

5 year mortality predictors in 498 103 UK Biobank participants: a prospective population-based study

Andrea Ganna, Erik Ingelsson



Summary

Background To our knowledge, a systematic comparison of predictors of mortality in middle-aged to elderly individuals has not yet been done. We investigated predictors of mortality in UK Biobank participants during a 5 year period. We aimed to investigate the associations between most of the available measurements and 5 year all-cause and cause-specific mortality, and to develop and validate a prediction score for 5 year mortality using only self-reported information.

Methods Participants were enrolled in the UK Biobank from April, 2007, to July, 2010, from 21 assessment centres across England, Wales, and Scotland with standardised procedures. In this prospective population-based study, we assessed sex-specific associations of 655 measurements of demographics, health, and lifestyle with all-cause mortality and six cause-specific mortality categories in UK Biobank participants using the Cox proportional hazard model. We excluded variables that were missing in more than 80% of the participants and all cardiorespiratory fitness test measurements because summary data were not available. Validation of the prediction score was done in participants enrolled at the Scottish centres. UK life tables and census information were used to calibrate the score to the overall UK population.

Findings About 500 000 participants were included in the UK Biobank. We excluded participants with more than 80% variables missing (n=746). Of 498 103 UK Biobank participants included (54% of whom were women) aged 37–73 years, 8532 (39% of whom were women) died during a median follow-up of 4.9 years (IQR 4.33–5.22). Self-reported health (C-index including age 0.74 [95% CI 0.73–0.75]) was the strongest predictor of all-cause mortality in men and a previous cancer diagnosis (0.73 [0.72–0.74]) was the strongest predictor of all-cause mortality in women. When excluding individuals with major diseases or disorders (Charlson comorbidity index >0; n=355 043), measures of smoking habits were the strongest predictors of all-cause mortality. The prognostic score including 13 self-reported predictors for men and 11 for women achieved good discrimination (0.80 [0.77–0.83] for men and 0.79 [0.76–0.83] for women) and significantly outperformed the Charlson comorbidity index (p<0.0001 in men and p=0.0007 in women). A dedicated website allows the interactive exploration of all results along with calculation of individual risk through an online questionnaire.

Interpretation Measures that can simply be obtained by questionnaires and without physical examination were the strongest predictors of all-cause mortality in the UK Biobank population. The prediction score we have developed accurately predicts 5 year all-cause mortality and can be used by individuals to improve health awareness, and by health professionals and organisations to identify high-risk individuals and guide public policy.

Funding Knut and Alice Wallenberg Foundation and the Swedish Research Council.

Introduction

Adequate identification and risk stratification of individuals with reduced life expectancy, especially in the middle-aged to elderly population, is an important public health priority¹ and a central issue in clinical decision making. Epidemiological studies that obtain several measurements through questionnaires, physical assessments, and biological samples can be done to compare the prognostic value of risk factors of short-term mortality and provide new hypotheses about health determinants. Moreover, these risk factors can be combined into a prognostic index to provide information about individual mortality risk or other health-related measures. Traditionally, association with mortality has been studied for one risk factor at a time,^{2–6} and the few studies that have investigated more than one risk factor

did not assess different causes of death and based their analyses on small study samples.⁷ Several prognostic indices for short-term mortality exist, but they have mainly been developed for and assessed in older individuals or in high-risk populations.^{8,9} Furthermore, the small sample sizes and the small number of risk factors investigated are the main limitations of all previous studies.

The UK Biobank project¹⁰ includes about 500 000 men and women aged 40–70 years. The participants underwent blood draw for biobanking and participated in detailed, questionnaire-based, physical, and biological measurements during 2007–10. Data from these assessments have been made available to all researchers worldwide after an approved application. We aimed to use these data to do a systematic and untargeted investigation of the associations between most of the

Published Online
June 4, 2015
[http://dx.doi.org/10.1016/S0140-6736\(15\)60175-1](http://dx.doi.org/10.1016/S0140-6736(15)60175-1)

See Online/Comment
[http://dx.doi.org/10.1016/S0140-6736\(15\)60578-5](http://dx.doi.org/10.1016/S0140-6736(15)60578-5)

Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden (A Ganna PhD); and Department of Medical Sciences, Molecular Epidemiology and Science for Life Laboratory, Uppsala University, Uppsala, Sweden (A Ganna, Prof E Ingelsson MD)

Correspondence to:
Prof Erik Ingelsson, Department of Medical Sciences, Molecular Epidemiology, Box 1115, Uppsala 75141, Sweden
erik.ingelsson@medsci.uu.se

For the website see <http://www.ubble.co.uk>

available measurements and 5 year all-cause and cause-specific mortality. We also aimed to create a prognostic index including only self-reported information to estimate individual mortality risk.

Methods

Study design and participants

Participants were enrolled in the UK Biobank from April, 2007, to July, 2010, from 21 assessment centres across England, Wales, and Scotland using standardised procedures. When participants agreed to take part in UK Biobank, they visited their closest assessment centre to provide baseline information, physical measures, and biological samples. In this prospective population-based study, we included all measurements available on April 10, 2014. We excluded variables that were missing in more than 80% of the participants and all cardiorespiratory fitness test measurements because summary data were not available. Thus, we included 655 measurements categorised into ten groups in the analyses: blood assays, cognitive function, early life factors, family history, health and medical history, lifestyle and environment, physical measures, psychosocial factors, sex-specific factors, and socio-demographics. In secondary analyses, we excluded all individuals with a Charlson comorbidity index^{11,12} of more than 0 (ie, those having any serious disease or disorder). All continuous variables were categorised into quintiles and constrained to have at least 20 deaths per category. If this was not possible, the categories were collapsed until this constraint was satisfied. The Charlson comorbidity index^{11,12} was calculated using self-reported diseases, obtained through a verbal interview by a trained nurse. The UK Biobank study was approved by the North West Multi-Centre Research Ethics Committee and all participants provided written informed consent to participate in the UK Biobank study. The UK Biobank protocol is available online.¹⁰

