DEMOGRAPHIC RESEARCH
A peer-reviewed, open-access journal of population sciences

DEMOGRAPHIC RESEARCH

VOLUME 50, ARTICLE 36, PAGES 1071–1112
PUBLISHED 22 MAY 2024
http://www.demographic-research.org/Volumes/Vol50/36/
DOI: 10.4054/DemRes.2024.50.36

Research Article

Incorporating subjective survival information in mortality and change in health status predictions: A Bayesian approach

Apostolos Papachristos

Dimitris Fouskakis

Contents

Demographic Research: Volume 50, Article 36
Research Article

Incorporating subjective survival information in mortality and change in health status predictions: A Bayesian approach

Apostolos Papachristos[1]

Dimitris Fouskakis[2]

Abstract

BACKGROUND
Subjective survival probabilities incorporate individuals' view about own future survival and they are associated with actual mortality patterns.

OBJECTIVE
The objective of this study is twofold. First, we apply a Bayesian methodology to incorporate the respondents' views about future survival, and second, we investigate whether subjective survival information is useful for predicting actual mortality and self-reported change in health.

METHODS
To achieve the above-mentioned objective, we adopt a two-step process. In the first step, we use a Bayesian linear regression model, under default priors, on the logit transformation of the subjective mortality probabilities to estimate the posterior distribution of the regression coefficients of the available explanatory variables. In the second step, we fit Bayesian logistic regression models on actual mortality and self-reported change in health, using a variety of priors derived from the posterior distributions of the first step Bayesian model. Data from the Health and Retirement Study (HRS) Waves 13 and 14 are used in this paper.

CONCLUSIONS
We conclude that the additional information incorporated via the subjective mortality probabilities is useful for predicting actual mortality but less useful for predicting self-reported change in health.

[1] Department of Mathematics, National Technical University of Athens, Greece.
Email: apapachristos@mail.ntua.gr.
[2] Department of Mathematics, National Technical University of Athens, Greece.
Email: fouskakis@math.ntua.gr.

CONTRIBUTION
The contribution of this study relates to the development of a procedure, which can be used to include prior information – based on subjective survival views – in hierarchical Bayesian regression models to improve the ability to predict mortality and self-reported change in health.

1. Introduction

Subjective survival probabilities are quantities that incorporate individuals' view on likely future survival, and they vary considerably across survey respondents (Hamermesh 1985). Gender is one of the main factors affecting survival expectations. Arpino, Bordone, and Scherbov (2018), using objective survival probabilities estimated from the US Health and Retirement Study, conclude that males tend to report higher survival expectations than females. Another important finding is that older people tend to report higher survival expectations compared to younger people. Several researchers have estimated a positive association between age and subjective life expectancy (Griffin, Loh, and Hesketh 2013; Mirowsky 1999; Ross and Mirowsky 2002).

Higher education and higher income are associated with higher subjective survival probabilities (Arpino, Bordone, and Scherbov 2018; Rappange, Brouwer, and van Exel 2016; Liu, Tsou, and Hammitt 2007; Mirowsky 1999; Balia 2014). Poor physical health is clearly associated with lower survival expectations (Liu, Tsou, and Hammitt 2007; Balia 2014; Hurd and McGarry 1995). In addition, several researchers find that the number of chronic diseases is associated with lower subjective survival probabilities (Rappange, Brouwer, and van Exel 2016). The evidence on the effect of smoking status on subjective survival probabilities is not conclusive. Khwaja, Sloan, and Chung (2007), using data from the Health and Retirement Study, note that smokers are optimistic whereas people who have never smoked are pessimistic in their survival predictions. Other researchers conclude also that current smokers overestimate survival (Liu, Tsou, and Hammitt 2007). On the other hand, it is well known that smoking is associated with heavier actual mortality (Doll et al. 1994). Race differentiates significantly subjective as well as actual survival expectations. Prior research notes that Black Americans tend to report higher survival expectations compared to White Americans of the same age, but they experience lower life expectancy at birth (Mirowsky 1999; Firebaugh et al. 2014).

One of the most interesting and challenging topics on subjective survival is the assessment of the accuracy of subjective survival expectations. On this topic, several researchers noted that actual in-sample mortality is associated with the subjective predictions. More specifically, Bago d'Uva et al. (2020) argue that the formation of accurate longevity expectations involves acquisition of health knowledge, perception of mortality risks and processing of information. They measure the accuracy of subjective predictions by comparing the subjective probabilities of living to 75 – as reported by Health and Retirement Study (HRS) – with an indicator of whether respondents survive to that age. Their findings confirm that subjective probabilities of survival to 75 predict respondents' survival to that age.

The predictive power of subjective survival probabilities about the actual survival from one survey wave to another has been investigated. Van Doorn and Kasl (1998) note that both self-rated health and self-rated life expectancy predict actual mortality up to the next wave, after controlling for health status and sociodemographic factors. They conclude that these two quantities contain different information about own future survival. Smith, Taylor Jr, and Sloan (2022) suggest that longevity expectations are good predictions of future mortality, because they are consistently updated with new health information but they do not incorporate all information relevant to future survival. Furthermore, Hurd and McGarry (2002) note that HRS respondents who were alive in HRS Waves 1 and 2 reported on average 50 % higher subjective survival probabilities than those who died between these waves. Other studies of the HRS and other longitudinal surveys demonstrate that subjective survival probabilities have predicted survival to a subsequent survey wave (Hurd 2009; Siegel, Bradley, and Kasl 2003). Elder (2013) conclude that the predictive ability of subjective survival probabilities depends at respondents' age. In particular, the subjective survival probabilities for respondents age 65 or less predict actual in-sample mortality well.

Subjective or self-rated health reflects individuals' perceptions about own health, and it has been shown that it is a strong predictor for mortality irrespective of objective health status, age, gender, and other sociodemographic factors (Mossey and Shapiro 1982). Higher subjective survival probabilities are associated with better self-rated health (Papachristos et al. 2020; Rappange, Brouwer, and van Exel 2016), and subjective survival probabilities are consistently updated after an improvement or a deterioration of self-rated health (Papachristos and Verropoulou 2022). Furthermore, the association of subjective health with physical and functional health were lower for older individuals compared to younger individuals (Pinquart 2001). In addition, previous studies note that the impact of serious health events on subjective health depends on age. In particular, younger individuals reduce their subjective health assessment after a serious health event more than older individuals (Wurm, Tomasik, and Tesch-Römer 2008). This could be related to the ageing process as health deterioration is normally expected in later life and older individuals are prepared (Heckhausen and Krueger 1993). Verropoulou (2014) notes that improvement in self-rated health is a more complex concept than deterioration and that is affected by behavioural risk factors and physical activity whereas older chronological age is related to lower self-rated health. On the one hand, worse subjective health assessment is strongly related to poor physical health, but on the other hand, better subjective health is a more complex concept, relating not only to the lack of illness but also to sociodemographic characteristics and self-image (Smith, Shelley, and Dennerstein 1994).

The objective of this study is twofold. First, we apply a Bayesian procedure to incorporate prior information, based on the respondents' views about future subjective survival views in Bayesian logistic regression models, to improve their ability to predict mortality. Our hypothesis is that the respondents' subjective survival views can be used to derive informative priors about the mean values of the coefficients of the predictors, which would provide supplementary information to the Bayesian mortality models. Second, we investigate whether the respondents' views about future survival include information useful for predicting future self-reported change in health.

The main contribution of this study relates to the development of a process which can be used to incorporate subjective survival views under informative priors in Bayesian mortality and (self-reported) change in health logistic regression models. The differential utility of subjective survival expectations on both mortality and perceived change in health is clearly demonstrated. One of the advantages of this study relates to the mismatch between the event that respondents are asked to forecast (i.e., survival to a given target age) and the evaluation of the actual outcome of this prediction (i.e., survival to the next survey wave) (see also Bago d'Uva et al. (2020); Hurd and McGarry (2002)). To address this issue, we propose a technique that adjusts self-reported subjective survival probabilities in order to reflect the two-year period up to the next wave. Then, this study focuses on the prior elicitation of subjective survival information using several choices for priors.

2. Case study

The study uses data from the HRS, which is a longitudinal household survey conducted by the Institute for Social Research at the University of Michigan. The HRS is sponsored by the National Institute on Aging (grant number NIA U01AG009740) and is conducted by the University of Michigan. It is an age-cohort-based longitudinal panel survey of persons aged 50 and older in the United States. The harmonised version of the longitudinal studies, RAND HRS, is provided by the Gateway to Global Aging Data. The RAND HRS Longitudinal File 2018 includes all 14 waves (HRS 2022). The data for HRS Wave 13 (W13) were collected in 2016 and consist of 20,147 respondents aged 50 or older, whereas the data for HRS Wave 14 (W14) were collected in 2018, two years later.

The W13 sample size is restricted to respondents who answered both survival probability questions. The main reason for this methodological decision is to address the mismatch between the time interval of HRS waves (about 2 years) and the time horizon which corresponds to subjective survival probabilities (on average 10 to 15 years). More specifically, as described in Section 3, we fit a Weibull model with two parameters, namely α and β, using non-linear optimization methods, in order to estimate the subjective survival probabilities for the next 2 years.

In terms of sample size, 2,257 respondents who did not provide a response at all, 96 respondents who answered only the 1st question, and 8,361 respondents who answered only the 2nd question are excluded. It is worth mentioning that based on HRS documentation the 1st question is asked to individuals aged 64 or younger only. This age-related threshold explains the large number of respondents who answered the 2nd subjective survival question only.

In addition, 296 respondents aged 50 or older who inconsistently answered both survival questions are excluded from the analysis (see Section 3). Finally, 174 cases with missing values in the explanatory variables are also excluded from the analysis. Additional statistical testing (Little 1988), as a sensitivity check, is performed to investigate if these missing values are missing completely at random (Rubin 1976). The results using the Fully Conditional Specification imputation method (Van Buuren et al. 2006) are presented in Appendix A-7. The final sample used for the mortality analysis consists of $n = 8,963$ respondents. The chronological age of these respondents ranges from 50 to 65 years old. It is worth emphasizing that our study targets the respondents whose chronological age ranges from 50 to 65 years old and our results are relevant to only this age group.

To investigate the first objective, a binary variable which indicates whether a respondent is alive (= 0) or dead (= 1) at Wave 14 is used as a response variable. In this sample, 104 respondents who participated in Wave 13 reported as dead in Wave 14, and the crude mortality rate is estimated by 2.56%. To investigate the second objective, a variable which indicates the self-reported change in health is used. In particular, at Wave 14 respondents were asked if their health is better, about the same, or worse since the last interview, at Wave 13. Based on this interview question, a binary variable taking the value of 0 if a respondent reported the same and somewhat better health status at Wave 14 – compared to the health status at Wave 13 – and the value of 1 if a respondent has a worse health status at Wave 14 – compared to the health status at Wave 13 – is used as a response variable. In terms of sample size, the starting point is the dataset used for the mortality analysis as described above. We noticed that 1,628 respondents who participated at Wave 13 – and they are included in the final sample used for the mortality analysis – did not provide a response for the self-reported change in health question at Wave 14.

Additional investigation about the missing values for the dependent variable, self-reported change in health, shows that about 90.5% of these correspond to respondents who participated in Wave 13 but not in Wave 14. The remaining 9.5% corresponds to respondents who died between the two waves, as well as those who participated but did not provide a response on their health change. Hence, the sample used for the change in health analysis reduces to $n = 7,335$ respondents. In this sample 1,333 respondents (18.2%) reported worse health status at Wave 14 compared to that at Wave 13. In addition, as a sensitivity check, in order to address the challenge of the missing values for the dependent variable, we assume that the missing mechanism is ignorable (Little and Rubin 2019), and by treating these missing values as unknown parameters, we estimate them (simultaneously with the rest of the parameters), under the Bayesian approach, using the posterior predictive distribution (see Appendix A-7).

2.1 Explanatory variables

The group of sociodemographic variables includes chronological age (in years), gender (factor with two levels – male = 0 and female = 1), race (factor with two levels – White American = 0 and Black American or other = 1), and years of education. Marital status (factor with two levels – married or partnered = 0 and separated, divorced, widowed, or never been married = 1) as well as the (standardized) total household income are also included. Regarding physical health, the number of limitations in activities of daily living (ADLs) out of a list of six basic/everyday tasks, the number of instrumental activities of daily living (IADLs) out of a list of three tasks (namely using the phone, managing money, and taking medications), and self-rated health (ranging from 1 = poor to 5 = excellent) are included in the models. In addition, the number of chronic conditions (ranging from 0 to 8) is also induced in the models. The eight chronic conditions included in the analysis are (1) high blood pressure or hypertension; (2) diabetes or high blood sugar; (3) cancer or a malignant tumor of any kind except skin cancer; (4) chronic lung disease except asthma, such as chronic bronchitis or emphysema; (5) heart attack, coronary heart disease, angina, congestive heart failure, or other heart problems; (6) stroke or transient ischemic attack; (7) emotional, nervous, or psychiatric problems; and (8) arthritis or rheumatism.

