Supplementary Materials for

Predictive modeling of U.S. health care spending in late life

Liran Einav, Amy Finkelstein*, Sendhil Mullainathan, Ziad Obermeyer

*Corresponding author. Email: afink@mit.edu

Published 29 June 2018, *Science* **360**, 1462 (2018)
DOI: 10.1126/science.aar5045

This PDF file includes:

Materials and Methods

Section A: Data and Variable Definitions

A.1. Data

We use administrative data from the Centers for Medicare and Medicaid Services (CMS) on Medicare claims for a 20% random sample of Medicare beneficiaries. Our baseline analysis relies on data from 2007 and 2008, although for some analyses we also use mortality data from subsequent years. We use the Master Beneficiary Summary File to provide the denominator of all beneficiaries, as well as basic demographics. We use the inpatient, outpatient, and carrier files to construct detailed measures of healthcare use, healthcare spending, and health. Finally, we use the Chronic Conditions file to define chronic conditions with which beneficiaries have been diagnosed by the end of a calendar year.

We begin with all Medicare beneficiaries in 2008 who are in the 20 percent denominator file (about 9.5 million beneficiaries). We exclude any individual who was in Medicare Advantage (i.e. not in fee-for-service Medicare) in any month in 2007 or 2008, because we do not observe healthcare claims for such individuals; this excludes about 2.8 million beneficiaries. We also exclude about 1.1 million beneficiaries who are not enrolled in Medicare Parts A and B in all months in 2007 and all months that they are alive in 2008; we do this so that we can have consistent and complete health and healthcare utilization records for our sample.

Our **baseline sample** therefore consists of about 5.6 million fee-for-service beneficiaries who were alive as of January 1, 2008 and who were continuously enrolled in Medicare Parts A and B for 2007 (as well as for any part of 2008 in which they were alive). We measure mortality over subsequent periods after January 1, 2008. Data from 2007 are used to form the mortality predictions from the vantage point of January 1, 2008.

A.2. Variables

All time-varying variables are coded relative to January 1, 2008. We construct measures over various time periods. Here we primarily define the measures. We discuss the time periods in more detail when we describe the features used in the prediction algorithm.

Mortality: The Medicare denominator file provides the date of patient death (if ever). We use this information to construct mortality over various periods after January 1, 2008. Our primary mortality measure is 12-month ("annual") mortality, although we also examine mortality over shorter horizons (a week, a month, and a quarter) and over longer horizons (3 and 5 years). In much of our discussion we refer to **"survivors"** as individuals who survive 12 months past January 1, 2008, and to "**decedents"** as individuals who died during the 12 months after January 1, 2008.

Medicare Spending: We analyze Medicare spending over various time periods. Medicare spending is defined as the sum of Medicare payments on the inpatient, outpatient, and carrier (a.k.a physician or other provider) files. This is a standard measure both in the end-of-life literature (e.g. (*13*)) and more generally in analysis of Medicare spending (e.g. (*22*)), but it excludes certain Medicare spending categories (such as home health services, and prescription drug expenditures for those who have Medicare Part D), as well as non-Medicare-covered expenditures such as nursing homes (see e.g. (*23*)).

For purposes of our prediction algorithm, we measure not only **total Medicare spending** but also **inpatient Medicare spending** and **outpatient Medicare spending,** which together sum to total. Inpatient spending is defined as any spending during the duration of an inpatient visit (from admission to discharge date). Outpatient spending is defined as spending that did not occur during an inpatient visit.

We also measure **out-of-pocket spending** (total out-of-pocket spending, as well as separately for inpatient and outpatient). Out-of-pocket spending reflects the amount for which the patient is responsible; it may be paid by the individual, paid by supplemental insurance, or left as unpaid medical debt. Finally, we include measures of payments made on behalf of a Medicare beneficiary by a non-Medicare primary payer.

When we analyze Medicare spending after January 1, 2008, our measures differ for survivors and decedents. For survivors, we measure spending in the 12 months after January 1, 2008. For decedents, we consider two different measures. Our "**backfilled"** spending measure follows the standard approach in the "end of life" literature (e.g. (*13*)) and measures spending in the 12 months prior to the date of death. Our **"unadjusted"** spending measure for decedents measures spending between January 1, 2008 and the date of death, which (by construction) occurs within the subsequent 12 months; this measure parallels the approach for survivors, but is truncated once the individual dies.

Healthcare utilization

Number of inpatient visits. We define an inpatient visit as an inpatient stay with a given admission and discharge date in the inpatient file. For the purpose of assigning inpatient visits to various time horizons, we use the date of admission.

Number of inpatient days. We measure the sum of the length of all the patient's inpatient visits to construct our measure of inpatient days. The length of a given inpatient visit is defined as the discharge date minus the admission date, plus one.

Number of inpatient procedures. Each claim (which essentially corresponds to a separate inpatient visit) on the inpatient files has up to six procedure codes associated with it. We count each non-missing procedure code as a procedure.

Number of inpatient ER visits. We use the inpatient file to define inpatient visits for which there was at least one charge billed to an Emergency Room. We identify these charges using the Revenue Center Code values of 0450-0459 and 0981 (see (*24*)).

Number of outpatient ER visits. We use the outpatient files to define outpatient ER visits. An outpatient ER visit is defined as a claim on the outpatient claims file with at least one charge billed to an Emergency Room; as with inpatient ER visits, we identify these charges using the Revenue Center Code values of 0450-0459 and 0981 (see (*24*)). We allow a maximum of one outpatient ER visit per day; we do this because we only have date of claims, not time of claims within the date, so as in most claims data, it's difficult to distinguish multiple visits from multiple claims that are associated with a single visit.

Number of physician visits. Physician visits are measured based on claims in the carrier file. We define physician visits as the sum of **primary care visits** and s**pecialty care visits**. We allow a maximum of one primary care visit per patient-day, and one specialist visit per patient-day for the same reason discussed above for capping outpatient ER visits. Following the approach in (*22*), our definition of primary care physicians and specialists follows the Dartmouth

Atlas (see (*25*), page 6). Specifically, we crosswalk the primary care and specialist definitions in the Dartmouth Atlas to the list of specialty codes in the CMS data.

Health measures

All of our health measures come from the claims data. Of course, as is well known in this literature, diagnoses are recorded only when healthcare utilization occurs; these commonly-used health measures therefore should be interpreted as reflecting *diagnosed* health problems as opposed to latent health (*26, 22, 27*).

Indicators for Gagne Conditions. We obtain from (*10*) a way of calculating – using ICD codes from the inpatient, outpatient, and carrier files – indicators for over 30 different conditions that they argue are helpful in predicting mortality in elderly patients. This is a combination of two commonly-used "comorbidity indices" (Charlson and Elixhauser). The full list of conditions we code is: alcohol abuse, anemia, cardiac arrhythmias, cerebrovascular disease, congestive heart failure, coagulopathy, complicated diabetes, dementia, depression, drug abuse, fluid and electrolyte disorders, hemiplegia, HIV/AIDS, hypertension, liver disease, metastatic cancer, myocardial infarction, neurodegenerative disorders, obesity, psychosis, pulmonary circulation disorders, chronic pulmonary disease, hypothyroidism, peripheral vascular disorder, renal failure, rheumatoid arthritis/collagen vascular diseases, any tumor, ulcer disease, uncomplicated diabetes, valvular disease, and weight loss. In addition, we coded ischemic stroke from ICD codes in a similar fashion to the coding of Gagne conditions.

Indicators for over 3,000 additional diagnoses. We map the approximately 15,000 ICD-9 diagnosis codes in the carrier, inpatient, and outpatient files to over 3,000 clinically salient diagnoses (e.g., pneumonia, back pain, fall). We use only the primary diagnosis listed on each claim for this mapping. The categorization system takes the AHRQ's Clinical Classifications Software (*28*) as a starting point. We then performed additional aggregations to engineer these categories, drawing on the experience of one of our co-authors, a physician with 10 years of clinical practice experience (ZO). This resulted in several improvements, such as creating categories to capture clinically specific entities (e.g., separating pulmonary embolism from other pulmonary heart disease), and breaking out frequently-occurring signs, symptoms, and ill-defined conditions often assigned as 'diagnoses' (e.g., shortness of breath, nausea, weakness). Further details can be found in (*11*). We also dropped features missing in over 99.9% of observations, to reduce sparsity and further focus the model on features that were likely to contribute to finding our rare outcome.

Indicators for Chronic Conditions. We use the 2007 Chronic Conditions segment of the Master Beneficiary Summary file to define the presence of 27 different chronic conditions by the end of 2007. The chronic conditions are defined by CMS; they are measured based on diagnoses coded in the past 1-3 years depending on the condition (see (*29*)). The conditions are Acquired Hypothyroidism (reference time period: 1 year), Acute Myocardial Infarction (1 year), Alzheimer's Disease and Related Disorders or Senile Dementia (3 years), Alzheimer's Disease (3 years), Anemia (1 year), Asthma (1 year), Atrial Fibrillation (1 year), Benign Prostatic Hyperplasia (1 year), Breast Cancer (1 year), Cataract (1 year), Chronic Kidney Disease (2 years), Chronic Obstructive Pulmonary Disease (1 year), Colorectal Cancer (1 year), Depression (1 year), Diabetes (2 years), Endometrial Cancer (1 year), Glaucoma (1 year), Heart Failure (2 years), Hip/Pelvic Fracture (1 year), Hyperlipidemia (1 year), Hypertension (1 year), Ischemic

Heart Disease (2 years), Lung Cancer (1 year), Osteoporosis (1 year), Prostate Cancer (1 year), Rheumatoid Arthritis / Osteoarthritis (2 years), and Stroke / Transient Ischemic Attack (1 year).

