Supplementary information

Polygenic basis and biomedical consequences of telomere length variation

In the format provided by the authors and unedited

Polygenic basis and biomedical consequences of telomere length variation

Supplementary Information

V. Codd et al.

Index

Investigation of rs334 and rs1609812

The Hgbu and Hgbd primers used for the multiplex qPCR telomere length (TL) measurement[15] bind within the *HBB* gene, where rs334 is located. Aligning the primer sequences to the genomic sequence with annotation of known SNPs using the UCSC genome browser (**Supplementary Figure 2**) reveals that rs334 is located 2bp from the 5' end of the reverse primer (Hgbd). Although the 5' end of primer sequences is less critical for primer binding, these primers also carry a large GC rich non-binding tail. Therefore reduction of the primer binding capacity from 18bp to 16bp, when there is already a large tail, could reduce efficiency of the reaction, leading to underestimation of the single copy repeat number calculated (S). Underestimating (S) would subsequently inflate the T/S ratio used as the leucocyte TL (LTL) measurement. The association we observed of the rs334 minor allele with longer LTL is consistent with this hypothesis.

We identified measurement plates that contained carriers of the minor allele for the SNP being analysed and extracted mean T and S values for all individuals on these plates. We compared carriers to non-carriers by fitting linear mixed models, fitting plate as a random effect to account for potential technical variation (our previous adjustments being at the level of the T/S ratio). We restricted this analysis to participants of black ethnicity. When analysing S measurements no further adjustments were made as this is a measure of DNA input into the assay and should not vary between individuals. For T measurement analyses, adjustments were made for age and sex to account for known biological variation.

An additional sentinel variant, rs1609812, was identified in the same locus as rs334. We tested this SNP as for rs334, however unlike rs334, rs1609812 is not restricted to non-European ancestries, so we adjusted the T measurement analysis for ethnicity also. To investigate this association further, we excluded rs334 and any participant with a diagnosis of thalassaemia, as other thalassaemia mutations in this region are located within the primer sequences (**Supplementary Figure 2**), and re-tested the association of rs1609812 with LTL. The original association (Beta=0.043, P=4.88x10^{-53}) is unaffected by rs334 carrier removal (Beta=0.044, P=5.32x10^{-54}) using the data prior to conditional analyses.

To investigate associations of rs1609812 with LTL further (particularly given the contrast between the strong association observed in our study and the lack of evidence from prior publications), we examined two separate large datasets, neither of which showed significant

associations (P=0.04 and P=0.01 in Li et al., 2020[3] and Dorajoo et al., 2019[4] respectively). Both previous studies used different single copy genes in the TL assay (36B4 and albumin).

Next, we specifically investigated whether rs334 or rs1609812 impacted the relative single copy gene copy number (S), suggesting a technical artefact, or the relative telomere copy number (T), suggesting a true association. Comparing carriers of rs334 (N=1015) to non-carriers (N=1576), carriers had significantly lower S values (beta -0.090 (-0.114, -0.067), p=2.4x10^{-13}) but no significant difference in T values (beta -0.006 (-0.027, 0.014), p=0.55). Separating carriers into heterozygotes (TA, N=999) and homozygotes (AA, N=16) shows an additive effect of the A allele on the relative S value (TA beta -0.089 (-0.113, -0.064), p=6.0x10^{-13}; AA beta -0.163 (-0.324, 0.001), p=0.048) compared to non-carriers. This supports the hypothesis that rs334 reduces the efficiency of primer binding, leading to lower S values and consequently elevated T/S ratios in rs334 carriers. Performing a similar analysis for rs1609812, we observed higher S values in carriers of the minor allele, consistent with this allele appearing to be associated with shorter LTL (**Supplementary Table 1**). Again this was approximately additive in effect (GA beta 0.007 (0.005, 0.009), p=9.0x10^{-16}; GG beta 0.011 (0.007, 0.016), p=1.3x10^{-6}). No effect was observed for the T measurement (GA beta -0.001 (-0.002, 0.001), p=0.27; GG beta -0.001 (-0.004, 0.002), p=0.61). Removal of rs334 carriers from this analysis did not change the effect estimates. Similarly, repeating the analysis of rs334 with rs1609812 minor allele carriers removed did not alter the findings for rs334. This data suggests that whilst rs1609812 is independent of rs334, the observed association with LTL is also false, potentially with a high LD SNP to rs1609812 potentially influencing the S measurement in a similar fashion to rs334.

Finally, we sought to investigate whether rs334 impacted the observed association between black ethnicity and longer LTL. Comparing LTL between participants of black and white ethnicities, LTL was significantly longer in black participants (beta 0.414 (0.391, 0.437), p=6.0x10^{-268}). Removal of rs334 carriers attenuated this effect, but LTL remained significantly longer in black participants (beta 0.349 (0.324, 0.374), p=1.0x10^{-167}).

Identifying conditionally independent variants associated with LTL

To identify independently associated variants at each locus we used GCTA (v1.25.2) .Starting with the most significantly associated SNPs, the model then adds SNPs iteratively in a forward stepwise manner, calculating conditional P-values for all SNPs. LD estimation within GCTA was estimated using a subset of 50,000 samples randomly selected from the UKB. Where a SNP shows evidence of collinearity with another ($R^2>0.9$), the conditional P-value is set to 1. This process is repeated until no further SNPs can be added to the model.

Variant annotation

Using VEP[54] gives information on the location of SNPs with respect to genes alongside several measures of functional consequence. Overall, each variant is placed into one of four impact categories. High impact variants include those predicted to ablate or truncate transcripts, alter the reading frame or alter splicing. Moderate variants are mostly predicted to alter protein sequence through addition, loss or change in amino acids, although this class also includes loss of regulatory region variants. Low impact variants include sequence changes that are unlikely to alter function (synonymous changes, start/stop retained) and modifier variants include those that may alter expression levels. CADD scores are returned for all variants and SIFT and PolyPhen annotations are given for variants that lead to amino acid changes. To aid interpretation all results were filtered to the canonical transcript of each gene.

Identifying likely-causal genes within conditionally independent GWS loci

To identify potential causal variants at each locus, we performed fine-mapping using a two-step approach in FINEMAP that allowed for more than one causal SNP, annotating the variants in the 95% credible sets to prioritise those with likely functional impact at loci that contained more than one independent sentinel (**Methods, Supplementary Table 2**). We then used the credible sets (a set of SNPs that have a 95% probability of containing the casual SNP, **Supplementary Table 2**) to identify the genes that are most likely influenced by the LTL associated variants. We first functionally annotated the credible set SNPs to identify those variants that are most likely to have an impact on gene function (**Methods**). In total, 72 variants, including 24 identified as being the most likely causal variant within a locus, were annotated as having protein altering consequences (**Supplementary Table 3**). These were

distributed across 44 loci, with the majority highlighting a single gene per locus. One exception is the RTEL1 locus, where four likely-causal SNPs are all located within both *RTEL1* and the *RTEL1-TNFRSF6B* readthrough transcript. However, these are most likely to only have a functional consequence within RTEL1 as the *RTEL1-TNFRSF6B* transcript is annotated as likely undergoing nonsense mediated decay and therefore not translated. Several genes contained more than one coding variant, most notably *SLX4* which contained seven, although only one was annotated as being damaging across multiple prediction tools (**Supplementary Table 3**).

To identify variants acting through altered gene expression, we searched for evidence of a shared causal variant influencing both LTL and gene expression using eQTL data from 48 tissues from the GTEx Consortium (**Methods**, **Supplementary Table 4**). We found evidence of colocalising eQTLs at 93 (68%) of the loci (**Supplementary Tables 4** and **5**). For several genes, there was strong evidence of a colocalised eQTL across a number of different tissues. Among the most striking were *RPA1*, *TEN1*, *POLI* and *AC011498*.6, which showed evidence of strong, ubiquitous colocalisation (**Supplementary Table 4**, **Figure 2**). There were also tissue specific signals for *IRF5, MPHOSPH6, AC007292.1, AC145285.2, LONP2, MCM4, CDK3, PPP1CB* and *TRMT61B*. *PPP1CB* and *TRMT61B* are within the same locus and show similar levels of colocalisation across different tissues. *AC011498*.6 and *AC007292*.1 are also within the same locus, as are *TEN1* and *CDK3*, with the strength of evidence pointing more towards the gene in the ubiquitous grouping (*TEN1* and *AC011498*.6).

Although many components of SHELTERIN and telomerase were within our conditionally independent GWS loci, not all were found. The two remaining SHELTERIN components *TINF2* and *TERF2IP* along with two further core telomerase genes *NOP10* and *GAR1* are within loci identified at <1%FDR (rs34354104, rs6564260, rs2615358 and rs73839158 respectively, **Supplementary Table 2**), Variants around *TINF2* have been associated with LTL previously, but only in a Singaporean Chinese population[5]. We also saw the previous *MOB1B* locus within the <1%FDR[4]. Although we believe there are loci genuinely associated with LTL within the FDR, this list may also contain more false-positive results than the estimated 1% due to the large number of variants being tested and the influence of LD on the FDR approach. We therefore have not undertaken further analyses on these variants at this stage and advise that this list is interpreted with caution.

Analysis of rare and ultra-rare variants in gene based tests

To understand the direction of association of genes associated with LTL in our gene based tests, we assessed the association of the contributing individual variants in single variant analyses. Where individual variants within a gene differed in direction of effect we see consistency with the effect of the gene score for variants most likely to be true loss of function (LoF, truncating or frameshift variants in early exons) and where minor allele counts are >10 (where estimates of individual variant effects are likely to be more reliable, **Supplementary Table 8**).

Variant annotation was extracted from VEP using the LOFTEE plugin[54,60] Variants were categorised into four groups and scored based on their predicted function. Firstly, ultra-rare potential inactivating variants (MAF<0.01%), i.e. those that were predicted with high confidence to truncate (STOP gain) or radically alter the protein sequence (splice site, frameshift) by the Loss-Of-Function Transcript Effect Estimator (LOFTEE) algorithm, were assigned a score of 1. This corresponds to a complete inactivation of the allele. The same score was applied to rare variants (MAF<0.1%) that were predicted to be pathogenic or likely pathogenic in ClinVar[79] (downloaded 17/12/2020), using the clinical significance annotation ("clinsig" field), where there was no conflicting annotation ('benign,' 'likely benign,' or 'uncertain significance').

We then incorporated ultra-rare variants with a predicted cryptic splice site by SpliceAI a score of 0.75, based on previous literature[80]. The SpliceAI algorithm, using default parameters, was used to estimate four delta scores, DS_AG (acceptor gain delta score), DS_AL (acceptor loss delta score), DS_DG (donor gain delta score), and DS_DL (donor loss delta score). Each of these scores has values ranging from 0 to 1. Variants with a value of any of these four scores ≥0.5 were designated as a cryptic splice site and assigned the score of 0.75.