In terms of mental health, depression is measured using the Center for Epidemiologic Studies Depression scale (ranging from 0 to 8; higher scores indicate more severe depression) (Radloff 1977; Lewinsohn et al. 1997). Regarding behavioral risk factors, the frequency of vigorous exercise (ranging from 1 = never to 5 = more than three times a week) and the smoking status (0 = never smoker, 1 = past smoker, and 2 = current smoker) are also included in the analysis.

2.2 Sample description

The sample characteristics are presented in Table 1. On average, subjective mortality probabilities, defined in Section 3, are higher for respondents who reported being alive at Wave 13 but not alive at Wave 14. Males are less likely to survive than females, and married or partnered people are more likely to survive than people who are divorced, or widowed, or have never been married. Black Americans (or other race) exhibit higher mortality than White Americans. Respondents with poor health, more disabilities, and more difficulties in performing everyday tasks as well as those who are more depressed tend to exhibit heavier in-sample mortality. Moreover, lower average household income and fewer years education are associated with higher in-sample mortality. On average respondents who do vigorous exercise and nonsmokers exhibit lower in-sample mortality.

Table 1: Sample characteristics

Variable	Reported alive at Wave 13 and Wave 14	Reported alive at Wave 13 but not alive at Wave 14	Total
Subjective mortality probability (mean)	1.06%	2.56%	1.07%
Female (% of total)	56.6%	53.8%	56.6%
Chronological age (mean)	57.1	58.0	57.2
Black Americans or other race (% of total)	42.2%	52.9%	42.3%
Years of education (mean)	13.1	12.6	13.1
Married or partnered (% of total)	65.7%	48.1%	65.4%
Household income (mean in $)	89,753	41,840	89,197
Self-rated health (mean)	3.1	2.1	3.1
Number of ADLs (mean)	0.3	0.8	0.3
Number of IADLs (mean)	0.1	0.3	0.1
Number of chronic conditions (mean)	1.7	3.0	1.7
Depression (mean)	1.6	3.2	1.6
Vigorous exercise (mean)	2.2	1.6	2.2
Smoking status (mean)	0.7	1.2	0.7

Notes: ADLs: Number of Activities of Daily Living for which respondents report difficulties. IADLs: Number of Instrumental Activities of Daily Living for which respondents report difficulties

3. Subjective mortality probabilities

In HRS there are two questions about future survival expectations. The first question relates to the self-reported probability of living to age 75 specifically (i.e., the target age is 75). The second question relates to self-reported probability of living to age 80 to 100. In this question the target age varies from 80 to 100 depending on respondent's chronological age. In particular, if a respondent is aged 65 or less the assigned target age is 85; if the respondent is aged 65 to 69 the assigned target age is 80; if the respondent is aged 70 to 74 the assigned target age is 85; if the respondent is aged 75 to 79 the assigned target age is 90; if the respondent is aged 80 to 84 the assigned target age is 95; and if the respondent is aged 85 to 89 the assigned target age is 100. The target age for all respondents is provided by HRS. Therefore, HRS respondents report subjective survival probabilities which on average refer to the next 10 to 15 years.

Self-reported subjective survival probabilities have three issues which need to be addressed (Perozek 2008). First, as mentioned in Section 2, about 296 respondents report lower probability of living to age 75 compared to the probability of living to age 80 to

100. These cases were excluded from the analysis. Second, for respondents that report equal probabilities of living to age 75, as well as to age 80 to 100, a constant was added to the first probability and subtracted from the second probability. This constant was set equal to 5%. For instance, if a respondent reports a 50% probability of living to age 75 and to age 80 to 100, the first probability is adjusted upwards to 55%, whereas the second probability is adjusted downwards to 45%. This adjustment ensures that the individual's subjective survival curve is strictly decreasing. Third, a floor of 1% and a cap of 99% are incorporated to ensure that self-reported subjective survival probabilities do not take the values of 0% or 100%. It is worth mentioning that in the documentation of RAND HRS it is stated that if a response to the first survival question is zero, then the response to second question is set to zero as well. These probabilities are adjusted to ensure that the individual's subjective survival curve is strictly decreasing, as described above.

HRS data are collected every two years. In particular, HRS Wave 13 data were collected in 2016, and HRS Wave 14 data were collected two years later, in 2018. In other words, the in-sample mortality and self-reported change in health data are available every two years, but the time horizon of the subjective survival probabilities refers to the next 10 to 15 years. To overcome this mismatch between the time interval of HRS waves (about 2 years) and the time interval which corresponds to subjective survival probabilities (SSP) (on average 10 to 15 years), we fit a Weibull model, with parameters α and β (Qian 1995). The conditional probability $S_t(x)$ of an individual aged x to survive to age $x+t$ can then be expressed as

$$S_t(x) = \frac{e^{-(\frac{x+t}{\alpha})^\beta}}{e^{-(\frac{x}{\alpha})^\beta}}.$$

We denote by $SSP_{t_1,x}$ the observed probability of a respondent aged x who reports the chance of own survival for the next t_1 years up to age 75 specifically and by $SSP_{t_2,x}$ the observed probability of a respondent aged x who reports the chance of own survival for the next t_2 years up to age 80 to 100. For each respondent i ($i = 1, \dots, n$), the Weibull model parameters can be estimated by solving the following system of two equations:

$$SSP_{t_1,x,i} = \frac{e^{-(\frac{x+t_1}{\alpha_i})^{\beta_i}}}{e^{-(\frac{x}{\alpha_i})^{\beta_i}}}$$

and

$$SSP_{t_2,x,i} = \frac{e^{-(\frac{x+t_2}{\alpha_i})^{\beta_i}}}{e^{-(\frac{x}{\alpha_i})^{\beta_i}}}.$$

The above system of non-linear equations is solved using the Broyden and Newton methods (Bouaricha and Schnabel 1997) in R (Package 'nleqslv'). For each respondent i ($i = 1, \ldots, n$), aged x, the subjective survival probability for the next 2 years is calculated then as follows:

$$SSP_{2,x,i} = \frac{e^{-(\frac{x+2}{\widehat{\alpha_i}})^{\widehat{\beta_i}}}}{e^{-(\frac{x}{\widehat{\alpha_i}})^{\widehat{\beta_i}}}}.$$

Subsequently, the subjective mortality probability (SMP) for the next 2 years, for individual i, aged x, is calculated as follows:

$$SMP_{2,x,i} = 1 - SSP_{2,x,i}.$$

4. Statistical modelling

The impact of the explanatory variables on actual mortality and on self-reported change in health is examined using Bayesian logistic regression models. The estimation process of the coefficients of the logistic regression models is separated in two steps. In the first step, we fit a Bayesian linear regression model, using default priors and all available explanatory variables, on the logit transformation of the subjective mortality probabilities. We derive the posterior densities which are used, after applying – for sensitivity reasons – several Bayesian methods, for constructing informative prior densities for the next step. In the second step, we fit Bayesian logistic regression models, using all the available explanatory variables and the informative priors (from the previous step) on the actual mortality as well as on the changes in health. In other words, the subjective survival information is introduced in the first step of the analysis, whereas the actual mortality and change in health information is incorporated in the second step of the analysis. Posterior summaries are obtained using MCMC methods with RStan (Stan Development Team 2022) and WinBUGS (Lunn et al. 2000). The final posterior means incorporate the information included on subjective mortality probabilities and the available data.

4.1 Prior elicitation based on subjective mortality probabilities

Prior elicitation is based on a Bayesian regression model, under default priors, fitted on the logit transformation of the subjective mortality probability for the next two years, of a respondent aged x. The explanatory variables described in Section 2.1 are included in this regression model. The posterior density of the model coefficients incorporate the effect of the explanatory variables on the subjective mortality probabilities. Thus, if we denote by $\boldsymbol{p}^* = (p_1^*, \ldots, p_n^*)^T$ the vector of the observed subjective mortality probabilities and by X_{ij} the value of the explanatory variable j ($j = 0, \ldots, p$) for each respondent i ($i = 1, \ldots, n$), with $X_{i0} = 1$, we fit the following model, using default priors,

$$\begin{aligned} \log\left(\frac{SMP_{2,x,i}}{1 - SMP_{2,x,i}}\right) &= \sum_{j=0}^{p} b_j X_{ij} + \epsilon_i, \;\; \epsilon_i \overset{indep}{\sim} N(0, \sigma^2), \\ p(\mathbf{b}, \sigma^2) &\propto \frac{1}{\sigma^2}, \end{aligned}$$

where by $\mathbf{b} = (b_0, b_1, \ldots, b_p)$ we denote the model coefficients and by σ^2 the error variance. The posterior distribution of $\mathbf{b}$, conditional on σ^2, is a multivariate normal distribution

$$p(\mathbf{b}|\boldsymbol{p}^*, \sigma^2) = N_{p+1}(\mathbf{b}^*, \sigma^2(\mathbf{X^T X})^{-1}),$$

where $\mathbf{X}$ denotes the design matrix and

$$\mathbf{b}^* = (\mathbf{X^T X})^{-1}\mathbf{X^T} \log\left(\frac{\boldsymbol{p}^*}{1 - \boldsymbol{p}^*}\right).$$

In addition, the posterior distribution of σ^2 is an inverse-gamma distribution

$$p(\sigma^2|\boldsymbol{p}^*) = IG\left(\frac{n-p-1}{2}, \frac{(n-p-1)s^2}{2}\right),$$

with mean

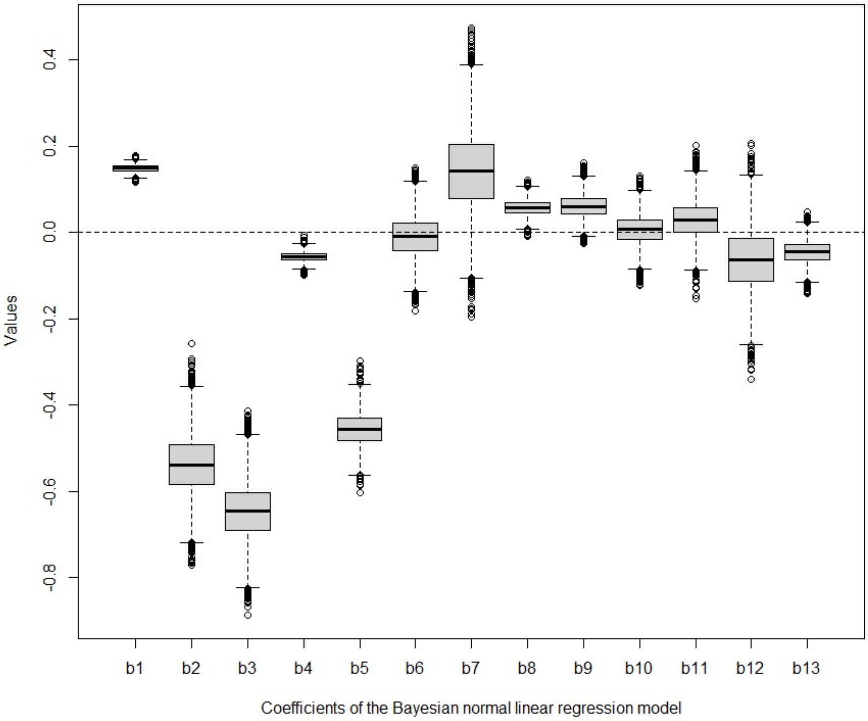
$$s^2 = \frac{1}{n-p-1}\left[\log\left(\frac{\boldsymbol{p}^*}{1 - \boldsymbol{p}^*}\right) - \mathbf{Xb}^*\right]^T \left[\log\left(\frac{\boldsymbol{p}^*}{1 - \boldsymbol{p}^*}\right) - \mathbf{Xb}^*\right].$$

The posterior means and standard deviations of the coefficients of the above model as well as the 95% credible intervals and the posterior distributions are presented in Table 2 and Figure 1, respectively. Higher subjective mortality probabilities are associated with older chronological age, lower educational attainment, worse self-rated health, and more chronic conditions. Black Americans (or other race) tend to report lower subjective mortality probabilities compared to White Americans, while more severe depression is associated with higher subjective mortality probabilities. On the other hand, females tend to report lower subjective mortality probabilities. The number of ADLs and IADLs, total household income, smoking status, marital status, and frequency of vigorous exercise are not significantly associated with the logit of subjective mortality probabilities. It is worth noticing that the posterior distributions of the coefficients of gender, race, number of IADLs, and marital status have the largest variability.