Procedures

Participant follow-up started at inclusion in the UK Biobank study and follow-up ended on Feb 17, 2014, or death, for all participants apart from those enrolled in Scotland, which had complete information up to Dec 31, 2012. All-cause mortality included all deaths occurring before Feb 17, 2014 (or Dec 31, 2012, for the participants enrolled in Scotland). Information about causes of death was obtained from National Health Service (NHS) Information Centre for participants from England and Wales, and from the NHS Central Register, Scotland for participants from Scotland. Detailed information about the linkage procedure is available online. We defined six cause-specific mortality categories using the International Classification of Diseases, edition 10 (ICD-10), classification as follows: neoplasms, C00–D48; diseases of the circulatory system, I05–I89; diseases of the respiratory system, J09–J99; diseases of

the digestive system, K20–K93; external causes of mortality and morbidity, V01–Y84; and other diseases, all remaining ICD-10 codes.

Statistical analysis

We imputed missing data separately for men and women, using the multiple imputation by chained equations approach, with five imputed datasets and ten iterations.¹³ For each variable, we specified a predictive mean matching model, including the ten most correlated predictors of the variable or of the missing status, the Nelson-Aalen estimate of cumulative hazard, the event indicator, the assessment centre, and self-reported health¹⁴ (appendix 1, p 4). All analysis results were aggregated with Rubin's rule after appropriate transformation.¹⁵ We checked whether the imputations were acceptable by comparison of plots of the distribution of recorded and imputed values for all measurements. All of the top 20 measurements most strongly associated with overall or cause-specific mortality, and all measurements included in the prediction score had a fraction of missing values lower than 9% (median 0·5% [IQR 0·3–1·0]).

We studied the sex-specific association of each variable with all-cause and cause-specific mortality using a Cox proportional hazard model or a cause-specific proportional hazard model¹⁶ with age as a timescale. The most common category within each variable was used as a reference. To model the age-dependent effect, we used an extended Cox model with three unit step functions for individuals younger than 53 years, between 53 and 62 years, and older than 62 years. These thresholds represent the tertiles of the age distribution in the population. Hence, hazard ratios were obtained for each age category. Both unstratified and age-dependent results were reported. The prediction model was developed in the entire dataset excluding participants enrolled at the Scottish centres, whose data were used for validation. For the prediction analyses, we did a sex-specific univariate analysis using time-in-study as the timescale. Age was added as a covariate in the model and an interaction with age was included if the hazard proportionality assumption was violated (a test based on Schoenfeld residuals had a p value lower than 0·0001). Discrimination was assessed on the basis of ten-fold cross-validated Harrell's C-index, accounting for competing risk.^{17,18} The Harrell's C-index is a generalisation of the area under the receiving characteristic curve for survival data. All C-indices reported include the effect of age in addition to the examined covariates.

We chose the 20 variables with the highest C-index for each cause-specific mortality category after excluding variables that were not self-reported, and hence unsuitable for inclusion in an online questionnaire. Self-reported measurements include all those variables that are directly obtained by asking the participants and not assessed through a medical specialist or a medical device. We used a backward stepwise variable selection approach with Akaike information criterion used to select independent variables

See Online for appendix

For the UK Biobank protocol see <http://www.ukbiobank.ac.uk/wp-content/uploads/2011/11/UK-Biobank-Protocol.pdf>

For more on the linkage procedure see <http://biobank.ctsu.ox.ac.uk/crystal/refer.cgi?id=115559>

to include in the final prediction model. The score was geographically validated in participants enrolled at the only two Scottish centres that were part of the UK Biobank.

Calibration was assessed using calibration plots and Hosmer-Lemeshow tests based on risk deciles. To obtain a 5-year mortality risk representative of the UK population, we reweighted the baseline hazard using life-tables from England and Wales from the years 2009–11. We further used census information from the year 2011. See appendix 1, p 2, for details about calibration. All analyses were done with R software version 3.1.0.

Role of the funding source

The funders of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. Both authors had access to all the data and were responsible for the decision to submit the manuscript for publication.

Results

Between April, 2007, and July, 2010, 498 849 participants were enrolled from 21 centres across England, Wales, and Scotland using standardised procedures specified

in the protocol. We excluded participants with more than 80% variables missing (n=746), resulting in 498 103 (54% of whom were women) participants included in our main analyses.

Of these 498 103 participants aged 37–73 years, 8532 (39% of whom were women) died during a median follow-up of 4·9 years (IQR 4·33–5·52; table 1). The most common causes of death were lung cancer in men (n=546) and breast cancer in women (n=489). The disease-free subcohort included 355 043 (55% of whom were women) participants. 67% of the variables had less than 5% missing participants, and 73% of the variables had less than 20% missing participants. Measures of eye function were obtained in only 120 000 participants.