To increase statistical power we also integrated ultra-rare missense mutations that were predicted to be damaging by all five computational algorithms used (SIFT, Polyphen2-HDIV, Polyphen2-HVAR, LRT, and MutationTaster) as previously described[81,82.] Annotation differs between these algorithms, for clarity we took the following annotations to be "damaging; damaging (SIFT) possibly or probably damaging (Polyphen2-HDIV, Polyphen2-HVAR), deleterious (LRT) and disease-causing-automatic or disease-causing (MutationTaster). As the effect of single amino acid changes are less likely to be as strong as inactivating variants, these

variants were assigned a gene-specific weight that took into account the cumulative frequencies of high-confidence LoF variants as compared to predicted damaging missense variants as previously described[82-84]. Given the cumulative allele frequency of all of the LOFTEE high confidence rare variants of a gene (G) as f_L, the cumulative allele frequency of all of the predicted damaging missense variants as f_M, the weight for the missense variants was estimated as:

$$\left(\frac{f_L \times (1 - f_L)}{f_M \times (1 - f_M)}\right)^{0.5}$$

For each participant an aggregated gene score was computed based on individual variant weightings with a maximum score of 1.

High confidence truncating or protein altering variants throughout RTEL1 were mostly associated with shorter LTL, consistent with data suggesting that the full length RTEL1 protein is required to facilitate telomere elongation by telomerase[29]. The direction of effect is also consistent with the known biological roles of TERF1, POT1, TERT and PARN. TERF1 and POT1 are both components of SHELTERIN and limit telomere elongation by restricting access of the telomere end to telomerase[72,73]. For both, LoF variants result in longer LTL. Reduction of TERT or PARN results in lower telomerase activity reflecting the shorter LTL effect observed (**Supplementary Table 8)**[24]. Our results suggest reduction in SAMHD1 results in longer LTL. Whilst reduction in SAMHD1 can result in telomere fragility, this has only been shown in the absence of TERF1[32]. A major role of SAMHD1 is in regulating cellular dNTP pools, especially dGTP levels. Loss of SAMHD1 results in elevated dGTP and dGTP levels have been seen to positively correlate with both telomerase activity and TL[74,75].

Within ATM there was less consistency in the direction of effect of single variants (**Supplementary Table 8**). Truncating variants located in early exons associated with both longer and shorter LTL. However, those variants with MAC>10, where we have more confidence in the single variant level data, were consistently associated with shorter LTL, which reflects previous studies showing loss of ATM resulting in short TL[76].

Loss of CTC1 results in diseases characterised as short TL syndromes, however this requires biallelic mutation, presumably resulting in complete LoF of CTC1. Our results suggest that

partial loss of CTC1 can result in moderate LTL lengthening. One possible explanation is that, like SHELTERIN, the CST complex limits telomerase activity at the telomere end and loss results in more accessibility and therefore lengthening[17].

A further 141 genes were nominally significant (p<0.05) and included both candidate genes (*TP53*, *PML*, *WRAP53*, *ACD*), one gene involved in nucleotide metabolism (*TYMS*) and *SMC1B*, a component of the cohesin complex that protects from telomere shortening during meiotic cell division[77].

Association of LTL with biomedical traits

Longer usual LTL was associated with a higher peak expiratory flow rate; there was a similar (but non-significant) association with genetically-determined LTL, suggesting longer telomeres enhance respiratory function (**Extended Data Figure 2, Supplementary Table 10**). A related issue of considerable interest is whether LTL is a determinant of physical fitness. We found associations of usual LTL with different measures of physical activity and with grip strength but no compelling genetic evidence to support a causal relationship (**Extended Data Figure 2, Supplementary Table 10**). Similarly, in women we found associations of longer usual LTL with several traits related to reproductive health, including age of first and last birth and age at menopause, but none of these were associated with genetically-determined LTL (**Extended Data Figure 2, Supplementary Table 10**). With respect to traits that may act as proxies for neurological/cognitive function, we observed an association between longer usual LTL and higher age at completion of education, with genetic evidence supportive of a causal association (**Figure 4**). By contrast, longer usual LTL was associated with prolonged snap button times (a measure of reaction time).

Given interest in associations of LTL with metabolic and endocrine factors that could mediate disease relationships[70], we explored association of LTL with biochemical traits available in UK Biobank (UKB). Genetically-determined LTL was associated with albumin, apolipoprotein A, high density lipoprotein, triglycerides insulin-like growth factor 1 (IGF1), and sex hormone binding globulin (**Figure 4**). In most cases, longer LTL was associated with lower values of these molecules, except for IGF1 and triglycerides.

Disease codes for health episodes statistics and death registry records

We defined 127 outcomes based on incident electronic health records data (primary and secondary ICD-10, ICD-19 and OPCS-4 codes). Precise outcome definitions are given in **Supplementary Table 11**. We grouped these outcomes into 10 broad categories: cancer, cardiovascular disease, digestive diseases, genitourinary diseases, immune/inflammatory diseases, infections, mental illness, musculoskeletal diseases, respiratory diseases, sensory disorders. We generated code lists using available UKB data from both self-reported data (UKB codes for the outcomes obtained from self-reported / verbal interview data gathered at the assessment centre visits) and electronic health records (ICD-9 and ICD-10 codes for use in both hospital episodes statistics (HES) and death registry data, and OPSC-3 and OPSC-4 codes to identify directly linked operations in HES).

A basic HES record is a finished consultant episode of care (the time spent under the care of one consultant). An admission, or spell, is defined as a continuous period of time spent as a patient within a trust and may include more than one episode. We defined first disease occurrence using within-hospital episodes and death records as follows:

Disease definition	ICD-9/10 codes identified from HES	OPCS3/4* operation codes identified from HES	ICD-9/10 codes identified in ONS records
First recorded hospitalisation or death with any diagnosis of disease of interest	Primary and secondary diagnosis during a hospital episode	Operation that is directly linked to the disease of interest	Underlying and related cause of death

*We did not differentiate between main and secondary operations

Definition of prevalent and incident cases

We defined the baseline date as the date of sample collection that LTL was measured in, and censored the end of follow-up in hospital health record data as 31/March/2020. This

corresponds to the latest UKB release update before the COVID-19 outbreak in the UK. We allocated censor dates on the basis of the location of a participant's assessment centre. Data about deaths is also subject to censoring using the same date of 31/March/2020 used for HES records.

Prevalent cases are defined as self-reported at baseline, or a recorded hospitalisation with any (primary or secondary) diagnosis of the disease of interest before the baseline date from the following sources:

- UKB self-report codes at baseline

- Relevant codes (ICD-9/10 and OPCS3/4) in primary or secondary diagnosis during a hospital episode HES

The rationale for including secondary fields from HES and ONS to define prevalent cases is because of the inclusion of less strict self-report codes from UKB.

Incident cases are defined as the first recorded hospitalisation or death with the disease of interest given as the primary or secondary cause occurring after the UKB baseline visit. Time-to-event is defined as the post-baseline date of the first incident hospitalisation or death, or otherwise censored at the end of study follow-up on 31/March/2020.

Investigating disease association non-concordance.

We assessed the proportional hazards assumption for all observational models by allowing for time-varying effects. Where proportional hazards were shown to deviate (time interaction P<0.05), we estimated the hazard ratios at baseline and at 10 years via linear combination. This analysis showed that the interpretation of our estimates do not change substantially, within the timeframe of the UKB study, even if allowed to vary over time.

For 26 diseases, we found Bonferroni-significant associations with usual LTL but non-significant associations with genetically-determined LTL (**Extended Data Figure 3**). For some of these conditions, the observational associations are likely to reflect residual confounding. For example, heart failure and atrial fibrillation (AF) are often secondary consequences of coronary artery disease (CAD). Adjusting for CAD made the observational association with AF non-significant (OR 1.01 (0.98-1.03), p=0.63) and substantially attenuated the association

with heart failure (OR 1.07 (1.03-1.10), p=5.5x10^{-5}). On the other hand, for other observational associations, such as with aortic stenosis, osteoporosis and chronic obstructive pulmonary disease, there was a trend towards a genetic association that was directionally consistent with the observational findings (**Extended Data Figure 3**), suggesting the need for greater power in MR analyses. Nevertheless, our power calculations suggest our MR analyses had power to detect an odds ratio of at least 1.10 for these diseases (**Supplementary Figure 13**).

For nine diseases, we found nominal-significant associations with genetically-determined LTL but no associations, or associations in the opposite direction, with usual LTL (**Figure 5**). In these situations, discrepancies could arise from either use of different case definitions (e.g., use of prevalent disease outcomes for genetic analysis vs incident disease outcomes for observational analysis) or residual confounding or both. As an example to illustrate this, we focused on hypertension, an outcome that was *inversely* associated with shorter LTL in MR analyses, yet *positively* associated with usual LTL. When restricting the MR analysis to incident disease outcomes, the inverse MR association attenuated to the null, suggesting that a disease history was has an effect on the LTL genetic associations. It is also possible that the change is due to a drop in statistical power. When using prevalent data, the positive observational association also attenuates to the null. After adjusting for several potential confounders, we observed an inverse association with usual LTL, in keeping with MR results (**Supplementary Table 13**).

Public health modelling of years' life lost

 We used three pieces of information to estimate differences in life expectancy associated with shorter measured LTL at baseline, categorised into 4 pre-defined groups, henceforth "exposure groups" namely, group 1 (<-1SD), group 2 (-1 to <0 SD), group 3 (0 to <1SD), and group 4 (≥1SD) [Reference group]:

(i)   age-at-risk specific hazard ratios for all-cause (and cause-specific) mortality in each exposure group versus the reference (derived from UKB);

(ii)   population all-cause (and cause-specific) mortality rates (derived from the EUROSTAT database (http://ec.europa.eu/eurostat/data/database) for UK and 28

EU countries as calculated averages during the three year period 2014-2016 (i.e. 2015 midpoint); and

(iii) prevalence of exposure groups in the population (derived from UKB).

We estimated population survival curves for each exposure group, utilising estimated age-at-risk specific hazard ratios for mortality by exposure groups in UKB and routine statistics on overall population mortality rates. We estimated differences in life-expectancy as differences in areas under any two survival curves compared.

Age-at-risk specific hazard ratios for mortality by exposure groups were estimated from UKB data. Specifically, a Cox regression model stratified by sex was fitted using a dataset in which participant ages-at-risk were deterministically updated by splitting the follow up times every 5-years and recalculating an age-at-risk variable at the beginning of each 5-year interval of follow up. Interactions between baseline exposure groups and linear and quadratic terms for the age-at-risk variable were included in the model to obtain smoothed hazard ratios. Thus, for participant in stratum *s* with exposure group indicator variable $E_{si(j)}$ (i.e. dummy variable equal to 1 if in exposure group *j* and zero otherwise) the log hazard rate at time *t* since baseline was modelled as:

$$\log\big(h_{si}(t)\big) = \log\big(h_{s0}(t)\big) + \sum_{j=1}^{3} \gamma_{0j}\, E_{si(j)} + \beta_1 agerisk_{si} + \beta_2 agerisk_{si}^2 + \sum_{j=1}^{3} \gamma_{1j}\, E_{si(j)} \times agerisk_{si} + \sum_{j=1}^{3} \gamma_{2j}\, E_{si(j)} \times agerisk_{si}^2 \qquad (1)$$

from which the age-at-risk specific hazard ratios (and 95% CIs) for mortality were obtained as linear combinations of the relevant estimated coefficients, with age-at-risk fixed at values corresponding to midpoints of 5-year age-groups from age 40 onwards (**Supplementary Figure 14**).