HCC Score. The HCC score is defined by the Centers for Medicare and Medicaid Services (CMS) for use in computing Medicare payments, and is designed to approximate predicted spending given demographics (including age, gender, and Medicaid eligibility) and diagnoses coded in the 12 months prior to January 1, 2008 in the inpatient, outpatient, and carrier claims data. Our HCC score derivation is based on (*30*).

Demographics

We use the Master Beneficiary Files to measure **age** (as of January 1, 2008), **race** (i.e. white or non white), and **gender**. We also use it to measure the individual's **geographic location** in 2007; specifically, we use information on the individual's ZIP code to map each individual to his or her Hospital Referral Region (HRR), as defined by the 1998 Dartmouth Atlas of Health Care. The 306 HRRs are collections of ZIP codes designed to approximate markets for tertiary hospital care (see (*31*) and (*32*).) Finally, we use these files to code an indicator variable for whether the individual was covered by **Medicaid** in any of the 12 months prior to January 1, 2008.

A.3 Summary Statistics

Table S1 shows some summary statistics for our baseline sample. The average age is 72, it is slightly over half female, and 14 percent non-white. Five percent of the sample dies over the year, 0.5 percent die within 30 days, and about one-quarter die over the subsequent five years. In the 12 months prior to January 1, 2008, average Medicare spending was about $6,000 and one-fifth of beneficiaries had an inpatient hospital admission. Those who end up dying in 2008 are noticeably older than survivors, and their 2007 healthcare spending is about three times higher.

Section B: Prediction Algorithm

In this Section we describe our predictors, our prediction algorithm, and its performance. To briefly summarize, we randomly set aside one-third of the baseline sample (the "test" sample), which we do not use as we optimize our prediction algorithm. For the remaining two-thirds (the "training" sample), we follow standard practice and tune the key parameters that govern the prediction model by cross-validation, finding those parameters that maximize the Area Under the Curve (AUC) criterion. Our optimal predictor gives rise to an AUC that is 0.867. One intuitive interpretation of AUC is that if we selected a random Medicare beneficiary who lives and one who dies, with 0.867 probability our model predicts a higher mortality risk for the beneficiary who ended up dying. We then apply the resulting prediction function to populate predicted mortalities for the Medicare beneficiaries in the test sample.

B.1. Potential predictors and feature selection

We construct thousands of potential predictors, which belong to three broad classes of variables: demographics, measures of prior healthcare utilization, and various indicators of health conditions.

First, we include demographic information on age (in years), sex, race (white or not), Medicaid status, and the residential location of the individual (using 306 HRR indicators).

Second, we use information on healthcare utilization over the 12 months prior to January 1 2008. This includes very detailed information on the beneficiaries' prior healthcare utilization, as well as the time path of this utilization, so that we can capture both the level and trend as potential predictors. Specifically, we summarize healthcare utilization by using total spending, inpatient spending, outpatient spending, out-of-pocket overall spending, out-of-pocket inpatient spending, out-of-pocket outpatient spending, number of inpatient visits, number of inpatient days, number of inpatient procedures, number of inpatient ER visits, number of outpatient ER visits, number of primary care visits, and number of specialty care visits. Importantly, because health (and therefore healthcare spending) is highly serially correlated, we try to capture the time path of recent healthcare experience by measuring the above variables at the quarter (rather than full year) level, and allowing each quarter to be a separate predictor. To emphasize the potential importance of recent health events, we also measure the healthcare variables described above for the day, 1-3 days, and 1-7 days prior to January 1, 2008, and allow these measures to serve as additional potential predictors.

Finally, we include a rich set of health measures. These include the approximately 3,000 diagnosis indicators described in Section A, over 30 Gagne conditions (measured separately by quarter, as with the healthcare utilization variables), 27 chronic condition indicators, and the beneficiary's HCC score. We measure the diagnosis indicators and Gagne conditions the day, 1-3 days, 1-7 days, 0-1 month, and 1-12 months prior to January 1, 2008. The chronic condition indicators and HCC score are measured as of the end of 2007.

B.2. Overview of the prediction procedure

Figure S1 provides a schematic way to understand the various components of the prediction algorithm. We first randomly draw a third of the sample to be part of the test sample. This group of beneficiaries is not used at all for prediction purposes in order to avoid overfitting. After

developing and optimizing the prediction algorithm, we apply the results to the beneficiaries in the test sample, and populate their predicted mortality. The main results in the paper are based on this test sample group of beneficiaries.

The remaining two-thirds of the baseline sample are used to develop our prediction algorithm. 90% of the data are used to train our predictions, 2.5% of the data are used to calculate ensemble weights, and 7.5% of the data are used to calibrate our predictions.

Normally, training the prediction algorithm would be straightforward. However, because mortality is not a frequent event in the baseline sample (only 5% of beneficiaries die over the 12 months after January 1, 2008), it is easier to train the prediction algorithm on a "balanced sample" in which half the beneficiaries die. We do so by randomly choosing only a fraction of the beneficiaries who survive, and training the algorithm on a selected sample that contains all the beneficiaries who die and a subset of the beneficiaries who survive. We randomly select approximately 1/19 = 5/95 of the survivors, so that the final balanced sample contains equal number of decedents and survivors. This final balanced sample is then randomly split into 5 equal-sized folds (with approximately 68,000 beneficiaries in each) on which our prediction algorithm is tuned.

Our final predictor is an ensemble of a random forest, gradient boosting regression trees, and LASSO. We calibrate our ensemble using Bayes' rule. This is necessary because while our predictions are fitted using a balanced sample, we then apply the resulting predictions to the test sample (which has the naturally-occurring, imbalanced proportion of decedents and survivors). Details on how we calibrate our predictions by Bayes' rule are below in Section B.5.

B.3. Tuning the prediction parameters

To obtain predictions we use an ensemble of a random forest, gradient boosting regression trees, and LASSO, which are all well-known and widely used machine learning techniques. We tune each individual algorithm using 5-fold cross-validation. For each vector of tuning parameters that we check, we estimate our algorithm five times. Each time we leave out one of the folds when we estimate the model and we use the left-out fold to calculate the performance of the tuning parameters. The performance measure that we use is area under the receiver operating characteristic curve (AUC). The use of AUC is a common metric in the machine learning literature in general and in the literature on mortality prediction in particular (e.g., *10-11, 18, 33-39*).

For the random forest, we tune four key parameters: (a) the number of prediction trees over which the random forest average is taken; (b) the number of observations in the bootstrap sample that each tree is using to generate predictions; (c) the number of distinct variables that are being considered for each split within a tree; and (d) the minimal number of observations (beneficiaries) in a node after which no additional splits are allowed. The number of trees and the size of the bootstrap sample did not make a large difference for tuning, so our parameter tuning focused on the latter two parameters. AUC is largest when nodes with fewer than 50 observations are not split any further and the number of variables considered at each split is 500.

We estimate gradient boosting regression trees using xgboost. We tune three parameters: (a) the number of trees used in the gradient boosting procedure; (b) the depth of each tree; and (c) the learning rate used to update between trees. The AUC is largest with 1,000 trees, a tree depth of 4, and a learning rate of 0.1. For LASSO, there is a single parameter to tune: the weight on the

penalty for large coefficient vectors in terms of the L1 norm (λ). The AUC is largest with a relative penalty of λ=0.00047.

B.4. Fitting the ensemble predictor

We combine our random forest, gradient boosting regression trees, and LASSO to create an ensemble predictor. Our ensemble predictor $\hat{p}_{ens}$ is given by the following linear combination of our individual predictors

$$\hat{p}_{ens} = \hat{\beta}_{rf}\hat{p}_{rf} + \hat{\beta}_{gb}\hat{p}_{gb} + \hat{\beta}_{lasso}\hat{p}_{lasso},$$

where $\hat{p}_x$ is the prediction from algorithm x and $\hat{\beta}_x$ is the associated weight. First, we estimate each of the individual models (the random forest, the gradient boosting regression trees, and LASSO) on the full balanced training sample. We then use these models to calculate predicted mortality in a separate sample that was not used for tuning any of the algorithms. Finally, we calculate the weights for our ensemble predictor by running an OLS regression of mortality on the predictions from each algorithm. We estimate this regression without a constant, so that our final ensemble is a linear combination of our three individual predictors. The gradient boosting regression trees get the largest weight of 0.807, the random forest gets a weight of 0.127, and the LASSO gets a weight of 0.066.

