0.0953723	1.0797	0.82089	1.42009	0.5248	0.099366	A	Yes	A	No	No	
	rs145733018	2	38567201	p22.2	ATL2	T	C	*0.02*	2.2	1.61,3.00	7.24E-07							rs145733018	C	0.788		T	C	0.0271128	1.12514	0.68436	1.84982	0.87155	0.050153	C	Yes	C	Yes	No	
	rs79368540	2	45189737	p21		C	T	*0.15*	1.09	1.06,1.13	6.13E-07							rs79368540	T	0.086		C	T	0.11936	0.92856	0.72533	1.18874	0.43639	-0.10937	T	Yes	C	No	No	
	rs11692700	2	67510377	p14	LINC01828	T	C	*0.03*	1.2	1.12,1.29	6.44E-07							rs11692700	C	0.182		T	C	0.0333711	0.89888	0.57577	1.40331	0.35187	0.250838	C	Yes	C	Yes	No	
	rs114928225	2	119449740	q14.2		T	A	*0.01*	1.65	1.35,2.01	7.24E-07		rs74787667	119362387	0.006	0.86	T=A,A=G	rs74787667	G	0.501		A	G	0.00693053	0.96151	0.36603	2.52577	0.80322	0.138688	G	Yes	G	Yes	No	
	rs7592999	2	140398327	q22.1		T	C	*0.04*	0.81	0.74,0.88	7.35E-07							rs7592999	C	-0.211		T	C	0.0301676	0.95623	0.59862	1.52748	0.18251	-0.35768	C	Yes	T	Yes	No	
	rs722864	2	173983204	q31.1	MAP3K20	G	A	*0.19*	0.93	0.90,0.96	5.53E-07							rs722864	A	-0.073		A	G	0.188466	0.86909	0.70715	1.06812	0.36186	-0.107	A	Yes	A	No	No	
	rs1866631	2	174075761	q31.1	MAP3K20	A	G	*0.4*	0.94	0.92,0.96	6.97E-07							rs1866631	G	-0.062		G	A	0.40789	0.98356	0.83538	1.15803	0.85269	-0.01707	G	Yes	G	No	No	
	rs185666783	4	67833774	q13.2	05377262	C	G	*0.29*	0.92	0.90,0.95	9.20E-08		rs2318540	67831581	0.45	1	C=T,G=G	rs2318540	G	-0.083		T	G	0.436954	0.89487	0.76108	1.05217	0.36126	-0.08646	T	No	T	Yes	No	
	rs7676823	4	164007992	q32.2		A	G	*0.34*	0.92	0.89,0.95	6.71E-07							rs7676823	G	-0.083		G	A	0.366063	0.90985	0.77007	1.07501	0.57624	-0.05178	G	Yes	G	No	No	
	rs78154696	5	1000156	p15.33		G	A	*0.03*	1.22	1.13,1.32	7.55E-07							rs78154696	A	0.199		G	A	0.0190886	1.20814	0.66743	2.18689	0.57874	0.177379	A	Yes	A	Yes	No	
	rs112333466	5	1249816	p15.33	TERT	C	T	*0.01*	1.55	1.32,1.81	8.11E-08							rs112333466	T	0.438		C	T	0.0127346	0.87603	0.42932	1.78755	0.55973	-0.22698	T	Yes	C	No	No	
	rs56345976	5	1276873	p15.33	TERT	A	G	*0.42*	1.1	1.07,1.13	2.60E-12							rs56345976	G	0.095	Graff	G	A	0.386287	1.06546	0.90365	1.25625	0.52741	0.056954	G	Yes	A	No	No	
	rs2853677	5	1287194	p15.33	TERT	A	G	*0.42*	1.12	1.09,1.15	2.66E-18							rs2853677	G	0.113	Jia, Graff	G	A	0.452979	0.90644	0.77141	1.0651	0.61236	-0.04581	G	Yes	G	Yes	No	
	rs112401627	5	1300269	p15.33		G	A	*0.03*	1.3	1.20,1.42	3.03E-10							rs112401627	A	0.262	Graff	G	A	0.0249509	1.10843	0.66083	1.85921	0.66324	-0.12357	A	Yes	G	No	No	
	rs6875416	5	90250631	q14.3	ADGRV1	A	T	*0.21*	0.83	0.78,0.90	6.99E-07							rs6875416	T	-0.186		A	T	0.189233	1.00235	0.81667	1.23024	0.74864	0.037183	T	Yes	T	No	No	
	rs114136906	5	150121458	q33.1	DCTN4	G	C	*0.02*	1.46	1.27,1.68	1.17E-07							rs114136906	C	0.378		G	C	0.0197113	1.4812	0.81527	2.69105	0.23539	0.384178	C	Yes	C	Yes	No	
	rs2316515	6	410848	p25.3	IRF4	G	A	*0.41*	0.92	0.89,0.95	1.42E-07							rs2316515	A	-0.083		A	G	0.376223	0.91044	0.77129	1.07468	0.39869	-0.07939	A	Yes	A	No	No	
	rs629444	6	25885814	p22.2	T1H2APS2	C	T	*0.1*	1.11	1.07,1.15	1.40E-08							rs629444	T	0.104		C	T	0.131829	1.45086	1.13822	1.84938	0.05695	0.25616	T	Yes	T	Yes	No	
	rs2179517	6	26198845	p22.2	HIST1H3D	G	C	*0.49*	0.94	0.92,0.96	3.99E-08							rs2179517	C	-0.062		G	C	0.446392	0.93379	0.79462	1.09732	0.0697	-0.16298	C	Yes	G	Yes	No	
	rs68141011	6	28217797	p22.1	ZKSCAN4	G	T	*0.13*	1.09	1.06,1.13	1.43E-07							rs68141011	T	0.086		G	T	0.160954	1.27011	1.01848	1.5839	0.13912	0.180917	T	Yes	T	Yes	No	
	rs114722608	6	29223493	p22.1	01929006	G	C	*0.09*	1.15	1.11,1.20	7.09E-12	rs3130825						rs3130825	C	0.140		G	C	0.121441	1.50925	1.17245	1.9428	0.00416	0.386218	C	Yes	C	Yes	Yes	
	rs115123779	6	29477821	p22.1	05375009	G	T	*0.18*	1.1	1.07,1.13	1.91E-09	rs1233478						rs1233478	T	0.095		G	T	0.196775	1.37146	1.11745	1.6832	0.00347	0.331315	T	Yes	T	Yes	Yes	