Population all-cause (and cause-specific) mortality rates per 100,000 were obtained in 5-year age-groups for the UK and 28 EU countries from the EUROSTAT database (http://ec.europa.eu/eurostat/data/database) as calculated averages during the three year period 2014-2016 (i.e. 2015 midpoint) (**Supplementary Figure 15**). Because the mortality rates were provided only up to age-group 80-84 years, but we desired to estimate the overall population survival curves, we used a piecewise cubic Hermite interpolation (PCHIP) method to smooth through the midpoints of 5-year age-groups and extrapolate the mortality rates to age 110 years (**Supplementary Figure 16**). Next, assuming exponential survival (i.e. constant

hazard) within each 5-year age group, we estimated the age-specific survival probability as $S_a = exp(-5 \times IR_a)$ and derived the overall population survival curves from age 35 onwards as the product of the relevant age-group specific survival probabilities (**Supplementary Figure 17**).

$$p(survival|agerisk \geq 35) = \prod_{agerisk \geq 35} S_a \qquad (2)$$

In order to infer population mortality rates appropriate for the reference exposure group used in our estimation of age-specific hazard ratios (i.e. group 4 with standardised LTL ≥1-SD longer than the population mean), we used ordinal logistic regression to model the age-specific prevalence of the four exposure groups in UKB by sex and decade of recruitment (**Supplementary Figure 18**). We used the age-specific prevalence estimates for the decade commencing in the year 2000 to infer the age-specific mortality rates appropriate for our reference group IR_{a0} as:[1]

$$IR_{a0} = \frac{IR_a}{p_{a0} + \sum_{j=1}^{3} p_{aj} \times RR_{aj}} \qquad (3)$$

where IR_a is the population mortality rate for age group a, p_{aj} is the age-specific prevalence of exposure group j, and RR_{aj} is the age-specific hazard ratio in comparison of exposure group j versus reference group ($j = 0$). The age-specific mortality rates in each of the non-reference exposure groups were then inferred in turn by multiplying the age-specific mortality rate for the reference group IR_{a0} by the age-specific hazard ratios RR_{aj} based on UKB data and equation *(2)* above used to infer the exposure group-specific population survival curves (**Supplementary Figure 19**). Finally, differences in life expectancy according to baseline exposure groups were estimated as difference in the areas under the survival curves for the reference group and each of the non-reference exposure groups in turn (**Supplementary Figure 20**). The areas under curves were calculated by numerical integration.

Monte Carlo simulation was used to calculate confidence intervals for the estimated reductions in life expectancy, taking into account uncertainty in the age-at-risk specific hazard ratios calculated from equation *(1)* above. In particular, new parameter estimates were randomly drawn from the multivariate normal distribution defined by the fitted model mean and covariance matrix, 200 times, and the above procedure repeated for each draw to calculate differences in life-expectancy for each index age of interest. Assuming asymptotic normality, the standard deviation of the 200 Monte Carlo estimates of differences in life

expectancy for each index age were used to calculate 95% confidence intervals around the originally estimated value. Histograms were inspected to judge that normality assumption was reasonable.
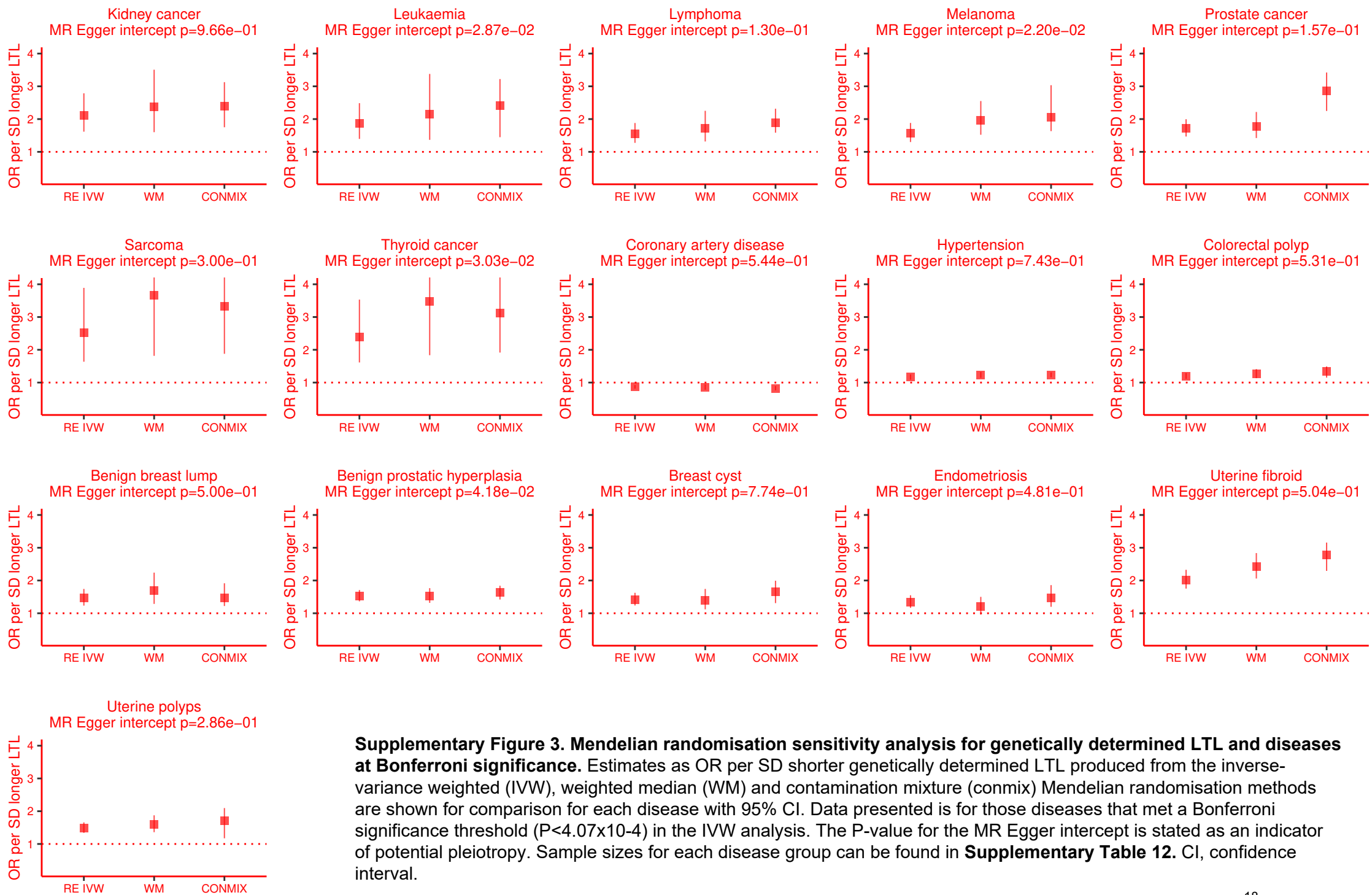
Supplementary references

72. van Steensel, B. and de Lange, T. Control of telomere length by the human telomeric protein TRF1. *Nature* **385**, 740-3 (1997).

73. Glousker, G., Briod A., Quadroni M., Lingner J. Human shelterin protein POT1 prevents severe telomere instability induced by homology-directed DNA repair. *EMBO J* **39**, e104500. (2020).

74. Franzolin, E., et al. The deoxynucleotide triphosphohydrolase SAMHD1 is a major regulator of DNA precursor pools in mammalian cells. *Proc Natl Acad Sci U S A*. 110, 14272–14277 (2013).

75. Gupta, A., et. al. Telomere length homeostasis responds to changes in intracellular dNTP pools. *Genetics* **193**, 1095-105 (2013).

76. Lee, S.S., Bohrson, C., Pike, A.M., Wheelan, S.J., Greider, C.W. ATM Kinase Is Required for Telomere Elongation in Mouse and Human Cells. *Cell Rep*. **13**, 1623-32 (2015).

77. Adelfalk, C., et al. Cohesin SMC1β protects telomeres in meiocytes. *J Cell Biol* **187**, 185–199 (2009).

78. Zhan, Y., et al. Exploring the Causal Pathway From Telomere Length to Coronary Heart Disease: A Network Mendelian Randomization Study.
Circ Res. **121**, 214-219 (2017).

79. Landrum, M.J., et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res*. **46**, D1062–7 (2018).

80. Jaganathan, K., et al. Predicting Splicing from Primary Sequence with Deep Learning. *Cell* **176**, 535-548 (2019).

81. NHLBI Exome Sequencing Project, Do, R., et al. Exome sequencing identifies rare LDLR and APOA5 alleles conferring risk for myocardial infarction. *Nature* **518**, 102–6 (2015).

82. Khera, A.V., et al. Diagnostic Yield and Clinical Utility of Sequencing Familial Hypercholesterolemia Genes in Patients with Severe Hypercholesterolemia. *J Am Coll Cardiol* **67**, 2578–89 (2016).

83. Do, R., et al. Exome sequencing identifies rare LDLR and APOA5 alleles conferring risk for myocardial infarction. *Nature* **518**, 102–6 (2015).

84. Khera, A.V., et al. Association of Rare and Common Variation in the Lipoprotein Lipase Gene with Coronary Artery Disease. *JAMA* **317**, 937–46 (2017).