Self-reported health was the strongest predictor of all-cause mortality in men (C-index 0·74 [95% CI 0·73–0·75]; figure 1, appendix 2). In women, a previous cancer diagnosis was the strongest mortality discriminator (0·73 [0·72–0·74]). Questionnaire-based measurements of health and medical history were the strongest predictors of cancer, cardiovascular, respiratory, digestive, and other disease-related mortality

	Men (n=227 074)	Women (n=271 029)
Age (years)	56·75 (8·20)	56·36 (8·00)
Number of deaths	5224	3308
Follow-up (years)	4·93 (4·32–5·53)	4·94 (4·35–5·52)
Number of cause-specific deaths*		
Neoplasms	2795/5219 (53%)	2288/3307 (69%)
Three most common death causes (ICD-10 code)	Lung cancer (C34), prostate cancer (C61), and oesophageal cancer (C15)	Breast cancer (C50), lung cancer (C34), and ovarian cancer (C56)
Diseases of the circulatory system	1352/5219 (26%)	442/3307 (13%)
Three most common death causes (ICD-10 code)	Chronic ischaemic heart disease (I25), acute myocardial infarction (I21), and aortic aneurysm and dissection (I71)	Acute myocardial infarction (I21), chronic ischaemic heart disease (I25), and subarachnoid haemorrhage (I60)
Diseases of the respiratory system	292/5219 (6%)	146/3307 (4%)
Three most common death causes (ICD-10 code)	Other chronic obstructive pulmonary disease (J44), other interstitial pulmonary diseases (J84), pneumonia, unspecified organism (J18)	Other chronic obstructive pulmonary disease (J44), pneumonia, unspecified organism (J18), and other interstitial pulmonary diseases (J84)
Diseases of the digestive system	238/5219 (5%)	103/3307 (3%)
Three most common death causes (ICD-10 code)	Alcoholic liver disease (K70), vascular disorders of intestine (K55), and fibrosis and cirrhosis of liver (K74)	Vascular disorders of intestine (K55), alcoholic liver disease (K70), and fibrosis and cirrhosis of liver (K74)
External causes of morbidity and mortality	192/5219 (4%)	90/3307 (3%)
Three most common death causes (ICD-10 code)	Intentional self-harm by hanging, strangulation, and suffocation (X70), unspecified fall (W19), and fall on and from stairs and steps (W10)	Unspecified fall (W19), intentional self-harm by hanging, strangulation, and suffocation (X70), and fall on and from stairs and steps (W10)
Other diseases	350/5219 (7%)	238/3307 (7%)
Three most common death causes (ICD-10 code)	Spinal muscular atrophy and related syndromes (G12), other ill-defined and unspecified causes of mortality (R99), and unspecified dementia (F03)	Spinal muscular atrophy and related syndromes (G12), other ill-defined and unspecified causes of mortality (R99), and other septicaemia (A41)

Data are mean (SD), median (IQR), or n/N (%). *Six deaths (five men and one woman) did not have information about the main cause of death and were therefore excluded from cause-specific calculations.