Hung (128)	rs138488080	6	29606761	p22.1	SUMO2P1	G	A	*0.15*	1.15	1.11,1.19	5.96E-18	rs3095268						rs3095268	A	0.140		G	A	0.192588	1.23874	1.00888	1.52096	0.08228	0.194992	A	Yes	A	Yes	No	
	rs114192654	6	29759750	p22.1		G	A	*0.34*	1.06	1.04,1.09	5.69E-07	rs1611203						rs1611203	A	0.058		G	A	0.340373	1.1331	0.95621	1.34271	0.1095	0.154377	A	Yes	A	Yes	No	
	rs116675020	6	29922740	p22.1	HLA-W	A	G	*0.38*	0.94	0.91,0.96	6.72E-07	rs16896742						rs16896742	G	-0.062		A	G	0.367725	0.92712	0.78507	1.09486	0.12162	-0.14333	G	Yes	A	Yes	No	
	rs115993819	6	30074163	p22.1	TRIM31	G	A	*0.24*	1.09	1.06,1.12	9.64E-10	rs2517600						rs2517600	A	0.086	Graff	G	A	0.234955	1.17317	0.96995	1.41897	0.25046	0.122419	A	Yes	A	Yes	No	
	rs116534499	6	30138162	p22.1	TRIM15	C	G	*0.43*	1.07	1.04,1.09	2.94E-08	rs1029239						rs1029239	G	0.068		C	G	0.472347	1.04377	0.88878	1.2258	0.21317	0.115315	G	Yes	G	Yes	No	
	rs116629156	6	30864829	p21.33	DDR1	T	C	*0.41*	1.07	1.05,1.10	1.52E-08	rs1049623						rs1049623	C	0.068		T	C	0.386855	1.01812	0.86344	1.2005	0.84131	0.018472	C	Yes	C	Yes	No	
	rs114103504	6	31002452	p21.33	MUC22	A	G	*0.49*	0.93	0.91,0.95	3.29E-09	rs3869098						rs3869098	G	-0.073	Graff	A	G	0.458067	0.97574	0.83061	1.14623	0.937	-0.00742	G	Yes	A	Yes	No	
	rs114052224	6	31067852	p21.33		A	G	*0.48*	1.06	1.04,1.09	1.26E-07	rs3130548						rs3130548	G	0.058		A	G	0.494275	0.98731	0.84091	1.15918	0.69834	-0.03555	G	Yes	A	No	No	
	rs2233959	6	31081065	p21.33	C6of15	T	C	*0.41*	1.07	1.05,1.10	5.14E-09							rs2233959	C	0.068		T	C	0.427701	1.02847	0.87447	1.2096	0.98378	0.001909	C	Yes	C	Yes	No	
	rs114689412	6	31117577	p21.33	CCHCR1	C	G	*0.13*	0.91	0.88,0.94	3.64E-07	rs1265113						rs1265113	G	-0.094		C	G	0.13878	0.9052	0.71808	1.14108	0.89206	-0.01764	G	Yes	C	Yes	No	
	rs2596499	6	31321429	p21.33	HLA-B	T	A	*0.3*	1.07	1.04,1.10	3.42E-07							rs2596499	A	0.068		T	A	0.333519	1.08602	0.91583	1.28783	0.48469	0.066408	A	Yes	A	Yes	No	
	rs2596496	6	31322782	p21.33	HLA-B	G	C	*0.36*	0.9	0.87,0.94	3.64E-07							rs2596496	C	-0.105		C	G	0.331486	1.09712	0.92534	1.3008	0.48742	0.067127	C	Yes	G	Yes	No	
	rs2596490	6	31324996	p21.33		C	G	*0.23*	0.87	0.83,0.92	2.82E-08							rs2596490	G	-0.139		G	C	0.200004	0.9991	0.8175	1.22104	0.87193	-0.0186	G	Yes	G	No	No	
	rs115176861	6	31412961	p21.33	HCP5	T	C	*0.48*	1.06	1.04,1.09	1.11E-07	rs2596464						rs2596464	C	0.058		T	C	0.461559	1.00954	0.85945	1.18583	0.96778	-0.00371	T	No	T	No	No	
	rs553108	6	31840455	p21.33	SLC44A4	G	A	*0.38*	1.07	1.04,1.09	6.63E-08							rs553108	A	0.068		A	G	0.420906	0.88328	0.75062	1.0394	0.04802	-0.18445	A	Yes	A	Yes	Yes	
	rs115200960	6	32335204	p21.32	C6of10	G	A	*0.17*	1.11	1.07,1.14	4.77E-11	rs2143462						rs2143462	A	0.104		G	A	0.20591	0.98473	0.80755	1.20078	0.70967	-0.0409	A	Yes	G	No	No	
	rs12722051	6	32609147	p21.32	HLA-DQA1	A	T	*0.18*	0.87	0.83,0.92	4.96E-07		rs2760995	32574358	0.21	0.7655	A=G,T=A	rs2760995	A	-0.139		G	A	0.193548	0.93112	0.76021	1.14044	0.66488	-0.04836	A	Yes	G	Yes	No	
	rs116767258	6	32757737	p21.32		A	G	*0.39*	0.94	0.92,0.96	9.10E-07	rs2857165						rs2857165	G	-0.062		G	A	0.381655	0.93558	0.79303	1.10376	0.31177	-0.09404	G	Yes	G	No	No	
	rs7383287	6	32783086	p21.32	HLA-DOB	A	G	*0.2*	1.1	1.07,1.13	1.12E-10							rs7383287	G	0.095	Graff	A	G	0.210484	0.94641	0.77749	1.15203	0.88216	-0.0165	G	Yes	A	No	No	
	rs117534741	6	72384541	q13		G	A	*0.02*	1.24	1.14,1.34	4.68E-07							rs117534741	A	0.215		G	A	0.021391	0.62865	0.35978	1.09847	0.09123	-0.56528	A	Yes	G	No	No	
	rs1321817	6	117734267	q22.1	ROS1	A	G	*0.37*	0.92	0.89,0.95	5.67E-07							rs1321817	G	-0.083		A	G	0.336818	0.82088	0.69283	0.972588	0.24896	-0.10795	G	Yes	A	Yes	No	
	rs6957511	7	130668618	q32.3	LINC-PINT	T	C	*0.4*	1.1	1.06,1.14	9.78E-07							rs6957511	C	0.095		C	T	0.381741	1.00109	0.84868	1.18087	0.8458	-0.01861	C	Yes	C	Yes	No	
	rs2565064	8	27327841	p21.2	CHRNA2	G	C	*0.29*	1.07	1.04,1.10	4.58E-07							rs2565064	C	0.068		C	G	0.318401	1.0108	0.85089	1.20076	0.71828	0.035317	C	Yes	G	No	No	
	rs67749759	8	27397087	p21.2	EPHX2	C	T	*0.07*	1.13	1.08,1.19	2.66E-07	rs4149254						rs4149254	T	0.122		C	T	0.0790957	0.81329	0.6047	1.09382	0.05211	-0.33894	T	Yes	C	No	No	
	rs111960002	8	144722420	q24.3	ZNF623	T	C	*0.05*	1.36	1.21,1.54	4.62E-07							rs111960002	C	0.307		T	C	0.039265	1.38294	0.90573	2.11159	0.21869	0.335211	C	Yes	C	Yes	No	
	rs10118776	9	6227418	p24.1	IL33	A	G	*0.06*	1.34	1.20,1.50	3.91E-07							rs10118776	G	0.293		G	A	0.0752974	1.10573	0.81646	1.49749	0.79448	0.044812	G	Yes	A	No	No	
	rs17185553	9	17934120	p22.2		G	C	*0.08*	1.29	1.17,1.43	8.94E-07							rs17185553	C	0.255		G	C	0.0672023	0.9875	0.71683	1.36038	0.76402	-0.05493	C	Yes	G	No	No	
	rs2518717	9	21959751	p21.3	1-145E5.5	T	C	*0.36*	1.09	1.06,1.13	3.35E-07		rs1985742	21961227	0.35	0.9913	T=A,C=T	rs1985742	T	0.086		A	T	0.360021	0.92893	0.78602	1.09782	0.59061	-0.05073	T	Yes	A	No	No	
	rs28557075	9	22066572	p21.3	KN2B-AS1	G	A	*0.09*	1.11	1.06,1.16	8.43E-07							rs28557075	A	0.104		G	A	0.0815304	0.84954	0.63423	1.13793	0.80103	-0.04035	A	Yes	G	No	No	
	rs1333040	9	22083404	p21.3	KN2B-AS1	T	C	*0.46*	1.1	1.06,1.14	7.02E-07							rs1333040	C	0.095	Graff	C	T	0.40567	1.1247	0.95521	1.32426	0.10078	0.148051	C	Yes	T	No	No	
	rs4879704	9	33427322	p13.3		A	C	*0.33*	0.92	0.89,0.95	9.16E-07							rs4879704	C	-0.083		A	C	0.296045	0.89289	0.74915	1.06422	0.11583	-0.15826	C	Yes	A	Yes	No	
	rs191205566	9	102587233	q22.33	NR4A3	C	T	*0.02*	1.4	1.24,1.59	1.17E-07							rs191205566	T	0.336		C	T	0.016834	0.62336	0.33262	1.16825	0.22648	-0.45564	T	Yes	C	No	No	
	rs75685923	9	136275229	q34.2	REXO4	C	T	*0.03*	1.38	1.22,1.57	6.16E-07							rs75685923	T	0.322		C	T	0.0508872	0.90087	0.62596	1.29653	0.85913	-0.03769	T	Yes	C	No	No	
	rs7897454	10	102011702	q24.31	CWF19L1	G	A	*0.04*	1.25	1.14,1.36	6.64E-07							rs7897454	A	0.223		G	A	0.0549547	1.05931	0.74426	1.50772	0.75682	0.061028	A	Yes	A	Yes	No	
	rs62621207	10	102672248	q24.31	SLF2	A	T	*0.05*	1.16	1.09,1.23	5.85E-07							rs62621207	T	0.148		A	T	0.0462662	1.30846	0.88683	1.93056	0.2183	0.281661	T	Yes	T	Yes	No	
	rs78853063	11	57250026	q12.1	SLC43A1	C	T	*0.08*	0.89	0.85,0.93	4.65E-07							rs78853063	T	-0.117		C	T	0.0793295	1.25614	0.92978	1.69704	0.20315	0.210517	T	Yes	T	No	No	
	rs78334599	11	115998756	q23.3		G	A	*0.04*	0.86	0.80,0.91	7.93E-07							rs78334599	A	-0.151		G	A	0.0241094	1.10806	0.65487	1.87485	0.68881	-0.11684	A	Yes	G	Yes	No	
	rs7487683	12	1036042	p13.33	RAD52	C	T	*0.04*	0.83	0.77,0.89	6.58E-08							rs7487683	T	-0.186		C	T	0.0328041	0.74379	0.47443	1.16607	0.36934	-0.22804	T	Yes	C	Yes	No	
	rs73351723	12	58831070	q14.1		G	A	*0.13*	1.09	1.06,1.13	3.81E-07							rs73351723	A	0.086		G	A	0.119777	1.25609	0.97819	1.61292	0.06102	0.258842	A	Yes	A	Yes	No	
	rs9668978	12	64913237	q14.2	-439H13.2	G	T	*0.29*	1.1	1.06,1.14	6.24E-07							rs9668978	T	0.095		G	T	0.278062	0.96981	0.81087	1.15991	0.32697	-0.1035	T	Yes	G	No	No	
	rs9602270	13	84281063	q31.1		A	T	*0.05*	1.27	1.16,1.39	3.28E-07							rs9602270	T	0.239		A	T	0.0335012	1.08229	0.69174	1.69335	0.61303	0.132687	T	Yes	T	Yes	No	
	rs8003466	14	34013721	q13.1	NPAS3	G	A	*0.18*	0.89	0.84,0.93	5.27E-07							rs8003466	A	-0.117		G	A	0.187852	0.85445	0.69614	1.04878	0.32128	-0.11276	A	Yes	G	Yes	No	
	rs8031813	15	49253961	q21.1	SHC4	A	C	*0.31*	0.91	0.87,0.94	4.29E-08							rs8031813	C	-0.094	Graff	C	A	0.310115	1.1813	0.9934	1.40473	0.02871	0.225341	C	Yes	A	Yes	Yes	
	rs6493361	15	49615952	q21.2	GALK2	C	G	*0.34*	1.09	1.05,1.12	7.09E-07							rs6493361	G	0.086		G	C	0.354089	0.86012	0.72693	1.01772	0.1868	-0.12645	G	Yes	G	Yes	No	
	rs11855650	15	70431773	q23		G	T	*0.38*	1.09	1.05,1.12	5.60E-07							rs11855650	T	0.086		G	T	0.392103	0.97314	0.82568	1.14692	0.71759	0.032526	T	Yes	T	Yes	No	
	rs79149102	15	75055819	q24.1		C	T	*0.03*	1.18	1.11,1.25	1.54E-07							rs79149102	T	0.166		C	T	0.0328552	1.03481	0.65929	1.6242	0.83352	0.055339	T	Yes	T	Yes	No	
	rs2229961	15	78880752	q25.1	CHRNA5	G	A	*0.02*	1.43	1.30,1.57	5.01E-14							rs2229961	A	0.358	Graff	G	A	0.021661	1.3566	0.77176	2.38464	0.24599	0.35796	A	Yes	A	Yes	No	
	rs8192479	15	78909398	q25.1	CHRNA3	C	T	*0.02*	1.28	1.17,1.40	2.29E-08							rs8192479	T	0.247	Graff	C	T	0.0205125	1.19374	0.67357	2.11563	0.2036	0.436389	T	Yes	T	Yes	No	
	rs2869551	15	78981423	q25.1	CHRNB4	A	G	*0.01*	0.75	0.67,0.84	5.09E-07							rs2869551	G	-0.288		A	G	0.00723877	1.08225	0.41839	2.79946	0.91503	-0.06283	G	Yes	A	Yes	No	
	rs12593207	15	78987225	q25.1	CHRNB4	G	A	*0.1*	0.85	0.82,0.89	2.88E-14							rs12593207	A	-0.163		G	A	0.0840969	1.09888	0.82198	1.46905	0.81298	-0.03842	A	Yes	G	Yes	No	
	rs189146505	15	79058730	q25.1	ADAMTS7	A	G	*0.04*	0.83	0.78,0.89	8.34E-08		rs138833245	79045868	0.04	0.8277	A=T,G=C	rs138833245	C	-0.186		T	C	0.0423519	1.16702	0.78082	1.74425	0.89756	0.029448	C	No	C	No	No	
	rs28450923	15	79065557	q25.1	ADAMTS7	A	G	*0.13*	0.88	0.84,0.92	7.76E-08		NA	NA	NA	NA	NA				Graff														No proxy SNP >0.5 R2