B.5. Addressing class imbalance

As mentioned earlier in this section, it is easier to train our predictors when the sample is balanced and has approximately equal number of decedents and survivors. Therefore, our resulting predicted mortality rates would be biased upwards, and need to be readjusted to the corresponding probabilities that would apply to the original, unbalanced sample. We follow the approach in (*39*) to use Bayes' rule to correct for the bias in the estimated probabilities.

We now describe this approach.[1] Let D be an indicator for dying, let S be an indicator for surviving, and let B be an indicator for being included in the balanced sample. Let $1/R$ be the ratio of decedents to survivors. By Bayes' rule

$$Prob(D|B) = \frac{Prob(B|D)Prob(D)}{Prob(B|D)Prob(D) + Prob(B|S)Prob(S)}.$$

Our ensemble predictor provides us with an estimate of $Prob(D|B)$ (in the balanced sample). To recover $Prob(D)$, which is our objective, we can replace $Prob(B|D) = 1$ and $Prob(B|S) = 1/R$. Thus we have

$$Prob(D|B) = \frac{Prob(D)}{Prob(D) + \frac{1}{R}(1 - Prob(D))},$$

And after rearranging we obtain

$$Prob(D) = \frac{\frac{1}{R}Prob(D|B)}{1 - \left(1 - \frac{1}{R}\right)Prob(D|B)}.$$

These three equations correspond, respectively, to equations (1), (3), and (4) in (*41*). The last equation describes the relationship between what we estimate, $Prob(D|B)$, and what we

[1] In this section, all probabilities are conditional on the features used to predict mortality, but we do not put this in our notation for ease of reading.

need, $Prob(D)$. In Fig. S2 we plot this relationship assuming that $R = 19$, which is close to the true R in the baseline sample.

Finally, because Bayes' rule is a theoretical relationship that may not hold exactly in the actual data, we use the 7.5% of the sample that we used to calibrate Bayes' rule, and fit a cubic relationship between the predicted mortality in the balanced sample and the actual mortality rates in the calibration sample; we use this empirical relationship rather than the theoretical relationship to map predicted mortality in the balanced sample to predicted mortality in the test sample.

That is, to construct our final analysis sample we take each beneficiary in the test sample, generate a mortality prediction based on the ensemble predictor described above, and then map this prediction to the actual prediction using the cubic relationship we obtained with the 7.5% calibration sample.

The resulting AUC from applying our predictor in the test sample is 0.867, which falls within the typical range of AUCs for this type of prediction exercise. As a benchmark, the papers we found and references at the end of Section B.3 that use machine learning to generate mortality predictions (*10-11, 19, 33-38*) obtain AUCs that range from 0.75 to 0.95, with some of them including predictors that are based electronic medical records and almost all of them predicting mortality within a given institution, where the sample is likely more homogeneous.

B.6. Performance

The resultant predictions are shown for the test sample in Fig. S3. It suggests that the model is well calibrated, as it is designed to be.

Table S2 shows summary statistics for predicted mortality for various sub-groups of the test sample. The predictions seem sensible, varying in expected ways with known risk factors such as age or cancer diagnosis.

Table S3 attempts to provide some guidance as to which predictors are the most important. To do this, we partition the predictors into different groups. We then report in column (2) the R-squared from regressing predicted mortality on each group of predictors separately, as a way to assess how much of the variation in predicted mortality that group alone can explain. While useful, we should be cautious about interpreting the R-squared statistics reported in column (2) as the importance of each group: many variables encode similar information, so it may be more important to assess the incremental predictive power of each group relative to a predictive model that includes all other predictors. This is what we report in column (3) of Table S3. For each group of predictors (shown in the left most column) we report the R-squared from regressing predicted mortality on all groups of predictors *except* the group indicated. The resulting R-squared should then be compared to the R-squared from including all predictors (0.815); the smaller the R-squared is when a given group is omitted, the more important that group is for prediction. According to this metric, it appears that medical expenditure variables are not important at all (excluding them reduces R-squared from 0.815 to 0.814), while chronic conditions and medical utilization are the type of variables that are most predictive of mortality (for example, excluding the chronic condition indicators reduced the R-squared from 0.815 to 0.737).

Supplementary Text

Section C: Potential impact on results from improved mortality prediction

A potential concern with our analysis is the possibility that with better prediction techniques and/or richer data sets, mortality would be more predictable. In this section we report an exercise to assess this concern.

Specifically, denote our predicted mortality for a given individual by *p* and denote by *d* an indicator that is equal to 1 if the individual died. We now construct a (hypothetically) improved predictor *p'* as a weighted average of *p* and *d*, such that $p' = 0.9p + 0.1d$.

We view the choice of 0.1 for the weight we put on the truth *(d)* as fairly aggressive in terms of a potential improvement in the prediction algorithm. To see this, note that the AUC measure discussed in Section B can be viewed as the probability that a random individual who ended up dying would have a higher predicted mortality than a random individual who end up surviving. Given our baseline AUC of $Pr(p_d > p_s) = 0.867$, it is easy to see that our revised predictor would perform much better, as it would be equal to $Pr(0.9p_d + 0.1 > 0.9p_s + 0) = Pr(p_d + 0.1/0.9 > 0.9p_s)$. Indeed, given that the vast majority of surviving individuals have a fairly low predicted mortality, the choice of 0.1 implies a vastly improved AUC of 0.963. We view this as aggressive because such high AUC is much higher than any AUC obtained in the literature for mortality predictions (see our discussion of the literature in Section B.5).

Figure S4 then replicates Fig. 3 using this vastly improved predictor. It shows that even with this substantial, hypothetical improvement in predictive power, high predicted mortality individuals still account for very little spending. For example, the highest risk percentile – who are still individuals with predicted mortality of about 47 percent – still accounts for only 5 percent of total spending. Relatedly, as shown in Fig. S5, even with this substantial hypothetical improvement in predictive power, high predicted mortality individuals remain rare.

We also examined how our results regarding spending differentials, adjusted for predicted mortality, would change with this hypothetical improved prediction. With our actual predictions, we estimated that adjusting survivors to have the same distribution of predicted mortality as decedents would eliminate 30-50 percent of the concentration of spending on decedents relative to survivors. With the hypothetical, improved predictor, we now eliminate 40-60% of the concentration of spending on decedents relative to survivors.

These exercises suggest that our primary conclusions are unlikely to change even if predicting mortality improves well beyond the current state of the art. Of course, our conclusions will change if prediction becomes perfect (or near perfect).

Section D: Analysis from the time point of hospital admission

Throughout the main text we predict mortality as of January 1, 2008. This captures the entire population of Medicare beneficiaries, and parallels the kind of statistics produced by the "end of life" literature which compares spending for all decedents over the prior 12 months to 12-month spending for survivors. However, a more meaningful analysis of the concentration of spending by ex-ante mortality probabilities would arguably generate these ex-ante mortality predictions at the time of an event that potentially triggers spending. In this section, therefore, we repeat our baseline analysis from the time point of an inpatient admission.

To do this, we focus on the subset of our baseline sample that had an inpatient admission in 2008. This consists of about 1.5 million beneficiaries, or about one quarter of our baseline sample. We refer to this as our "inpatient admission sample". For the inpatient admission sample, we measure mortality over subsequent periods after their inpatient admission, which we

refer to as the "index event." Individuals may have more than one inpatient admission in a calendar year; in such cases, we randomly select an inpatient admission to be the "index event." We exclude any patient who was in Medicare Advantage in the 12 months before or after the index event or who did not have Parts A and B coverage in the 12 months before or after the index event. Table S4 presents some summary statistics for the final inpatient admission sample, which consists of about 1.2 million beneficiaries. Annual mortality is now 21 percent, compared to 5 percent in our baseline sample.

We follow the same prediction procedure for the inpatient admission sample as we do in our primary analysis (see Section B). The main difference is that whereas in our primary analysis, our measures were all made from a common vantage point (of January 1, 2008), they are now made from the vantage point of the index admission, which varies across individuals. All potential predictors, mortality outcomes, and spending are as described above for our baseline sample, but are now measured in time periods relative to the index event (i.e. the date of hospital admission) rather than relative to January 1, 2008. In addition, since we can measure the primary diagnosis recorded at admission for the index admission, we use this to generate a set of additional predictors. To do so, we use Clinical Classifications Software classifier (*28*) to classify diagnosis code at admission to over 250 different diagnosis indicators. We also generate both 1-year and 30-day mortality predictions for the inpatient admission sample.

Our optimal predictor gives rise to an AUC of 0.844 for one-year mortality prediction. For the 30-day mortality prediction in the inpatient subsample, our optimal predictor gives rise to an AUC of 0.842. Table S5 presents a summary of the mortality predictions for both annual and 30-day mortality. Figs. S6-S9 replicate the four exhibits in the main text for annual mortality for the inpatient admission sample. Overall, while the inpatient admission sample naturally represents sicker and higher-mortality patients, the primary qualitative insights remain the same.

Section E: Spending differentials, adjusted for predicted mortality

In this section we use our mortality predictions to investigate how much of the concentration of spending on the dead is accounted for by a simple mechanical fact: sicker individuals are associated with higher healthcare spending and also with higher mortality. Fig. 4 showed that, not surprisingly, spending is increasing in predicted mortality. Since those who end up dying have higher predicted mortality than those who end up living, at least some of the spending difference between decedents and survivors naturally reflects these underlying health differences.