Table 2: Posterior means (sds) and 95% credible intervals for the coefficients of the Bayesian regression model on subjective mortality probabilities

Explanatory variable	Coefficient	95% Credible intervals
b_0 (Intercept)	–9.957 (0.522)	(–10.950, –8.921)
b_1 (Chronological age)	0.149 (0.008)	(0.132, 0.165)
b_2 (Female)	–0.538 (0.068)	(–0.670, –0.404)
b_3 (Black Americans or other)	–0.646 (0.067)	(–0.776, –0.512)
b_4 (Years of education)	–0.055 (0.011)	(–0.077, –0.033)
b_5 (Self-rated health)	–0.457 (0.039)	(–0.533, –0.380)
b_6 (Number of ADLs)	–0.009 (0.047)	(–0.102, 0.083)
b_7 (Number of IADLs)	0.142 (0.091)	(–0.038, 0.319)
b_8 (Depression)	0.057 (0.018)	(0.022, 0.093)
b_9 (Number of chronic conditions)	0.061 (0.027)	(0.009, 0.114)
b_{10} (Household income)	0.007 (0.034)	(–0.060, 0.073)
b_{11} (Smoking status)	0.029 (0.043)	(–0.056, 0.115)
b_{12} (Divorced, widowed, or never married)	–0.063 (0.072)	(–0.200, 0.079)
b_{13} (Vigorous exercise)	–0.045 (0.026)	(–0.096, 0.004)

Figure 1: Posterior distributions for the coefficients of the Bayesian regression model on subjective mortality probabilities



4.2 Subjective Bayesian logistic regression models on actual mortality

Several Bayesian approaches are incorporated in order to examine whether subjective mortality probabilities affect the predictive ability of logistic regressions on actual mortality. Let Y_i be a binary variable taking the value of 0 if the respondent i is alive and the value of 1 otherwise $(i = 1, \dots, n)$; see Section 2. $\mathbf{b}^*$ is the posterior mean of the coefficients from the Bayesian regression model on subjective mortality probabilities and s^2 is the posterior mean of σ^2 from the Bayesian regression model on subjective mortality probabilities; see Section 4.1.

The first model (M1) is built using the concept of g-priors (Zellner 1986). Liang et al. (2008) introduce the hyper-g-priors by placing a prior distribution to the parameter g of the g-prior. If we denote by $\mathbf{b} = (b_0, b_1, \dots, b_p)$ the coefficients, on all available explanatory variables (see Section 2.1), the proposed model has the form

$$\begin{aligned} Y_i|p_i &\overset{indep}{\sim} Bernouli(p_i), \\ \log\left(\frac{p_i}{1-p_i}\right) &= \sum_{j=0}^{p} b_j X_{ij}, \\ p(\mathbf{b}) &= N_{p+1}(\mathbf{b}^*, gs^2(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}), \\ g &\sim p(g) = \frac{a-2}{2}(1+g)^{-\frac{a}{2}},\ g > 0. \end{aligned}$$

According to the above model formulation, the prior used for the model coefficients is the posterior distribution from the Bayesian regression model on subjective mortality probabilities of Section 4.1, with its covariance matrix multiplied by g. Therefore the parameter g controls the influence of the prior in the final posterior; smaller values of g indicate a less informative prior. Under the prior used for g, the induced prior on the shrinkage factor $w = \frac{g}{g+1} \in (0, 1)$ is a $Beta(1, \frac{a}{2} - 1)$ distribution; values of w closer to one indicate a more informative prior. Regarding the selection of values for the hyperparameter a, we adopt the recommendation of Liang et al. (2008), who suggest a range of reasonable options, $2 < a \leqslant 4$. Therefore, we focus on two options, a = 3 and a = 4. The prior distribution of the shrinkage factor for a = 4 is uniform, whereas for a = 3 it places most of the probability mass near 1. In Sections 4.3 and 5 we use a = 3, while in Appendix A-6 we perform a sensitivity analysis using a = 4.

The second Bayesian approach is based on the idea of power priors (Ibrahim and Chen 2000). The elicitation of prior knowledge can be based on historical data; however, the uncertainty of this data is quantified by a random scalar a_0. In this study prior knowledge about own future survival is incorporated via the subjective mortality probabilities. The second model (M2), using the recommended setup by Ibrahim and Chen (2000) for a_0, has the following structure:

$$\begin{aligned} Y_i|p_i &\overset{indep}{\sim} Bernouli(p_i), \\ \log\left(\frac{p_i}{1-p_i}\right) &= \sum_{j=0}^{p} b_j X_{ij}, \\ p(\mathbf{b}) &= N_{p+1}(\mathbf{b}^*, a_0^{-1}s^2(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}), \\ p(a_0) &\sim Beta(3, 3). \end{aligned}$$

Thus, for once more, the prior used for the model coefficients is the posterior distribution from the Bayesian regression model on subjective mortality probabilities of Section 4.1, with a covariance matrix multiplied by a_0^{-1}. Therefore large values of a_0 indicate a more informative prior.

The third Bayesian approach is based again on the idea of the g-prior, with fixed g this time, where an overall or averaged information from subjective mortality probabilities is included in the model based on the methodology of Hanson, Branscum, and Johnson (2014). The third model (M3) has the following structure:

$$\begin{aligned} Y_i|p_i &\overset{indep}{\sim} Bernouli(p_i), \\ \log\left(\frac{p_i}{1-p_i}\right) &= \sum_{j=0}^{p} b_j X_{ij}, \\ p(\mathbf{b}) &= N_{p+1}(b_g \mathbf{e}_1, gn(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}), \end{aligned}$$

where the first element of the $(p+1)$-dimensional vector $\mathbf{e}_1$ is equal to one and all of its other elements are equal to zero, yielding to a non-zero prior mean of $\mathbf{b}$ for the intercept-only term. Following Hanson, Branscum, and Johnson (2014) we set $b_g = \delta(\alpha_\pi) - \delta(\beta_\pi)$ and $g = \frac{\delta'(\alpha_\pi) - \delta'(\beta_\pi)}{p+1}$, where δ is the digamma function and δ' is the trigamma function. The parameters α_π and β_π correspond to a beta distribution, fitted on the observed data of the subjective mortality probabilities. Using the available data, we estimate α_π = 0.3414 and β_π = 5.2584, yielding to b_g = –4.6138 and g = 0.6754.

The fourth Bayesian approach (M4) is based on fitting a logistic regression model and incorporating prior information on the coefficients via independent normal distributions:

$$\begin{aligned} Y_i|p_i &\overset{indep}{\sim} Bernouli(p_i), \\ \log\left(\frac{p_i}{1-p_i}\right) &= \sum_{j=0}^{p} b_j X_{ij}, \\ p(b_j) &= N\left(\tilde{b_j}, \tilde{\sigma}^2\right), \;\; j = 0, 1, \ldots, p. \end{aligned}$$

The mean and standard deviation of the independent prior normal distributions are estimated by WinBUGS (Lunn et al. 2000) after fitting the Bayesian regression model on subjective mortality probabilities of Section 4.1. No shrinkage parameter is used in this model. In addition to the above models, we have also fitted models M1 and M4 using non-informative priors in order to better understand the impact of prior information on the coefficients' posterior means. Under model M1, we have used a fixed value for g equal to

$4n$ (see Fouskakis, Ntzoufras, and Draper (2009)), while under model M4 we have used vague independent normal priors with mean equal to 0 and variance equal to 10,000.

Models M1–M3 under the informative priors, as well as model M1 under the non-informative prior, were fitted in RStan. Model M4, under the informative and non-informative prior setups, was fitted in WinBUGS. In the Appendix A-1, A-2, and A-3 we present results regarding the convergence of the MCMC samplers under models M1–M3 with the informative priors. No convergence issues were detected. Regarding model M4 with the informative prior and the two models with the non-informative priors, again no convergence issues were detected (results are omitted for brevity reasons). The leave-one-out information criterion is used for model selection (Vehtari, Gelman, and Gabry 2017).

4.3 Results

The posterior means, standard deviations, and 95% credible intervals for the coefficients of models M1–M4 (see Section 4.2) are presented in Table 3. Older chronological age, male gender, smoking status, more chronic conditions and deteriorating self-rated health are associated with increased risk of mortality. Moreover, individuals who are divorced or widowed or have never been married exhibit higher mortality risk than those who are married or partnered.

On the other hand, our results indicate that respondents who do vigorous exercise, those with higher total household income, and those with less severe depression are negatively associated with the risk of mortality. Fewer years of education and more difficulties with ADLs and IADLs are negatively associated with the risk of mortality in models M1–M3. Nevertheless, by observing the values of the 95% credible intervals all these associations are not significant. It is worth noting that lower educational attainment (Brown et al. 2012) and deteriorating functional health status (Scott et al. 1997) are significant predictors of mortality.

Our findings on the association of race with actual mortality illustrate the incorporation of subjective survival information in the model. More specifically, the results of models M1 and M3 indicate that White Americans tend to have lower mortality compared to Black Americans (or other race), whereas the results of models M2 and M4 indicate the opposite. By observing the 95% credible intervals, however, only the association under model M4 is significant. According to the results of the Bayesian regression model on subjective mortality probabilities, Black Americans tend to report lower subjective mortality probabilities compared to White Americans, but they experience higher mortality compared to White Americans of the same age, and this finding is in line with the literature (Mirowsky 1999; Firebaugh et al. 2014). Since the posterior mean value of the race coefficient under the non-informative g-prior setup (see Appendix A-4) indicates a stronger positive association of Black Americans (or other race) with mortality, we conclude that the incorporation of subjective survival information affects the value of the race coefficient materially.

Table 3: Posterior means (sds) [95% credible intervals] of the coefficients of the Bayesian logistic regression mortality models

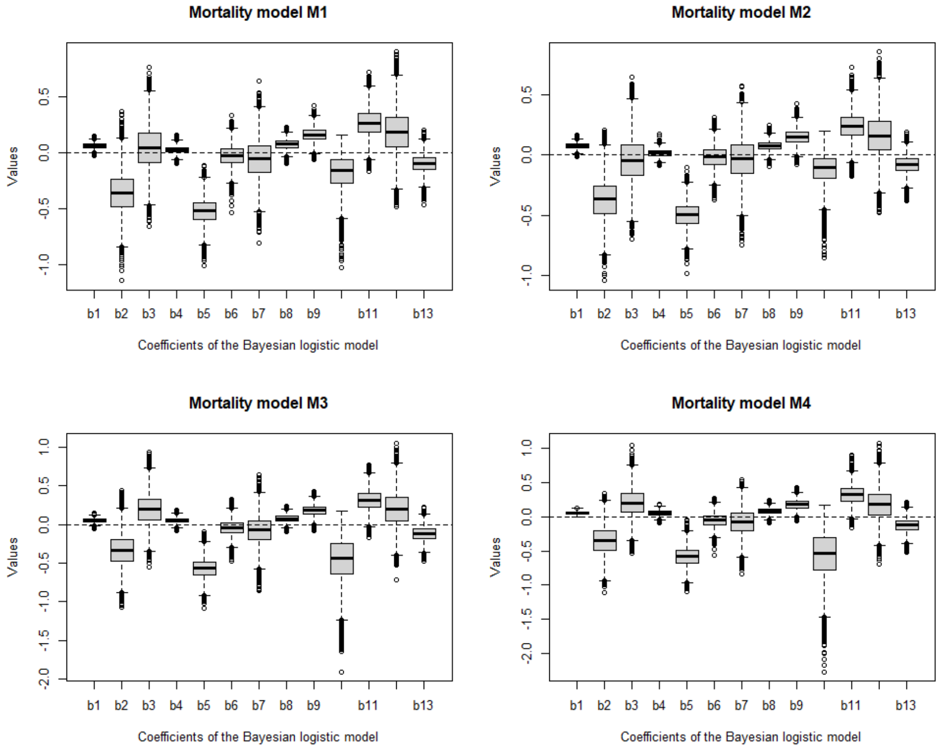
Coefficients	Logistic regression using hyper-g-prior(M1)	Logistic regression using power prior (M2)
b_0 (Intercept)	–7.405 (1.443) [–10.250, –4.568]	–7.695 (1.345) [–10.320, –4.993]
b_1 (Chronological age)	0.066 (0.023) [0.019, 0.111]	0.076 (0.022) [0.030, 0.118]
b_2 (Female)	–0.354 (0.183) [–0.714, 0.006]	–0.367 (0.183) [–0.694, –0.022]
b_3 (Black Americans or other)	0.050 (0.189) [–0.314, 0.429]	–0.042 (0.187) [–0.398, 0.327]
b_4 (Years of education)	0.029 (0.032) [–0.032, 0.095]	0.018 (0.031) [–0.038, 0.082]
b_5 (Self-rated health)	–0.520 (0.114) [–0.750, –0.306]	–0.500 (0.104) [–0.714, –0.301]
b_6 (Number of ADLs)	–0.026 (0.093) [–0.213, 0.150]	–0.018 (0.088) [–0.196, 0.150]
b_7 (Number of IADLs)	–0.054 (0.042) [–0.408, 0.278]	–0.035 (0.174) [–0.381, 0.295]
b_8 (Depression)	0.078 (0.042) [–0.005, 0.160]	0.075 (0.041) [–0.006, 0.155]
b_9 (Number of chronic conditions)	0.165 (0.064) [0.038, 0.290]	0.154 (0.062) [0.035, 0.277]
b_{10} (Household income)	–0.176 (0.160) [–0.545, 0.072]	–0.120 (0.129) [–0.412, 0.081]
b_{11} (Smoking status)	0.270 (0.119) [0.037, 0.504]	0.238 (0.113) [0.020, 0.461]
b_{12} (Divorced, widowed, or never married)	0.188 (0.192) [–0.185, 0.571]	0.162 (0.179) [–0.191, 0.521]
b_{13} (Vigorous exercise)	–0.092 (0.080) [–0.248, 0.065]	–0.082 (0.072) [–0.226, 0.059]
LOO information criterion	1029.4	1031.8