	rs28624856	15	79075233	q25.1	ADAMTS7	T	C	*0.24*	0.91	0.88,0.94	1.06E-09		rs2046144	79074562	0.25	0.823	C=A,T=G	rs2046144	A	-0.094		A	G	0.244456	1.04973	0.8711	1.265	0.29156	0.11562	A	Yes	G	Yes	No	
	rs77719127	15	79110783	q25.1	MORF4L1	C	T	*0.17*	1.14	1.10,1.18	1.79E-13							rs77719127	T	0.131	Graff	C	T	0.185129	0.94242	0.76673	1.15836	0.5011	-0.0767	T	Yes	C	No	No	
	rs76164573	15	79198760	q25.1		T	G	*0.04*	0.86	0.81,0.91	8.83E-07							rs76164573	G	-0.151		T	G	0.0536822	1.00087	0.70107	1.42887	0.37517	-0.18269	G	Yes	T	Yes	No	
	rs78442819	16	10740982	p13.13	TEKT5	G	C	*0.19*	0.89	0.85,0.93	6.48E-07							rs78442819	C	-0.117		G	C	0.199626	0.81449	0.66668	0.995055	0.00836	-0.33065	C	Yes	G	Yes	Yes	
	rs9926896	16	26980646	p12.1		T	C	*0.01*	2.7	1.90,3.83	3.06E-08		rs183161830	26746618	0.001	1	T=A,C=G	rs183161830	G	0.993		A	G	0.00110691	1.05237	0.09366	11.824	0.79266	-0.41375	G	Yes	A	No	No	
	rs17181550	17	70299958	q24.3		T	G	*0.43*	0.94	0.92,0.96	1.98E-07							rs17181550	G	-0.062		T	G	0.422128	0.8544	0.7263	1.00508	0.11404	-0.14584	G	Yes	T	Yes	No	
	rs79421398	18	20741135	q11.2	CABLES1	T	C	*0.05*	1.23	1.14,1.34	2.75E-07							rs79421398	C	0.207		T	C	0.0508811	1.22249	0.84483	1.76898	0.17988	0.296591	C	Yes	C	Yes	No	
	rs66500423	19	41195170	q13.2	NUMBL	T	C	*0.3*	1.07	1.04,1.09	2.23E-07							rs66500423	C	0.068		T	C	0.314675	0.99077	0.83357	1.17761	0.93224	-0.00811	C	Yes	T	No	No	
	rs4803356	19	41207206	q13.2	ADCK4	C	G	*0.07*	0.89	0.85,0.93	8.31E-07							rs4803356	G	-0.117		C	G	0.0724676	0.7159	0.5255	0.975272	0.02537	-0.3924	G	Yes	C	Yes	Yes	
	rs11881918	19	41334199	q13.2	490E21.12	G	A	*0.09*	0.86	0.82,0.91	4.10E-09							rs11881918	A	-0.151		G	A	0.0586096	1.10084	0.78107	1.55151	0.56929	0.11097	A	Yes	A	No	No	
	rs2258380	19	41338988	q13.2	490E21.12	C	G	*0.23*	1.08	1.05,1.11	9.15E-07							rs2258380	G	0.077		C	G	0.232318	1.11823	0.92416	1.35307	0.08192	0.187367	G	Yes	G	Yes	No	
	rs67210567	19	41357457	q13.2	CYP2A6	G	T	*0.03*	0.79	0.73,0.86	2.96E-08		rs1801272	41354533	0.03	0.9696	G=A,T=T	rs1801272	T	-0.236	Graff	A	T	0.0262501	0.51896	0.31053	0.867308	0.00264	-0.83473	T	Yes	A	Yes	Yes	
	rs184589612	19	41412192	q13.2	490E21.13	T	C	*0.02*	0.77	0.70,0.85	3.23E-07							rs184589612	C	-0.261		T	C	0.0182608	0.99282	0.54536	1.80742	0.9548	0.018957	C	Yes	C	No	No	
	rs12981718	19	54567858	q13.42	VSTM1	G	A	*0.07*	1.36	1.21,1.53	2.13E-07							rs12981718	A	0.307		G	A	0.0672276	1.05311	0.76386	1.4519	0.80183	-0.04785	A	Yes	G	No	No	
	rs13036436	20	61988382	q13.33	CHRNA4	A	G	*0.2*	1.08	1.05,1.12	9.03E-07		rs6011779	61984317	0.19	0.9237	G=C,A=T	rs6011779	C	0.077		C	T	0.20196	0.87991	0.71989	1.07551	0.22419	-0.13632	C	Yes	C	Yes	No	
	rs61541144	20	62527305	q13.33	DNAJC5	G	A	*0.07*	0.89	0.85,0.93	5.61E-07							rs61541144	A	-0.117		G	A	0.0773723	1.24133	0.9159	1.6824	0.05509	0.329426	A	Yes	A	No	No	
	rs71658797	1				T	A	0.10	1.14									rs71658797	A	0.131	Hung-35, Fritsche-14, Jia	T	A	0.132595	1.0729	0.84633	1.36013	0.76778	0.039568	A	Yes	A	Yes	No	
	rs13080835	3				T	G	0.51	1.12									rs13080835	G	0.111	Hung-35, Fritsche-14, Jia	G	T	0.498428	1.00669	0.85743	1.18193	0.98531	-0.00168	G	Yes	G	Yes	No	
	rs71603396	4				G	A	0.14	1.92				rs35622894	17929747	0.11	0.7735	G=T,A=C	rs35622894	C	0.652		T	C	0.149946	1.02182	0.81602	1.27952	0.64865	-0.05914	C	No	T	No	No	
	rs13156167	5				T	C	0.13	1.14				rs13167280	1280477	0.13	0.6088	T=G,C=A	rs13167280	A	0.132		G	A	0.130554	0.9374	0.73904	1.18901	0.61096	-0.06877	A	No	G	No	No	
	rs56345976	5				A	G	0.42	1.10									rs56345976	G	0.093	Hung-128	G	A	0.386287	1.06546	0.90365	1.25625	0.52741	0.056954	G	Yes	A	No	No	
	rs33961405	5				A	G	0.49	1.10									rs33961405	G	0.098		G	A	0.495485	0.9998	0.85156	1.17385	0.57738	0.051428	G	Yes	A	No	No	
	rs7705526	5				C	A	0.34	1.12									rs7705526	A	0.117	Hung-35, Fritsche-14, Jia	C	A	0.344027	1.01008	0.85308	1.19597	0.45172	-0.07169	A	Yes	C	No	No	
	rs2853677	5				A	G	0.42	1.12									rs2853677	G	0.111	Hung-128, Jia	G	A	0.452979	0.90644	0.77141	1.0651	0.61236	-0.04581	G	Yes	G	Yes	No	
	rs148297846	5				G	A	0.08	1.21									rs148297846	A	0.194		G	A	0.0900245	1.09844	0.82895	1.45556	0.6306	0.075284	A	Yes	A	Yes	No	
	rs62329683	5				C	T	0.16	1.19									rs62329683	T	0.178		C	T	0.145719	0.92099	0.73397	1.15566	0.94839	0.008163	T	Yes	T	Yes	No	
	rs112401627	5				G	A	0.03	1.30									rs112401627	A	0.265	Hung-128	G	A	0.0249509	1.10843	0.66083	1.85921	0.66324	-0.12357	A	Yes	G	No	No	
	rs4635969	5				A	G	0.81	1.16									rs4635969	G	0.145		G	A	0.170552	0.77686	0.62794	0.961088	0.09891	-0.19395	A	No	G	Yes	No	
	rs380286	5				A	G	0.58	1.15									rs380286	G	0.141	Fritsche-14	G	A	0.389351	0.83436	0.70782	0.983518	0.0321	-0.19414	A	No	G	Yes	Yes	
	rs410805	5				C	T	0.17	1.14									rs410805	T	0.133		T	C	0.18262	0.8557	0.69437	1.0545	0.32016	-0.11438	T	Yes	T	Yes	No	
	rs55901723	5				C	T	0.79	1.14			rs391314	NA	NA	NA	NA	NA																		No proxy SNP >0.5 R2
	rs40183	5				A	G	0.91	1.13									rs40183	G	0.125		G	A	0.0976285	1.08945	0.83058	1.42902	0.71711	0.054926	A	No	A	No	No	
	rs6889886	5				T	C	0.74	1.09									rs6889886	C	0.085		C	T	0.248282	0.90326	0.75038	1.08729	0.0746	-0.18668	T	No	C	Yes	No	
	rs806794	6				G	A	0.70	1.08									rs806794	A	0.079		A	G	0.26456	0.94166	0.78519	1.12932	0.59265	-0.05426	G	No	A	Yes	No	
	rs3999544	6				T	C	0.44	1.08									rs3999544	C	0.076	Fritsche-14	C	T	0.47932	0.87522	0.74526	1.02785	0.01706	-0.21434	C	Yes	C	Yes	Yes	
	rs9467711	6				A	C	0.10	1.19				NA	NA	NA	NA	NA																		No proxy SNP >0.5 R2
	rs6912292	6				C	A	0.78	1.08				rs1535275	26520767	0.25	1	C=G,A=C	rs1535275	C	0.080		G	C	0.204292	1.05074	0.86131	1.28184	0.1585	0.155821	G	No	C	Yes	No	
	rs116461399	6				G	C	0.10	1.15			rs3131337						rs3131337	C	0.144		G	C	0.131365	1.66171	1.29868	2.12621	0.00032	0.466923	C	Yes	C	Yes	Yes	
	rs116045192	6				C	T	0.19	1.14			rs3117426						rs3117426	T	0.134		C	T	0.215483	1.17304	0.96404	1.42735	0.11657	0.172669	T	Yes	T	Yes	No	
	rs114924489	6				A	G	0.25	1.08			rs1233487						rs1233487	G	0.076		A	G	0.290349	1.09722	0.9191	1.30985	0.38231	0.086261	G	Yes	G	Yes	No	
	rs114050064	6				T	C	0.28	1.08			rs1611350						rs1611350	C	0.076		T	C	0.302186	1.23478	1.0358	1.47198	0.04473	0.20086	C	Yes	C	Yes	Yes	
	rs115813114	6				T	C	0.27	1.09			rs1632999						rs1632999	C	0.085		T	C	0.284281	1.20983	1.01172	1.44674	0.0378	0.219134	C	Yes	C	Yes	Yes	
	rs111597677	6				A	C	0.26	1.12			rs1611713						rs1611713	C	0.110		A	C	0.286535	1.14201	0.95575	1.36456	0.10182	0.160676	C	Yes	C	Yes	No	
	rs115960997	6				G	A	0.31	1.09			rs1611314	NA	NA	NA	NA	NA																		No proxy SNP >0.5 R2
	rs116405142	6				A	G	0.27	1.08			rs916570						rs916570	G	0.080		A	G	0.281015	1.1435	0.95598	1.36781	0.0764	0.179391	G	Yes	G	Yes	No	
	rs115993819	6				G	A	0.24	1.09			rs2517600						rs2517600	A	0.088	Hung-128	G	A	0.234955	1.17317	0.96995	1.41897	0.25046	0.122419	A	Yes	A	Yes	No	
	rs142696802	6				C	T	0.28	1.08			rs2517602						rs2517602	T	0.081		C	T	0.309371	1.1731	0.98552	1.39639	0.01341	0.246223	T	Yes	T	Yes	Yes	
	rs116074010	6				A	T	0.25	1.08			rs3130127						rs3130127	T	0.081		A	T	0.247813	1.11564	0.92591	1.34426	0.28383	0.111912	T	Yes	T	Yes	No	
	rs115915654	6				C	T	0.24	1.08			rs1264372						rs1264372	T	0.078		C	T	0.266649	1.09262	0.91095	1.31052	0.46438	0.075665	T	Yes	T	Yes	No	
	rs116552815	6				G	A	0.41	1.07			rs1264323						rs1264323	A	0.068		G	A	0.386855	1.01812	0.86344	1.2005	0.84131	0.018472	A	Yes	A	Yes	No	
	rs114103504	6				G	A	0.51	1.07			rs3869098						rs3869098	A	0.070	Hung-128	A	G	0.458067	0.97574	0.83061	1.14623	0.937	-0.00742	G	No	A	Yes	No	
	rs115669859	6				G	A	0.36	1.13			rs2508015						rs2508015	A	0.121		G	A	0.398918	0.96439	0.81866	1.13605	0.36309	-0.08559	A	Yes	G	No	No	
	rs182364552	6				T	A	0.28	1.08			rs9263573	rs2233986	31079076	0.34	0.9912	T=G,A=A	rs2233986	A	0.080		G	A	0.329411	1.034	0.87164	1.2266	0.44176	0.075165	A	Yes	A	Yes	No	
	rs3130552	6				G	A	0.30	1.08									rs3130552	A	0.078		G	A	0.355241	0.99829	0.84419	1.1805	0.98271	-0.00202	A	Yes	G	No	No	
	rs116651383	6				T	C	0.19	1.11			rs3094191	rs3132514	31142758	0.15	0.7565	T=T,C=C	rs3132514	C	0.108		T	C	0.227033	0.97989	0.80915	1.18665	0.8213	-0.02442	C	Yes	T	No	No	
	rs115681412	6				C	G	0.34	1.12			rs1265158						rs1265158	G	0.113		C	G	0.376662	1.01816	0.86275	1.20157	0.91405	0.010179	G	Yes	G	Yes	No	
	rs141707415	6				T	A	0.39	1.10			rs9264669	rs2524119	31229404	0.38	0.8828	A=T,T=C	rs2524119	T	0.097		T	C	0.451106	0.90821	0.77287	1.06725	0.32741	-0.08668	T	No	T	Yes	No	
	rs111324505	6				C	T	0.45	1.07			rs7764413						rs7764413	T	0.069		T	C	0.484745	0.92184	0.78509	1.08241	0.97085	-0.00328	C	No	T	Yes	No	
	rs115494074	6				C	T	0.24	1.10			rs2394982	rs2507997	31314781	0.22	0.9718	C=A,T=G	rs2507997	G	0.095		A	G	0.311907	1.01241	0.85138	1.20391	0.66658	-0.04158	G	No	A	No	No	
	rs139850307	6				T	C	0.49	1.12			rs1050556	rs2523589	31327334	0.44	0.6526	C=G,T=T	rs2523589	G	0.109		G	T	0.480033	0.92281	0.78588	1.0836	0.54761	-0.05533	T	No	G	Yes	No	