In Table S6 we provide one way to quantify how much of the spending difference between decedents and survivors in our primary analysis is accounted for by the fact that those who subsequently die are ex-ante sicker. It shows average annual spending for decedents – both unadjusted and backfilled. It also shows average annual spending for survivors (column 1), and how it changes in column 2 if we reweight the population of survivors, such that they have the same ex-ante mortality distribution as the decedents.

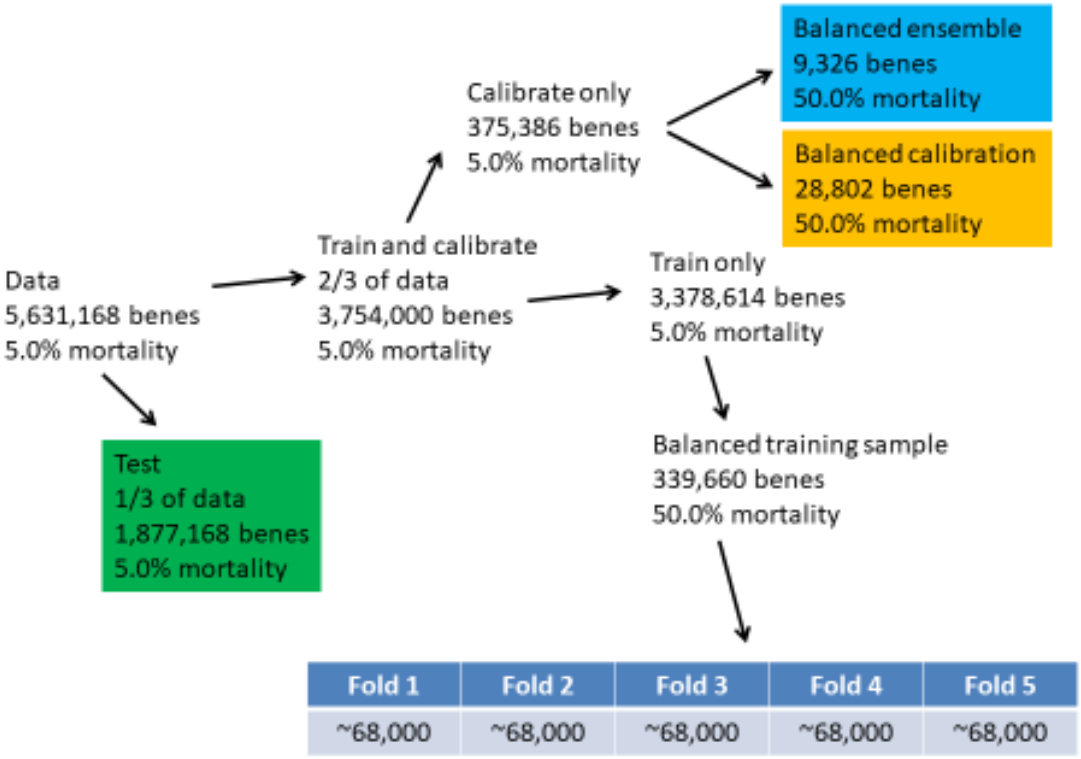
Survivors have lower mortality probabilities than decedents; the reweighting procedure therefore effectively gives more weight to those survivors with higher mortality probabilities (and also higher spending). Adjusting for differences in ex-ante mortality probabilities in this way reduces the disparity in spending on the dead relative to the living. Indeed, the first row of Table S6 indicates that when we reweight survivors to have the same distribution of ex-ante mortality as decedents, spending on survivors more than doubles. This eliminates a substantial portion of the overall spending difference. Since annual spending on decedents is between three

and five times higher than on survivors – depending on whether decedent spending is "backfilled" or "unadjusted" – our analysis indicates that accounting for ex-ante health can eliminate between 30 and 50 percent of the overall difference.

These results suggest that a non-trivial share of the concentration of spending on the ex-post dead reflects the fact that they are ex-ante sicker. However, the flip side of this finding should also be emphasized: a non-trivial difference remains even after accounting for the fact that those who die are sicker than those who live.

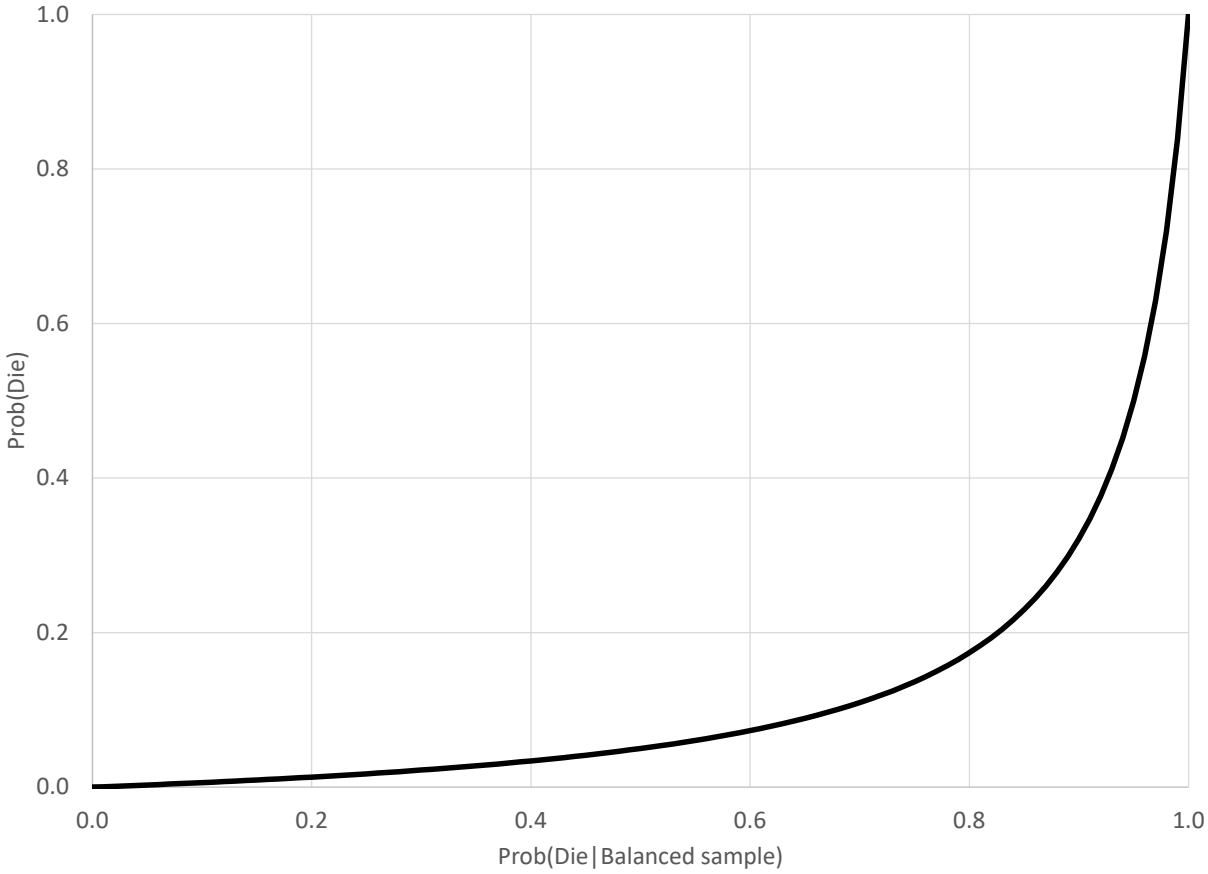
Why is death expensive, even conditional on ex-ante health? The remaining rows of Table S6 provide some initial clues. The next two rows show that the remaining difference in spending between decedents and (ex-ante) similar survivors is almost entirely attributable to differences in inpatient spending; outpatient spending is small and – once the distribution of predicted mortality for survivors is re-weighted to match that of decedents – fairly similar for survivors and decedents. Differences in the inpatient experience between decedents and ex-ante similar survivors in turn appears to be reflected largely in their number of inpatient admissions (last row) – other characteristics of the inpatient experience, such as length of stay, use of the intensive care unit, or number of procedures are more similar between decedents and (ex-ante similar) survivors (not shown).

Fig. S1.