λ=1.145
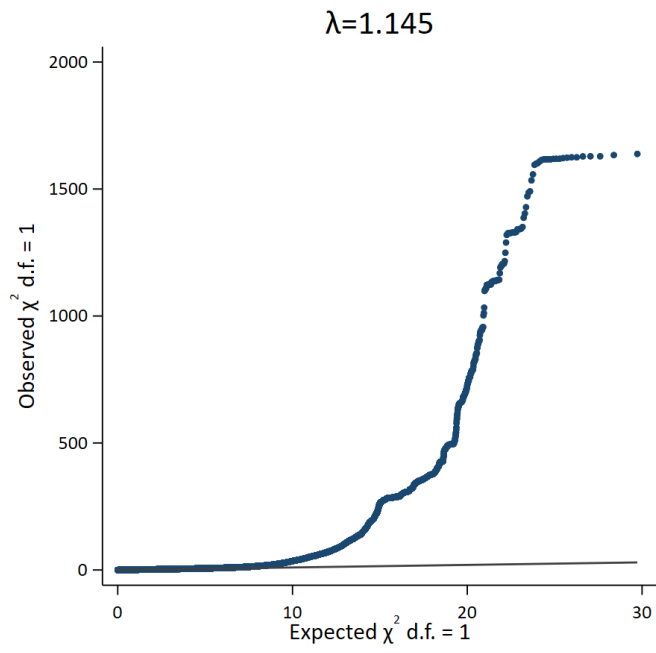
Supplementary Figure 1. Quantile-quantile plot. Observed versus expected results from the GWAS is plotted. The genomic inflation factor (λ) is also shown.
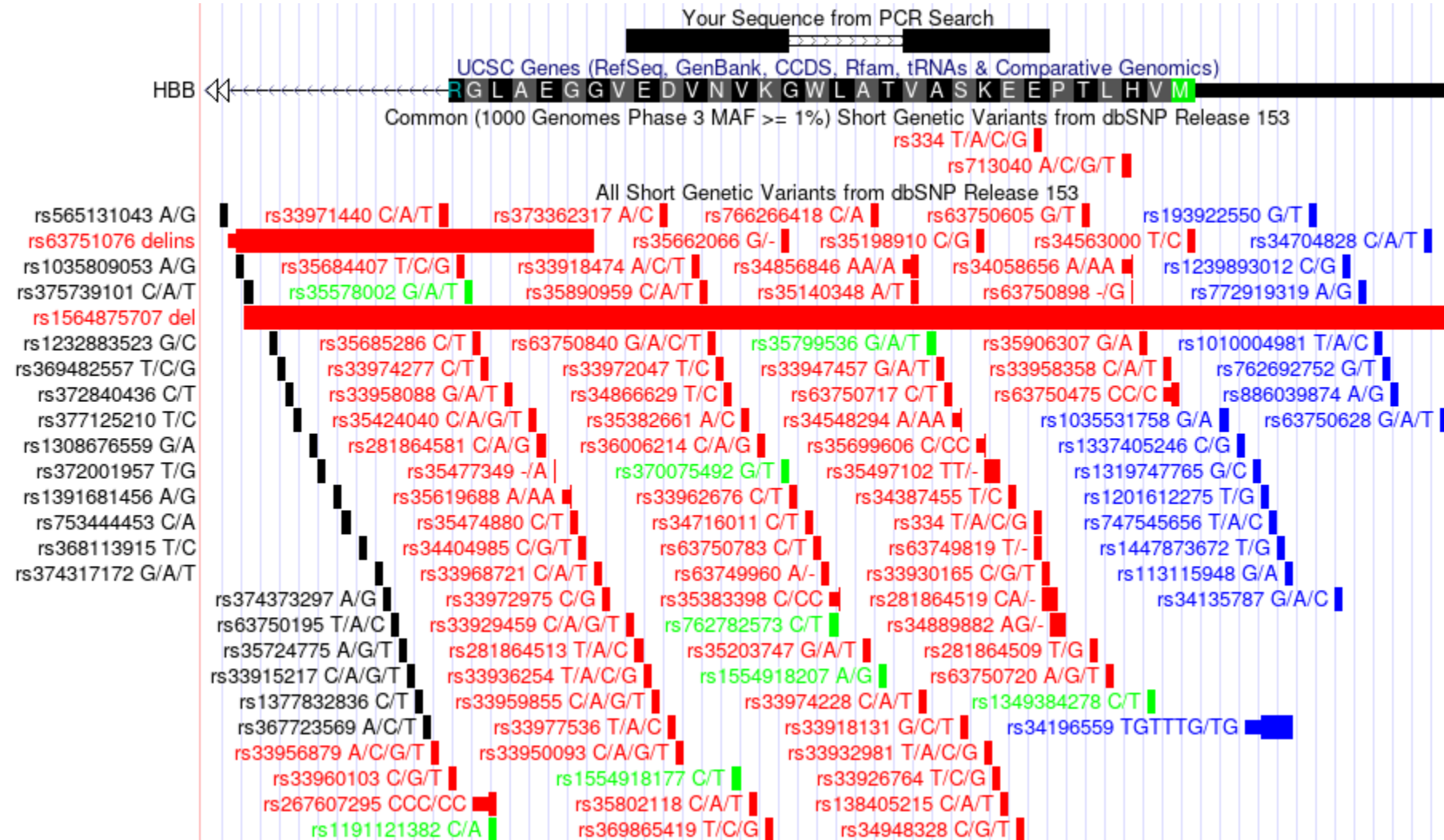
Supplementary Figure 2. Alignment of single copy PCR gene primers within *HBB*. Primer sequences are shown as black bars at the top with SNP positions illustrated below. The position of rs334 can be seen to sit just within the reverse primer binding sequence. Another SNP rs713040 (R^2>0.7 to rs1609812) sits just outside the sequence. Primers sequences and SNP positions are visualised using the *In-silico* PCR tool within the UCSC genome browser (https://genome.ucsc.edu)

Supplementary Figure 3. Mendelian randomisation sensitivity analysis for genetically determined LTL and diseases at Bonferroni significance. Estimates as OR per SD shorter genetically determined LTL produced from the inverse-variance weighted (IVW), weighted median (WM) and contamination mixture (conmix) Mendelian randomisation methods are shown for comparison for each disease with 95% CI. Data presented is for those diseases that met a Bonferroni significance threshold (P<4.07x10-4) in the IVW analysis. The P-value for the MR Egger intercept is stated as an indicator of potential pleiotropy. Sample sizes for each disease group can be found in **Supplementary Table 12.** CI, confidence interval.

Supplementary Figure 4. Mendelian randomisation sensitivity analysis for genetically determined LTL and diseases at nominal significance. Estimates as OR per SD shorter genetically determined LTL produced from the inverse-variance weighted (IVW), weighted median (WM) and contamination mixture (conmix) Mendelian randomisation methods are shown for comparison for each disease with 95% CI. Data presented is for those diseases that met a nominal significance threshold (P< 0.05) in the IVW analysis. The P-value for the MR Egger intercept is stated as an indicator of potential pleiotropy. Sample sizes for each disease group can be found in Supplementary Table 12. CI, confidence interval.

Supplementary Figure 5. Assessing potential non-linear disease associations with LTL for diseases associated with LTL at Bonferroni significance in MR analyses. Fractional polynomial models were fitted for each disease, adjusting for age, sex, WBC and ethnic group. The continuous shape of each association (relative to a reference value of 0) is shown for each disease as a hazard ratio (HR) with 95% CI (shaded). Sample sizes for each disease are given in (**Supplementary Table 12**). CI, confidence interval.

Supplementary Figure 6. Assessing potential non-linear disease associations with LTL for diseases associated with LTL at nominal significance in MR analyses. Fractional polynomial models were fitted for each disease, adjusting for age, sex, WBC and ethnic group. The continuous shape of each association (relative to a reference value of 0) is shown for each disease as a hazard ratio (HR) with 95% CI (shaded). Sample sizes for each disease are given in (**Supplementary Table 12**). CI, confidence interval.

| Event name	Cases	MR analysis with incident and prevalent cases (Odds Ratio)	P	I²	Cases	MR analysis with incident cases (Odds Ratio)	P	I²	Cases	Observational analysis with incident cases (Hazard Ratio)	P	Observational analysis with incident cases (Odds Ratio)	P	Cases	Observational analysis with incident and prevalent cases (Odds Ratio)	P
Cancer																
Breast	18269/235702	1.11 (1.00, 1.24)	4×10^{-2}	35%	8293/218135	1.18 (1.02, 1.36)	2×10^{-2}	26%	8293/218135	1.02 (0.99, 1.05)	3×10^{-1}	1.02 (0.99, 1.06)	2×10^{-1}	17541/227383	0.99 (0.96, 1.01)	3×10^{-1}
Colorectal	7904/437060	1.14 (1.00, 1.30)	4×10^{-2}	7%	5086/419691	1.14 (0.98, 1.33)	9×10^{-2}	0%	5086/419691	1.00 (0.96, 1.04)	1	1.01 (0.96, 1.05)	8×10^{-1}	7630/422235	1.02 (0.98, 1.05)	3×10^{-1}
Bladder	3595/437060	1.28 (1.03, 1.57)	2×10^{-2}	21%	2287/421057	1.14 (0.89, 1.48)	3×10^{-1}	16%	2287/421057	0.94 (0.88, 1.00)	5×10^{-2}	0.95 (0.89, 1.01)	8×10^{-2}	3465/422235	0.97 (0.92, 1.02)	2×10^{-1}
Lymphoma	4623/437060	1.55 (1.27, 1.89)	1×10^{-5}	31%	2824/420621	1.59 (1.25, 2.03)	1×10^{-4}	26%	2824/420621	1.00 (0.94, 1.06)	9×10^{-1}	1.00 (0.95, 1.06)	1	4438/422235	0.91 (0.87, 0.96)	7×10^{-5}
Leukaemia	1939/437060	1.86 (1.39, 2.50)	2×10^{-5}	26%	1214/421574	2.44 (1.72, 3.46)	3×10^{-7}	17%	1214/421574	0.82 (0.76, 0.90)	6×10^{-6}	0.83 (0.76, 0.90)	1×10^{-5}	1875/422235	0.79 (0.74, 0.85)	6×10^{-12}
Cardiovascular																
Hypertension	278315/437060	1.17 (1.09, 1.25)	2×10^{-5}	73%	11388/164776	1.06 (0.93, 1.19)	4×10^{-1}	21%	11388/164776	0.93 (0.90, 0.95)	1×10^{-7}	0.93 (0.90, 0.96)	2×10^{-6}	268847/422235	1.01 (1.00, 1.02)	1×10^{-1}
Endocrine																
Hyperthyroidism	5954/437060	0.76 (0.64, 0.90)	2×10^{-3}	34%	1917/418384	0.93 (0.69, 1.25)	6×10^{-1}	28%	1917/418384	0.99 (0.92, 1.06)	7×10^{-1}	0.99 (0.92, 1.06)	8×10^{-1}	5768/422235	0.95 (0.91, 0.99)	8×10^{-3}
Genitourinary																
Endometriosis	7537/235702	1.34 (1.16, 1.55)	5×10^{-5}	21%	1215/221295	1.56 (1.12, 2.16)	7×10^{-3}	5%	1215/221295	1.02 (0.94, 1.11)	6×10^{-1}	1.02 (0.94, 1.11)	6×10^{-1}	7303/227383	1.02 (0.98, 1.06)	3×10^{-1}
Respiratory																
Pleurisy	2351/437060	0.76 (0.61, 0.96)	2×10^{-2}	2%	445/420422	1.19 (0.70, 2.01)	5×10^{-1}	0%	445/420422	0.93 (0.80, 1.07)	3×10^{-1}	0.93 (0.81, 1.07)	3×10^{-1}	2258/422235	0.90 (0.85, 0.96)	2×10^{-3}
Sensory																
Retinal detachment	5564/437060	1.25 (1.08, 1.45)	3×10^{-3}	0%	2853/419700	1.13 (0.91, 1.41)	2×10^{-1}	7%	2853/419700	1.03 (0.97, 1.09)	3×10^{-1}	1.04 (0.98, 1.10)	2×10^{-1}	5388/422235	1.04 (1.00, 1.09)	5×10^{-2}

Panels 1–2: Odds ratio (95% CI) per SD longer genetically-predicted LTL. Panel 3: Hazard ratio (95% CI) per SD longer usual LTL. Panels 4–5: Odds ratio (95% CI) per SD longer usual LTL.