Coefficients	Logistic regression using g-prior (population average inference M3)	Logistic regression using independent normal priors (M4)
b_0 (Intercept)	–7.026 (1.637) [–10.225, –3.850]	–9.947 (0,543) [–10.950, –8.798]
b_1 (Chronological age)	0.051 (0.026) [0.001, 0.102]	0.139 (0.008) [0.123, 0.155]
b_2 (Female)	–0.337 (0.206) [–0.736, 0.069]	–0.506 (0.065) [–0.631, –0.378]
b_3 (Black Americans or other)	0.194 (0.198) [–0.189, 0.557]	–0.556 (0.065) [–0.682, –0.431]
b_4 (Years of education)	0.055 (0.037) [–0.014, 0.131]	–0.048 (0.010) [–0.069, –0.028]
b_5 (Self-rated health)	–0.568 (0.129) [–0.823, –0.318]	–0.492 (0.036) [–0.564, –0.421]
b_6 (Number of ADLs)	–0.044 (0.098) [–0.244, 0.143]	0.011 (0.042) [–0.072, 0.093]
b_7 (Number of IADLs)	–0.077 (0.189) [–0.465, 0.276]	0.130 (0.080) [–0.025, 0.286]
b_8 (Depression)	0.078 (0.046) [–0.013, 0.166]	0.072 (0.017) [0.041, 0.137]
b_9 (Number of chronic conditions)	0.189 (0.069) [0.047, 0.315]	0.089 (0.025) [0.041, 0.137]
b_{10} (Household income)	–0.458 (0.291) [–1.109, 0.015]	–0.009 (0.033) [–0,074, 0.057]
b_{11} (Smoking status)	0.317 (0.129) [0.069, 0.571]	0.072 (0.041) [–0.010, 0.151]
b_{12} (Divorced, widowed, or never married)	0.199 (0.218) [–0.226, 0.625]	0.011 (0.068) [–0.118, 0.146]
b_{13} (Vigorous exercise)	–0.118 (0.093) [–0.310, 0.060]	–0.060 (0.025) [–0.108, –0.011]
LOO information criterion	1031.5	1066.2

Notes: ADLs: Number of Activities of Daily Living for which respondents report difficulties. IADLs: Number of Instrumental Activities of Daily Living for which respondents report difficulties. LOO information criterion: Leave-One-Out information.

The posterior means are similar across models M1–M3, but they differ for model M4 due to the prior independence assumption and the absence of a shrinkage factor. The results indicate that prior elicitation was performed in a similar manner under models M1–M3. The estimated (posterior mean) hyper-parameters for model M1 are $w = 0.977$ and $g = 53.52$; the latter is equal to 0.15% of the value of $g = 4n$ under the non-informative g-prior setup, whereas for model M2 is $a_0 = 0.046$. In terms of the estimated (posterior mean) hyper-parameters, similar results, as the ones under model M1, are obtained using a different hyper-prior for the shrinkage factor (see Appendix A-6). Regarding the missing values for the explanatory variables sensitivity analysis based on the fully conditional specification imputation method is conducted, and the results are presented in Appendix A-7. No major changes in the beta coefficients are observed for the mortality model M1.

Figure 2: Posterior distributions of the coefficients of the Bayesian logistic regression models M1–M4 on actual mortality



The results in this section indicate that prior survival information affects the predictive ability of mortality models. For example, the posterior means of several coefficients, under the model with a non-informative g-prior (see Appendix A-4), are quite different compared to those under model M1. In addition, the posterior variability of the model coefficients varies across models M1–M4. For instance, the estimated posterior distributions based on model M4 have less variability, while the ones under models M1 and M2 have more. The boxplots of the posterior distributions of the Bayesian logistic regression coefficients for models M1–M4 are presented in Figure 2. Finally, the mortality model M1 exhibits the lowest leave-one-out information criterion and we conclude that this is the best performing model regarding its predictive ability. In addition, if we compare it with the one under a non-informative g-prior (see Appendix A-4), we observe that it has a better predictive performance.

5. Subjective Bayesian logistic regression models for self-reported change in health

For brevity reasons, we have fitted only model M1, using this time as a response the binary variable Y_i, taking the value of 0 if a respondent reported the same or somewhat better health status at Wave 14 and the value of 1 otherwise; see Section 2. As in Section 4.2, $\mathbf{b}^*$ is the posterior mean of the coefficients from the Bayesian regression model on subjective mortality probabilities, and s^2 is the posterior mean of σ^2 from the Bayesian regression model on subjective mortality probabilities; see Section 4.1. The posterior means, standard deviations, and 95% credible intervals of the coefficients of the Bayesian logistic regression model M1 on the self-reported change in health are presented in Table 4, and the associated posterior distributions of the coefficients are presented in Figure 3.

Black Americans report, on average, better change in health compared to White Americans, while individuals with poor self-rated health at Wave 13 tend to report worse change in health at Wave 14. In addition, people with more years of education, more severe depression, more chronic conditions, more ADLs, and higher total household income tend to report worse change in health. Finally, smoking status is associated with tendency to report worse change in health. The effects of the remaining variables are not significant. The estimated (posterior mean) hyper-parameters for the change in health model M1 are $w = 0.998$ and $g = 466.2$. It is worth noting that the estimated parameter g has a lower value (in particular 1.59%) compared to the value of g $(= 4n)$ under the non-informative prior setup.

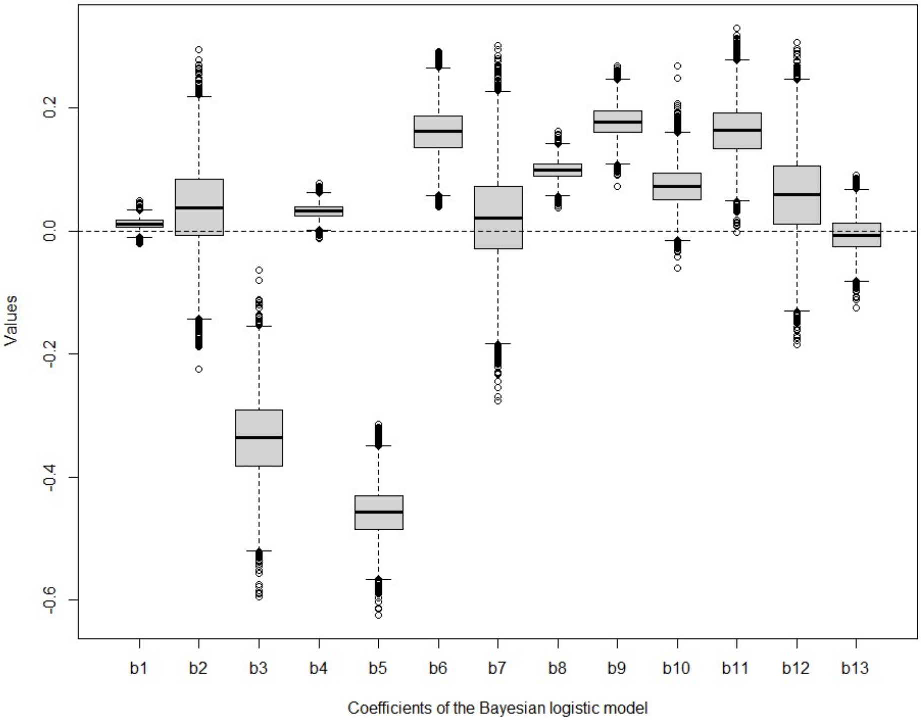
Table 4: Posterior means (sds) [95% credible intervals] of the coefficients of the Bayesian logistic regression model M1 on self-reported change in health

Coefficients	Logistic regression using hyper-g-prior (M1)
b_0 (Intercept)	–1.940 (0.523) [–2.937, –0.923]
b_1 (Chronological age)	0.012 (0.008) [–0.005, 0.028]
b_2 (Female)	0.038 (0.068) [–0.096, 0.171]
b_3 (Black Americans or other)	–0.337 (0.067) [–0.472, –0.206]
b_4 (Years of education)	0.032 (0.011) [0.009, 0.055]
b_5 (Self-rated health)	–0.458 (0.041) [–0.538, –0.377]
b_6 (Number of ADLs)	0.162 (0.039) [0.086, 0.238]
b_7 (Number of IADLs)	0.021 (0.076) [–0.128, 0.169]
b_8 (Depression)	0.099 (0.016) [0.067, 0.130]
b_9 (Number of chronic conditions)	0.177 (0.025) [0.128, 0.227]
b_{10} (Household income)	0.073 (0.033) [0.011, 0.141]
b_{11} (Smoking status)	0.163 (0.043) [0.080, 0.248]
b_{12} (Divorced, widowed, or never married)	0.059 (0.071) [–0.081, 0.198]
b_{13} (Vigorous exercise)	–0.007 (0.001) [–0.060, 0.047]
LOO information criterion	6193.7

Notes: ADLs: Number of Activities of Daily Living for which respondents report difficulties. IADLs: Number of Instrumental Activities of Daily Living for which respondents report difficulties. LOO information criterion: Leave-One-Out information.

For sensitivity analysis reasons we have also fitted the same model with a different prior on the shrinkage factor; results are almost identical (see Appendix A-6) and the two models have almost the same predictive ability. Finally, the results Table 4 are this time quite similar with those under the non-informative prior (see Appendix A-5), and the model with the non-informative prior has a slightly better predictive performance. Regarding the missing values for the dependent variable ‘self-reported change in health’ (see Section 2), as a sensitivity analysis, and by assuming that the missing mechanism is ignorable, we have treated these missing values as unknown parameters and simultaneously estimate them (together with the rest of the model parameters), using the posterior predictive distribution under model M1 on self-reported change in health. The results are presented in Appendix A-7; it is worth noting that the impact of household income on self-reported change in health is not now significant, but apart from this, no other material changes in the beta coefficients between the two M1 health models are observed. Therefore we conclude that our results provide less evidence that subjective mortality probabilities incorporate supplementary information which is useful for the prediction of self-reported change in health.

Figure 3: Posterior distributions of the coefficients of the Bayesian logistic regression model M1 on self-reported change in health



6. Discussion

Subjective survival probabilities are quantities that incorporate individuals' view on likely future survival, and they are linked to actual mortality patterns (Papachristos et al. 2020). The objective of this study is twofold: first, to develop a Bayesian process to incorporate the respondents' views about future survival on in-sample actual mortality models; and second, to investigate whether the respondents' views about future survival incorporate useful information for predicting self-reported change in health. To achieve this aim, we adopt a two-step process. In the first step, the prior means and the covariance matrix for the logistic regression coefficients are estimated by fitting a Bayesian linear regression model on the logit transformation of subjective mortality probabilities. It is worth noticing that the subjective mortality probability for each respondent is adjusted using a Weibull survival model to reflect the next two-year period. In the second step, we fit Bayesian logistic regression models on actual mortality using the prior distributions from the first step.