Table 1: Baseline characteristics

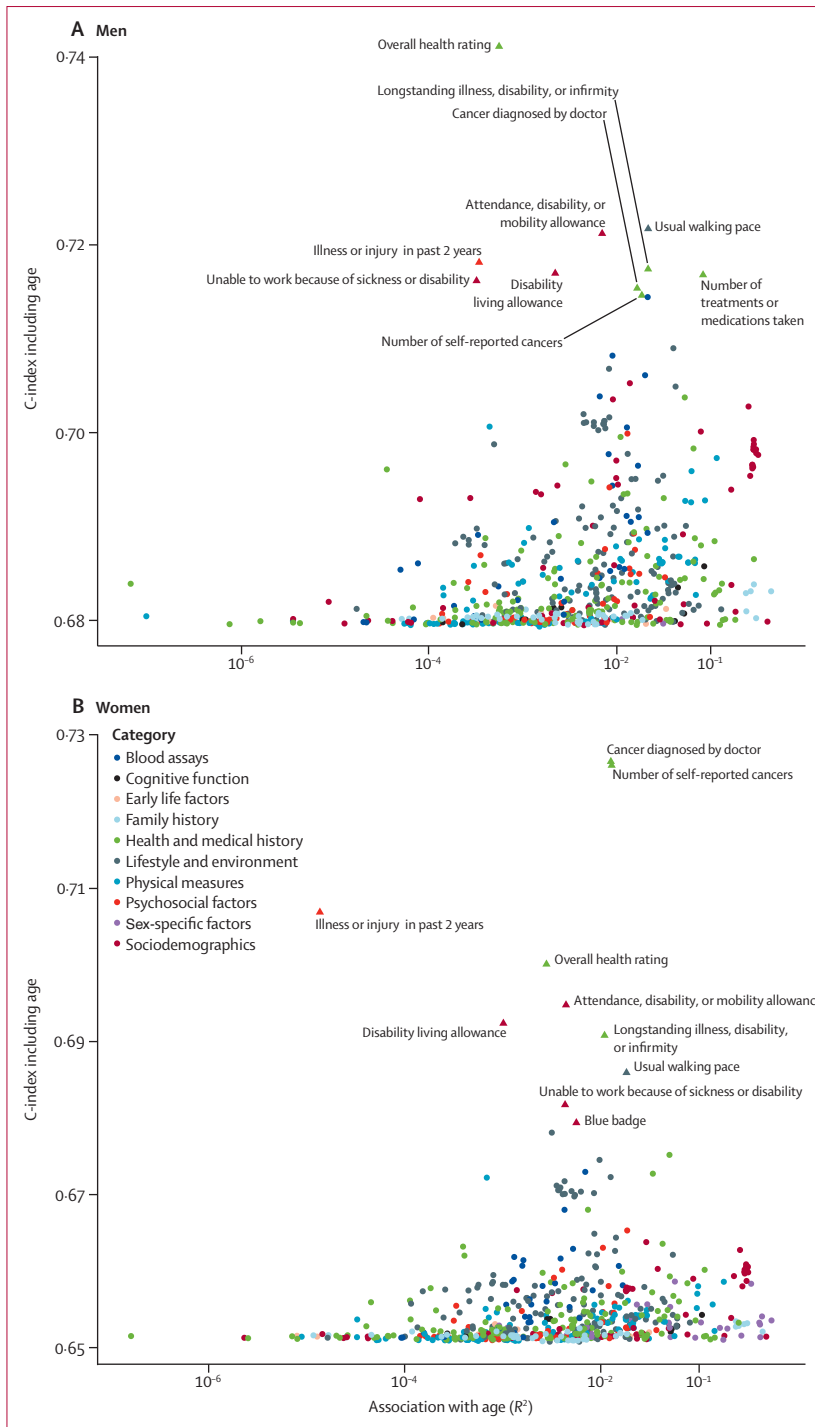


Figure 1: Ability to predict 5-year mortality for 655 measurements in men (A) and women (B)
 Each dot represents a measurement from the UK Biobank ordered by the ability to discriminate all-cause mortality (C-index, y-axis) and association with age (R^2 , x-axis). We report the C-index from models, including age-measurement interaction only if the test based on Schoenfeld residuals had a p value lower than 0.0001. Measurements with higher C-index values are better discriminators of overall mortality. The association with age can be used to identify age-dependent or age-independent measurements. For example, in men, overall health rating has a C-index of 0.74. This value estimates the probability that, given two participants, one alive at 5 years and one that died, the alive participant has a lower predicted risk of dying than the one that actually died; where the risk is obtained from the participants' age and overall health rating by fitting to a Cox proportional hazard model. Triangles represent the ten measurements with largest C-indices.

(appendix 1, pp 5–8, 10). Questions related to psychological and sociodemographic factors achieved the largest discrimination of mortality because of external causes (eg, suicide and accidental falling; appendix 1, p 9).

Self-reported walking pace was a strong predictor of mortality in both men (C-index 0.72 [95% CI 0.71–0.73]) and women (0.69 [0.68–0.70]); stronger than smoking habits and other lifestyle measurements. For example, men aged 40–52 years reporting a slow pace to the question “how would you describe your usual walking pace?” had a 3.7 (95% CI 2.8–4.8) times higher risk of death than those reporting a steady average pace (appendix 2). From blood assays, the coefficient of variation of red blood cell size was the strongest predictor of all-cause and cardiovascular mortality. Red cell distribution width is a measurement of anisocytosis, which is commonly in anaemia and other blood conditions. From physical measures (eg, spirometry, bone density, blood pressure, anthropometry, and whole-body bioimpedance measures), pulse rate and forced expiratory volume in 1 s were the strongest predictors of all-cause mortality.

We used a rank-based score to identify measurements that were consistently good discriminators of future mortality across the cause-specific mortality categories (appendix 2). Self-reported health and disability allowance had the highest score—ie, being consistently associated with mortality across several causes of death—whereas previous cancer diagnosis ranked low since it was a specific (but very strong) predictor of cancer-related mortality. Several sociodemographic measurements were strong predictors across categories. For example, although average total household income before tax was only the 33rd best discriminator of overall mortality in men, it had the seventh best ranking-based score, showing a large degree of consistency across mortality categories.

When comparing men and women (figure 2), we observed overall better prediction for men, mainly because of different underlying causes of deaths and stronger age effect in men. However, presence or self-reported illness and injuries in the past 2 years were, in relative terms, better discriminators of all-cause and cancer mortality in women than in men (figure 2, appendix 1, p 11). Similarly, smoking habits were stronger predictors of mortality due to diseases of the circulatory and respiratory systems in women than in men (appendix 1, pp 12, and 13). Differences between women and men were pronounced for mortality due to diseases of the digestive system (appendix 1, p 14).

The subsample excluding individuals with serious diseases or disorders (Charlson comorbidity index >0) consisted of 355 043 (55% women) participants, 3678 of whom died during the 4.9 years of median (IQR 4.35–5.52) follow-up. When comparing mortality predictors with those of the entire study population, we observed a reduction in discrimination for overall health and medical history measurements, whereas measures

of smoking habits were similar or stronger, effectively being the strongest predictors of all-cause mortality in both men and women (figure 3). All results can be explored interactively (panel 1).

We created a prediction score based on 13 questions for men and 11 questions for women. Variables included in the score and the coefficients for the score's calculation are available in the appendix 2. We did a geographical validation of the prediction scores by assessing their performances in the participants from the two Scottish centres (714 of 35 810 died), which were not used to develop the prediction score. The score had good discrimination abilities (C-index 0.80 [95% CI 0.77–0.83] for men and 0.79 [0.76–0.83] for women; table 2) and substantially improved prediction compared with age alone (0.68 [0.65–0.71] for men and 0.67 [0.64–0.71] for women; $p < 0.0001$). These results were similar to a score obtained with lasso penalised Cox's regression (0.80 [0.77–0.83] for men and 0.80 [0.76–0.83] for women; table 2 and appendix 1, p 4), which consisted of an increased number of measurements, including those not assessable through an online questionnaire. Our prediction score with 13 or 11 questions showed significantly better discrimination than the Charlson comorbidity index (0.75 [0.72–0.77] for men and 0.76 [0.73–0.80] for women; for comparison with the main prediction model obtained by bootstrapping $p < 0.0001$ and $p = 0.0007$). This finding is not surprising in view of the low prevalence of major comorbidities in study participants. When the Charlson index was added to the prediction score, the discrimination improvement was modest (C-index improvement 0.001; table 2).

The calibration was good for women ($p = 0.28$) but poor for men ($p = 0.0402$) with the Hosmer and Lemeshow test; table 2 and appendix 1, p 15), resulting in a general underestimation of the true risk in men. The reason for the poorly calibrated prediction score in Scottish men was their higher mortality (appendix 1, p 16). The discrimination ability of the prediction score decreased with age. For example, the C-index was 0.84 and 0.81 for men and women aged 40–50 years, but only 0.72 for men and women aged 60–70 years (appendix 1, p 17).