Supplementary Figure 7. MR and observational analyses with different case definitions and statistical analysis approaches. Left to right: (i) MR analysis with prevalent and incident analysis (main MR analysis, included here as a reference), (ii) MR analysis restricted to incident cases, (iii) observational Cox regression analysis restricted to incident cases (main observational analysis), (iv) observational logistic regression analysis restricted to incident cases, (v) observational logistic regression with incident and prevalent cases.
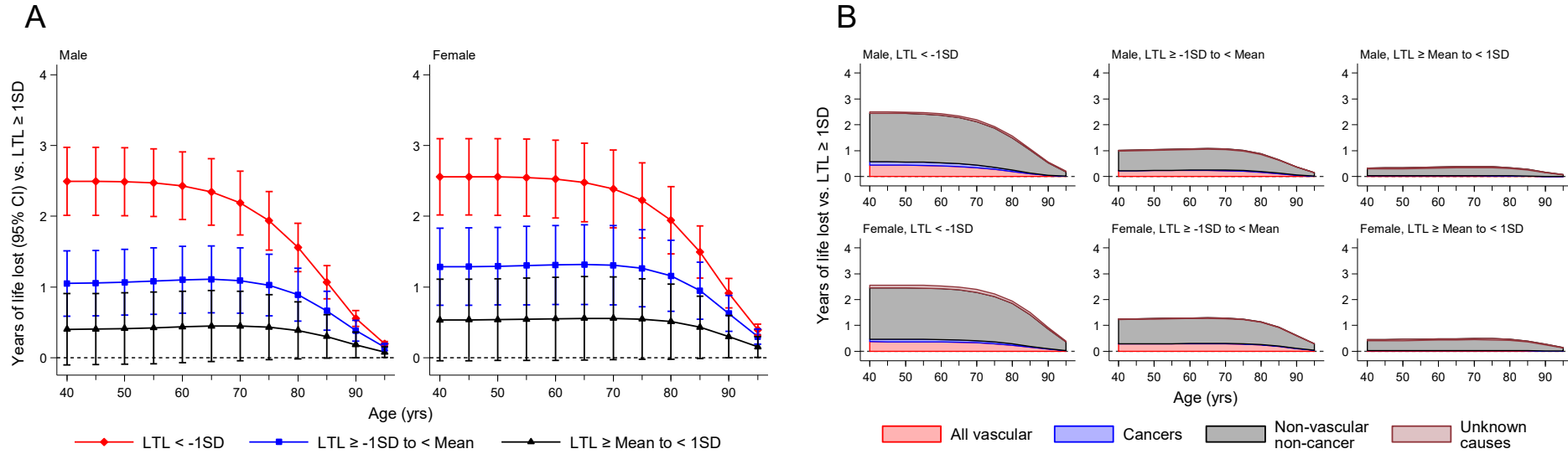
Supplementary Figure 8. Years of life lost using EU 2015 mortality rates. Estimates of years of life lost were estimated by applying hazard ratios (HRs) for cause-specific mortality calculated from UK Biobank data (specific to age-at-risk and stratified by sex) to population mortality rates for 28 European Union (EU) countries during year 2015 (by sex and 5-year age groups). These were estimated for four standardised LTL groups (group 1 (<-1SD), group 2 (-1 to <0 SD), group 3 (0 to <1SD), and group 4 (≥1SD)) from 40-95 years of age. Group 4 were used as the reference group. Data is shown for males and females separately. This was performed for all cause mortality (A) and disease specific mortality (B). UK Biobank data included 458,309 participants and 28,345 deaths (comprising 5,984 vascular deaths, 14,916 cancer deaths, 7,244 non-vascular non-cancer deaths, and 201 deaths of unknown causes). CI, confidence interval.
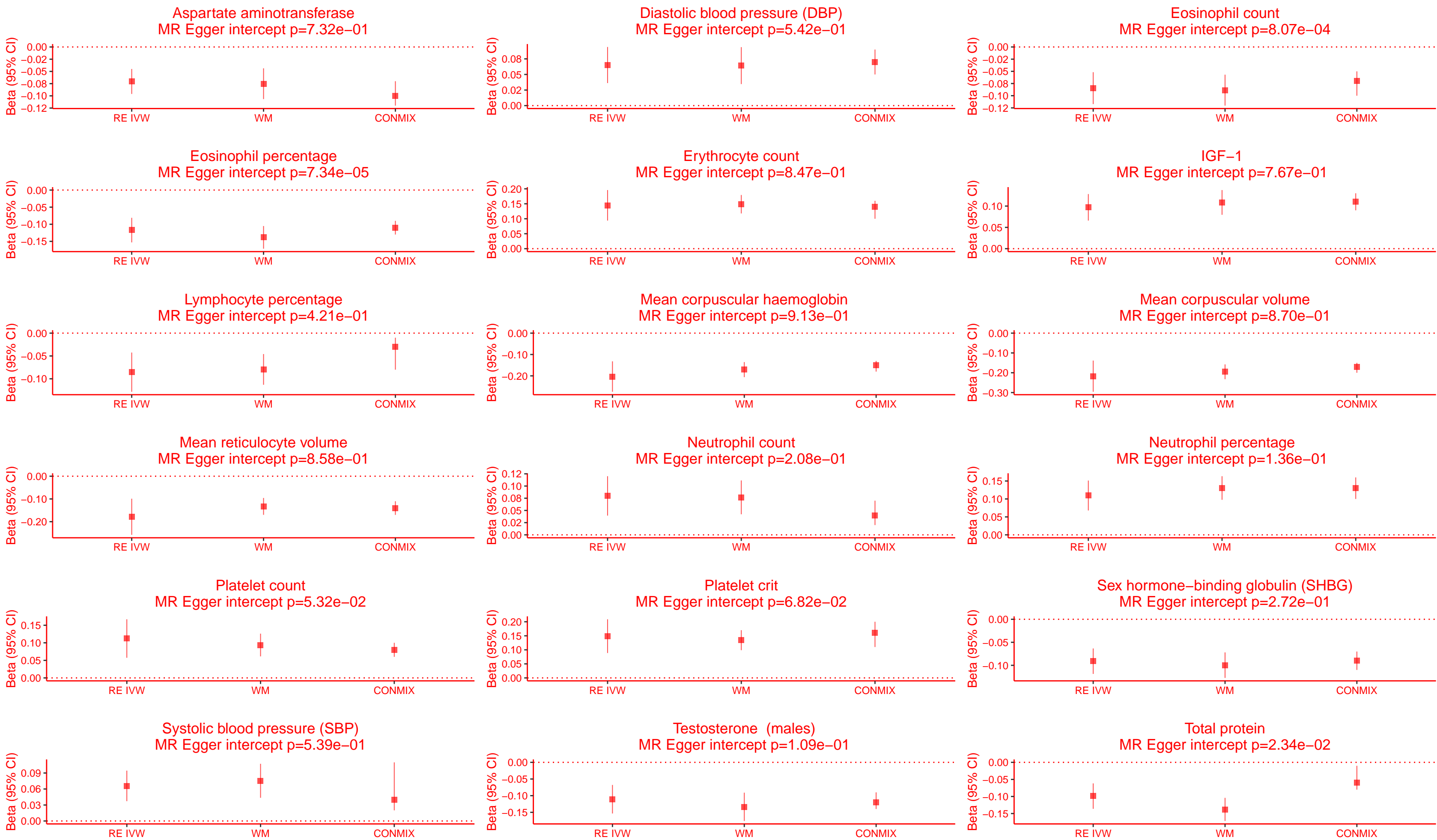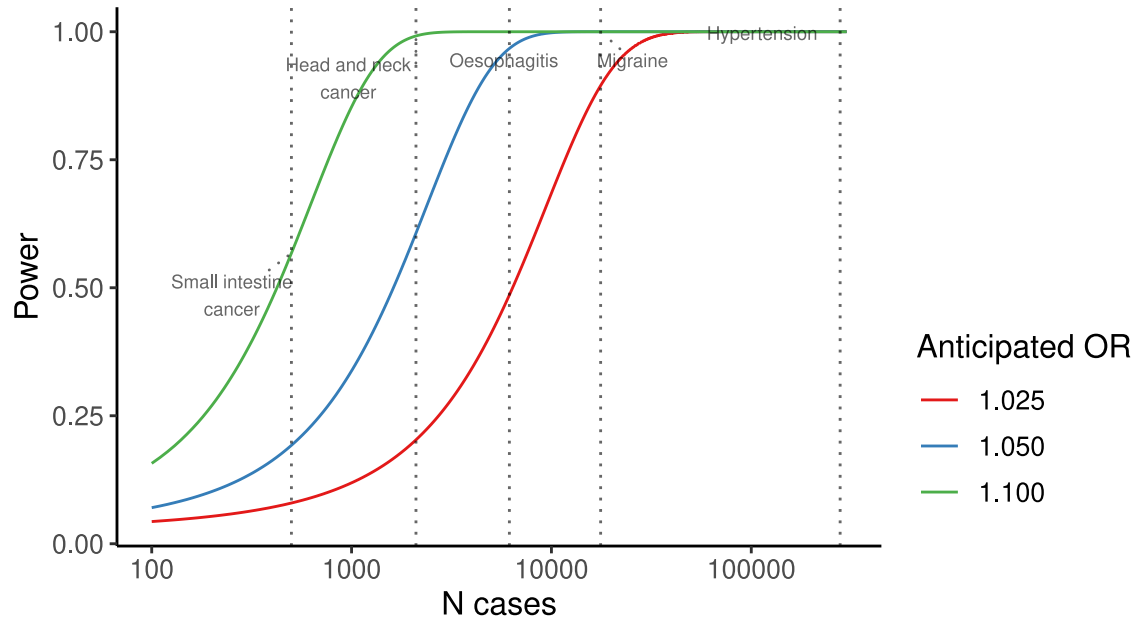
Supplementary Figure 9. Mendelian randomisation sensitivity analyses for genetically determined LTL associations with biomedical traits at Bonferroni significance. Estimates as beta per SD shorter genetically determined LTL produced from the inverse-variance weighted (IVW), weighted median (WM) and contamination mixture (conmix) Mendelian randomisation methods are shown for comparison for each trait with 95% CI. Data presented is for those traits that met a Bonferroni significance threshold (P<5.38x10-4) in the IVW analysis (**Supplementary Table 10**). The P-value for the MR Egger intercept is stated as an indicator of potential pleiotropy. Sample sizes for each trait analysed can be found in **Supplementary Table 10.** CI, confidence interval.