Our results show that male gender, older individuals, smoking status, more chronic conditions, and individuals with poor self-rated health are associated with higher risk of in-sample mortality. Black Americans (or other race) exhibit a positive association with mortality; however, the incorporation of subjective survival information affects materially the magnitude and direction of the association. In addition, White Americans, individuals with poor self-rated health, more years of education, severe depression, more chronic conditions, more ADLs, and higher total household income tend to report worse change in health compared to the previous wave. Finally, smoking status is also associated with tendency to report worse change in health.

These results are broadly in line with the literature. Self-rated health is a strong predictor of mortality, even after including other covariates known to predict mortality (Idler and Benyamini 1997); this also holds for the number of chronic diseases (Verropoulou 2014). Depression and smoking are associated with higher mortality (Wulsin, Vaillant, and Wells 1999; Ezzati and Lopez 2003); this comes to an agreement with our results, although the effect of depression is marginally non-significant. Race differentiates actual mortality patterns, as Black Americans experience lower life expectancy than White Americans of the same age (Firebaugh et al. 2014). On the other hand, married individuals exhibit lower mortality risk than those who are not married (Johnson et al. 2000). In this study the estimated posterior means of the models using informative priors (see Table 3) and the posterior means of M1 using a non-informative g-prior indicate a weak association with the risk of in-sample mortality (see Appendix A-4).

Several researchers noted the so-called gender paradox: older women report poorer health than older men, but women live longer (Arber and Cooper 1999). This comes to an agreement with our results, although the effect on self-reported change in health is not significant. Better educational attainment is associated with better self-rated health for men and women aged 25 years or older (Subramanian, Huijts, and Avendano 2010), but the exact mechanism by which education affects current health status is less clear (Cutler and Lleras-Muney 2006). In contrast, our results indicate that respondents with higher educational attainment tend to report worse change in health compared to that reported in the previous wave. A possible explanation of this contradiction is that our study targets respondents whose chronological age ranges from 50 to 65 years old and our results are relevant to only this age group. In addition, the response variable in our models is the self-reported change in health and not the current self-rated health status, as in other studies. Further investigation based on a broader age group may be useful for understanding the impact of educational attainment on future self-reported change in health.

Regarding the additional information incorporated via subjective mortality probabilities, our findings suggest that they are useful for predicting mortality but less useful for predicting self-reported change in health. This is expected as subjective mortality probabilities contain information particularly relevant for own future survival but not necessarily relevant for future change in health status.

A few limitations of this study as well as future research ideas should be mentioned. First, this study uses data from two consecutive HRS Waves (W13 and W14). However, additional waves covering a lengthier time period can be incorporated as part of future research. Second, alternative choices for informative priors as well as additional techniques for calculating the adjusted two-year subjective survival probability per respondent can be investigated. Possible future research steps would also include the investigation of innovative ways to incorporate subjective mortality probabilities in Bayesian models as well as the expansion of model universe, for instance by adding random effects to better handle longitudinal datasets.

References

Arber, S. and Cooper, H. (1999). Gender differences in health in later life: The new paradox? *Social Science & Medicine* 48(1): 61–76. doi:10.1016/S0277-9536(98)00289-5.

Arpino, B., Bordone, V., and Scherbov, S. (2018). Smoking, education and the ability to predict own survival probabilities. *Advances in Life Course Research* 37: 23–30. doi:10.1016/j.alcr.2018.06.001.

Bago d'Uva, T., Erdogan-Ciftci, E., O'Donnell, O., and Van Doorslaer, E. (2020). Who can predict their own demise? Heterogeneity in the accuracy and value of longevity expectations. *The Journal of the Economics of Ageing* 17: 100135. doi:10.1016/j.jeoa.2017.10.003.

Balia, S. (2014). Survival expectations, subjective health and smoking: Evidence from SHARE. *Empirical Economics* 47: 753–780. doi:10.1007/s00181-013-0750-1.

Bouaricha, A. and Schnabel, R.B. (1997). Algorithm 768: TENSOLVE: A software package for solving systems of nonlinear equations and nonlinear least-squares problems using tensor methods. *ACM Transactions on Mathematical Software (TOMS)* 23(2): 174–195. doi:10.1145/264029.264032.

Brown, D.C., Hayward, M.D., Montez, J.K., Hummer, R.A., Chiu, C.T., and Hidajat, M.M. (2012). The significance of education for mortality compression in the United States. *Demography* 49(3): 819–840. doi:10.1007/s13524-012-0104-1.

Cutler, D. and Lleras-Muney, A. (2006). Education and health: Evaluating theories and evidence. National Bureau of Economic Research. Working paper w12352 .

Doll, R., Peto, R., Wheatley, K., Gray, R., and Sutherland, I. (1994). Mortality in relation to smoking: 40 years' observations on male British doctors. *British Medical Journal* 309: 901–911. doi:10.1136/bmj.309.6959.901.

Elder, T.E. (2013). The predictive validity of subjective mortality expectations: Evidence from the health and retirement study. *Demography* 50(2): 569–589. doi:10.1007/s13524-012-0164-2.

Ezzati, M. and Lopez, A.D. (2003). Estimates of global mortality attributable to smoking in 2000. *The Lancet* 362(9387): 847–852. doi:10.1016/S0140-6736(03)14338-3.

Firebaugh, G., Acciai, F., Noah, A.J., Prather, C., and Nau, C. (2014). Why the racial gap in life expectancy is declining in the United States. *Demographic Research* 31(32): 975–1006. doi:10.4054/DemRes.2014.31.32.

Fouskakis, D., Ntzoufras, I., and Draper, D. (2009). Bayesian variable selection using cost-adjusted BIC, with application to cost-effective measurement of quality of health care. *The Annals of Applied Statistics* 3(2): 663 – 690. doi:10.1214/08-AOAS207.

Griffin, B., Loh, V., and Hesketh, B. (2013). A mental model of factors associated with subjective life expectancy. *Social Science & Medicine* 82: 79–86. doi:10.1016/j.socscimed.2013.01.026.

Hamermesh, D.S. (1985). Expectations, life expectancy, and economic behavior. *The Quarterly Journal of Economics* 100(2): 389–408. doi:10.2307/1885388.

Hanson, T.E., Branscum, A.J., and Johnson, W.O. (2014). Informative *g*-priors for logistic regression. *Bayesian Analysis* 9(3): 597–612. doi:10.1214/14-BA868.

Heckhausen, J. and Krueger, J. (1993). Developmental expectations for the self and most other people: Age grading in three functions of social comparison. *Developmental Psychology* 29(3): 539–548. doi:10.1037//0012-1649.29.3.539.

HRS (2022). Health and Retirement Study. RAND HRS Longitudinal File 2018 public use dataset. Produced and distributed by the University of Michigan with funding from the National Institute on Aging (grant number NIA U01AG009740). Ann Arbor, MI.

Hurd, M.D. (2009). Subjective probabilities in household surveys. *Annual Review of Economics* 1: 543–562. doi:10.1146/annurev.economics.050708.142955.

Hurd, M.D. and McGarry, K. (1995). Evaluation of the subjective probabilities of survival in the health and retirement study. *Journal of Human Resources* 30: S268–S292. doi:10.2307/146285.

Hurd, M.D. and McGarry, K. (2002). The predictive validity of subjective probabilities of survival. *The Economic Journal* 112(482): 966–985. doi:10.1111/1468-0297.00065.

Ibrahim, J.G. and Chen, M.H. (2000). Power prior distributions for regression models. *Statistical Science* 15(1): 46–60. doi:10.1214/ss/1009212673.

Idler, E.L. and Benyamini, Y. (1997). Self-rated health and mortality: A review of twenty-seven community studies. *Journal of Health and Social Behavior* 38(1): 21–37. doi:10.2307/2955359.

Johnson, N.J., Backlund, E., Sorlie, P.D., and Loveless, C.A. (2000). Marital status and mortality: The national longitudinal mortality study. *Annals of Epidemiology* 10(4): 224–238. doi:10.1016/S1047-2797(99)00052-6.

Khwaja, A., Sloan, F., and Chung, S. (2007). The relationship between individual expectations and behaviors: Mortality expectations and smoking decisions. *Journal of Risk and Uncertainty* 35: 179–201. doi:10.1007/s11166-007-9019-4.

Lewinsohn, P.M., Seeley, J.R., Roberts, R.E., and Allen, N.B. (1997). Center for epidemiologic studies depression scale (CES-D) as a screening instrument for depression among community-residing older adults. *Psychology and Aging* 12(2): 277. doi:10.1037//0882-7974.12.2.277.

Liang, F., Paulo, R., Molina, G., Clyde, M.A., and Berger, J.O. (2008). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association* 103(481): 410–423. doi:10.1198/016214507000001337.

Little, R.J. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American statistical Association* 83(404): 1198–1202. doi:10.1080/01621459.1988.10478722.

Little, R.J. and Rubin, D.B. (2019). *Statistical analysis with missing data*. Hoboken: John Wiley & Sons. doi:10.1002/9781119482260.

Liu, J.T., Tsou, M.W., and Hammitt, J.K. (2007). Health information and subjective survival probability: Evidence from Taiwan. *Journal of Risk Research* 10(2): 149–175. doi:10.1080/13669870701191802.

Lunn, D., Thomas, A., Best, N., and Spiegelhalter, D. (2000). WinBUGS – A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing* 10: 325–337. doi:10.1023/A:1008929526011.

Mirowsky, J. (1999). Subjective life expectancy in the US: Correspondence to actuarial estimates by age, sex and race. *Social Science & Medicine* 49(7): 967–979. doi:10.1016/S0277-9536(99)00193-8.

Mossey, J.M. and Shapiro, E. (1982). Self-rated health: A predictor of mortality among the elderly. *American Journal of Public Health* 72(8): 800–808. doi:10.2105/AJPH.72.8.800.

Papachristos, A. and Verropoulou, G. (2022). Gender, health and socio-demographic influences on updating subjective survival probabilities. In: Skiadas, C.H. and Skiads, C. (eds.). *Quantitative methods in demography: Methods and related applications in the Covid-19 era*. Cham: Springer: 245–259. doi:10.1007/978-3-030-93005-916.

Papachristos, A., Verropoulou, G., Ploubidis, G.B., and Tsimbos, C. (2020). Factors incorporated into future survival estimation among Europeans. *Demographic Research* 42(2): 15–56. doi:10.4054/DemRes.2020.42.2.

Perozek, M. (2008). Using subjective expectations to forecast longevity: Do survey respondents know something we don't know? *Demography* 45(1): 95–113. doi:10.1353/dem.2008.0010.

Pinquart, M. (2001). Correlates of subjective health in older adults: A meta-analysis. *Psychology and Aging* 16(3): 414–426. doi:10.1037//0882-7974.16.3.414.

Qian, J. (1995). A Bayesian weibull survival model [PhD Thesis]. Durham: Duke University.

Radloff, L.S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement* 1(3): 385–401. doi:10.1177/014662167700100306.

Rappange, D.R., Brouwer, W.B., and van Exel, J. (2016). Rational expectations? An explorative study of subjective survival probabilities and lifestyle across Europe. *Health Expectations* 19(1): 121–137. doi:10.1111/hex.12335.

Ross, C.E. and Mirowsky, J. (2002). Family relationships, social support and subjective life expectancy. *Journal of Health and Social Behavior* 43(4): 469–489. doi:10.2307/3090238.

Rubin, D.B. (1976). Inference and missing data. *Biometrika* 63(3): 581–592. doi:10.1093/biomet/63.3.581.

Scott, W.K., Macera, C.A., Cornman, C.B., and Sharpe, P.A. (1997). Functional health status as a predictor of mortality in men and women over 65. *Journal of Clinical Epidemiology* 50(3): 291–296. doi:10.1016/S0895-4356(96)00365-4.

Siegel, M., Bradley, E.H., and Kasl, S.V. (2003). Self-rated life expectancy as a predictor of mortality: Evidence from the HRS and AHEAD surveys. *Gerontology* 49(4): 265–271. doi:10.1159/000070409.

Smith, A.M., Shelley, J.M., and Dennerstein, L. (1994). Self-rated health: Biological continuum or social discontinuity? *Social Science & Medicine* 39(1): 77–83. doi:10.1016/0277-9536(94)90167-8.

Smith, V., Taylor Jr, D., and Sloan, F. (2022). Longevity expectations and death: Can people predict their own demise? In: *The Economics of Environmental Risk*. Northampton: Edward Elgar Publishing: 146–154.

Stan Development Team (2022). RStan: The R interface to Stan. R package version 2.21.5. https://mc-stan.org.

Subramanian, S.V., Huijts, T., and Avendano, M. (2010). Self-reported health assessments in the 2002 world health survey: How do they correlate with education? *Bulletin of the World Health Organization* 88(2): 131–138. doi:10.2471/BLT.09.067058.