Graff	rs2249935	6				G	A	0.20	1.08									rs2249935	A	0.081		G	A	0.236689	1.11144	0.91972	1.34312	0.29923	0.109884	A	Yes	A	Yes	No	
	rs115471405	6				T	C	0.29	1.08			rs2844571						rs2844571	C	0.072		T	C	0.328619	1.01903	0.85897	1.20892	0.84346	0.018876	C	Yes	C	Yes	No	
	rs115595430	6				G	T	0.58	1.12			rs2523627						rs2523627	T	0.114		G	T	0.378536	0.97614	0.82726	1.1518	0.53827	-0.05676	G	No	G	No	No	
	rs114965247	6				T	G	0.25	1.08			rs2844535						rs2844535	G	0.077		T	G	0.263758	1.10242	0.91849	1.32318	0.20945	0.130324	G	Yes	G	Yes	No	
	rs139089278	6				C	T	0.32	1.13			rs2442715	rs2844518	31372569	0.29	1	C=C,T=T	rs2844518	T	0.127		C	T	0.364524	1.10927	0.93868	1.31086	0.20744	0.115426	T	Yes	T	Yes	No	
	rs114279165	6				A	G	0.30	1.08			rs3131621						rs3131621	G	0.078		A	G	0.366057	0.93482	0.79147	1.10414	0.16675	-0.12798	G	Yes	A	No	No	
	rs116822326	6				A	G	0.16	1.15			rs3094604						rs3094604	G	0.140	Hung-35	A	G	0.191348	1.19677	0.97457	1.46964	0.12205	0.179965	G	Yes	G	Yes	No	
	rs115441327	6				T	C	0.25	1.08			rs3130048						rs3130048	C	0.078		T	C	0.289525	1.05404	0.88296	1.25826	0.65073	0.044517	C	Yes	C	Yes	No	
	rs2763981	6				A	T	0.29	1.08									rs2763981	T	0.079		T	A	0.349255	0.90295	0.76285	1.06877	0.14693	-0.13717	T	Yes	T	Yes	No	
	rs652888	6				A	G	0.20	1.10									rs652888	G	0.095		A	G	0.255512	0.97508	0.8113	1.17193	0.71718	0.037435	G	Yes	G	Yes	No	
	rs114676850	6				C	T	0.28	1.13			rs424232						rs424232	T	0.119		C	T	0.290373	0.99666	0.83519	1.18935	0.66423	-0.04281	T	Yes	C	No	No	
	rs9270868	6				A	G	0.20	1.11				rs2097432	32590771	0.22	0.78	A=T,G=C	rs2097432	C	0.108		T	C	0.235424	1.11925	0.92581	1.3531	0.21553	0.131961	C	No	C	Yes	No	
	rs116080888	6				A	C	0.59	1.09			rs11751024						rs11751024	C	0.086		C	A	0.388854	0.84349	0.71554	0.994316	0.04274	-0.18406	A	No	C	Yes	Yes	
	rs113574092	6				A	G	0.49	1.13			rs9271364						rs9271364	G	0.123		A	G	0.489601	0.9764	0.8316	1.14641	0.43989	-0.06881	A	No	A	No	No	
	rs35867827	6				C	T	0.90	1.16									rs35867827	T	0.151		T	C	0.0802392	0.8408	0.62638	1.12863	0.22421	-0.20199	C	No	T	Yes	No	
	rs190788477	6				G	A	0.28	1.17			rs9271971	rs9271611	32591609	0.4	0.7399	G=A,A=G	rs9271611	G	0.155		A	G	0.386963	1.03444	0.87727	1.21977	0.73806	0.031553	G	No	G	Yes	No	
	rs115566240	6				T	C	0.74	1.09				rs35656734	32600153	0.32	0.8276	C=C,T=T	rs35656734	C	0.085		C	T	0.24376	0.85044	0.70575	1.02481	0.07078	-0.18987	T	No	C	Yes	No	
	rs9272307	6				G	T	0.22	1.09				rs9272306	32603880	0.17	1	G=C,T=T	rs9272306	T	0.087		C	T	0.11327	0.97761	0.75908	1.25906	0.9054	-0.02127	T	Yes	C	No	No	
	rs1048372	6				T	C	0.59	1.16				NA	NA	NA	NA	NA																		No proxy SNP >0.5 R2
	rs9273429	6				G	A	0.66	1.14				rs1049053	32634405	0.25	0.6382	G=C,A=T	rs1049053	T	0.134		C	T	0.333622	0.91908	0.77506	1.08986	0.20906	-0.12393	C	No	C	No	No	
	rs114830752	6				C	A	0.15	1.11			rs9275602						rs9275602	A	0.101		C	A	0.202522	1.01268	0.82934	1.23656	0.99376	-0.00088	A	Yes	C	No	No	
	rs116036768	6				C	G	0.60	1.12			rs2621408						rs2621408	G	0.114		C	G	0.369963	0.95421	0.80802	1.12685	0.40923	-0.07716	C	No	C	No	No	
	rs7383287	6				A	G	0.20	1.10									rs7383287	G	0.091	Hung-128	A	G	0.210484	0.94641	0.77749	1.15203	0.88216	-0.0165	G	Yes	A	No	No	
	rs239935	6				A	G	0.48	1.07									rs239935	G	0.067		G	A	0.489351	0.94992	0.80905	1.11532	0.18272	-0.1211	A	No	G	Yes	No	
	rs11780471	8				A	G	0.94	1.15									rs11780471	G	0.141	Hung-35, Fritsche-19, Jia	G	A	0.0553898	1.06462	0.7489	1.51344	0.36496	0.178825	A	No	A	No	No	
	rs4236709	8				A	G	0.22	1.13									rs4236709	G	0.124	Hung-35, Fritsche-14, Jia, Dai	G	A	0.23445	0.98229	0.81274	1.18722	0.88974	0.014861	G	Yes	A	No	No	
	rs885518	9				A	G	0.10	1.17									rs885518	G	0.157	Hung-35, Fritsche-19, Jia	A	G	0.100562	1.39647	1.0633	1.83404	0.05472	0.286665	G	Yes	G	Yes	No	
	rs62560775	9				A	G	0.11	1.18									rs62560775	G	0.166	Fritsche-19, Jia	A	G	0.100047	1.00097	0.7661	1.30786	0.37769	0.130727	G	Yes	G	Yes	No	
	rs1333040	9				T	C	0.41	1.14									rs1333040	C	0.131	Hung-128	C	T	0.40567	1.1247	0.95521	1.32426	0.10078	0.148051	C	Yes	T	No	No	
	rs11591710	10				A	C	0.14	1.16									rs11591710	C	0.148	Hung-35, Fritsche-14	A	C	0.132841	0.9333	0.73711	1.18172	0.6147	-0.06745	C	Yes	A	No	No	
	rs7943272	11				A	C	0.20	1.12									rs7943272	C	0.115		A	C	0.202947	1.13212	0.92659	1.38323	0.23759	0.133484	C	Yes	C	Yes	No	
	rs17121997	11				C	T	0.48	1.10									rs17121997	T	0.098		T	C	0.496978	0.91242	0.7771	1.0713	0.40578	-0.07746	T	Yes	T	Yes	No	
	rs7953330	12				C	G	0.69	1.09									rs7953330	G	0.087	Hung-35, Fritsche-14, Jia	G	C	0.294417	0.73843	0.61933	0.880422	0.00055	-0.34006	C	No	G	Yes	Yes	
	rs11064614	12				G	A	0.43	1.13									rs11064614	A	0.118		G	A	0.447011	1.15988	0.98679	1.36333	0.01228	0.226406	A	Yes	A	Yes	Yes	
	rs56404467	13				G	A	0.02	1.51									rs56404467	A	0.409		G	A	0.0253935	1.06453	0.63825	1.77553	0.25276	0.319437	A	Yes	A	Yes	No	
	rs748404	15				C	T	0.77	1.15									rs748404	T	0.140		T	C	0.19669	0.99911	0.81649	1.22259	0.65007	0.052181	C	No	C	No	No	
	rs66759488	15				G	A	0.36	1.07									rs66759488	A	0.068	Hung-35, Shi, Jia	G	A	0.367899	1.32582	1.12143	1.56747	0.00271	0.274821	A	Yes	A	Yes	Yes	
	rs8031813	15				C	A	0.69	1.10									rs8031813	A	0.099	Hung-128	C	A	0.310115	1.1813	0.9934	1.40473	0.02871	0.225341	C	No	A	Yes	Yes	
	rs7359317	15				T	G	0.38	1.10									rs7359317	G	0.096		G	T	0.393073	0.88289	0.74893	1.04082	0.23004	-0.11405	G	Yes	G	Yes	No	
	rs146782107	15				C	T	0.74	1.15									rs146782107	T	0.142		T	C	0.236086	0.84134	0.69678	1.01591	0.06049	-0.20266	C	No	T	Yes	No	
	rs4887059	15				C	T	0.62	1.19									rs4887059	T	0.171		C	T	0.368706	1.22748	1.03952	1.44942	0.00036	0.345734	C	No	T	Yes	Yes	
	rs3885951	15				A	G	0.11	1.22									rs3885951	G	0.199	Hung-35	A	G	0.0944913	1.19085	0.90296	1.57053	0.04471	0.318058	G	Yes	G	Yes	Yes	
	rs55781567	15				C	G	0.37	1.30									rs55781567	G	0.262	Hung-35, Fritsche-14, Jia	C	G	0.371797	1.24259	1.05177	1.46803	0.00106	0.306993	G	Yes	G	Yes	Yes	
	rs56219480	15				T	C	0.96	1.22									rs56219480	C	0.202		C	T	0.0353428	0.60812	0.39241	0.942423	0.01285	-0.60452	T	No	C	Yes	Yes	
	rs2229961	15				G	A	0.02	1.43									rs2229961	A	0.359	Hung-128	G	A	0.021661	1.3566	0.77176	2.38464	0.24599	0.35796	A	Yes	A	Yes	No	
	rs3743078	15				C	G	0.78	1.26									rs3743078	G	0.233		C	G	0.215303	1.31028	1.07834	1.59211	0.00027	0.399877	C	No	G	Yes	Yes	
	rs77438700	15				G	A	0.94	1.24									rs77438700	A	0.217		A	G	0.0679359	0.91707	0.6672	1.26053	0.3902	-0.15268	G	No	A	Yes	No	
	rs8192479	15				C	T	0.02	1.28									rs8192479	T	0.248	Hung-128	C	T	0.0205125	1.19374	0.67357	2.11563	0.2036	0.436389	T	Yes	T	Yes	No	
	rs17487514	15				C	T	0.30	1.18									rs17487514	T	0.164		C	T	0.328791	1.05908	0.89264	1.25655	0.23028	0.116372	T	Yes	T	Yes	No	
	rs143527274	15				T	C	0.97	1.31									rs143527274	C	0.267		C	T	0.0259093	1.11221	0.66928	1.84827	0.71764	-0.10465	T	No	C	Yes	No	
	rs148832167	15				G	A	0.62	1.13			rs1808065						rs1808065	A	0.119		G	A	0.353725	1.05844	0.895	1.25172	0.32986	0.092316	G	No	A	Yes	No	
	rs12905116	15				T	A	0.72	1.13									rs12905116	A	0.126		T	A	0.231894	1.0567	0.87392	1.27771	0.39933	0.092973	T	No	A	Yes	No	
	rs3861180	15				A	G	0.94	1.22									rs3861180	G	0.199		A	G	0.0591544	0.86106	0.61115	1.21317	0.83446	-0.04419	A	No	A	No	No	
	rs112827102	15				C	T	0.92	1.18				NA	NA	NA	NA	NA																		No proxy SNP >0.5 R2
	rs28450923	15				G	A	0.86	1.19				NA	NA	NA	NA	NA				Hung-128														No proxy SNP >0.5 R2
	rs36061084	15				G	A	0.44	1.16				rs2004038	79071406	0.41	0.96	G=G,A=A	rs2004038	A	0.150		G	A	0.463801	1.07274	0.91325	1.26009	0.06922	0.165587	A	Yes	A	Yes	No	
	rs7182809	15				G	C	0.83	1.10									rs7182809	C	0.096		G	C	0.154928	0.92992	0.74455	1.16145	0.82058	-0.02969	G	No	G	No	No	
	rs77719127	15				C	T	0.17	1.14									rs77719127	T	0.133	Hung-128	C	T	0.185129	0.94242	0.76673	1.15836	0.5011	-0.0767	T	Yes	C	No	No	
	rs11072811	15				C	A	0.49	1.10									rs11072811	A	0.091		A	C	0.475511	0.98756	0.84098	1.15969	0.13383	-0.13517	C	No	A	Yes	No	
	rs1551821	18				C	A	0.86	1.23									rs1551821	A	0.211		A	C	0.181331	0.965	0.78372	1.18821	0.28128	-0.12953	C	No	A	Yes	No	
	rs11879413	19				T	C	0.91	1.16									rs11879413	C	0.147		C	T	0.0586679	1.09748	0.77886	1.54644	0.5866	0.105972	T	No	T	No	No	
	rs56113850	19				T	C	0.56	1.14									rs56113850	C	0.128	Hung-35, Fritsche-14, Jia	T	C	0.420448	1.03358	0.87854	1.21598	0.35254	0.088223	T	No	C	Yes	No	
	rs67210567	19				T	G	0.97	1.27				rs1801272	41354533	0.03	0.97	G=A,T=T	rs1801272	A	0.237	Hung-128	A	T	0.0262501	0.51896	0.31053	0.867308	0.00264	-0.83473	T	No	A	Yes	Yes	
	rs41309931	20				G	T	0.12	1.17									rs41309931	T	0.157	Hung-35	G	T	0.132071	1.10521	0.87119	1.40209	0.59422	0.073295	T	Yes	T	Yes	No	
	rs504022	21				C	A	0.06	1.32									rs504022	T	0.278		T	G	0.0644839	0.85087	0.612	1.18297	0.46875	-0.16818	T	No	T	Yes	No	Allele T subsituted for A to correct stand mismatch
Dai	rs17038564	2	65496058	2p14	ACTR2	A	G		1.11									rs17038564	G	0.102		A	G	0.0253565	0.93207	0.56012	1.55102	0.38081	-0.25625	G	NA	A	No	No	
	rs2293607	3	169482335	3q26.2	N-LRRC34	C	T		1.12									rs2293607	T	0.117		T	C	0.228383	1.00197	0.82764	1.21303	0.3725	0.096262	C	NA	C	No	No	
	rs11375254	3	189343242	3q28	TP63	TA	T		1.21				rs12696594	189357616	0.5	0.89	A DEL=A,A=G	rs12696594	A	0.19		A	G	0.498054	1.00308	0.85436	1.17769	0.95605	-0.00503	A	NA	A	Yes	No	
	rs13167280	5	1280477	5p15	TERT	G	A		1.30									rs13167280	A	0.265		G	A	0.130554	0.9374	0.73904	1.18901	0.61096	-0.06877	A	NA	G	No	No	
	rs401681	5	1322087	5p15	CLPTM1L	T	C		1.16									rs401681	C	0.145	Hung-35	C	T	0.390184	0.82812	0.70256	0.976107	0.02791	-0.19884	T	NA	C	Yes	Yes	
	rs2517873	6	29875992	6p21.3	HLA-A	G	A		1.20				rs376316	29878190	0.18	0.8813	G=T,A=C	rs376316	C	0.183		T	C	0.223089	1.25238	1.03117	1.52105	0.02402	0.246276	C	NA	C	Yes	Yes	
	rs3817963	6	32368087	6p21.3	23-BTNL2	T	C		1.06									rs3817963	C	0.061		T	C	0.313228	0.9811	0.82528	1.16635	0.93077	0.008395	C	NA	C	Yes	No	
	rs1853837	6	41497035	6p21.1	-328M4.2	C	A		1.15									rs1853837	A	0.14		C	A	0.106707	0.9884	0.76226	1.28162	0.79819	-0.0368	A	NA	C	No	No	
	rs5879422	6	117784658	6q22.1	S1-GOPC	TTG	T		1.09				rs9374663	117782634	0.5	1	TG DEL=A,TG=G	rs9374663	A	0.084		A	G	0.464495	0.90208	0.76802	1.05954	0.5321	-0.05648	G	NA	A	Yes	No	
	rs4236709	8	32410110	8p12	NRG1	A	G		1.17									rs4236709	G	0.156	Graff, Hung-35, Fritsche-14, Jia	G	A	0.23445	0.98229	0.81274	1.18722	0.88974	0.014861	G	NA	A	No	No	
	rs10429489	9	21787521	9p21.3	MTAP	G	A		1.11									rs10429489	A	0.1		G	A	0.093522	1.43587	1.08257	1.90449	0.03623	0.323636	A	NA	A	Yes	Yes	
	rs35201538	9	33422488	9p13.3	AQP3	CT	C		1.10				rs10758201	33419770	0.3	0.97	T DEL=G,T=T	rs10758201	G	0.095		G	T	0.292788	0.90859	0.76187	1.08356	0.12975	-0.15133	T	NA	G	Yes	No	
	rs4573350	9	124955115	9q33.2	MORN5	C	T		1.09									rs4573350	T	0.09		T	C	0.311502	0.96955	0.81537	1.15288	0.63815	-0.04745	C	NA	T	Yes	No	
	rs12265047	10	114487925	10q25.2	9.1-VTI1A	A	G		1.14									rs12265047	G	0.134		G	A	0.0261799	0.89012	0.53683	1.47591	0.45024	-0.22326	G	NA	G	Yes	No	
	rs55768116	11	118108331	11q23.3	L2-MPZL3	A	C		1.13									rs55768116	C	0.118		C	A	0.483593	0.9282	0.7905	1.08988	0.5648	-0.05396	A	NA	C	Yes	No	
	rs11610143	12	52349071	12q13.13	ACVR1B	G	C		1.07									rs11610143	C	0.065		C	G	0.196465	1.12523	0.91873	1.37814	0.18447	0.152447	G	NA	G	No	No	
	rs1200399	14	35293185	14q13.1	BAZ1A	T	C		1.11									rs1200399	C	0.1		C	T	0.0432949	0.86306	0.58	1.28425	0.30941	-0.2217	C	NA	C	Yes	No	
	rs77468143	15	49376624	15q21.1	S2-GALK2	G	T		1.08									rs77468143	T	0.077	Hung-35, Fritsche-14, Jia	T	G	0.222111	0.82449	0.68005	0.999597	0.06449	-0.20643	G	NA	T	Yes	No	
	rs200595745	17	65915289	17q24.2	TF-KPNA2	TAATAAT	A		1.12				rs11079710	65885911	0.27	0.8678	AATAATAAT=C,DEL=T	rs11079710	T	0.114		C	T	0.257682	1.15891	0.96391	1.39336	0.34222	0.102207	T	NA	T	Yes	No	
Young	rs2273953	1				T	C		0.65	0.49-0.85	0.001							rs2273953	G	-0.431		G	A	0.217024	0.90276	0.74334	1.09637	0.54267	-0.06856	A	NA	G	Yes	No	Assumed minus-strand designation and corrected the
	rs763110	1				C	T		0.72	0.49-1.1	0.09							rs763110	T	-0.329		C	T	0.35219	0.91542	0.77397	1.08273	0.70785	-0.03578	T	NA	C	Yes	No	
	rs16944	2				A	G		1.2	0.9-1.6	0.14							rs16944	G	0.182		A	G	0.356099	1.01315	0.85686	1.19794	0.88885	0.013191	A	NA	G	Yes	No	
	rs3087386	2				T	C		0.79	0.59-1	0.1							rs3087386	G	-0.236		A	G	0.45261	0.97467	0.82955	1.14519	0.97121	0.003257	A	NA	G	No	No	Assumed minus-strand designation and corrected the
	rs5743836	3				T	C		2.2	0.8-6.6	0.12							rs5743836	G	0.788		A	G	0.16419	0.76848	0.61915	0.953844	0.00643	-0.34342	G	NA	A	No	Yes	Assumed minus-strand designation and corrected the
	rs4073	4				A	T		1.5	1.1-2.1	0.005							rs4073	T	0.405		A	T	0.474616	0.99818	0.85001	1.17218	0.86388	-0.01571	A	NA	A	No	No	
	rs1799895	4				C	G		0.28	0.1-0.9	0.02							rs1799895	G	-1.273		C	G	0.010525	2.41057	0.98887	5.87623	0.01557	1.19386	G	NA	G	No	Yes	
	rs6413429	5				G	T		1.5	1-2.3	0.04							rs6413429	A	0.405		C	A	0.0777329	0.90241	0.66927	1.21677	0.65838	-0.07456	A	NA	C	No	No	
	rs1799930	8				A	G		1.4	1.1-1.9	0.006							rs1799930	G	0.336		G	A	0.293677	1.07593	0.90189	1.28356	0.85711	-0.018	A	NA	G	Yes	No	
	rs10115703	9				G	A		1.4	0.9-2	0.1							rs10115703	A	0.336		G	A	0.0742745	0.88786	0.65439	1.20464	0.21537	-0.21119	A	NA	G	No	No	
	rs2031920	10				C	T		2.1	1-4.3	0.03							rs2031920	T	0.742		C	T	0.0282862	1.4663	0.88969	2.4166	0.1759	0.36609	T	NA	T	Yes	No	
	rs360721	11				G	C		1.4	1.1-1.9	0.009							rs360721	C	0.336		G	C	0.308935	0.90606	0.76178	1.07768	0.27986	-0.11009	C	NA	G	No	No	
	rs1799732	11				C	DEL		0.68	0.48-0.96	0.02		rs11214613	113335259	0.08	1	DEL=A,G=G	rs11214613	A	-0.386		A	G	0.10042	1.194	0.91501	1.55805	0.45399	0.11205	A	NA	G	Yes	No	
	rs1139417	12				G	A		1.3	1-1.8	0.05	rs767455						rs767455	T	0.262		T	C	0.417721	1.10963	0.94285	1.30591	0.64112	0.042841	C	NA	C	No	No	Assumed minus-strand designation and corrected the
	rs4934	14				A	G		1.6	1.2-2.2	0.004							rs4934	G	0.470		G	A	0.477137	0.97233	0.82804	1.14178	0.90508	0.011044	A	NA	A	No	No	
	rs16969968	15				G	A		1.8	1.2-2.7	0.004							rs16969968	A	0.588		G	A	0.366056	1.27657	1.07978	1.50922	0.00031	0.33659	A	NA	A	Yes	Yes	
	rs2306022	15				G	A		1.6	0.9-7.6	0.04							rs2306022	T	0.470		C	T	0.0716077	1.02936	0.75381	1.40564	0.8717	-0.02889	T	NA	C	No	No	Assumed minus-strand designation and corrected the
	rs2317676	17				A	G		0.59	0.39-0.89	0.008							rs2317676	G	-0.528		A	G	0.0574176	0.80988	0.57427	1.14214	0.13385	-0.28587	G	NA	A	Yes	No	
	rs2279115	18				C	A		0.71	0.53-0.97	0.03							rs2279115	T	-0.342		G	T	0.465976	1.03731	0.8832	1.21832	0.48977	0.062583	G	NA	T	No	No	Assumed minus-strand designation and corrected the
	rs13181	19				T	G		0.74	0.5-1.1	0.11							rs13181	G	-0.301		T	G	0.358759	0.93085	0.78754	1.10023	0.26757	-0.1036	G	NA	T	Yes	No	