Participants from UK Biobank showed lower mortality than the general population (appendix 1). We therefore calibrated our prediction score using UK life tables and census information (appendix 1, p 2). The data used to obtain the 5-year mortality risk is reported in the appendix 2, together with an example. The calculations are also implemented in an online questionnaire.

Discussion

In this large, contemporary prospective cohort study, we did an extensive analysis of associations of more than 600 measurements with 5 year all-cause and cause-specific mortality in UK Biobank participants). In this report, we have presented only a small part of our findings; however, all our results are available in an

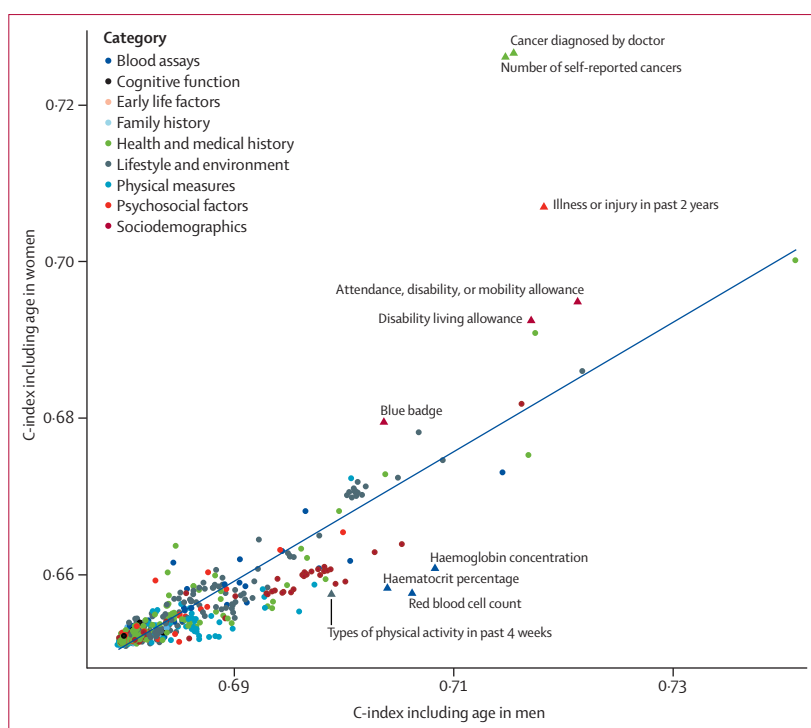


Figure 2: Comparison between men and women of the ability to predict 5-year mortality

Each dot represents a measurement from the UK Biobank ordered by the ability to discriminate all-cause mortality (C-index) in men versus women. We report the C-index from models, including age-measurement interaction only if the test based on Schoenfeld residuals had a p value lower than 0.0001. Measurements deviating from the regression line (shown in blue) have a C-index that differs between men and women more than expected in view of the overall association between the two sexes. For example, even if the C-index for illness or injury in the last 2 years is lower in women than in men (C-index 0.71 vs 0.72), it substantially deviates from the regression line, suggesting a stronger relative importance in women than in men. Each colour represents a different measurement category. Triangles represent the ten measurements that are more strongly deviating from the regression line.

interactive website where the observed associations can be explored in detail to generate new research hypotheses. Several key messages can be deduced from this study. First, measures that can simply be obtained by verbal interview without physical examination are the strongest predictors of all-cause mortality in middle-aged to elderly individuals. Self-reported health and walking pace were the strongest predictors in both sexes and across different causes of deaths. Second, several risk factors have different discrimination abilities in men and women. These differences can be explained by different diseases underlying the observed mortality, sex-specific reporting biases, or true biological differences. Third, in previously healthy individuals, smoking habits remains the strongest category of mortality predictors.

Most of the adult population in developed countries seeks out health information online.¹⁹ Although this approach could result in overdiagnosis and anxiety,^{20,21} online information has the potential to improve self-awareness and understanding of health determinants in the public, help with patient–physician interaction,²² and increase shared decision making.²³ As a way to disseminate our findings, we developed a prediction