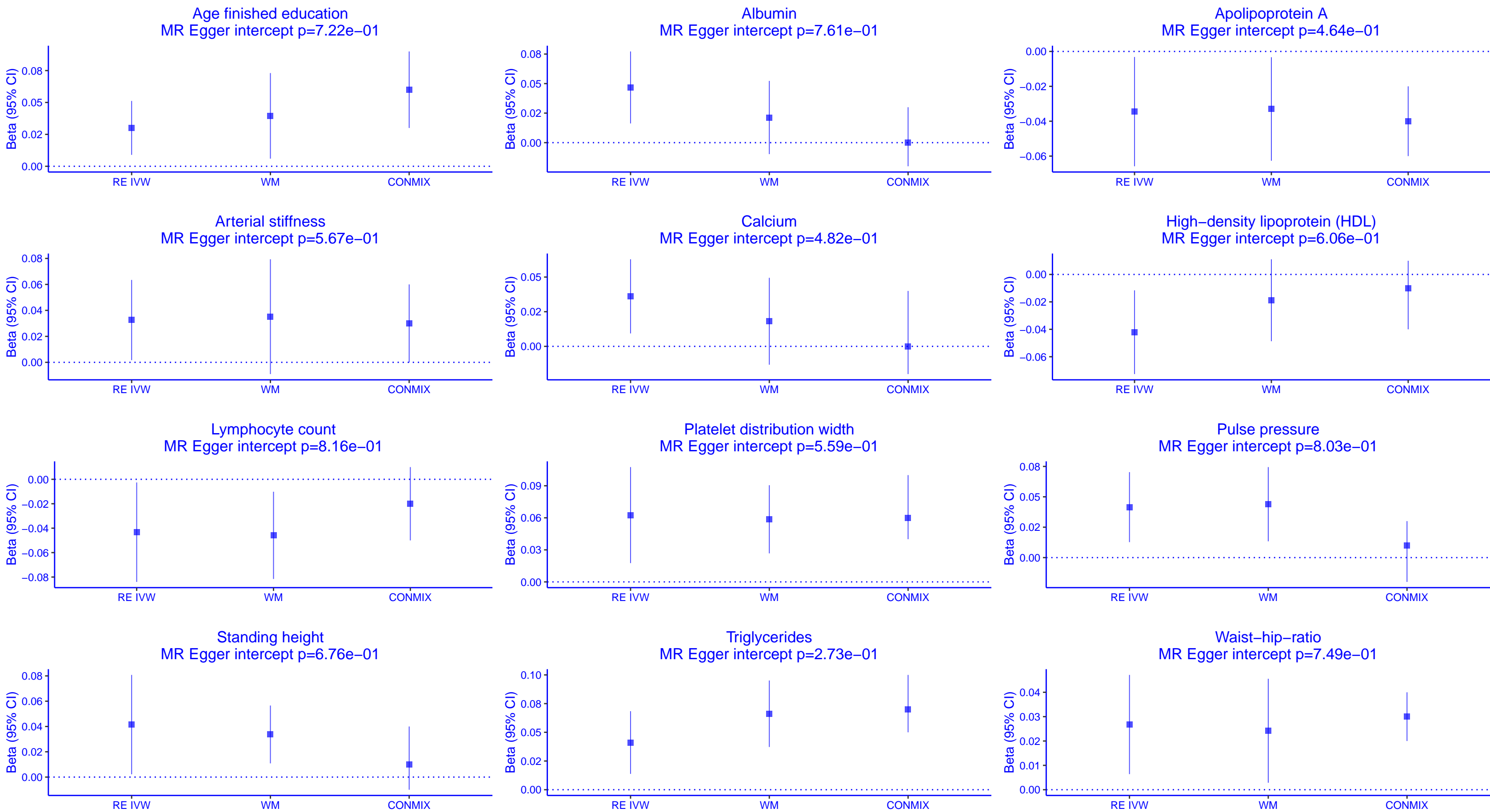
Supplementary Figure 10. Mendelian randomisation sensitivity analyses for genetically determined LTL associations with biomedical traits at nominal significance. Estimates as beta per SD shorter genetically determined LTL produced from the inverse-variance weighted (IVW), weighted median (WM) and contamination mixture (conmix) Mendelian randomisation methods are shown for comparison for each trait with 95% CI. Data presented is for those traits that met a nominal significance threshold (P<0.05) in the IVW analysis (**Supplementary Table 10**). The P-value for the MR Egger intercept is stated as an indicator of potential pleiotropy. Sample sizes for each trait analysed can be found in **Supplementary Table 10**. CI, confidence interval.
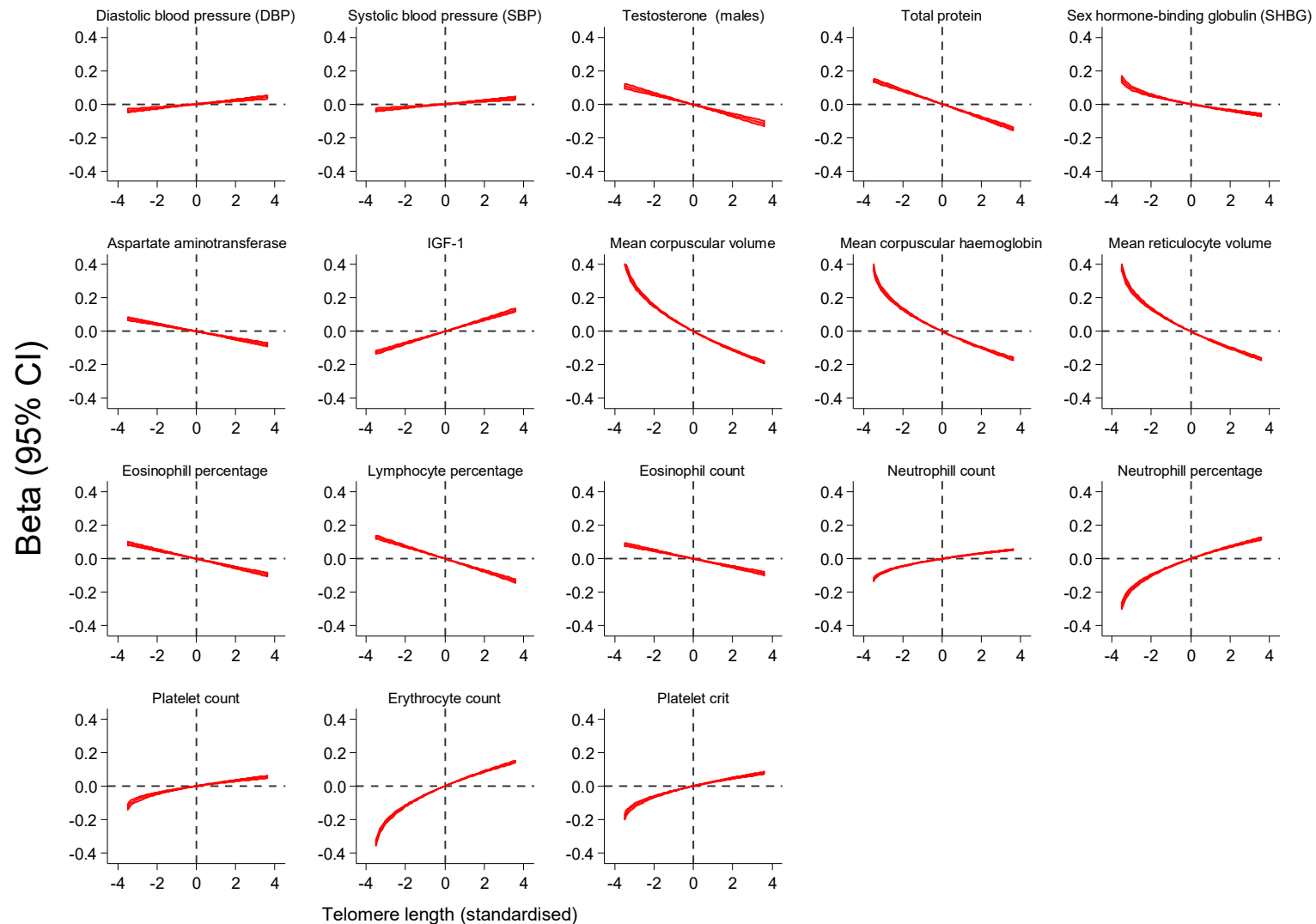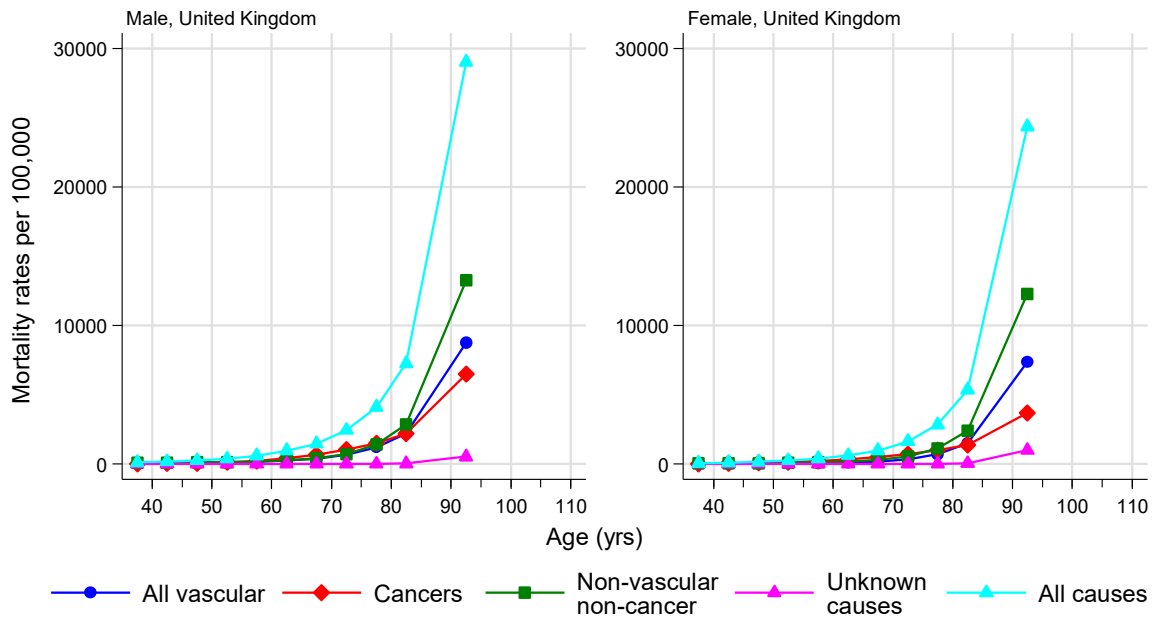
Supplementary Figure 11. Assessing potential non-linear biomedical trait associations with LTL associated with LTL at Bonferroni significance in MR analyses. Fractional polynomial models were fitted for each biomedical trait, adjusting for age, sex, WBC and ethnic group. The continuous shape of each association (relative to a reference value of 0) is shown for each disease. Data for biomedical traits that were associated with genetically determined LTL in the MR analyses at a Bonferroni level of significance (P<5.38x10^{-4}, **Supplementary Table 10**) are shown as the effect size (beta) with 95% CI (shaded). Sample sizes for each trait are given in (**Supplementary Table 10**). CI, confidence interval.
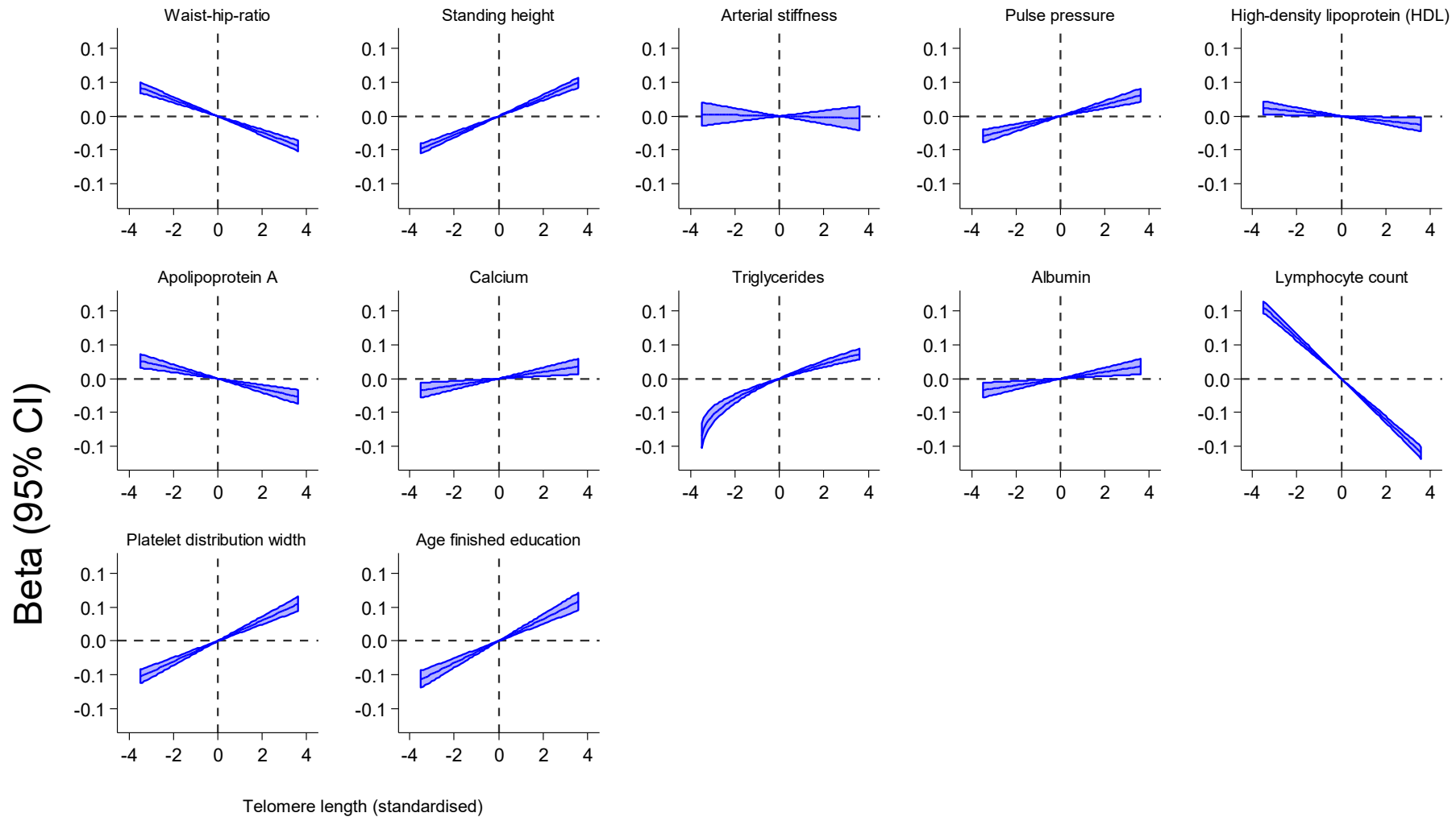
Supplementary Figure 12. Assessing potential non-linear biomedical trait associations with LTL associated with LTL at nominal significance in MR analyses. Fractional polynomial models were fitted for each biomedical trait, adjusting for age, sex, WBC and ethnic group. The continuous shape of each association (relative to a reference value of 0) is shown for each disease. Data for biomedical traits that were associated with genetically determined LTL in the MR analyses at a Bonferroni level of significance (P<0.05, **Supplementary Table 10**) are shown as the effect size (beta) with 95% CI (shaded). Sample sizes for each trait are given in (**Supplementary Table 10**). CI, confidence interval.