8.3 Appendix 2 – Full GWAS Results for Significant SNPs in Manchester Cohort

uniqID	rsID	chr	pos	ref	effect	MAF	P	or	beta	se	r2	IndSigSNP	Genomic Locus	Nearest Gene	dist	func	CADD	RDB
1:8985036:C:T	rs10864370	1	8985036	C	T	0.2396	8.00E-06	0.700706	-0.46962	0.105173	1	rs10864370	1	RP3-477M7.6	15072	intergenic	1.095	5
1:14233836:A:T	rs144716332	1	14233836	T	A	0.01889	2.54E-06	3.01568	1.4583	0.30997	1	rs144716332	2	PRDM2	82261	intergenic	7.681	7
1:37521551:A:C	rs581878	1	37521551	C	A	0.4423	2.57E-05	1.33914	0.399564	0.094946	0.622217	rs6676142	3	GRIK3	21820	intergenic	0.869	NA
1:37521740:A:G	rs6676142	1	37521740	A	G	0.4384	6.38E-06	1.38642	0.412207	0.091332	1	rs6676142	3	GRIK3	22009	intergenic	1.857	7
1:37521938:A:G	rs6687485	1	37521938	G	A	0.4354	0.000161	1.3097	0.342353	0.09074	0.847621	rs6676142	3	GRIK3	22207	intergenic	0.401	5
1:37521983:G:T	rs6679188	1	37521983	T	G	0.4274	0.000271	1.30019	0.330606	0.090788	0.773466	rs6676142	3	GRIK3	22252	intergenic	13.3	5
1:37522030:A:G	rs6687588	1	37522030	G	A	0.4294	0.000271	1.30027	0.330564	0.090783	0.773392	rs6676142	3	GRIK3	22299	intergenic	3.197	5
1:37522629:C:T	rs822904	1	37522629	T	C	0.4205	NA	NA	NA	NA	0.760688	rs6676142	3	GRIK3	22898	intergenic	0.585	NA
1:37522637:G:T	rs6426006	1	37522637	T	G	0.4622	NA	NA	NA	NA	0.642341	rs6676142	3	GRIK3	22906	intergenic	0.056	6
1:37522641:G:T	rs202023532	1	37522641	T	G	0.4563	NA	NA	NA	NA	0.647889	rs6676142	3	GRIK3	22910	intergenic	0.767	6
1:37522645:G:T	rs6690521	1	37522645	T	G	0.4324	NA	NA	NA	NA	0.755882	rs6676142	3	GRIK3	22914	intergenic	0.588	6
1:37522916:A:T	rs656826	1	37522916	T	A	0.4254	0.00017	1.30663	0.342192	0.091002	0.752033	rs6676142	3	GRIK3	23185	intergenic	2.031	7
1:37522957:A:G	rs527907	1	37522957	G	A	0.4254	0.00018	1.30483	0.340938	0.09104	0.752033	rs6676142	3	GRIK3	23226	intergenic	1.786	NA
1:37524007:G:GA	rs10714160	1	37524007	G	GA	0.3877	NA	NA	NA	NA	0.61497	rs6676142	3	GRIK3	24276	intergenic	0.909	NA
1:37525639:A:G	rs550273	1	37525639	A	G	0.4245	0.000343	1.29993	0.317932	0.088794	0.748949	rs6676142	3	GRIK3	25908	intergenic	1.406	5

1:37525676:A:C	rs613720	1	37525676	C	A	0.4245	0.00017	1.30638	0.342114	0.090999	0.748949	rs6676142	3	GRIK3	25945	intergenic	2.131	NA
1:37525698:C:T	rs549417	1	37525698	T	C	0.4235	0.000275	1.29565	0.331157	0.091022	0.745057	rs6676142	3	GRIK3	25967	intergenic	3.046	NA
1:37526481:C:T	rs521696	1	37526481	T	C	0.4125	0.00025	1.30038	0.334698	0.091399	0.788436	rs6676142	3	GRIK3	26750	intergenic	4.785	NA
1:37530753:A:G	rs218406	1	37530753	G	A	0.4235	0.000434	1.28268	0.318729	0.090588	0.745897	rs6676142	3	GRIK3	31022	intergenic	0.482	NA
1:37534313:C:T	rs849942	1	37534313	T	C	0.4274	0.000519	1.28367	0.319097	0.091941	0.702869	rs6676142	3	GRIK3	34582	intergenic	0.508	NA
1:37534673:C:T	rs218424	1	37534673	T	C	0.4245	0.000588	1.28655	0.31794	0.0925	0.674994	rs6676142	3	GRIK3	34942	intergenic	8.689	NA
4:30168938:G:T	rs76640173	4	30168938	G	T	0.02584	3.46E-06	0.348977	-1.70726	0.367815	1	rs76640173	4	RP11-174E22.2	158999	intergenic	2.907	7
4:30168960:C:T	rs115363044	4	30168960	C	T	0.02584	3.48E-06	0.349104	-1.70675	0.367838	1	rs76640173	4	RP11-174E22.2	159021	intergenic	1.073	7
4:30169405:A:AAAG	rs148111093	4	30169405	AAAG	A	0.02584	NA	NA	NA	NA	1	rs76640173	4	RP11-174E22.2	159466	intergenic	1.409	NA
4:30169961:A:G	rs77296350	4	30169961	G	A	0.02584	3.68E-06	0.350378	-1.69464	0.366128	1	rs76640173	4	RP11-174E22.2	160022	intergenic	0.002	7
4:30171304:A:G	rs114414502	4	30171304	G	A	0.02584	3.66E-06	0.350165	-1.69587	0.366286	1	rs76640173	4	RP11-174E22.2	161365	intergenic	0.862	7
4:30171934:G:T	rs116585012	4	30171934	G	T	0.02584	3.85E-06	0.35021	-1.68841	0.365496	1	rs76640173	4	RP11-174E22.2	161995	intergenic	0.199	7
4:30174024:A:G	rs115852532	4	30174024	G	A	0.02386	1.69E-05	0.358837	-1.62412	0.377451	0.919223	rs76640173	4	RP11-174E22.2	164085	intergenic	2.102	7
4:30179734:G:T	rs75573218	4	30179734	G	T	0.02584	4.54E-06	0.349184	-1.67948	0.366316	1	rs76640173	4	RP11-174E22.2	169795	intergenic	2.182	6