For an **interactive representation of results** see <http://www.ubble.co.uk>

For more on the **questionnaire** see <http://www.ubble.co.uk>

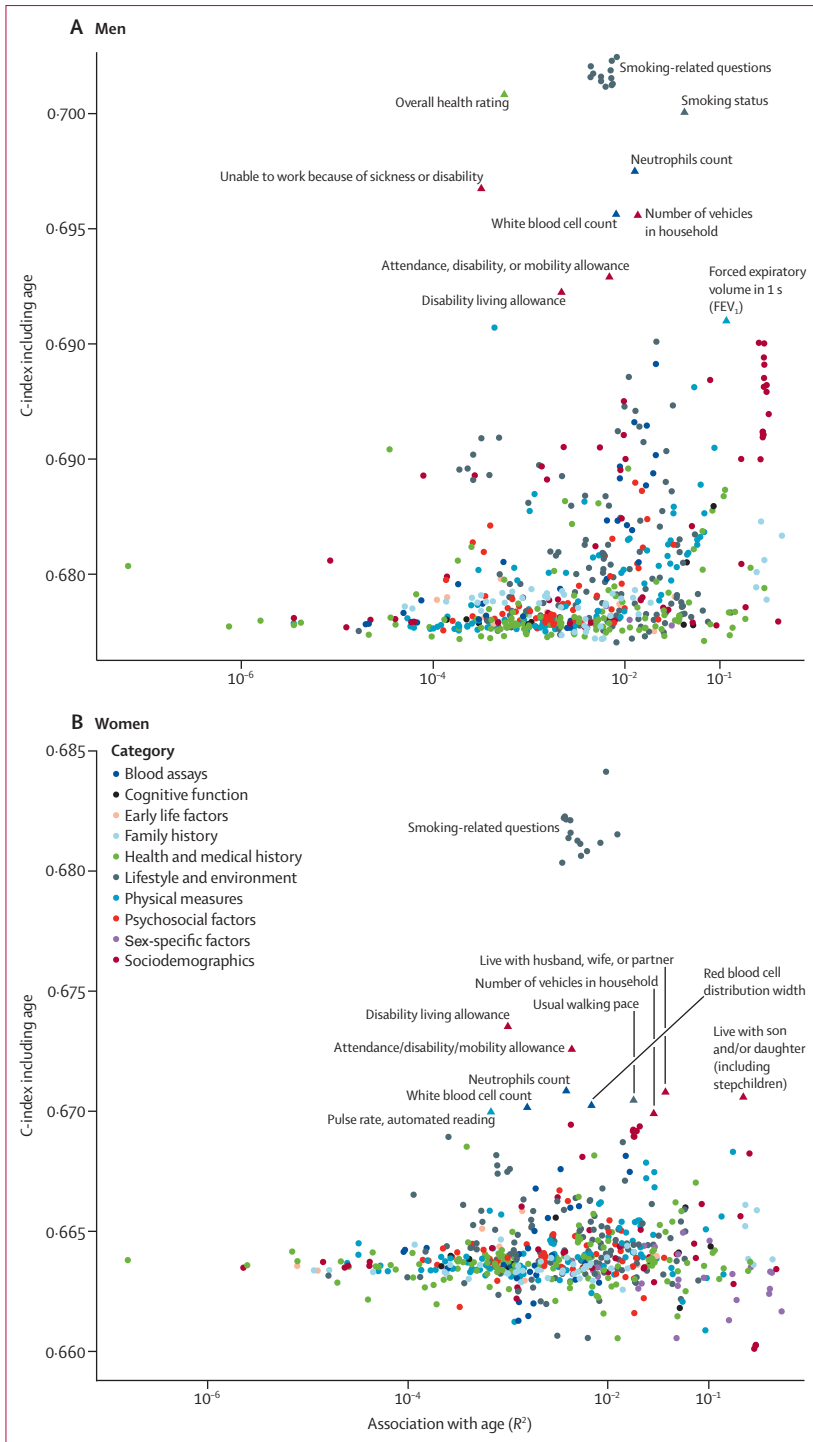


Figure 3: Ability to predict 5-year mortality for 655 measurements in men (A) and women (B) without previous major disease
 Each dot represents a measurement from the UK Biobank ordered by the ability to discriminate all-cause mortality (C-index, y-axis) and association with age (R^2 , x-axis) in individuals with a Charlson index equal to zero. We report the C-index from models including age-measurement interaction only if the test based on Schoenfeld residuals had a p value lower than 0.0001. Measurements with a higher C-index are better discriminators of overall mortality. The association with age can be used to identify age-dependent or age-independent measurements. Measures of smoking habits were the strongest predictors of all-cause mortality in both men and women. Each colour represents a different measurement category. Triangles represent the ten measurements with the largest C-index.

Panel 1: Ubble website

All results can be explored interactively at the Ubble website. Information is presented in separate plots for men and women, where each measurement is shown by a dot. When clicking on the dots, information about the association with mortality in terms of hazard ratios and p values are shown. A point-and-click interface below the plots allows selection of the cause-specific mortality categories or visualisation of only one specific category of measurements. A web-based questionnaire is available for the calculation of the personalised 5 year mortality risk.

	Men		Women	
	p value	C-index	p value	C-index
Discrimination				
Age	..	0.679	..	0.672
Age + Charlson score	..	0.746	..	0.762
Age + prediction model	..	0.800	..	0.794
Age + Charlson score + final prediction model	..	0.801	..	0.795
Lasso regression	..	0.803	..	0.796
Calibration: Hosmer-Lemeshow test				
Age + prediction model	0.0402	..	0.28	..

Table 2: Calibration and discrimination characteristics of the prediction model for 5-year mortality in participants from the Glasgow and Edinburgh centres (geographical validation)

score based on self-reported information that can be used to calculate a personalised 5 year mortality risk. The score is calibrated for individuals aged 40–70 years living in the UK, but generalisable to other populations under two main assumptions. First, the associations observed in the UK Biobank for the risk factors included in the score should be generalisable. This assumption is sufficient to claim the generalisability of the discrimination performance of the score. Second, the distribution of the risk factors and the mortality rates should be similar to those in the UK as a whole. If the second assumption is not satisfied, applying an approach such as the one used in this study, the score can be recalibrated with external information from national statistics, census data, or survey studies.

The risk factors included in the prediction model were obtained with an automatic variable selection process with the aim to maximise prediction. Few of the risk factors are actionable on the individual level and most do not directly cause disease; however, this should not be interpreted as an impediment to improvement of health status. Increased physical activity, smoking cessation, and a healthy diet can improve lifespan and reduce prevalence of serious diseases.^{24,25} The proposed prediction score has several potential applications at the individual, clinical, and public policy levels. At the individual level, it can be

used to improve self-awareness of the health status, providing incentives for lifestyle changes. Clinicians might use this score to identify patients at high risk of mortality to target with specific interventions. Finally, government and health organisations can use this information to prioritise public policy to decrease the burden of specific risk factors.

To our knowledge, no previous studies have examined such an extensive number of measurements from a large epidemiological study (panel 2). Several of the observed associations have already been reported in the scientific literature;^{4,30–32} however, our approach allows the ranking of these measurements in terms of discrimination ability, providing information about the relative importance of each variable as predictor of all-cause and cause-specific mortality.