Van Buuren, S., Brand, J.P., Groothuis-Oudshoorn, C.G., and Rubin, D.B. (2006). Fully conditional specification in multivariate imputation. *Journal of statistical computation and simulation* 76(12): 1049–1064. doi:10.1080/10629360600810434.

Van Doorn, C. and Kasl, S.V. (1998). Can parental longevity and self-rated life expectancy predict mortality among older persons? Results from an Australian cohort. *The Journals of Gerontology: Series B: Psychological Sciences and Social Sciences* 53B(1): S28–S34. doi:10.1093/geronb/53B.1.S28.

Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing* 27: 1413–1432. doi:10.1007/s11222-016-9696-4.

Verropoulou, G. (2014). Specific versus general self-reported health indicators predicting mortality among older adults in Europe: Disparities by gender employing SHARE longitudinal data. *International Journal of Public Health* 59: 665–678. doi:10.1007/s00038-014-0563-9.

Wulsin, L.R., Vaillant, G.E., and Wells, V.E. (1999). A systematic review of the mortality of depression. *Psychosomatic Medicine* 61(1): 6–17. doi:10.1097/00006842-199901000-00003.

Wurm, S., Tomasik, M.J., and Tesch-Römer, C. (2008). Serious health events and their impact on changes in subjective health and life satisfaction: The role of age and a positive view on ageing. *European Journal of Ageing* 5: 117–127. doi:10.1007/s10433-008-0077-5.

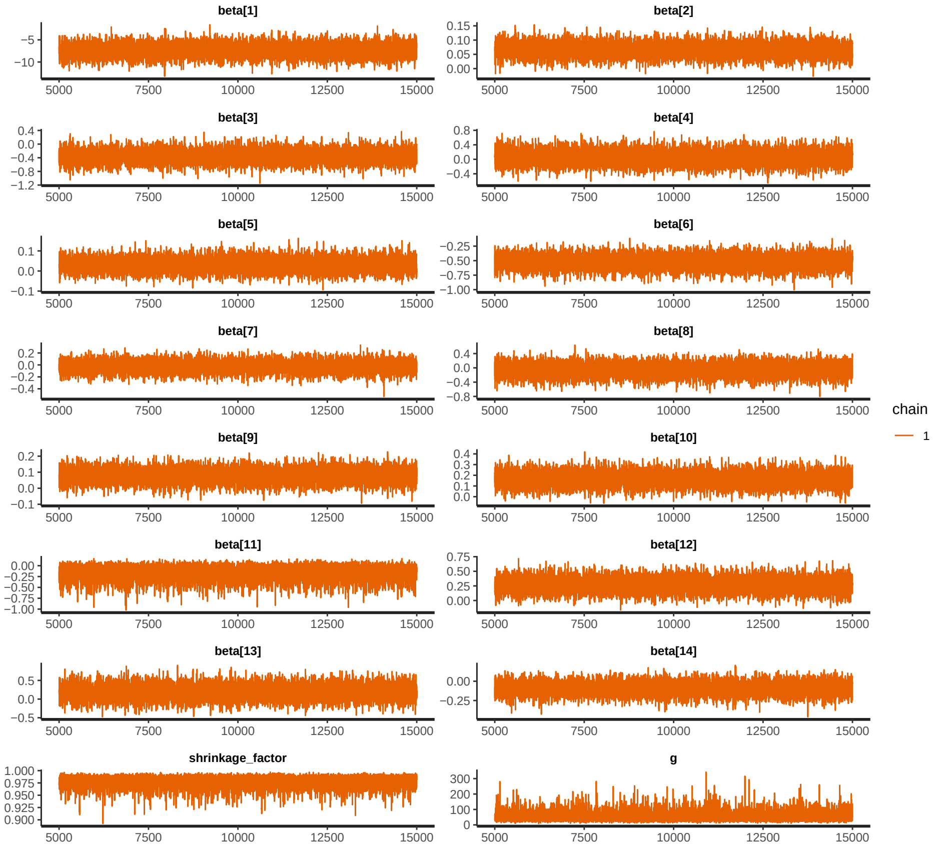
Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis using g-prior distributions. In: Goel, P. and Zellner, A. (eds.). *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*. Amsterdam: North-Holland: 233–243.

Appendices

A-1. Mortality model M1 – Logistic regression using hyper g-prior

The simulations illustrated in Figure A-1.1 suggest that MCMC algorithms for the mortality model M1 have converged.

Figure A-1.1: MCMC trace plots



The autocorellation structure of M1 logistic regression coefficients and their posterior distributions are presented in Figures A-1.2 and A-1.3, respectively. It is worth noting that the distributions of parameters g and w have fat tails and that the coefficients have positive autocorellation.

Figure A-1.2: Autocorrelation plots

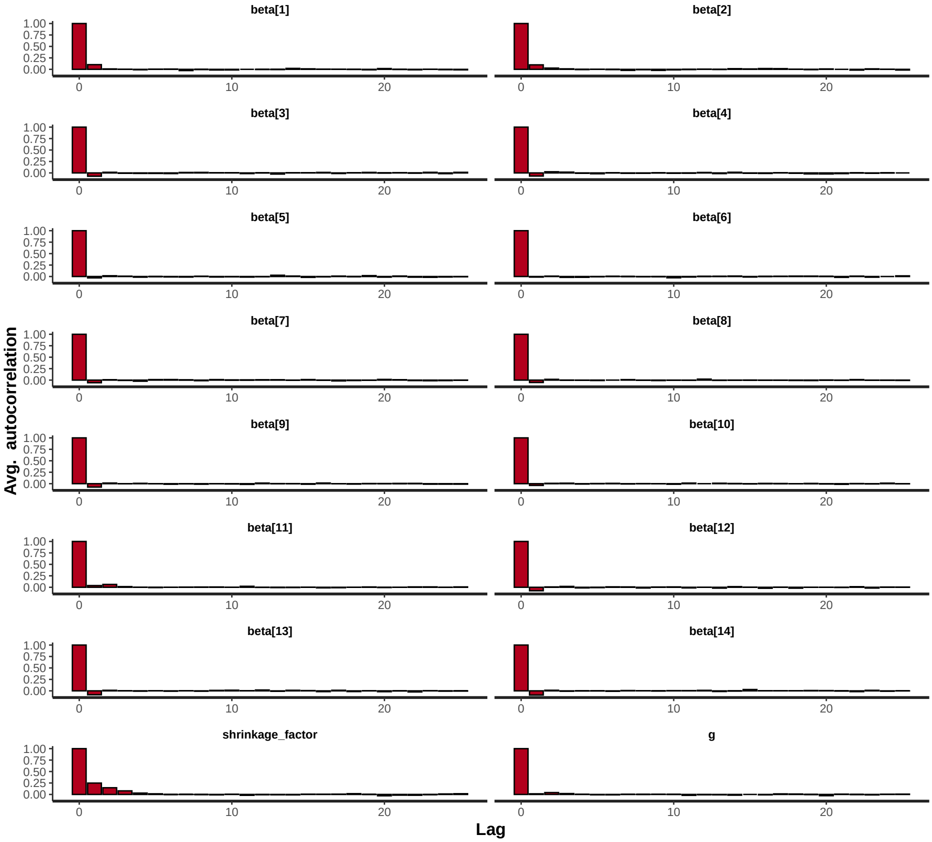
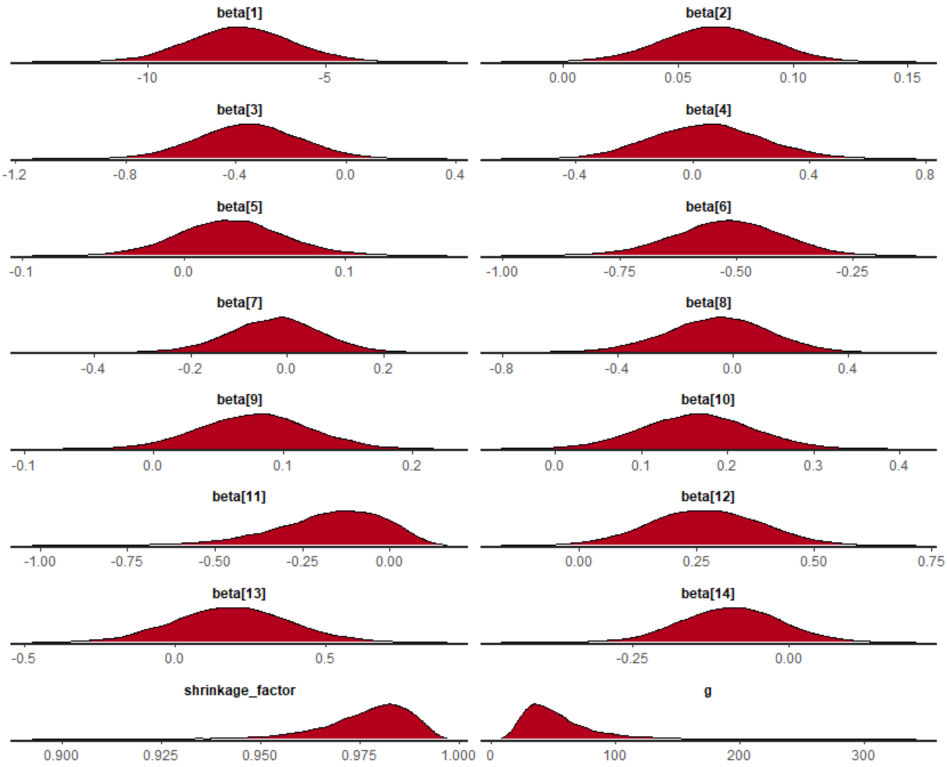


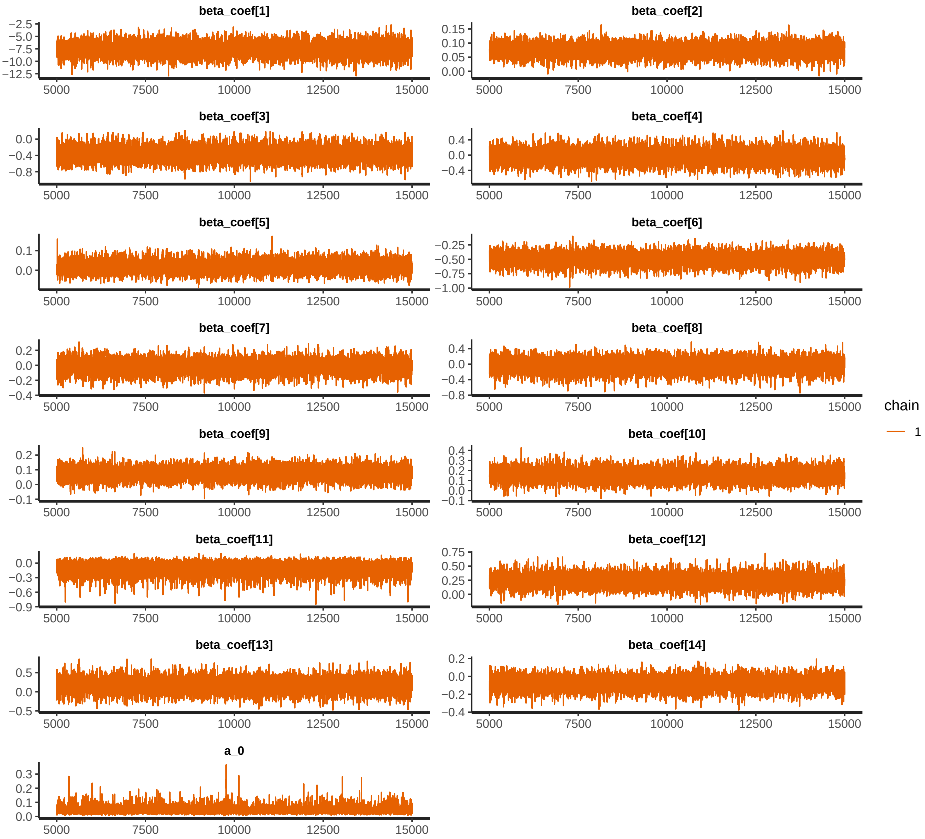
Figure A-1.3: Posterior distributions of the coefficients



A-2. Mortality model M2 – Logistic regression using power prior

The simulations illustrated in Figure A-2.1 suggest that MCMC algorithms for the mortality model M2 have converged.

Figure A-2.1: MCMC trace plots



The autocorellation structure of M2 logistic regression coefficients and their posterior distributions are presented in Figures A-2.2 and A-2.3, respectively. It is worth noting that the distribution of parameter a_0 has fat tails and that the coefficients have no autocorellation.