4:30181119:C:CTT	rs564195700	4	30181119	C	CTT	0.02485	NA	NA	NA	NA	0.959527	rs76640173	4	RP11-174E22.2	171180	intergenic	3.978	NA
4:30185359:C:G	rs80092455	4	30185359	G	C	0.02584	NA	NA	NA	NA	1	rs76640173	4	RP11-174E22.2	175420	intergenic	0.459	7
4:30187639:C:T	rs75356750	4	30187639	T	C	0.02485	NA	NA	NA	NA	0.959527	rs76640173	4	RP11-174E22.2	177700	intergenic	0.378	5
4:30195644:G:T	rs73223718	4	30195644	G	T	0.01889	NA	NA	NA	NA	0.642476	rs76640173	4	RP11-174E22.2	185705	intergenic	0.726	7
4:30205385:C:T	rs73223722	4	30205385	T	C	0.01889	NA	NA	NA	NA	0.642476	rs76640173	4	RP11-174E22.2	195446	intergenic	0.838	7
4:30210737:A:C	rs6819391	4	30210737	C	A	0.01889	NA	NA	NA	NA	0.642476	rs76640173	4	RP11-174E22.2	200798	intergenic	0.668	5
4:30212116:A:G	rs60401852	4	30212116	G	A	0.01889	NA	NA	NA	NA	0.642476	rs76640173	4	RP11-174E22.2	202177	intergenic	0.912	6
4:30212194:A:C	rs57562711	4	30212194	A	C	0.01889	NA	NA	NA	NA	0.642476	rs76640173	4	RP11-174E22.2	202255	intergenic	0.569	7
4:30212685:A:G	rs73223724	4	30212685	A	G	0.01889	NA	NA	NA	NA	0.642476	rs76640173	4	RP11-174E22.2	202746	intergenic	5.923	7
4:30213404:A:T	rs11934322	4	30213404	A	T	0.01889	NA	NA	NA	NA	0.642476	rs76640173	4	RP11-174E22.2	203465	intergenic	0.139	7
4:30213444:A:T	rs11930025	4	30213444	T	A	0.01889	NA	NA	NA	NA	0.642476	rs76640173	4	RP11-174E22.2	203505	intergenic	0.344	6
4:30213600:A:G	rs73223725	4	30213600	A	G	0.01889	NA	NA	NA	NA	0.642476	rs76640173	4	RP11-174E22.2	203661	intergenic	0.536	6
4:30213690:C:G	rs73223726	4	30213690	G	C	0.01889	NA	NA	NA	NA	0.642476	rs76640173	4	RP11-174E22.2	203751	intergenic	0.108	6
4:30214160:A:G	rs11935524	4	30214160	A	G	0.01889	NA	NA	NA	NA	0.642476	rs76640173	4	RP11-174E22.2	204221	intergenic	2.164	5
4:30214281:C:T	rs11944617	4	30214281	C	T	0.01889	NA	NA	NA	NA	0.642476	rs76640173	4	RP11-174E22.2	204342	intergenic	0.131	6
4:30214571:A:C	rs73223727	4	30214571	A	C	0.01889	NA	NA	NA	NA	0.642476	rs76640173	4	RP11-174E22.2	204632	intergenic	13.47	7
4:30214741:G:T	rs11943659	4	30214741	T	G	0.01889	NA	NA	NA	NA	0.642476	rs76640173	4	RP11-174E22.2	204802	intergenic	12.16	6
4:30214753:C:T	rs57400401	4	30214753	C	T	0.01889	NA	NA	NA	NA	0.642476	rs76640173	4	RP11-174E22.2	204814	intergenic	0.06	7

4:30215110:C:T	rs6836992	4	30215110	T	C	0.01889	NA	NA	NA	NA	0.642476	rs76640173	4	RP11-174E22.2	205171	intergenic	4.908	7
4:30215507:C:T	rs6448699	4	30215507	T	C	0.01889	NA	NA	NA	NA	0.642476	rs76640173	4	RP11-174E22.2	205568	intergenic	0.215	5
4:30216257:A:G	rs7662079	4	30216257	G	A	0.01889	NA	NA	NA	NA	0.642476	rs76640173	4	RP11-174E22.2	206318	intergenic	2.607	6
4:30217539:A:G	rs141957700	4	30217539	G	A	0.01889	NA	NA	NA	NA	0.642476	rs76640173	4	RP11-174E22.2	207600	intergenic	0.085	6
4:30218508:A:G	rs73223730	4	30218508	A	G	0.01889	NA	NA	NA	NA	0.642476	rs76640173	4	RP11-174E22.2	208569	intergenic	3.41	5
4:30218695:A:G	rs7680660	4	30218695	G	A	0.01889	NA	NA	NA	NA	0.642476	rs76640173	4	RP11-174E22.2	208756	intergenic	0.029	7
4:30219313:C:T	rs7681560	4	30219313	C	T	0.01889	NA	NA	NA	NA	0.642476	rs76640173	4	RP11-174E22.2	209374	intergenic	2.2	5
4:30219615:A:C	rs73223731	4	30219615	C	A	0.01889	NA	NA	NA	NA	0.642476	rs76640173	4	RP11-174E22.2	209676	intergenic	1.53	7
4:30219617:A:G	rs73223732	4	30219617	G	A	0.01889	NA	NA	NA	NA	0.642476	rs76640173	4	RP11-174E22.2	209678	intergenic	0.057	7
4:30221873:A:G	rs73223733	4	30221873	G	A	0.01889	NA	NA	NA	NA	0.642476	rs76640173	4	RP11-174E22.2	211934	intergenic	1.725	7
4:30222207:C:T	rs73223734	4	30222207	T	C	0.01889	NA	NA	NA	NA	0.642476	rs76640173	4	RP11-174E22.2	212268	intergenic	1.559	7
4:30222603:A:G	rs16883109	4	30222603	G	A	0.01889	NA	NA	NA	NA	0.642476	rs76640173	4	RP11-174E22.2	212664	intergenic	0.406	7
4:30223587:C:T	rs73223735	4	30223587	C	T	0.01889	NA	NA	NA	NA	0.642476	rs76640173	4	RP11-174E22.2	213648	intergenic	1.153	7
4:30223966:C:T	rs7691913	4	30223966	T	C	0.01889	NA	NA	NA	NA	0.642476	rs76640173	4	RP11-174E22.2	214027	intergenic	2.548	7
4:30224887:C:G	rs78977954	4	30224887	C	G	0.01889	NA	NA	NA	NA	0.642476	rs76640173	4	RP11-174E22.2	214948	intergenic	0.247	7
4:30225317:A:G	rs73223736	4	30225317	A	G	0.01889	NA	NA	NA	NA	0.642476	rs76640173	4	RP11-174E22.2	215378	intergenic	3.001	6
4:30225508:G:T	rs34052174	4	30225508	G	T	0.01889	NA	NA	NA	NA	0.642476	rs76640173	4	RP11-174E22.2	215569	intergenic	1.752	5
4:30226180:A:G	rs59144212	4	30226180	G	A	0.01889	NA	NA	NA	NA	0.642476	rs76640173	4	RP11-174E22.2	216241	intergenic	2.802	6

4:30228019:C:T	rs7672946	4	30228019	C	T	0.01889	NA	NA	NA	NA	0.642476	rs76640173	4	RP11-174E22.2	218080	intergenic	1.358	7
4:30229098:C:T	rs73223737	4	30229098	T	C	0.01889	NA	NA	NA	NA	0.642476	rs76640173	4	RP11-174E22.2	219159	intergenic	1.785	7
4:30229686:A:G	rs73223738	4	30229686	G	A	0.01889	NA	NA	NA	NA	0.642476	rs76640173	4	RP11-174E22.2	219747	intergenic	1.184	7
4:30229977:G:T	rs7686382	4	30229977	G	T	0.01889	NA	NA	NA	NA	0.642476	rs76640173	4	RP11-174E22.2	220038	intergenic	0.893	7
4:30231440:A:G	rs73223739	4	30231440	A	G	0.01889	NA	NA	NA	NA	0.642476	rs76640173	4	RP11-174E22.2	221501	intergenic	1.325	6
4:30231557:A:T	rs73223741	4	30231557	A	T	0.01789	NA	NA	NA	NA	0.603447	rs76640173	4	RP11-174E22.2	221618	intergenic	1.98	7
4:30233080:A:T	rs73223743	4	30233080	T	A	0.01889	NA	NA	NA	NA	0.642476	rs76640173	4	RP11-174E22.2	223141	intergenic	6.779	6
6:124413195:A:G	rs17629528	6	1.24E+08	G	A	0.02684	7.04E-06	0.401602	-1.16411	0.259126	1	rs17629528	5	NKAIN2	0	intronic	2.135	6
6:124438622:A:G	rs17630088	6	1.24E+08	G	A	0.02684	1.13E-05	0.408363	-1.12453	0.256139	0.923256	rs17629528	5	NKAIN2	0	intronic	4.634	7
6:131852291:C:T	rs12204890	6	1.32E+08	T	C	0.04871	3.52E-06	0.504759	-0.96511	0.2081	1	rs12204890	6	ARG1	41992	intergenic	1.336	5
7:78464211:C:T	rs1928912	7	78464211	C	T	0.4066	4.57E-07	0.683777	-0.5085	0.10082	0.813404	rs17389497	7	MAGI2	0	intronic	2.253	NA
7:78464280:C:T	rs1928911	7	78464280	T	C	0.4085	4.59E-07	0.684005	-0.50844	0.100828	0.815197	rs17389497	7	MAGI2	0	intronic	5.988	NA
7:78464645:A:T	rs12670051	7	78464645	T	A	0.4066	4.86E-07	0.686267	-0.50781	0.100925	0.820285	rs17389497	7	MAGI2	0	intronic	7.218	5
7:78464855:C:T	rs7794352	7	78464855	T	C	0.4066	4.20E-07	0.683947	-0.50951	0.100702	0.820285	rs17389497	7	MAGI2	0	intronic	12.09	6
7:78465200:C:G	rs7808431	7	78465200	C	G	0.4066	5.01E-07	0.685864	-0.5067	0.100816	0.820285	rs17389497	7	MAGI2	0	intronic	1.423	7

7:78465349:A:G	rs9791886	7	78465349	G	A	0.3608	1.20E-07	0.672483	-0.55148	0.104182	1	rs17389497	7	MAGI2	0	intronic	5.708	5
7:78465639:A:G	rs1411533	7	78465639	A	G	0.4066	1.41E-07	0.674966	-0.53147	0.100964	0.820285	rs17389497	7	MAGI2	0	intronic	1.165	NA
7:78465646:C:T	rs1411532	7	78465646	T	C	0.4066	1.41E-07	0.674966	-0.53147	0.100964	0.820285	rs17389497	7	MAGI2	0	intronic	6.224	NA
7:78465922:A:T	rs17454991	7	78465922	A	T	0.3608	1.16E-07	0.672152	-0.5518	0.10413	1	rs17389497	7	MAGI2	0	intronic	6.42	7
7:78466130:C:G	rs17389497	7	78466130	C	G	0.3608	1.00E-07	0.667014	-0.54598	0.102515	1	rs17389497	7	MAGI2	0	intronic	0.393	7
7:78468394:G:T	rs2052405	7	78468394	T	G	0.4433	0.000239	0.764781	-0.35771	0.097361	0.679628	rs17389497	7	MAGI2	0	intronic	3.315	7
7:78473591:C:T	rs961928	7	78473591	C	T	0.4066	4.02E-07	0.679518	-0.50212	0.099078	0.820285	rs17389497	7	MAGI2	0	intronic	0.227	7
7:78474408:C:T	rs1330495	7	78474408	C	T	0.3509	1.74E-07	0.672118	-0.5385	0.103061	0.957009	rs17389497	7	MAGI2	0	intronic	6.672	5
7:78480565:A:AG	rs11314712	7	78480565	AG	A	0.4066	NA	NA	NA	NA	0.820285	rs17389497	7	MAGI2	0	intronic	8.741	NA
7:78481391:A:G	rs11770408	7	78481391	G	A	0.3519	4.03E-07	0.675496	-0.52141	0.10289	0.952626	rs17389497	7	MAGI2	0	intronic	2.54	6
7:78497786:C:T	rs4730552	7	78497786	T	C	0.3598	5.00E-07	0.668744	-0.49632	0.098744	0.995614	rs17389497	7	MAGI2	0	intronic	5.948	7
7:78506474:A:G	rs12669340	7	78506474	G	A	0.3509	3.02E-06	0.685782	-0.47635	0.102006	0.923168	rs17389497	7	MAGI2	0	intronic	1.241	7
7:78511912:C:T	rs10260996	7	78511912	T	C	0.3956	7.20E-05	0.735529	-0.3854	0.097086	0.774126	rs17389497	7	MAGI2	0	intronic	2.342	6