All analysis are sex-stratified and adjusted only for age. The reasons for this approach were that the main purpose of this study was not to address causality, but rather prediction; and for consistency across measurements (as different risk factors would have needed different adjustments if we wanted to assess causes). Hence, observed associations are likely to be confounded and researchers should consider proper adjustments before claiming causality. Moreover, discrimination abilities of the single measurements and of the risk score should be assessed in an age-dependent context. Overall, we observed a lower discrimination for older participants in the present study, and we could not investigate discrimination in individuals older than 70 years using data from the UK Biobank. In view of the large number of statistical tests done, approaches to take multiple testing into account should be considered when interpreting the significance of the reported associations. The UK Biobank had a response rate of about 5·5%. Even if the distribution of measurements in the UK Biobank is different from the general population, the generalisability of the associations is guaranteed as long as sufficient heterogeneity across measurements exists, as previously discussed.^{33,34}

For example, in men, the hazard ratio for association between poor or fair versus excellent or good self-reported health was 2·8 (95% CI 2·7–3·0), which is similar to what was observed in a large survey sample from the USA (OR [odds ratio] 3·2 [3·1–3·3]) in white participants), which has been designed to be representative of the US population aged 18 years or older.⁴

Nevertheless, response bias might have affected the generalisability of the observed associations and to a lesser extent (because of the multivariable nature of the model) that of the prediction score. Moreover, self-reported variables are always subject to misclassification bias and this aspect should be considered when interpreting the results from the association analyses. Generally, mortality data from death certificates are subject to inaccurate reporting with misclassification of death causes.³⁵ In the UK, routine checks are done by the Office for National

Panel 2: Research in context

Systematic review

We searched PubMed for published studies in all languages with the term “(overall or all-cause) and mortality and (prognostic or prediction)” from Nov 1, 1994, to Nov 15, 2014. We excluded studies not done in a population or community-based setting and studies published before 2000. Most of the studies explored only one risk factor^{2–6} or only a few risk factors.^{26–28} Walter and colleagues⁷ examined the predictive abilities of 162 measurements in 5974 participants from the Rotterdam study. Baer and colleagues²⁹ explored the association between 18 risk factors and cause-specific mortality by using a competing risk approach similar to ours. Prognostic indices for mortality in older adults (>60 years old) have been reviewed by Yourman and colleagues.⁸ Six indices were developed in community-dwelling patients, but all studies had lower sample sizes and used fewer predictors for score development than did our study. None of the previous studies have been done in the UK. A collection of calculators in elderly people is available online.

Interpretation

To the best of our knowledge, this is the first study to provide a detailed analysis of predictors of mortality in a large, contemporary prospective cohort study. Moreover, we provide the first prognostic index for 5-year mortality calibrated for the UK population between 40 and 70 years old. Our findings suggest that measures that can be simply obtained by verbal interview without physical examination (eg, self-reported health and walking pace) are the strongest predictors of all-cause mortality. Moreover, we could rank the measurements in terms of their predictive abilities, provide new insights about their relative importance, and show some unexpected associations. The most predictive self-reported measurements were combined in a prediction score that could be used by non-professionals to improve self-awareness of their health status, by clinicians to identify patients at high risk of mortality to target with specific interventions, and by governmental and health organisations to decrease the burden of certain risk factors.

Statistics to increase the quality of mortality data³⁶ and validation studies cross-referencing outcomes across death, cancer, primary, and secondary care datasets are continuing in the UK Biobank. According to a large study in Finland,³⁷ the largest misclassification is expected between deaths from diseases of the circulatory system and diseases of the respiratory system.

In summary, we present an untargeted, large-scale examination of all-cause and cause-specific mortality predictors in middle-aged to elderly individuals. Our results are easily accessible at a dedicated website to help with the necessary exploration and generation of new research hypotheses. Potential extensions of our study include the reporting of associations of additional urine, blood, and DNA biomarkers, the extension of the prediction score to a longer follow-up, and validation in other populations.

Contributors

AG did the statistical analyses. Both authors developed the study design, planned the analyses, contributed to the interpretation of data, and wrote the manuscript.

Declaration of interests

We declare no competing interests.

Acknowledgments

This work was funded by Knut and Alice Wallenberg Foundation and the Swedish Research Council. This research has been done using the UK Biobank resources. We thank Robert Karlsson and Stefan Gustafsson for their work with design of the website and Therese Andersson,

For the **Ubble website** see <http://www.ubble.co.uk>

For **calculators in elderly people** see <http://eprognoisis.ucsf.edu/default.php>

Mark Clements, and Alexander Ploner for the useful discussions about the statistical methods. We would also like to thank Sense About Science for providing invaluable input regarding the content of the website. The computations were done using resources provided by Swedish National Infrastructure for Computing through Uppsala Multidisciplinary Center for Advanced Computational Science under Project b2011036.