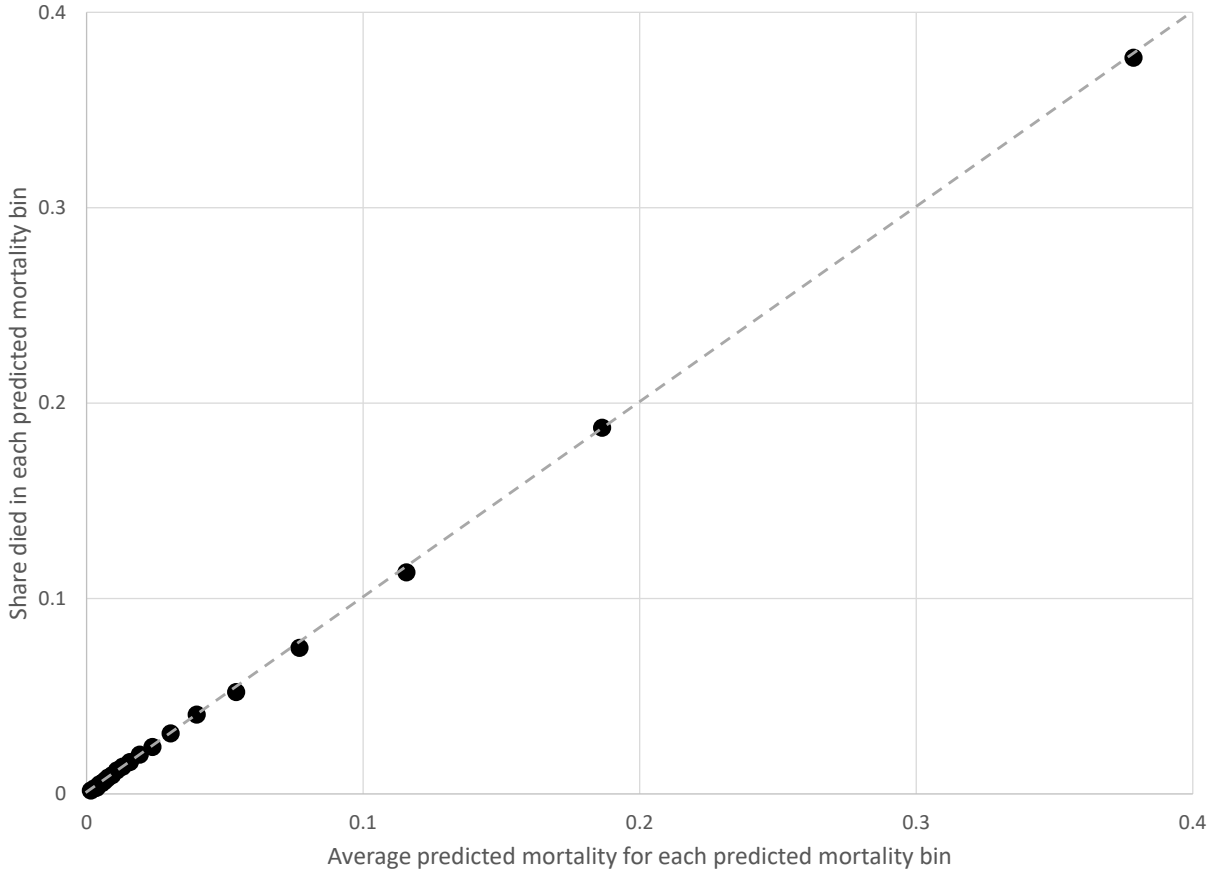
Schematic of prediction algorithm. Figure shows the sub-samples set aside to implement our prediction algorithm ("test" sample) and the sub-samples used in various ways to train and calibrate our prediction algorithm. We set aside one-third of the baseline sample as the test sample. From the two-thirds of the remaining baseline sample that is used to develop our prediction algorithm, we use 90% of the sample to train our predictors, 2.5% to calibrate the ensemble weights, and 7.5% to calibrate our predictions from the balanced training sample.

Fig. S2.



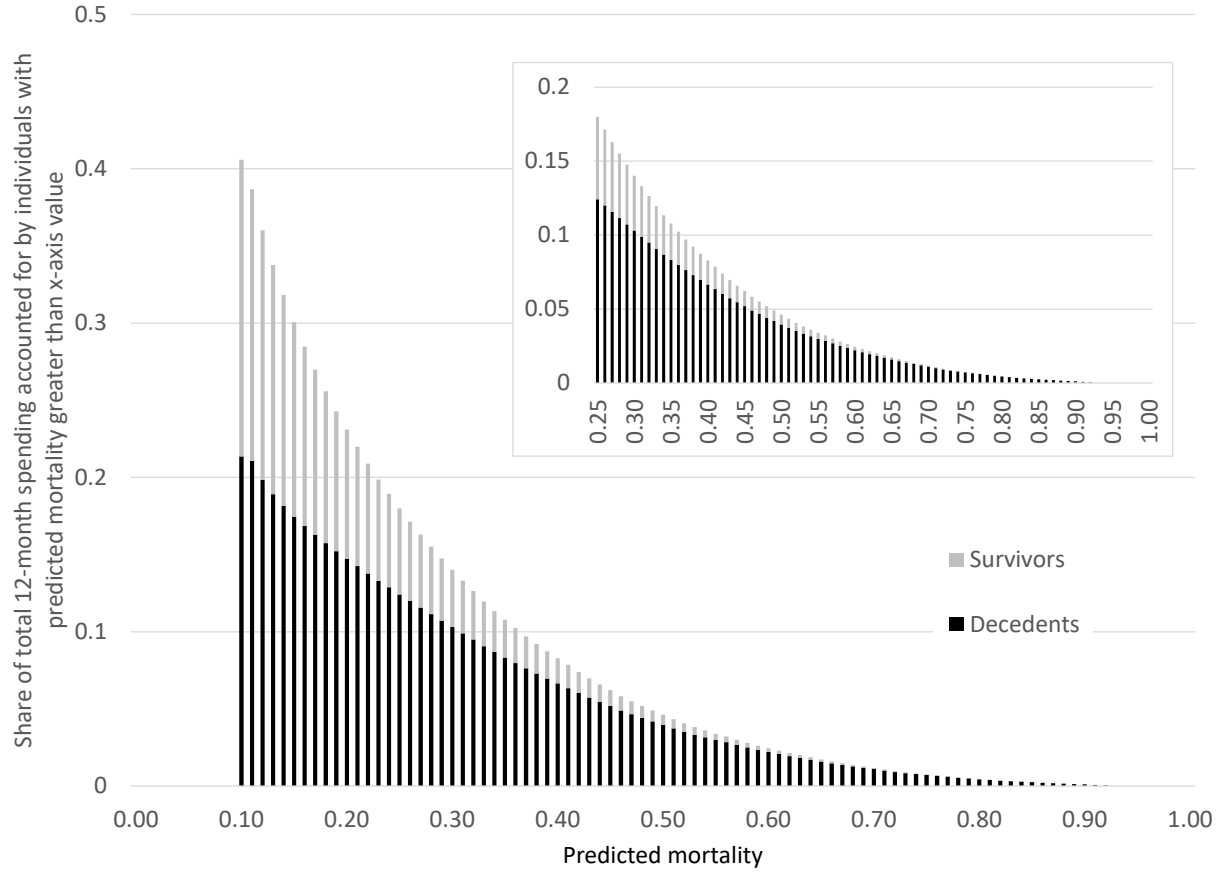
Using Bayes' rule to transition from balanced sample to impbalanced sample. Figure shows the theoretical relationship between $Prob(D)$ and $Prob(D|B)$ assuming that $R = 19$. See Section B.5 for more details.

Fig. S3.



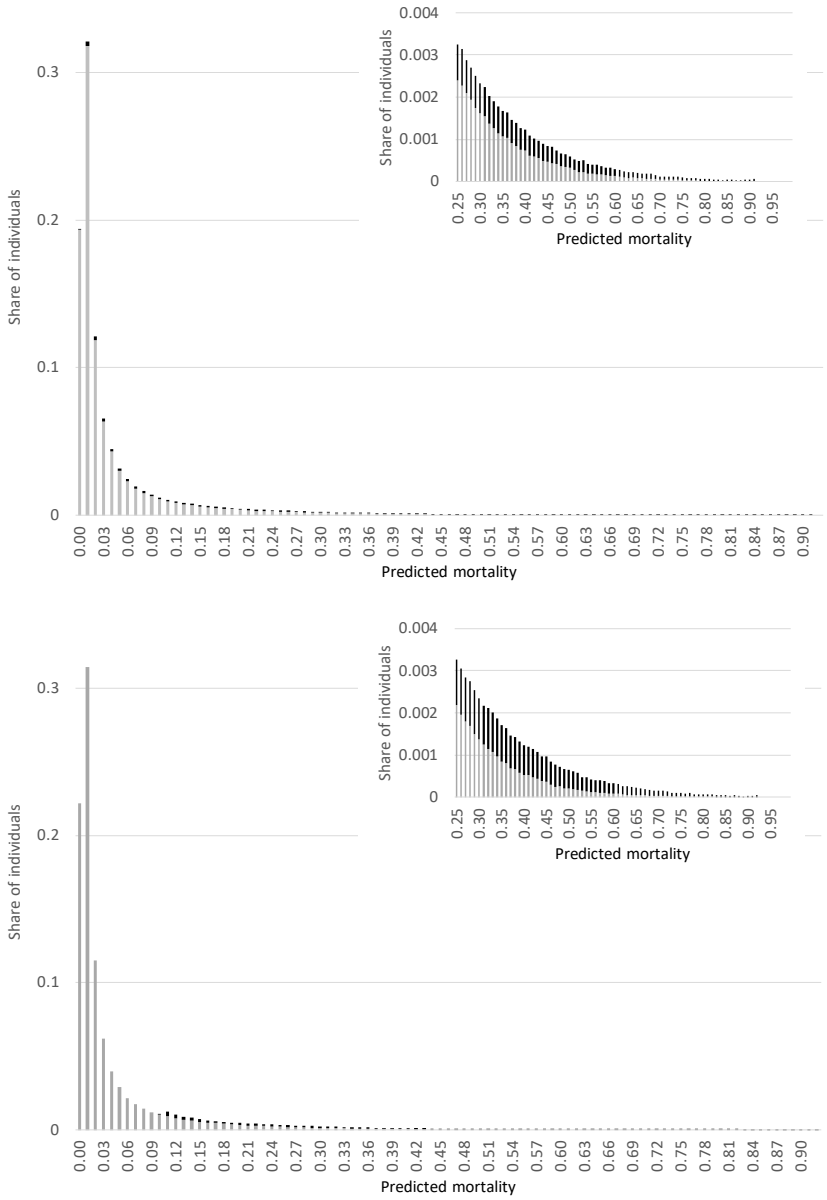
Model fit. Figure shows our final predictions on the horizontal axis against the actual mortality rate on the vertical axis for bins of beneficiaries in the test sample. To construct this figure, we sorted all individuals in the test sample by their predicted mortality, and divided them into 20 equal-sized bins. Within each bin we compute the average predicted mortality (horizontal axis) and the mortality share (vertical axis). The model seems to be well calibrated.