7:78513820:A:G	rs7788393	7	78513820	G	A	0.3956	0.000116	0.741687	-0.37517	0.097343	0.774126	rs17389497	7	MAGI2	0	intronic	0.065	6
7:78514198:A:T	rs34542090	7	78514198	T	A	0.3579	2.36E-06	0.682286	-0.482	0.102117	0.902077	rs17389497	7	MAGI2	0	intronic	2.312	7
7:78517379:C:T	rs2080724	7	78517379	C	T	0.4085	3.13E-05	0.728472	-0.40281	0.096742	0.640827	rs17389497	7	MAGI2	0	intronic	1.237	5
7:78518132:C:G	rs6949646	7	78518132	G	C	0.4145	NA	NA	NA	NA	0.613701	rs17389497	7	MAGI2	0	intronic	0.827	6
7:78518191:C:T	rs112160495	7	78518191	C	T	0.4145	NA	NA	NA	NA	0.613701	rs17389497	7	MAGI2	0	intronic	0.196	7
7:78518420:C:T	rs1642898	7	78518420	T	C	0.4145	6.36E-05	1.34736	0.383429	0.095882	0.613701	rs17389497	7	MAGI2	0	intronic	1.906	6
7:78519102:A:G	rs1642902	7	78519102	A	G	0.4145	NA	NA	NA	NA	0.613701	rs17389497	7	MAGI2	0	intronic	1.263	7
7:78519111:C:T	rs1642903	7	78519111	C	T	0.4145	NA	NA	NA	NA	0.613701	rs17389497	7	MAGI2	0	intronic	2.085	7
7:78519306:C:T	rs150064128	7	78519306	T	C	0.4145	NA	NA	NA	NA	0.613701	rs17389497	7	MAGI2	0	intronic	0.115	7
7:78519949:A:T	rs191224678	7	78519949	T	A	0.4145	NA	NA	NA	NA	0.613701	rs17389497	7	MAGI2	0	intronic	1.163	7
7:78520049:C:T	rs113606840	7	78520049	C	T	0.4145	NA	NA	NA	NA	0.613701	rs17389497	7	MAGI2	0	intronic	4.365	7
7:78520787:A:C	rs4534082	7	78520787	A	C	0.4145	NA	NA	NA	NA	0.613701	rs17389497	7	MAGI2	0	intronic	0.792	6
7:78522482:A:C	rs202116580	7	78522482	C	A	0.4175	NA	NA	NA	NA	0.605809	rs17389497	7	MAGI2	0	intronic	2.303	6
7:78522489:C:T	rs113911324	7	78522489	T	C	0.4175	NA	NA	NA	NA	0.605809	rs17389497	7	MAGI2	0	intronic	1.387	6
7:78522616:A:G	rs6970742	7	78522616	A	G	0.4145	NA	NA	NA	NA	0.613701	rs17389497	7	MAGI2	0	intronic	1.257	6
7:78522759:C:T	rs6972051	7	78522759	T	C	0.4145	NA	NA	NA	NA	0.613701	rs17389497	7	MAGI2	0	intronic	7.09	6
7:78522848:A:T	rs6961796	7	78522848	A	T	0.4145	NA	NA	NA	NA	0.613701	rs17389497	7	MAGI2	0	intronic	3.271	6

7:78522889:A:C	rs12538009	7	78522889	A	C	0.4145	NA	NA	NA	NA	0.613701	rs17389497	7	MAGI2	0	intronic	5.045	7
7:78523544:C:T	rs1642904	7	78523544	T	C	0.4155	5.77E-05	1.34834	0.386109	0.096002	0.612104	rs17389497	7	MAGI2	0	intronic	2.172	7
7:78524295:C:G	rs1799019	7	78524295	G	C	0.4155	6.85E-05	1.34491	0.381984	0.095941	0.612104	rs17389497	7	MAGI2	0	intronic	2.507	5
7:78524370:C:G	rs4727776	7	78524370	C	G	0.3658	5.74E-07	0.667498	-0.50517	0.101039	0.789852	rs17389497	7	MAGI2	0	intronic	2.119	5
7:78524614:C:T	rs1799017	7	78524614	T	C	0.4155	6.22E-05	1.35833	0.377871	0.094366	0.612104	rs17389497	7	MAGI2	0	intronic	4.961	NA
7:78524726:C:T	rs1799016	7	78524726	C	T	0.4155	5.35E-05	1.35085	0.387568	0.09594	0.612104	rs17389497	7	MAGI2	0	intronic	0.218	NA
7:78524838:T:TA	rs5885100	7	78524838	TA	T	0.4185	NA	NA	NA	NA	0.602679	rs17389497	7	MAGI2	0	intronic	8.167	NA
7:78525163:A:T	rs1642905	7	78525163	T	A	0.4155	6.75E-05	1.34569	0.382297	0.095938	0.612104	rs17389497	7	MAGI2	0	intronic	9.004	6
7:78526104:C:G	rs1642906	7	78526104	C	G	0.4155	6.75E-05	1.34572	0.382316	0.095938	0.612104	rs17389497	7	MAGI2	0	intronic	0.265	6
7:78527758:C:G	rs1642908	7	78527758	C	G	0.4155	5.60E-05	1.35281	0.386865	0.096021	0.612104	rs17389497	7	MAGI2	0	intronic	0.282	7
7:78527994:A:G	rs1642909	7	78527994	G	A	0.4235	6.77E-05	1.35109	0.380308	0.095454	0.646321	rs17389497	7	MAGI2	0	intronic	0.056	7
7:78528140:C:T	rs1799012	7	78528140	T	C	0.4235	6.87E-05	1.35077	0.379946	0.095443	0.646321	rs17389497	7	MAGI2	0	intronic	6.207	NA
7:78529998:C:T	rs4255073	7	78529998	C	T	0.4105	3.42E-05	0.728053	-0.401	0.096777	0.630306	rs17389497	7	MAGI2	0	intronic	3.275	7
7:78530733:G:T	rs10233975	7	78530733	T	G	0.4155	3.34E-05	0.728021	-0.40154	0.096787	0.631013	rs17389497	7	MAGI2:MAGI2-IT1	00:00	ncRNA_exonic	2.219	7
7:78530979:C:T	rs2052407	7	78530979	T	C	0.4245	NA	NA	NA	NA	0.66274	rs17389497	7	MAGI2:MAGI2-IT1	00:00	ncRNA_intronic	0.822	5
7:78531682:A:G	rs1799008	7	78531682	G	A	0.4195	9.07E-05	1.34256	0.375763	0.095997	0.60277	rs17389497	7	MAGI2:MAGI2-IT1	00:00	ncRNA_intronic	8.89	NA

7:78531705:A:G	rs4730561	7	78531705	A	G	0.3936	1.19E-05	0.71839	-0.43423	0.099173	0.699507	rs17389497	7	MAGI2:MAGI2-IT1	00:00	ncRNA_intronic	4.283	5
7:78532462:A:C	rs60317714	7	78532462	A	C	0.3658	9.79E-07	0.699878	-0.50784	0.10373	0.600068	rs17389497	7	MAGI2:MAGI2-IT1	00:00	ncRNA_intronic	1.48	3a
8:21500728:G:T	rs10503711	8	21500728	T	G	0.4483	1.08E-05	0.732043	-0.41947	0.095315	0.962317	rs12549783	8	GFRA2	47186	intergenic	1.148	7
8:21501464:C:T	rs6995510	8	21501464	T	C	0.4463	1.21E-05	0.734649	-0.41821	0.095569	0.962007	rs12549783	8	GFRA2	46450	intergenic	0.7	7
8:21501471:A:G	rs62492192	8	21501471	G	A	0.4394	1.06E-05	0.729716	-0.42303	0.096059	0.988452	rs12549783	8	GFRA2	46443	intergenic	1.154	7
8:21502545:G:T	rs12549783	8	21502545	G	T	0.4384	6.08E-06	1.37607	0.438173	0.096868	1	rs12549783	8	GFRA2	45369	intergenic	0.162	7
8:21505259:A:T	rs4134428	8	21505259	T	A	0.4493	9.31E-06	1.3682	0.427379	0.096417	0.95829	rs12549783	8	GFRA2	42655	intergenic	7.242	7
8:21506198:A:T	rs13274654	8	21506198	A	T	0.4702	0.000239	1.29962	0.352977	0.09607	0.871865	rs12549783	8	GFRA2	41716	intergenic	12.22	5
8:21510088:A:T	rs17427734	8	21510088	A	T	0.4006	1.32E-05	0.728542	-0.41905	0.096199	0.84431	rs12549783	8	GFRA2	37826	intergenic	1.691	6
8:21510244:C:G	rs17499620	8	21510244	G	C	0.4085	1.11E-05	0.730898	-0.41985	0.095552	0.817949	rs12549783	8	GFRA2	37670	intergenic	0.86	7
8:21510700:A:T	rs12681966	8	21510700	T	A	0.4016	1.03E-05	0.726681	-0.42338	0.096	0.839988	rs12549783	8	GFRA2	37214	intergenic	0.985	5
8:21510901:C:T	rs12679408	8	21510901	C	T	0.4085	1.23E-05	0.734201	-0.41766	0.09553	0.817949	rs12549783	8	GFRA2	37013	intergenic	1.155	6
8:21512427:A:G	rs1383573	8	21512427	G	A	0.4493	1.53E-05	1.38886	0.395897	0.091539	0.958429	rs12549783	8	GFRA2	35487	intergenic	0.313	6

8:21513913:A:G	rs11988772	8	21513913	G	A	0.4871	0.000316	1.29919	0.347415	0.09646	0.815202	rs12549783	8	GFRA2	34001	intergenic	0.672	5
9:90139505:C:T	rs79863888	9	90139505	C	T	0.1272	8.95E-06	0.64694	-0.62571	0.140888	0.907278	rs4878090	9	DAPK1	0	intronic	5.558	5
9:90139865:C:T	rs77962276	9	90139865	C	T	0.1272	8.66E-06	0.646341	-0.62643	0.140826	0.907278	rs4878090	9	DAPK1	0	intronic	1.301	6
9:90140925:A:G	rs4877361	9	90140925	A	G	0.1372	1.13E-05	0.666472	-0.59335	0.135127	0.984472	rs4878090	9	DAPK1	0	intronic	0.38	6
9:90141002:C:T	rs4877362	9	90141002	C	T	0.1372	1.37E-05	0.66659	-0.58709	0.135003	0.984472	rs4878090	9	DAPK1	0	intronic	3.766	7
9:90141193:A:G	rs4878088	9	90141193	G	A	0.1362	1.05E-05	0.664452	-0.59554	0.135177	0.992188	rs4878090	9	DAPK1	0	intronic	4.305	6
9:90141324:C:G	rs12685665	9	90141324	G	C	0.1402	1.38E-05	0.666878	-0.58678	0.135002	0.961877	rs4878090	9	DAPK1	0	intronic	0.157	5
9:90143089:A:C	rs17477827	9	90143089	C	A	0.1352	1.14E-05	0.666646	-0.59309	0.135141	1	rs4878090	9	DAPK1	0	intronic	1.721	7
9:90143928:A:G	rs4878090	9	90143928	A	G	0.1352	7.83E-06	0.662386	-0.60434	0.135207	1	rs4878090	9	DAPK1	0	intronic	3.422	5
9:90144609:C:T	rs17399459	9	90144609	C	T	0.1352	1.38E-05	0.670964	-0.58519	0.134624	1	rs4878090	9	DAPK1	0	intronic	4.399	5
9:90146173:C:T	rs4878093	9	90146173	C	T	0.1352	1.50E-05	0.662776	-0.56357	0.130189	1	rs4878090	9	DAPK1	0	intronic	2.915	7
9:90148464:A:G	rs12686192	9	90148464	G	A	0.1382	2.01E-05	0.668161	-0.55391	0.12991	0.976849	rs4878090	9	DAPK1	0	intronic	0.548	4

9:90148887:C:T	rs12686443	9	90148887	C	T	0.1382	2.01E-05	0.668162	-0.55389	0.129909	0.976849	rs4878090	9	DAPK1	0	intronic	4.584	7
9:90154407:C:T	rs55873303	9	90154407	T	C	0.1392	NA	NA	NA	NA	0.953606	rs4878090	9	DAPK1	0	intronic	2.985	7
9:90154789:A:T	rs77992676	9	90154789	A	T	0.1362	1.73E-05	0.663085	-0.55928	0.130166	0.976469	rs4878090	9	DAPK1	0	intronic	0.131	6
9:90156846:C:T	rs1041326	9	90156846	C	T	0.1392	2.65E-05	0.669851	-0.54388	0.129431	0.938306	rs4878090	9	DAPK1	0	intronic	5.404	3a
9:90158881:A:G	rs78445378	9	90158881	A	G	0.1372	2.59E-05	0.669893	-0.54521	0.129606	0.984472	rs4878090	9	DAPK1	0	intronic	0.024	6
9:90159340:A:C	rs12685761	9	90159340	A	C	0.1372	2.57E-05	0.669648	-0.54535	0.129591	0.984472	rs4878090	9	DAPK1	0	intronic	0.319	7
9:90162495:A:T	rs2378748	9	90162495	T	A	0.1571	NA	NA	NA	NA	0.840898	rs4878090	9	DAPK1	0	intronic	0.315	7
9:90162497:A:T	rs2378749	9	90162497	T	A	0.1153	NA	NA	NA	NA	0.835729	rs4878090	9	DAPK1	0	intronic	0.035	7
9:90164559:A:G	rs12683332	9	90164559	A	G	0.1362	2.86E-05	0.670238	-0.53966	0.128965	0.992188	rs4878090	9	DAPK1	0	intronic	6.674	7
9:90170834:C:T	rs12685588	9	90170834	C	T	0.1362	3.32E-05	0.671565	-0.53182	0.128131	0.992188	rs4878090	9	DAPK1	0	intronic	3.233	5
9:90171415:C:T	rs7038971	9	90171415	C	T	0.1342	2.61E-05	0.666462	-0.53901	0.128184	0.976441	rs4878090	9	DAPK1	0	intronic	7.212	4
9:90172040:A:C	rs17479142	9	90172040	C	A	0.1322	3.38E-05	0.668777	-0.52405	0.126384	0.960881	rs4878090	9	DAPK1	0	intronic	2.313	4
9:90172917:A:G	rs57196281	9	90172917	A	G	0.1332	4.49E-05	0.676183	-0.5218	0.127866	0.968641	rs4878090	9	DAPK1	0	intronic	4.108	5