References

- Oortwijn W, Nelissen E, Adamini S, van den Heuvel S, Geuskens G, Burdof L. Social determinants state of the art reviews—health of people of working age—full report. Luxembourg: European Commission Directorate General for Health and Consumers, 2011.
- Banks E, Joshy G, Abhayaratna WP, et al. Erectile dysfunction severity as a risk marker for cardiovascular disease hospitalisation and all-cause mortality: a prospective cohort study. *PLoS Med* 2013; **10**: e1001372.
- Hu FB, Willett WC, Li T, Stampfer MJ, Colditz GA, Manson JE. Adiposity as compared with physical activity in predicting mortality among women. *N Engl J Med* 2004; **351**: 2694–703.
- McGee DL, Liao Y, Cao G, Cooper RS. Self-reported health status and mortality in a multiethnic US cohort. *Am J Epidemiol* 1999; **149**: 41–46.
- Thun MJ, Peto R, Lopez AD, et al. Alcohol consumption and mortality among middle-aged and elderly U.S. adults. *N Engl J Med* 1997; **337**: 1705–14.
- Tota-Maharaj R, Blaha MJ, McEvoy JW, et al. Coronary artery calcium for the prediction of mortality in young adults <45 years old and elderly adults >75 years old. *Eur Heart J* 2012; **33**: 2955–62.
- Walter S, Mackenbach J, Voko Z, et al. Genetic, physiological, and lifestyle predictors of mortality in the general population. *Am J Public Health* 2012; **102**: e3–10.
- Yourman LC, Lee SJ, Schonberg MA, Widera EW, Smith AK. Prognostic indices for older adults: a systematic review. *JAMA* 2012; **307**: 182–92.
- Eagle KA, Lim MJ, Dabbous OH, et al. A validated prediction model for all forms of acute coronary syndrome: estimating the risk of 6-month postdischarge death in an international registry. *JAMA* 2004; **291**: 2727–33.
- Palmer LJ. UK Biobank: bank on it. *Lancet* 2007; **369**: 1980–82.
- Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J Chronic Dis* 1987; **40**: 373–83.
- Quan H, Sundararajan V, Halfon P, et al. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. *Med Care* 2005; **43**: 1130–39.
- White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. *Stat Med* 2011; **30**: 377–99.
- White IR, Royston P. Imputing missing covariate values for the Cox model. *Stat Med* 2009; **28**: 1982–98.
- Marshall A, Altman DG, Holder RL, Royston P. Combining estimates of interest in prognostic modelling studies after multiple imputation: current practice and guidelines. *BMC Med Res Methodol* 2009; **9**: 57.
- Prentice RL, Kalbfleisch JD, Peterson AV Jr, Flournoy N, Farewell VT, Breslow NE. The analysis of failure times in the presence of competing risks. *Biometrics* 1978; **34**: 541–54.
- Wolbers M, Blanche P, Koller MT, Witteman JC, Gerds TA. Concordance for prognostic models with competing risks. *Biostatistics* 2014; **15**: 526–39.
- Harrell FE Jr, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. *JAMA* 1982; **247**: 2543–46.
- Hesse BW, Nelson DE, Kreps GL, et al. Trust and sources of health information: the impact of the Internet and its implications for health care providers: findings from the first Health Information National Trends Survey. *Arch Intern Med* 2005; **165**: 2618–24.
- Horvitz E. Cyberchondria: studies of the escalation of medical concerns in web search. *ACM TransManag Inf Syst* 2009; **27**: 231–37.
- Sillence E, Briggs P, Harris PR, Fishwick L. How do patients evaluate and make use of online health information? *Soc Sci Med* 2007; **64**: 1853–62.
- White RW, Horvitz E. Experiences with web search on medical concerns and self diagnosis. *AMIA Annu Symp Proc* 2009; **2009**: 696–700.
- Sheridan SL, Harris RP, Woolf SH. Shared Decision-Making Workgroup of the USPSTF. Shared decision making about screening and chemoprevention. a suggested approach from the U.S. Preventive Services Task Force. *Am J Prev Med* 2004; **26**: 56–66.
- Rizzuto D, Orsini N, Qiu C, Wang HX, Fratiglioni L. Lifestyle, social factors, and survival after age 75: population based study. *BMJ* 2012; **345**: e5568.
- Li K, Husing A, Kaaks R. Lifestyle risk factors and residual life expectancy at age 40: a German cohort study. *BMC Med* 2014; **12**: 59.
- Pavlik VN, de Moraes SA, Szklo M, Knopman DS, Mosley TH Jr, Hyman DJ. Relation between cognitive function and mortality in middle-aged adults: the atherosclerosis risk in communities study. *Am J Epidemiol* 2003; **157**: 327–34.
- Varadhan R, Yao W, Matteini A, et al. Simple biologically informed inflammatory index of two serum cytokines predicts 10 year all-cause mortality in older adults. *J Gerontol A Biol Sci Med Sci* 2014; **69**: 165–73.
- Newman AB, Sachs MC, Arnold AM, et al. Total and cause-specific mortality in the cardiovascular health study. *J Gerontol A Biol Sci Med Sci* 2009; **64**: 1251–61.
- Baer HJ, Glynn RJ, Hu FB, et al. Risk factors for mortality in the nurses' health study: a competing risks analysis. *Am J Epidemiol* 2011; **173**: 319–29.
- Doll R, Peto R, Boreham J, Sutherland I. Mortality in relation to smoking: 50 years' observations on male British doctors. *BMJ* 2004; **328**: 1519.
- Elbaz A, Sabia S, Brunner E, et al. Association of walking speed in late midlife with mortality: results from the Whitehall II cohort study. *Age* 2013; **35**: 943–52.
- Jouven X, Empana JP, Schwartz PJ, Desnos M, Courbon D, Ducimetiere P. Heart-rate profile during exercise as a predictor of sudden death. *N Engl J Med* 2005; **352**: 1951–58.
- Manolio TA, Collins R. Enhancing the feasibility of large cohort studies. *JAMA* 2010; **304**: 2290–91.
- Collins R. What makes UK Biobank special? *Lancet* 2012; **379**: 1173–74.
- Maudsley G, Williams EM. Inaccuracy' in death certification—where are we now? *J Public Health Med* 1996; **18**: 59–66.
- Office for National Statistics. Mortality Statistics: Metadata. 2014. <http://www.ons.gov.uk/ons/guide-method/user-guidance/health-and-life-events/index.html> (accessed Nov 15, 2014).
- Lahti RA, Penttila A. The validity of death certificates: routine validation of death certification and its effects on mortality statistics. *Forensic Sci Int* 2001; **115**: 15–32.