Fig. S4.



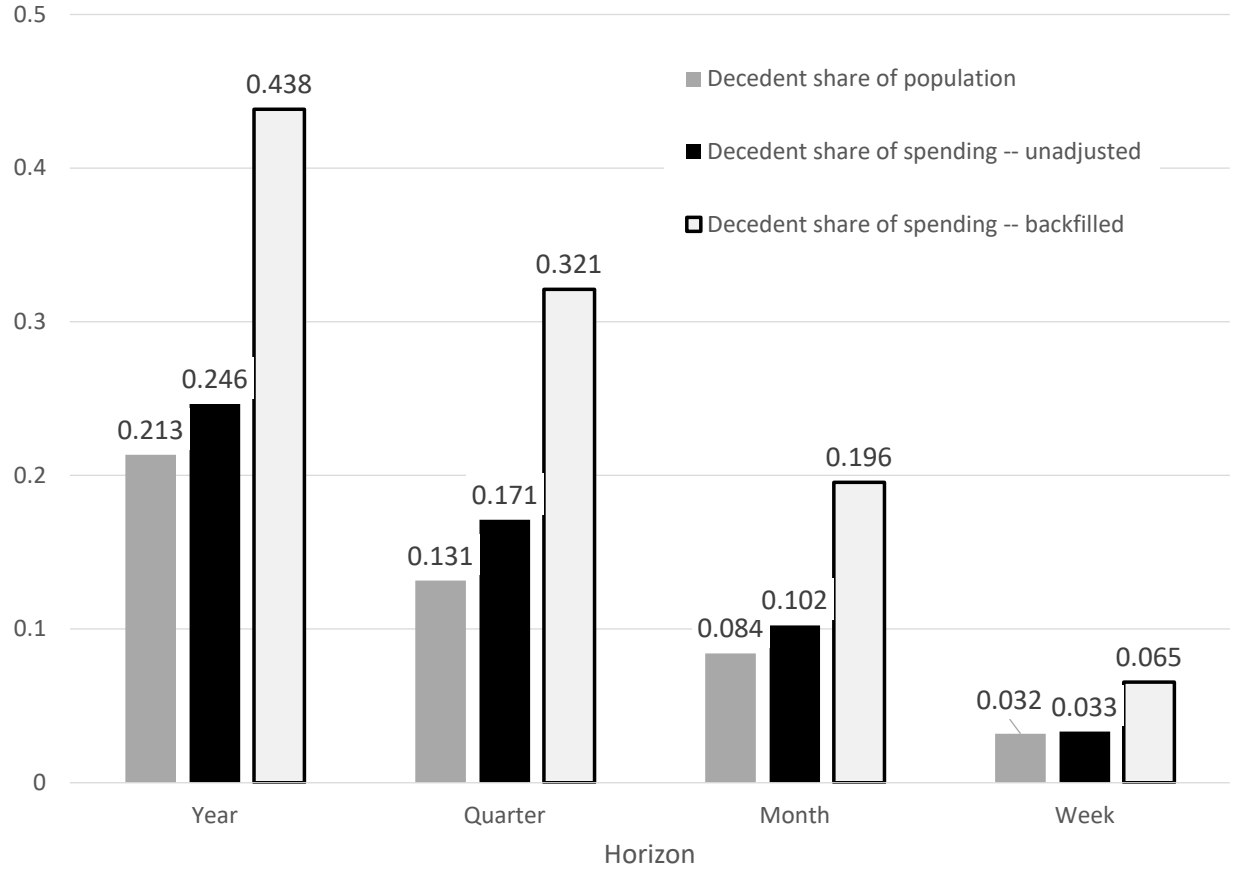
Spending by predicted mortality, for ex-post decedents and survivors (hypothetical improved predictor). This figure parallels Fig. 3 in the main text, but instead of using the mortality predictions for the test sample, it uses a (hypothetical) improved predictor that, as described in Section C, is a weighted average of the mortality prediction and whether ex-post, the individual actually died. For each level of this (hypothetical) improved predicted annual mortality (x-axis), exhibit shows the share of total annual Medicare spending that is accounted for by individuals with predicted mortality of that value or greater. It separately shows the share accounted for by decedents (black) and for survivors (gray). All results are based on the "backfilled" measure of decedent spending.

Fig. S5.



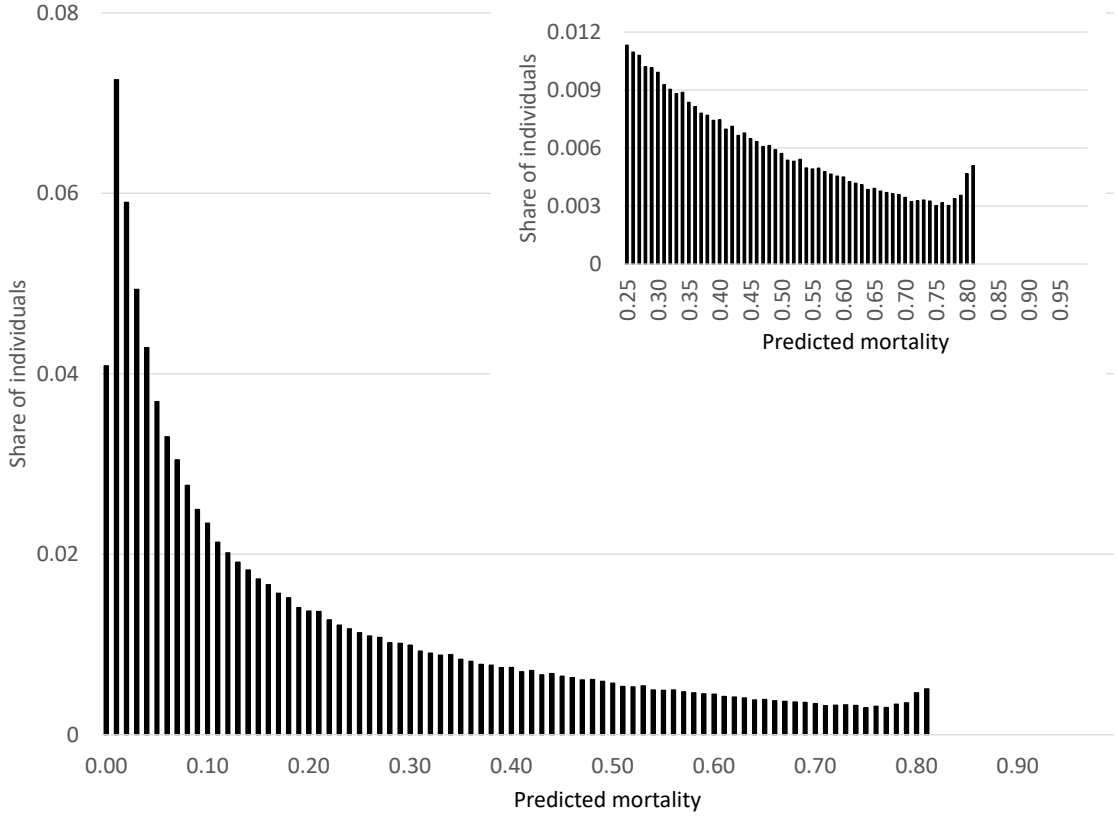
Distribution of predicted mortality, for actual and hypothetical improved predictor. The top panel replicates Fig. 2 in the main text, but indicates survivors and decedents separately. The bottom panel repeats the same figure, but for the (hypothetical) improved predictor described in Section C, which (by design) has greater share of decedents in the high end of the predicted mortality distribution.

Fig. S6.