9:90172939:A:T	rs57813441	9	90172939	T	A	0.1332	4.49E-05	0.676221	-0.52176	0.127861	0.968641	rs4878090	9	DAPK1	0	intronic	6.521	4
9:90173374:A:G	rs17479184	9	90173374	A	G	0.1332	4.50E-05	0.67627	-0.52173	0.127869	0.968641	rs4878090	9	DAPK1	0	intronic	3.306	5
9:90174759:A:G	rs75337578	9	90174759	A	G	0.1322	0.000137	0.69473	-0.48952	0.128347	0.960844	rs4878090	9	DAPK1	0	intronic	4.013	4
9:90176617:A:G	rs76796844	9	90176617	G	A	0.1322	0.000114	0.689997	-0.49563	0.128414	0.960844	rs4878090	9	DAPK1	0	intronic	6.24	5
9:90176981:A:G	rs74884476	9	90176981	G	A	0.1322	0.000151	0.696536	-0.48565	0.128159	0.960844	rs4878090	9	DAPK1	0	intronic	1.042	5
9:90178806:C:G	rs928114	9	90178806	C	G	0.1312	0.000114	0.689997	-0.49563	0.128414	0.953091	rs4878090	9	DAPK1	0	intronic	1.587	5
9:90183188:A:G	rs36233193	9	90183188	A	G	0.1322	3.89E-05	0.673093	-0.52657	0.127996	0.960881	rs4878090	9	DAPK1	0	intronic	6.504	5
9:90184270:A:G	rs34006274	9	90184270	G	A	0.1541	0.001016	0.750153	-0.39087	0.118952	0.804615	rs4878090	9	DAPK1	0	intronic	1.93	4
9:90184967:C:T	rs36203437	9	90184967	C	T	0.1322	4.01E-05	0.673921	-0.52596	0.128062	0.960881	rs4878090	9	DAPK1	0	intronic	4.838	4
9:90185139:C:G	rs36203441	9	90185139	C	G	0.1322	5.02E-05	0.678791	-0.51884	0.127955	0.960881	rs4878090	9	DAPK1	0	intronic	12.93	2b
9:90186711:C:G	rs4878103	9	90186711	C	G	0.1342	6.36E-05	0.682954	-0.51128	0.12785	0.960818	rs4878090	9	DAPK1	0	intronic	0.935	5
9:90194227:A:G	rs12685116	9	90194227	G	A	0.1461	0.001041	0.766722	-0.42007	0.1281	0.794343	rs4878090	9	DAPK1	0	intronic	12.55	5

9:90194242:A:C	rs12683054	9	90194242	A	C	0.1272	0.000106	0.719045	-0.52788	0.136205	0.922472	rs4878090	9	DAPK1	0	intronic	9.892	5
9:90194645:A:G	rs12685372	9	90194645	G	A	0.1461	0.001025	0.766376	-0.42054	0.128069	0.794343	rs4878090	9	DAPK1	0	intronic	3.241	6
9:90195542:A:G	rs36204768	9	90195542	A	G	0.1282	0.000131	0.724112	-0.52029	0.136032	0.930068	rs4878090	9	DAPK1	0	intronic	1.724	6
9:90197840:A:G	rs78934970	9	90197840	A	G	0.1252	0.000179	0.72543	-0.51986	0.138739	0.892326	rs4878090	9	DAPK1	0	intronic	7.296	5
9:106275350:C:T	rs145244544	9	1.06E+08	C	T	0.1143	0.001941	1.2531	0.43027	0.138834	0.658791	rs79091275	10	RP11-436F21.1	83175	intergenic	0.117	6
9:106313968:A:G	rs79091275	9	1.06E+08	A	G	0.0825	5.94E-06	1.51502	0.720836	0.159178	1	rs79091275	10	RP11-436F21.1	121793	intergenic	1.143	7
9:106336114:A:G	rs74932565	9	1.06E+08	A	G	0.0825	2.28E-05	1.47754	0.725909	0.171368	0.923891	rs79091275	10	RNA5SP291	115640	intergenic	0.215	6
9:106377751:A:G	rs79746057	9	1.06E+08	A	G	0.1163	0.000638	1.29146	0.46559	0.136338	0.664101	rs79091275	10	RNA5SP291	74003	intergenic	1.787	7
9:106423369:A:G	rs79895321	9	1.06E+08	G	A	0.07952	3.01E-05	1.3961	0.708984	0.169898	0.811489	rs79091275	10	RNA5SP291	28385	intergenic	2.89	3a
9:106429018:A:G	rs77728130	9	1.06E+08	G	A	0.07952	3.93E-05	1.38406	0.699478	0.170128	0.811489	rs79091275	10	RNA5SP291	22736	intergenic	10.77	7
9:133023677:A:G	rs79366318	9	1.33E+08	G	A	0.0497	1.91E-06	2.46274	1.0697	0.224608	1	rs79366318	11	HMCN2	23204	intergenic	7.269	5
10:5551233:G:T	rs35859955	10	5551233	T	G	0.2256	6.99E-06	0.683634	-0.50859	0.113167	1	rs35859955	12	CALML3-AS1	4973	intergenic	0.177	4
10:21056587:A:C	rs4747421	10	21056587	C	A	0.1034	3.09E-05	0.665382	-0.52603	0.126251	0.869691	rs12244585	13	NEBL	12314	intergenic	1.073	7
10:21064851:A:G	rs10828130	10	21064851	G	A	0.1282	0.001256	0.730439	-0.39108	0.121231	0.706133	rs12244585	13	NEBL	4050	intergenic	5.515	6

10:21078477:A:G	rs12244585	10	21078477	G	A	0.1093	7.50E-06	0.652192	-0.56167	0.125398	1	rs12244585	13	NEBL	0	intronic	2.841	7
11:81264068:C:G	rs77234487	11	81264068	C	G	0.0328	0.002675	0.608593	-0.67691	0.225421	0.744897	rs80240928	14	RP11-664H7.1	509	downstream	11.91	6
11:81366365:C:G	rs76323555	11	81366365	G	C	0.02485	6.36E-06	0.428993	-1.17093	0.259394	1	rs80240928	14	RP11-664H7.2	98897	intergenic	1.634	7
11:81372090:C:T	rs80240928	11	81372090	C	T	0.02485	5.96E-06	0.427131	-1.17317	0.259103	1	rs80240928	14	RP11-664H7.2	104622	intergenic	1.422	7
12:31469921:A:G	rs67228087	12	31469921	G	A	0.06461	8.91E-06	0.518916	-0.8586	0.193287	1	rs67228087	15	FAM60A	0	intronic	4.093	3a
15:101658292:C:T	rs12898233	15	1.02E+08	C	T	0.05567	8.02E-06	1.89751	1.03855	0.232612	1	rs12898233	16	RP11-424I19.1	540	downstream	0.207	7
15:101658669:A:C	rs12437607	15	1.02E+08	A	C	0.05567	1.20E-05	1.80744	1.01151	0.231086	1	rs12898233	16	RP11-424I19.1	917	downstream	0.016	5
15:101661778:G:T	rs60398588	15	1.02E+08	G	T	0.05467	NA	NA	NA	NA	0.982107	rs12898233	16	RP11-424I19.1	4026	intergenic	1.488	2b
15:101662001:C:T	rs61094855	15	1.02E+08	C	T	0.05268	1.31E-05	1.88188	0.970114	0.222564	0.946558	rs12898233	16	RP11-424I19.1	4249	intergenic	2.189	4
15:101662165:A:T	rs59101697	15	1.02E+08	A	T	0.05268	1.33E-05	1.8807	0.969433	0.222565	0.946558	rs12898233	16	RP11-424I19.1	4413	intergenic	2.934	2b
15:101662626:A:C	rs2412070	15	1.02E+08	A	C	0.05467	3.76E-05	0.711731	-0.91044	0.220894	0.911825	rs12898233	16	RP11-424I19.1	4874	intergenic	2.93	NA
15:101664912:T:TG	rs71458732	15	1.02E+08	TG	T	0.05268	NA	NA	NA	NA	0.911241	rs12898233	16	RP11-424I19.1	7160	intergenic	0.219	NA
15:101664914:A:T	rs201171595	15	1.02E+08	A	T	0.05268	NA	NA	NA	NA	0.911241	rs12898233	16	RP11-424I19.1	7162	intergenic	1.332	5
15:101665261:C:T	rs12913337	15	1.02E+08	C	T	0.05169	1.68E-05	1.8308	0.953494	0.221537	0.9289	rs12898233	16	RP11-424I19.1	7509	intergenic	1.286	4
15:101667226:C:T	rs4965789	15	1.02E+08	T	C	0.05268	6.33E-05	0.729853	-0.87841	0.219597	0.911241	rs12898233	16	RP11-424I19.1	9474	intergenic	5.31	4

19:35610725:C:G	rs4806091	19	35610725	C	G	0.06759	8.71E-08	0.536598	-0.93973	0.175588	1	rs4806091	17	FXYD3	0	intronic	0.28	4

8.4 Appendix 3 – Lung Cancer Risk Survey

Participant ID ___________

Lung Cancer Risk Survey

1) What do you think the risk of developing lung cancer is **for any person** in the general population of the UK? *Please circle one answer.*
   a) 1 chance in 2 (50%)
   b) 1 chance in 4 (25%)
   c) 1 chance in 8 (13%)
   d) 1 chance in 14 (7%)
   e) 1 chance in 30 (3%)
   f) 1 chance in 100 (1%)

2) In an imaginary room filled with 100 people *exactly like you*, how many do you think will get lung cancer in the next 6 years? *Please fill in the blank.*

   _____out of 100

3) What do you think your chances of developing lung cancer are compared to other people your age in the UK population? *Please circle one answer.*
   a) Much less likely
   b) Less likely
   c) The same
   d) More likely
   e) Much more likely

*If you are an **ex-smoker**, please **skip** question 4:*

4) What do you think your chances of developing lung cancer are compared to other smokers? *Please circle one answer.*
   a) Much less likely
   b) Less likely
   c) The same
   d) More likely
   e) Much more likely

Participant ID ___________

*If you are a **current smoker**, please **skip** question 5:*

5) What do you think your chances of developing lung cancer are compared to other ex-smokers?
   a) Much less likely
   b) Less likely
   c) The same
   d) More likely
   e) Much more likely

6) Have you ever been told a score or percentage indicating your personal risk of lung cancer? *Please circle one answer.*
   a) No
   b) Yes – by my GP
   c) Yes – I did an online test
   d) Yes – Another source. *Please write what this was:* _________
   e) I'm not sure

7) Do you think these factors *increase* a person's risk of lung cancer?
   *Please circle **Yes** or **No** or **Not Sure***
   a) Age - **Yes/No/Not Sure**
   b) Air Pollution - **Yes/No/Not Sure**
   c) Asbestos Exposure - **Yes/No/Not Sure**
   d) Asthma - **Yes/No/Not Sure**
   e) Being Overweight - **Yes/No/Not Sure**
   f) Changes or mutations in your genes - **Yes/No/Not Sure**
   g) Chronic Obstructive Pulmonary Disease (COPD) - **Yes/No/Not Sure**
   h) Family History of Lung Cancer - **Yes/No/Not Sure**
   i) Poor Diet - **Yes/No/Not Sure**
   j) Radon Exposure **- Yes/No/Not Sure**
   k) Smoking – **Yes/No/Not Sure**
   l) Second-hand smoke - **Yes/No/Not Sure**

Participant ID ___________

8) Think about **all** the people who are diagnosed with lung cancer in the UK in one year.

d) What percentage do you think will survive for 1 year after diagnosis?
_____%

e) What percentage do you think will survive for 5 years after diagnosis?
_____%

f) What percentage do you think will survive for 10 years after diagnosis?
_____%

9) In the last two weeks, how often have you worried about developing lung cancer? *Please circle one answer.*

a) Not at all
b) Several days
c) More than half the days
d) Nearly every day

10) How much does any worry about getting lung cancer impact your mood
(cause you distress or upset)? Pl*ease circle one answer.*

a) Not at all
b) A small amount
c) Quite a lot
d) A lot

11)

Over the last 2 weeks, how often have you been bothered by the following problems? *(Use "✓" to indicate your answer)*	Not at all	Several days	More than half the days	Nearly every day
1. Feeling nervous, anxious or on edge	0	1	2	3
2. Not being able to stop or control worrying	0	1	2	3
3. Little interest or pleasure in doing things	0	1	2	3

4. Feeling down, depressed, or hopeless	0	1	2	3

If anything in this questionnaire has caused you concern, please discuss it with the nurse during your Lung Health Check.

Thank you for participating in our research!