Supplementary Figure 13. Power calculations for Mendelian randomization disease analysis. Power to detect anticipated odds ratios (OR, line chart) for annotated examples that reflect distribution of case counts across all prevalents and incident diseases (box plot, total cases 1,882,779): minimum number of cases (small intestine cancer, N=500), bottom quartile (head and neck cancer=2,098), median (oesophagitis, N=6,163), upper quartile (migraine, N=17,627), and maximum (hypertension, N=278,315). The y axis is shared for both plots. Box represents inter-quartile range with median line, whisker indicate minimum/maximum.
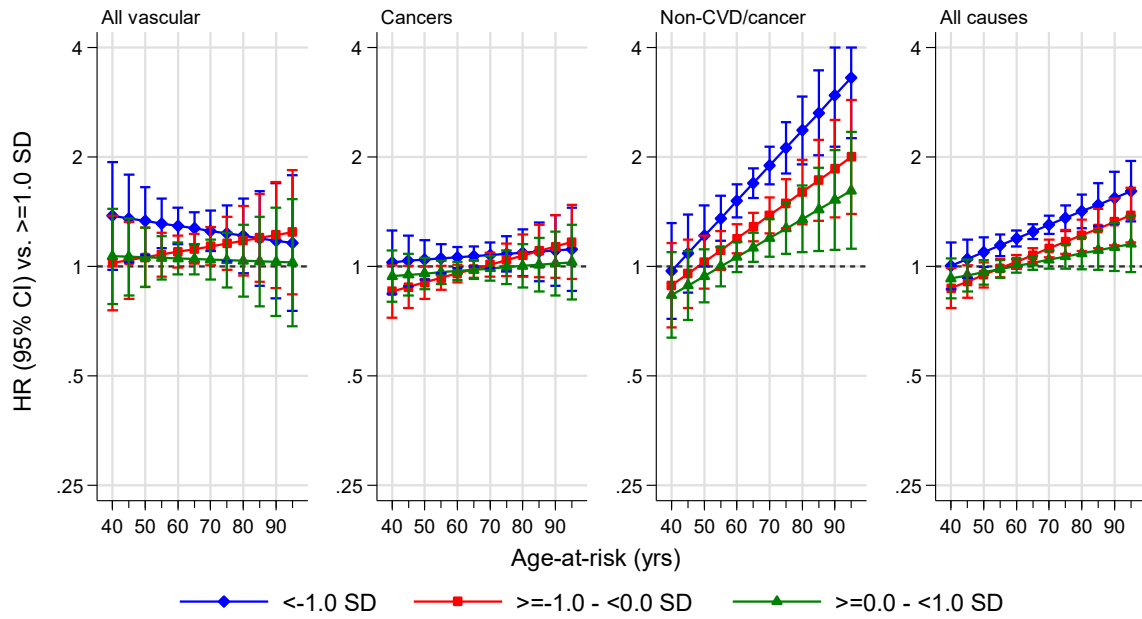
Supplementary Figure 14. Age-at-risk specific hazard ratios for all-cause mortality by LTL exposure groups estimated from UKB data. Data presented are HR with 95% CI estimated for four standardised LTL groups (group 1 (<-1SD), group 2 (-1 to <0 SD), group 3 (0 to <1SD), and group 4 (≥1SD)) from 40-95 years of age. Group 4 were used as the reference group. UK Biobank data included 458,309 participants and 28,345 deaths (comprising 5,984 vascular deaths, 14,916 cancer deaths, 7,244 non-vascular non-cancer deaths, and 201 deaths of unknown causes). CI, confidence interval.