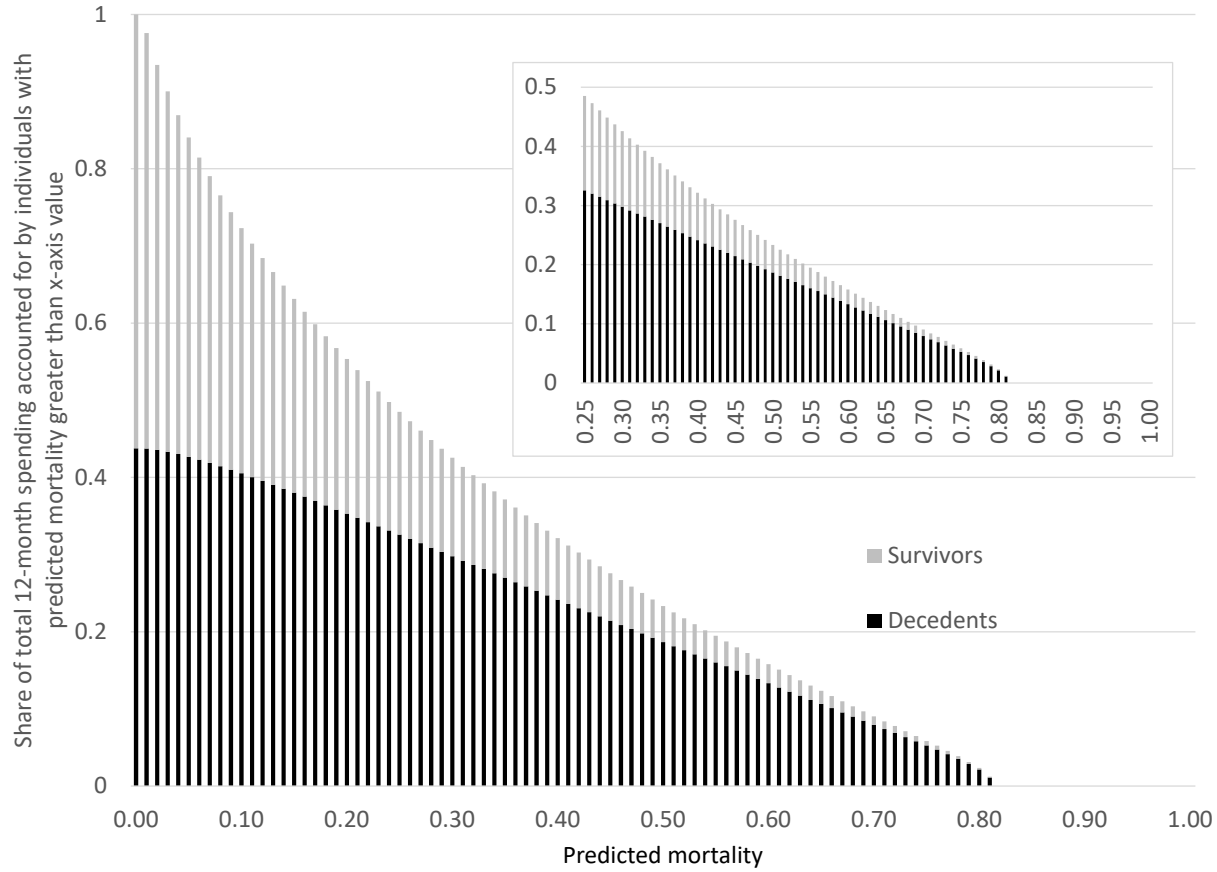
Concentration of spending on the ex-post dead (inpatient admission sample). This figure parallels Fig. 1 in the main text, but uses the inpatient admission sample (N=1,249,938) rather than the baseline sample. Figure shows mortality rates and decedent share of total Medicare spending for various time intervals after the "index event" (date of hospital admission). Spending for survivors is always measured in the time interval since the index event. For decedents, we report two spending measures: The "backfilled" approach measures spending looking backwards from the point of death for the length of the relevant interval (for example for the one-year measure, we measure spending over the 12 months prior to death); the "unadjusted" approach measures spending looking forward over the relevant time interval since the index event.

Fig. S7.



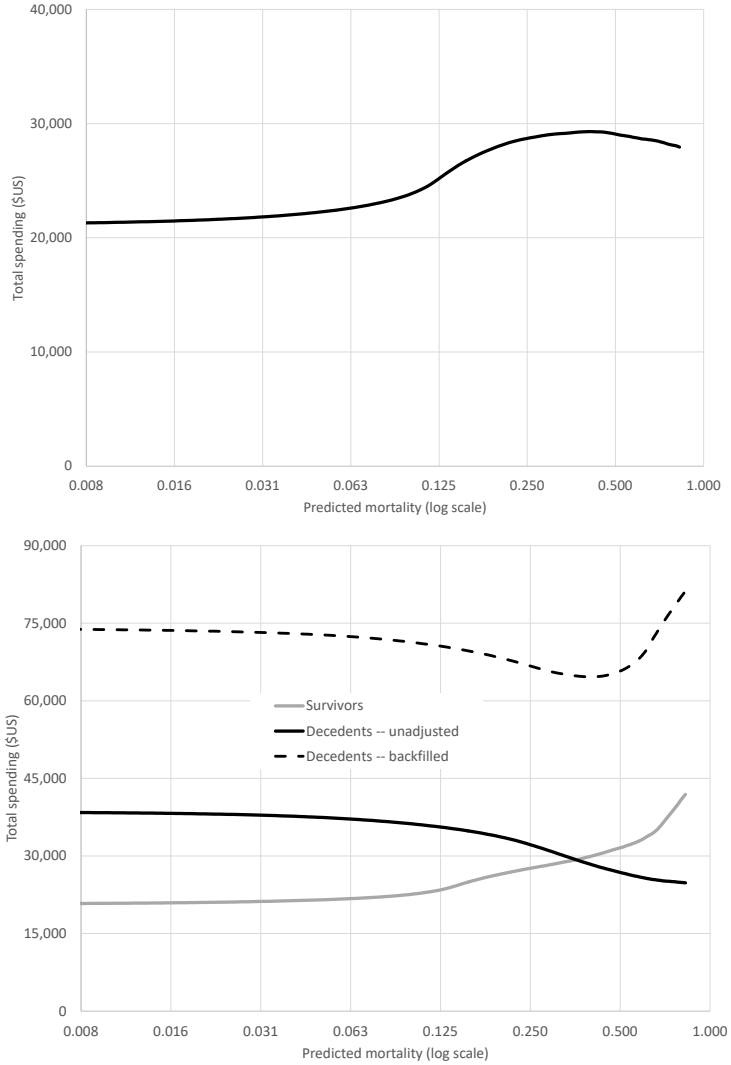
Distribution of predicted mortality (inpatient admission sample). This figure parallels Fig. 2 in the main text, but uses the inpatient admission sample rather than the baseline sample. It shows distribution of predicted annual mortality from the index event. All data are from the test subsample (N=416,787).

Fig. S8.



Concentration of spending by ex-ante mortality (inpatient admission sample). This figure parallels Fig. 3 in the main text, but uses the inpatient admission sample rather than the baseline sample. For each level of predicted annual mortality (x-axis), figure shows the share of total annual Medicare spending that is accounted for by individuals with predicted mortality of that value or greater. It separately shows the share accounted for by decedents (black) and for survivors (gray). All results are based on the "backfilled" measure of decedent spending. All data are from the test subsample.

Fig. S9.



Spending by predicted mortality (inpatient admission sample). This figure parallels Fig. 4 in the main text, but uses the inpatient admission sample rather than the baseline sample. Top panel shows kernel density of total Medicare spending in the 12 months after the index event, against predicted annual mortality (log scale). Bottom panel presents it separately for survivors and decedents, using two different measures of spending for decedents as defined in Fig. S6. All data are from the test subsample.

Table S1.

	Full sample	Survivors	Decedents
Demographics:			
Female	0.56	0.56	0.55
Non-white	0.14	0.14	0.13
Age on January 1, 2008	72.1	71.6	80.1
Mortality post January 1, 2008:			
30 days	0.005	0.000	0.091
One year	0.050	0.000	1.000
Five years	0.236	0.196	1.000
Spending and utilization in 12 months prior to January 1, 2008:			
Average spending ($)	6,235	5,644	17,426
Any hospital admission	0.20	0.18	0.47
Beneficiaries	5,631,168	5,348,434	282,734

Summary statistics of baseline sample. This table uses Medicare data from 2007 and 2008 to present summary statistics for our baseline sample: a 20% random sample of fee-for-service Medicare beneficiaries alive on January 1, 2008. We report mortality at various time intervals after January 1, and Medicare spending and utilization in the 12 months prior to January 1. "Decedents" are defined as individuals who die within 12 months of January 1, 2008.

Table S2.

Sample	N	Mortality rate	Percentiles of predicted mortality			Share of dead w/ pred. mort. > 0.5
			75th	95th	99th	
All	1,877,168	0.050	0.046	0.242	0.466	0.089
Age >85	207,277	0.158	0.230	0.450	0.638	0.117
Cancer	148,473	0.094	0.109	0.396	0.647	0.190
Metastatic Cancer	15,034	0.337	0.411	0.681	0.846	0.320

Distribution of annual mortality predictions for the test sample. Mortality outcomes are measured in the 12 months after January 1, 2008. In each row, we report the number of beneficiaries, their annual mortality rate, and their annual mortality probability. Specifically, we report the 75^{th}, 95^{th} and 99^{th} percentiles of their distribution of predicted mortality; in addition, for those who (ex-post) die within the subsequent 12 months, we report the share of decedents who had a predicted annual mortality above 50 percent.