Figure A-2.2: Autocorrelation plots

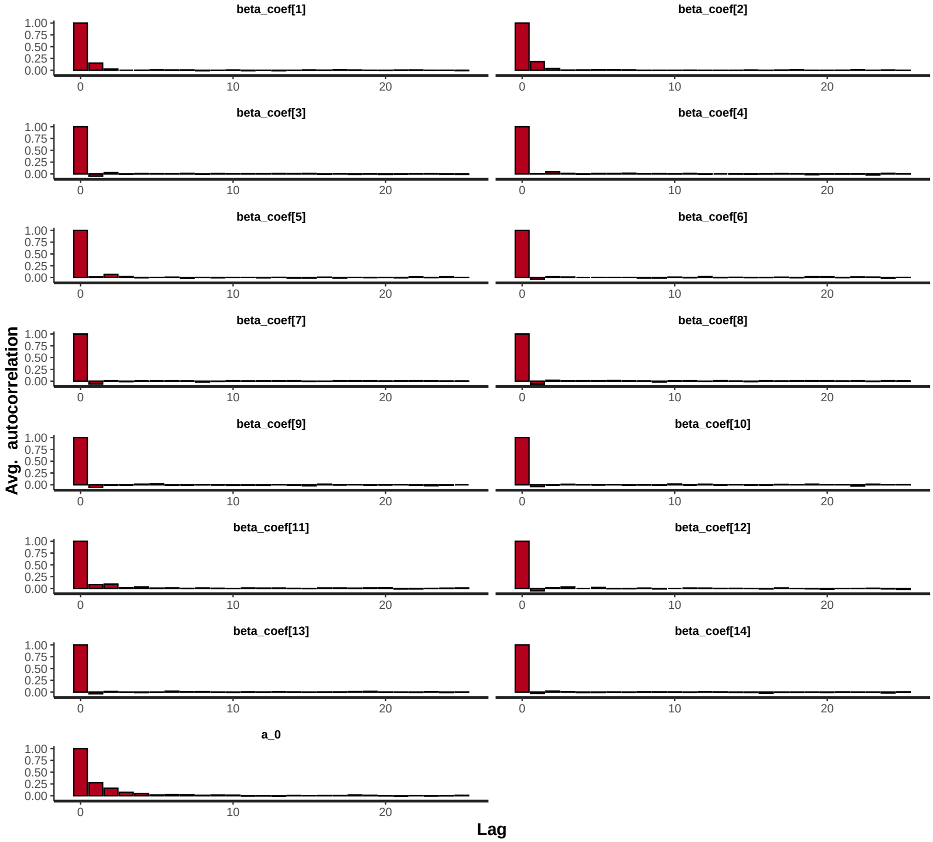
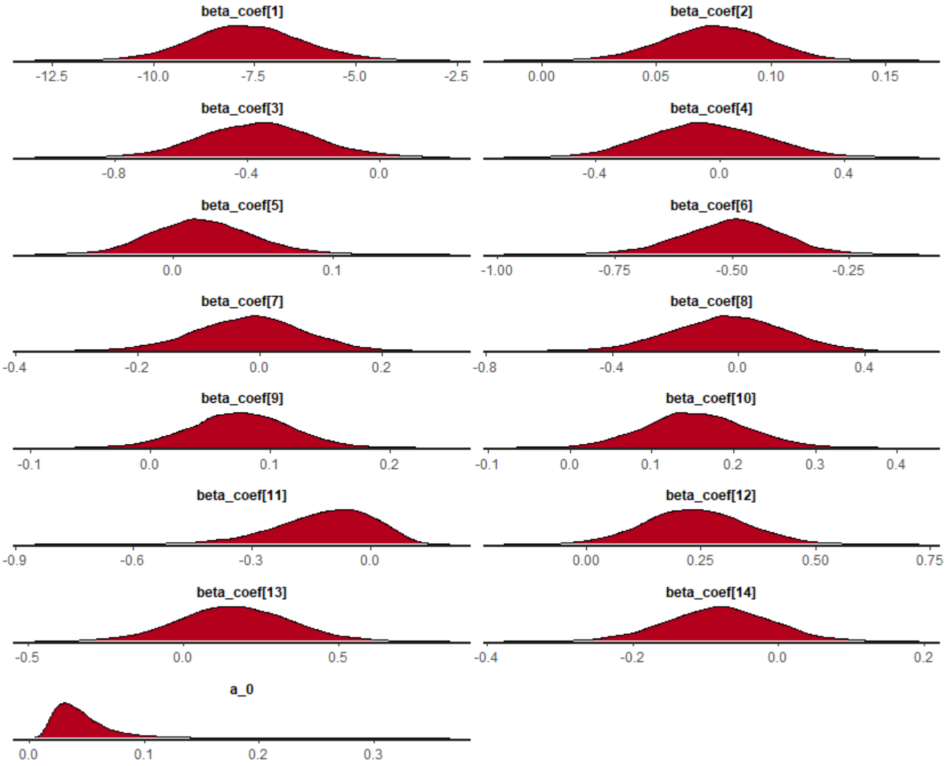


Figure A-2.3: Posterior distributions of the coefficients



A-3. Mortality model M3 – Logistic regression using g-prior (population average inference)

The simulations illustrated in Figure A-3.1 suggest that MCMC algorithms for the mortality model M3 have converged. The autocorellation structure of M3 logistic regression coefficients is presented in Figure A-3.2. It is worth noting that the coefficients have no autocorellation.

Figure A-3.1: MCMC trace plots

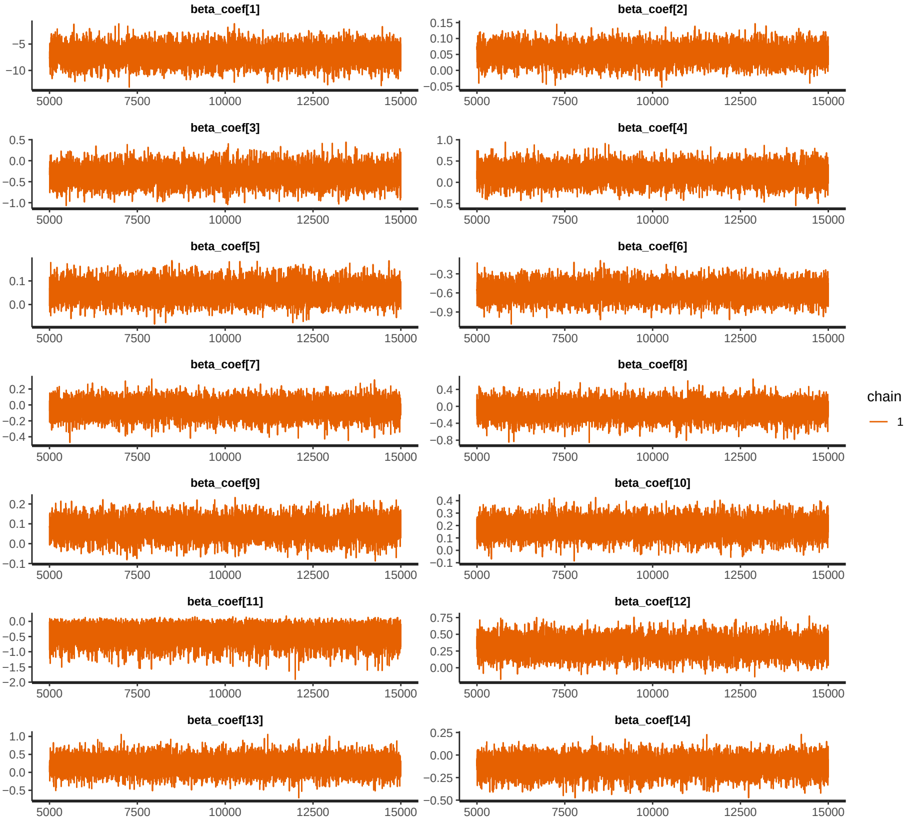
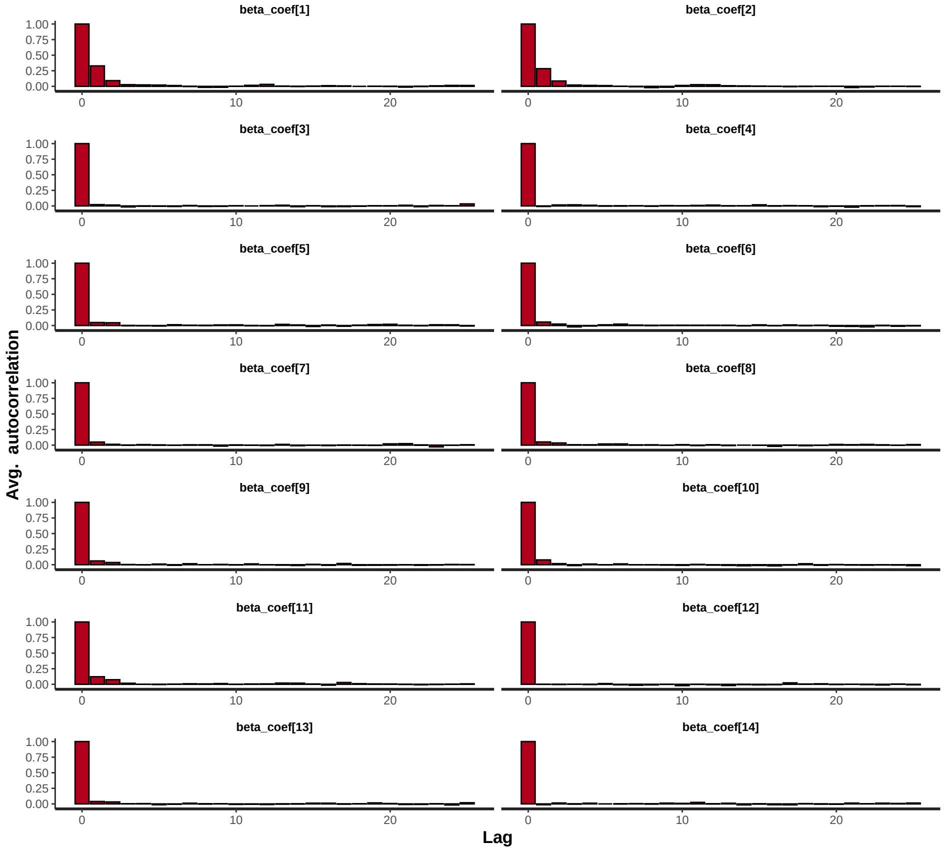


Figure A-3.2: Autocorrelation plots



A-4. Logistic regression mortality models using non-informative prior

The posterior means, standard deviations, and 95% credible intervals from the logistic regression coefficients using a non-informative g-prior ($g = 4n = 35,852$) for the mortality model M1 are presented in Table A-4.1. We observe differences between the posterior mean values of the coefficients with the corresponding ones under the informative g-prior (see Table 3) as well as a worse fit.

Table A-4.1: Posterior means (sds) [95% credible intervals] of the coefficients of the Bayesian logistic regression mortality model M1 using a non-informative g-prior

Coefficients	Logistic regression using non-informative g-prior (M1)
b_0 (Intercept)	–7.113 (1.686) [–10.453, –3.828]
b_1 (Chronological age)	0.052 (0.026) [0.001, 0.104]
b_2 (Female)	–0.340 (0.210) [–0.752, 0.068]
b_3 (Black Americans or other)	0.198 (0.206) [–0.202, 0.606]
b_4 (Years of education)	0.057 (0.038) [–0.015, 0.135]
b_5 (Self-rated health)	–0.583 (0.132) [–0.844, –0.327]
b_6 (Number of ADLs)	–0.048 (0.098) [–0.241, 0.1140]
b_7 (Number of IADLs)	–0.086 (0.194) [–0.485, 0.274]
b_8 (Depression)	0.078 (0.045) [–0.009, 0.167]
b_9 (Number of chronic conditions)	0.184 (0.069) [0.047, 0.322]
b_{10} (Household income)	–0.552 (0.336) [–1.298, –0.001]
b_{11} (Smoking status)	0.322 (0.133) [0.061, 0.584]
b_{12} (Divorced, widowed, or never married)	0.189 (0.219) [–0.234, 0.609]
b_{13} (Vigorous exercise)	–0.121 (0.097) [–0.317, 0.062]
LOO information criterion	1033.2

Notes: ADLs: Number of Activities of Daily Living for which respondents report difficulties. IADLs: Number of Instrumental Activities of Daily Living for which respondents report difficulties. LOO information criterion: Leave-One-Out information.