Male, United Kingdom

Female, United Kingdom

Legend: All vascular · Cancers · Non-vascular non-cancer · Unknown causes · All causes

Supplementary Figure 15. UK population mortality rates during year 2015 downloaded from EUROSTAT online database. To maintain consistency with analyses conducted in UKB, the mortality rate for non-vascular non-cancer causes was recalculated as the difference of all-cause mortality and the sum of vascular mortality (I00-I99), cancer mortality (C00-D48), and unknown causes of mortality (R00-R99).
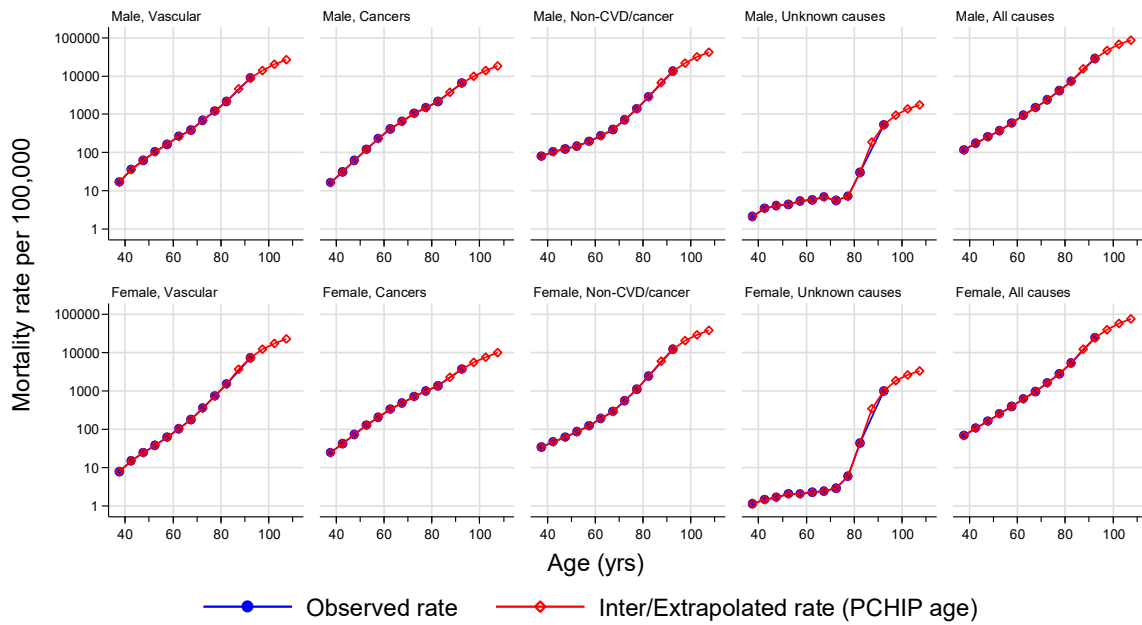
Supplementary Figure 16. Assessment of adequacy of a piecewise cubic Hermite interpolation (PCHIP) method to smooth and extrapolate UK population mortality rates during 2015. Data was downloaded from EUROSTAT online database beyond the database's upper bound age cut-off of 84 years to facilitate modelling.
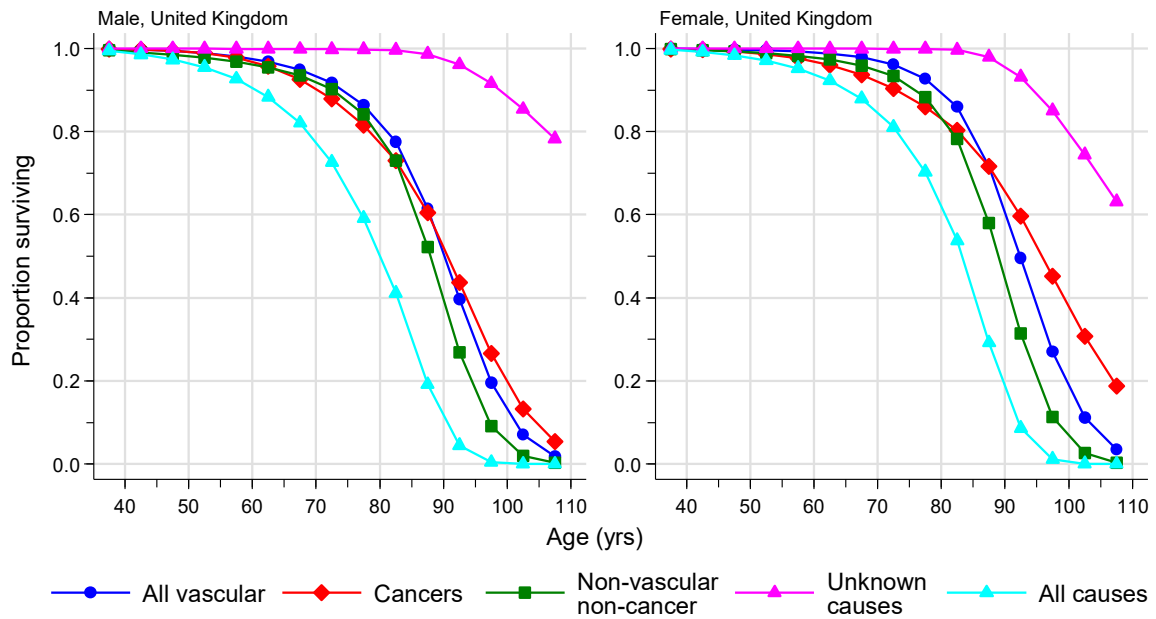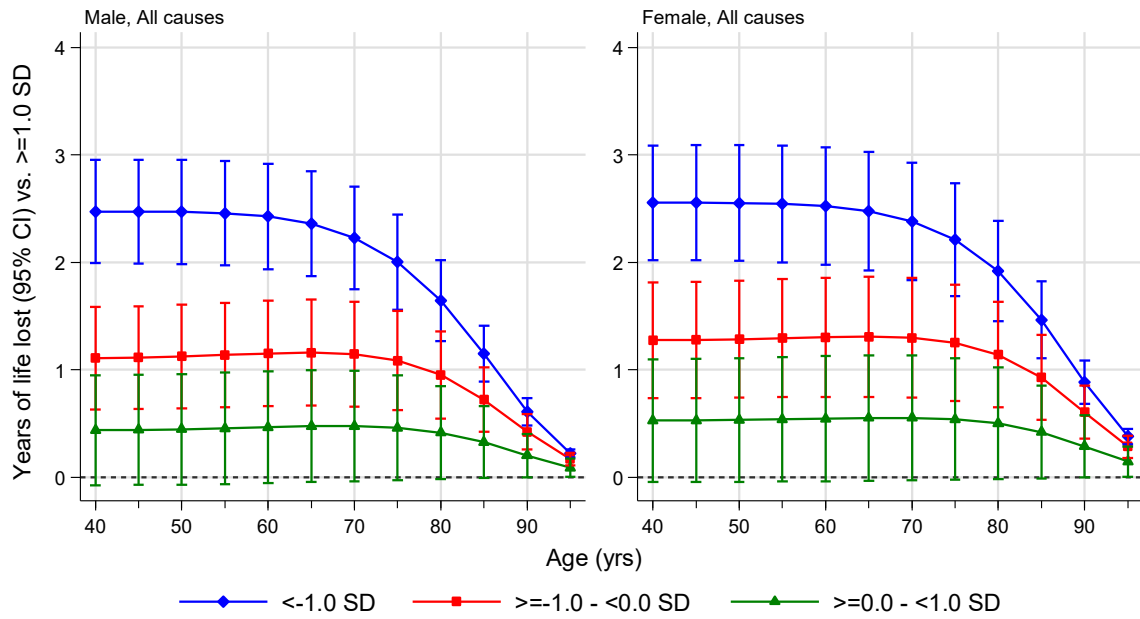
Supplementary Figure 17. Derived population survival curves for all-cause and cause-specific mortality from age 35 years based on smoothed and extrapolated UK population mortality rates during 2015.
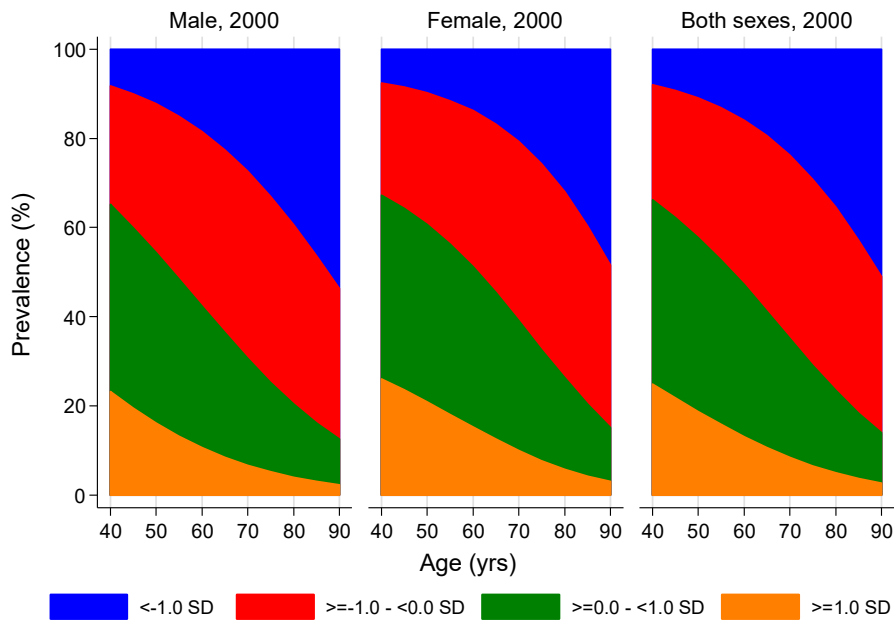
Supplementary Figure 18. Modelled age-specific prevalence of LTL exposure groups in UKB by sex and decade of recruitment.
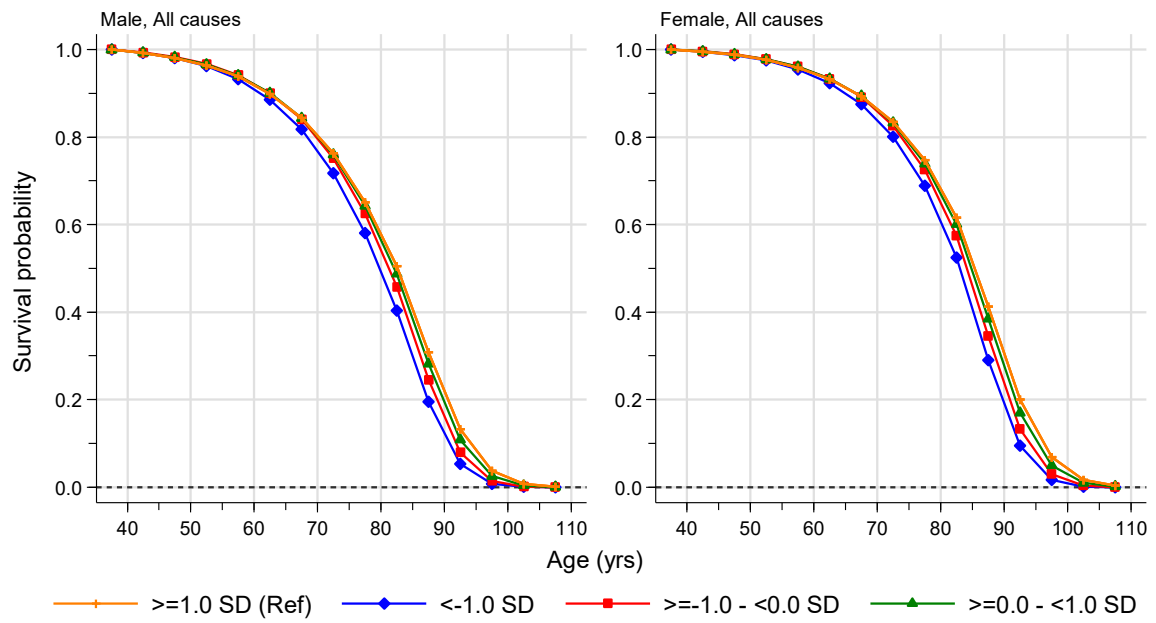
Supplementary Figure 19. Inferred survival curves for UK population by sex and LTL exposure groups.

Supplementary Figure 20. Estimated sex-specific reductions in life expectancy in the UK population according to LTL exposure groups. Estimates of sex-specific years of life lost were estimated by applying hazard ratios (HRs) to population mortality rates for United Kingdom (UK) during year 2015. Data are presented for four standardised LTL groups; group 1 (<-1SD), group 2 (-1 to <0 SD), group 3 (0 to <1SD), and group 4 (≥1SD) from 40-95 years of age. Group 4 was used as the reference group. Data is shown for males and females separately. The UK Biobank data included 458,309 participants and 28,345 deaths (comprising 5,984 vascular deaths, 14,916 cancer deaths, 7,244 non-vascular non-cancer deaths, and 201 deaths of unknown causes). CI, confidence interval.