A-5. Logistic regression self-reported change in health model using non-informative prior

The posterior means, standard deviations, and 95% credible intervals from the logistic regression coefficients using a non-informative g-prior ($g = 4n = 29,340$) for model M1 on the self-reported change in health are presented in Table A-5.1. We observe similar results as those on Table 3, under the informative g-prior, and a slightly better fit.

Table A-5.1: Posterior means (sds) [95% credible intervals] of the coefficients of the Bayesian logistic regression self-reported change in health model M1, using a non-informative g-prior

Coefficients	Logistic regression using non-informative g-prior (M1)
b_0 (Intercept)	–1.907 (0.525) [–2.998, –0.879]
b_1 (Chronological age)	0.011 (0.008) [–0.005, 0.027]
b_2 (Female)	0.039 (0.067) [–0.096, 0.169]
b_3 (Black Americans or other)	–0.336 (0.068) [–0.465, –0.204]
b_4 (Years of education)	0.031 (0.011) [0.009, 0.053]
b_5 (Self-rated health)	–0.456 (0.041) [–0.537, –0.376]
b_6 (Number of ADLs)	0.162 (0.039) [0.086, –0.239]
b_7 (Number of IADLs)	0.021 (0.076) [–0.128, 0.168]
b_8 (Depression)	0.099 (0.016) [0.067, 0.130]
b_9 (Number of chronic conditions)	0.178 (0.025) [0.127, 0.229]
b_{10} (Household income)	0.074 (0.034) [0.010, 0.142]
b_{11} (Smoking status)	0.162 (0.042) [0.078, 0.244]
b_{12} (Divorced, widowed, or never married)	0.061 (0.070) [–0.077, 0.203]
b_{13} (Vigorous exercise)	–0.007 (0.027) [–0.059, 0.046]
LOO information criterion	6193.2

Notes: ADLs: Number of Activities of Daily Living for which respondents report difficulties. IADLs: Number of Instrumental Activities of Daily Living for which respondents report difficulties. LOO information criterion: Leave-One-Out information.

A-6. Sensitivity analysis – Logistic regression mortality and self-reported change in health models for a = 4

As part of the sensitivity analysis for model M1, the coefficients for the mortality and self-reported change in health models are calculated using different prior for g under the hyper g-prior setup. In particular, we use a = 4, which means that the shrinkage factor $w = \frac{g}{g+1} \sim Beta(1, 1)$. The results for mortality and self-reported change in health models are presented in Table A-6.1. The estimated (posterior mean) hyper-parameters for the mortality model are $w = 0.973$ and $g = 46.71$; the latter is equal to 0.13% of the value of $g = 4n$ under the non-informative g-prior setup, whereas for the change in health model the estimated (posterior mean) hyper-parameters are $w = 0.998$ and $g = 512.4$. These values indicate that prior survival information affects the inferential procedure of the mortality model only. It is worth noticing that the posterior mean values of the coefficients under model M1 on actual mortality, as well as on self-reported change in health, are similar to those presented in Table 3 and Table 4, respectively, for a = 3. Regarding the predictive ability of the models, mortality model M1 with $a = 4$ has a slightly better fit, while model M1 on self-reported change in health with $a = 4$ a sligthly worse fit.

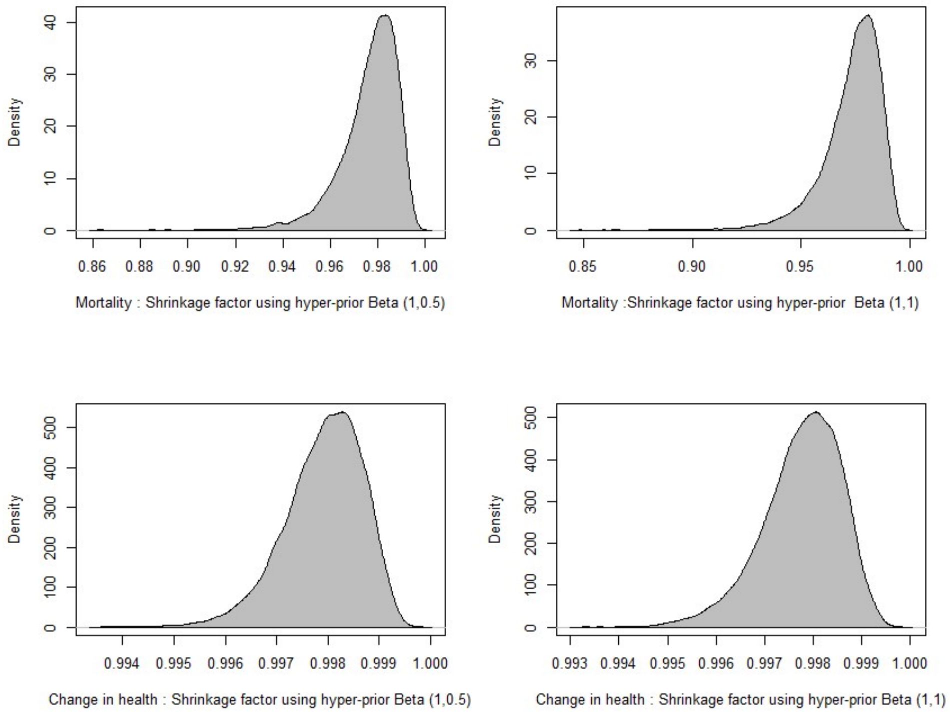
Table A-6.1: Posterior means (sds) [95% credible intervals] of the coefficients of the Bayesian logistic regression mortality and self-reported change in health models (M1), using a Beta(1,1) prior distribution on the shrinkage factor *w*

Coefficients	Mortality model (M1)	Change in health model (M1)
b_0 (Intercept)	–7.412 (1.426) [–10.233, –4.600]	–1.930 (0.522) [–2.946, –0.907]
b_1 (Chronological age)	0.011 (0.008) [0.022, 0.112]	0.012 (0.008) [–0.005, 0.028]
b_2 (Female)	–0.357 (0.178) [–0.704, –0.010]	0.038 (0.067) [–0.094, 0.171]
b_3 (Black Americans or other)	–0.336 (0.068) [–0.465, –0.204]	–0.337 (0.068) [–0.472, –0.203]
b_4 (Years of education)	0.027 (0.032) [–0.031, 0.091]	0.032 (0.011) [0.010, 0.053]
b_5 (Self-rated health)	–0.517 (0.108) [–0.735, –0.308]	–0.457 (0.041) [–0.538, –0.377]
b_6 (Number of ADLs)	–0.027 (0.092) [–0.214, 0.150]	0.163 (0.038) [0.088, 0.237]
b_7 (Number of IADLs)	–0.050 (0.180) [–0.413, 0.296]	0.022 (0.075) [–0.123, 0.169]
b_8 (Depression)	0.077 (0.041) [–0.005, 0.156]	0.099 (0.016) [0.068, 0.130]
b_9 (Number of chronic conditions)	0.162 (0.064) [0.036, 0.289]	0.178 (0.025) [0.128, 0.228]
b_{10} (Household income)	–0.162 (0.153) [–0.512, 0.075]	0.073 (0.033) [0.010, 0.140]
b_{11} (Smoking status)	0.262 (0.114) [0.043, 0.485]	0.163 (0.043) [0.077, 0.246]
b_{12} (Divorced, widowed, or never married)	0.182 (0.190) [–0.186, 0.555]	0.060 (0.071) [–0.081, 0.199]
b_{13} (Vigorous exercise)	–0.091 (0.078) [–0.247, 0.055]	–0.007 (0.027) [–0.060, 0.046]
LOO information criterion	1029.7	6193.5

Notes: ADLs: Number of Activities of Daily Living for which respondents report difficulties. IADLs: Number of Instrumental Activities of Daily Living for which respondents report difficulties. LOO information criterion: Leave-One-Out information.

The plots of the marginal posterior distributions of the shrinkage factor, using $a = 3$ and $a = 4$, for the M1 models on mortality and change in health, are depicted in Figure A-6.1. As expected, for either value of a, from the results in the main body of the manuscript, the density plots under the mortality models have fat left tails and the distributions are more concentrated around the mean (0.97). On the other hand, the distributions for the M1 models on change in health are even more concentrated near one.

Figure A-6.1: Marginal posterior distributions of the shrinkage factor, under the different values of the hyper-parameter a



A-7. Imputation techniques for the missing values

Regarding the missing values for the explanatory variables used for the mortality analysis (174 respondents), first we investigate if these missing values are completely random (missing completely at random, MCAR) or not. To do this we have performed the Little's test (Little 1988) (Test statistic = 446, p-value $<1\%$). This result indicates that the 174 missing values for the explanatory variables cannot be treated as missing completely at random. Next, we further assume that the missing mechanism is missing at random and not missing not at random and we impute these 174 missing values using the fully conditional specification method (Van Buuren et al. 2006). The beta coefficients for the mortality M1 model using the imputed sample of size $n = 9,137$ are presented in Table A-7.1 (second column) and the estimated (posterior mean) hyper-parameters are $w = 0.979$ and $g = 59.52$. We do not observe any worth mentioned differences between the results of the M1 'imputed' model on mortality and the ones under the M1 mortality model using the reduced dataset (see Table 3).

Table A-7.1: Posterior means (sds) [95% credible intervals] of the coefficients of Bayesian logistic regression model M1 on mortality (second column) and Bayesian logistic regression model M1 on self-reported change in health (third column)

Coefficients	Mortality model (M1)	Change in health model (M1)
b_0 (Intercept)	–7.357 (1.452) [–10.174, –4.497]	–1.846 (0.445) [–2.698, –0.935]
b_1 (Chronological age)	0.040 (0.023) [0.017, 0.108]	0.009 (0.007) [–0.005, 0.024]
b_2 (Female)	–0.315 (0.185) [–0.678, 0.045]	0.058 (0.066) [–0.067, 0.189]
b_3 (Black Americans or other)	0.061 (0.187) [–0.301, 0.432]	–0.336 (0.068) [–0.467, –0.202]
b_4 (Years of education)	0.037 (0.032) [–0.024, 0.104]	0.035 (0.011) [0.015, 0.058]
b_5 (Self-rated health)	–0.534 (0.112) [–0.758, –0.319]	–0.451 (0.042) [–0.534, –0.373]
b_6 (Number of ADLs)	–0.034 (0.112) [–0.218, 0.141]	0.157 (0.038) [0.082, 0.232]
b_7 (Number of IADLs)	–0.069 (0.179) [–0.430, 0.274]	0.003 (0.074) [–0.144, 0.146]
b_8 (Depression)	0.071 (0.041) [–0.011, 0.152]	0.101 (0.016) [0.070, 0.131]
b_9 (Number of chronic conditions)	0.174 (0.064) [0.048, 0.297]	0.169 (0.025) [0.118, 0.216]
b_{10} (Household income)	–0.201 (0.175) [–0.600, 0.063]	0.040 (0.030) [–0.019, 0.101]
b_{11} (Smoking status)	0.276 (0.118) [0.047, 0.512]	0.152 (0.042) [0.067, 0.233]
b_{12} (Divorced, widowed, or never married)	0.168 (0.191) [–0.203, 0.547]	0.044 (0.070) [–0.092, 0.183]
b_{13} (Vigorous exercise)	–0.103 (0.080) [–0.263, 0.049]	–0.015 (0.026) [–0.065, 0.037]

Notes: In the first case (second column) we have used imputation methods for the missing values of the explanatory variables, while in the second case (third column) we have used Bayesian imputation methods to treat the missing values in the dependent variable. ADLs: Number of Activities of Daily Living for which respondents report difficulties. IADLs: Number of Instrumental Activities of Daily Living for which respondents report difficulties.

Regarding the missing values for the dependent variable 'self-reported change in health' (1,628 respondents), we noticed that the majority of the respondents, about 90.5%, participated in Wave 13 but not in Wave 14. The remaining 9.5% are respondents who died between the two waves as well as those who participated but did not provide a response on their health change. To address this challenge, as a sensitivity check, we compare the results of the M1 health model based on the reduced sample ($n = 7,335$) (see Table 4) to the ones using the M1 health model under the imputed dataset ($n = 9,137$). In the second case, by assuming that the missing mechanism is ignorable, the missing values for the dependent variable 'self-reported change in health' are treated as unknown parameters and simultaneously estimated (together with the rest of the model parameters) using the posterior predictive distribution. The beta coefficients for the imputed dataset, are presented in Table A-7.1 (third column) and the estimated (posterior mean) hyper-parameters are $w = 0.998$ and $g = 672.3$. In this imputed dataset the impact of household income on the dependent variable is now not significant, but apart from this, no other material changes in the beta coefficients between the two M1 health models are observed.