Table S3.

	Number of distinct predictors w/in each group	R^2 (from each group alone)	R^2 (from all groups except the group indicated)
	(1)	(2)	(3)
Chronic Condition indicators	27	0.536	0.737
Medical utilization variables	70	0.567	0.741
Demographics	310	0.192	0.785
Indicators for additional diagnoses	869	0.501	0.797
Gagne indicators	256	0.356	0.811
Medical expenditure variables	90	0.222	0.814
All included	1,622	0.815	

Predictive importance of different groups of predictors. This table reports the R^2 from regressing predicted mortality in the test sample on various groups of predictors described in the left-most column. The regression is at the beneficiary level. Column (2) reports the R^2 from a given group of predictors shown in the left-most column. Column (3) reports the R^2 from including all of the predictors *except* the group shown in the left-most column.

Table S4.

	Inpatient Admission Sample		
	Full sample	Survivors	Decedents
Demographics:			
Female	0.57	0.58	0.55
Non-white	0.15	0.15	0.14
Age at index event	74.3	72.9	79.7
Mortality post index event:			
30 days	0.084	0.000	0.394
One year	0.213	0.000	1.000
Five years	0.484	0.344	1.000
Spending and utilization in 12 months prior to index event:			
Average spending ($)	14,356	12,151	22,480
Any hospital admission	0.41	0.37	0.56
Beneficiaries	1,249,938	983,174	266,764

Summary statistics for inpatient admission sample. This table uses Medicare data from 2007, 2008, and 2009 to present a set of summary statistics for our inpatient admission sub-sample that are parallel to those presented in Table S1 for the baseline sample. Instead of measuring mortality and spending for the beneficiaries relative to January 1, 2008 as in the baseline sample, these are all measured relative to the "index event" – i.e. the date of a (randomly selected) inpatient admission in 2008. Because the index admission can occur at any point throughout 2008, we need to measure spending and mortality into 2009.

Table S5.

Panel A. One-year predicted mortality

Sample	N	Mortality rate	Percentiles of predicted mortality			Share of dead w/ pred. mort. > 0.5	Share of dead w/ pred. mort. > 0.8
			75th	95th	99th		
All	416,787	0.214	0.331	0.674	0.794	0.399	0.031
Age >85	75,775	0.386	0.547	0.755	0.805	0.519	0.035
Cancer	41,709	0.306	0.492	0.766	0.807	0.565	0.061
Metastatic Cancer	10,471	0.628	0.745	0.806	0.809	0.818	0.124

Panel B. 30-day predicted mortality

Sample	N	Mortality rate	Percentiles of predicted mortality			Share of dead w/ pred. mort. > 0.5
			75th	95th	99th	
All	416,787	0.084	0.115	0.319	0.472	0.032
Age >85	75,775	0.161	0.231	0.421	0.503	0.039
Cancer	41,709	0.115	0.171	0.393	0.500	0.051
Metastatic Cancer	10,471	0.256	0.340	0.495	0.513	0.100

Distribution of mortality for inpatient admission sample. Panel A parallels Table S2, but uses the inpatient admission sample. Panel B shows results for 30-day mortality in the inpatient admission sample. All data are from the test subsample.

Table S6.

	Survivors		Decedents	
	Unweighted	Weighted	Unadjusted	Backfilled
	(1)	(2)	(3)	(4)
Total spending	6,581	13,810	22,886	33,839
Outpatient spending	3,078	5,806	3,799	7,269
Inpatient spending	3,503	8,004	19,087	26,570
Inpatient visits	0.34	0.78	1.40	2.05

Survivor spending adjusted for predicted mortality. This table shows – in the test sample – spending and utilization separately for those who survive 12 months post January 1, 2008 ("survivors") and those who die within 12 months of January 1, 2008 ("decedents"). Columns 1 and 2 present two spending and utilization measures for survivors: both measure the outcome in the 12 months from January 1, 2008, but in column 2 we reweight the survivors so that they have the same distribution of annual mortality probabilities as the decedents. Columns 3 and 4 present two spending and utilization measures for decedents: the "backfilled measure" based on the 12 months prior to death and the "unadjusted" measure based on the 12 months starting from January 1, 2008.

References and Notes

1. G. F. Riley, J. D. Lubitz, Long-term trends in Medicare payments in the last year of life. *Health Serv. Res.* **45**, 565–576 (2010). doi:10.1111/j.1475-6773.2010.01082.x Medline

2. A. Gawande, "Letting go," *New Yorker*, 2 August 2010; www.newyorker.com/magazine/2010/08/02/letting-go-2.

3. E. Porter, "Rationing health care more fairly," *New York Times*, 21 August 2012; www.nytimes.com/2012/08/22/business/economy/rationing-health-care-more-fairly.html.

4. E. J. Emanuel, L. L. Emanuel, The economics of dying—The illusion of cost savings at the end of life. *N. Engl. J. Med.* **330**, 540–544 (1994). doi:10.1056/NEJM199402243300806 Medline

5. Medicare Payment Advisory Commission, "Report to the Congress: Selected Medicare Issues," June 1999; https://babel.hathitrust.org/cgi/pt?id=mdp.39015046749704;view=1up;seq=3.

6. A. A. Scitovsky, "The high cost of dying": What do the data show? *Milbank Q.* **83**, 825–841 (2005). doi:10.1111/j.1468-0009.2005.00402.x Medline

7. J. P. Newhouse, An iconoclastic view of health cost containment. *Health Aff. (Millwood)* **12** (suppl. 1), 152–171 (1993). doi:10.1377/hlthaff.12.suppl_1.152 Medline

8. A. S. Detsky, S. C. Stricker, A. G. Mulley, G. E. Thibault, Prognosis, survival, and the expenditure of hospital resources for patients in an intensive-care unit. *N. Engl. J. Med.* **305**, 667–672 (1981). doi:10.1056/NEJM198109173051204 Medline

9. M. M. Desai, S. T. Bogardus Jr., C. S. Williams, G. Vitagliano, S. K. Inouye, Development and validation of a risk-adjustment index for older patients: The high-risk diagnoses for the elderly scale. *J. Am. Geriatr. Soc.* **50**, 474–481 (2002). doi:10.1046/j.1532-5415.2002.50113.x Medline

10. J. J. Gagne, R. J. Glynn, J. Avorn, R. Levin, S. Schneeweiss, A combined comorbidity score predicted mortality in elderly patients better than existing scores. *J. Clin. Epidemiol.* **64**, 749–759 (2011). doi:10.1016/j.jclinepi.2010.10.004 Medline

11. M. Makar, M. Ghassemi, D. M. Cutler, Z. Obermeyer, Short-term mortality prediction for elderly patients using Medicare claims data. *Int. J. Mach. Learn. Comput.* **5**, 192–197 (2015). doi:10.7763/IJMLC.2015.V5.506 Medline

12. L. C. Yourman, S. J. Lee, M. A. Schonberg, E. W. Widera, A. K. Smith, Prognostic indices for older adults: A systematic review. *JAMA* **307**, 182–192 (2012). doi:10.1001/jama.2011.1966 Medline

13. C. Hogan, J. Lunney, J. Gabel, J. Lynn, Medicare beneficiaries' costs of care in the last year of life. *Health Aff. (Millwood)* **20**, 188–195 (2001). doi:10.1377/hlthaff.20.4.188 Medline

14. S. K. Inouye, S. T. Bogardus Jr., G. Vitagliano, M. M. Desai, C. S. Williams, J. N. Grady, J. D. Scinto, Burden of illness score for elderly persons: Risk adjustment incorporating the cumulative impact of diseases, physiologic abnormalities, and functional impairments. *Med. Care* **41**, 70–83 (2003). doi:10.1097/00005650-200301000-00010 Medline

15. P. Glare, K. Virik, M. Jones, M. Hudson, S. Eychmuller, J. Simes, N. Christakis, A systematic review of physicians' survival predictions in terminally ill cancer patients. *BMJ* **327**, 195–198 (2003). doi:10.1136/bmj.327.7408.195 Medline

16. J. R. Lakin, M. G. Robinson, R. E. Bernacki, B. W. Powers, S. D. Block, R. Cunningham, Z. Obermeyer, Estimating 1-year mortality for high-risk primary care patients using the "surprise" question. *JAMA Intern. Med.* **176**, 1863–1865 (2016). doi:10.1001/jamainternmed.2016.5928 Medline

17. A. Elfiky, M. Pany, R. Parikh, Z. Obermeyer, A machine learning approach to predicting short-term mortality risk in patients starting chemotherapy. bioRxiv 204081 [Preprint]. 18 October 2017. https://doi.org/10.1101/204081.

18. A. Rajkomar, E. Oren, K. Chen, A. M. Dai, N. Hajaj, P. J. Liu, X. Liu, M. Sun, P. Sundberg, H. Yee, K. Zhang, G. E. Duggan, G. Flores, M. Hardt, J. Irvine, Q. Le, K. Litsch, J. Marcus, A. Mossin, J. Tansuwan, J. W. De Wang, J. Wilson, D. Ludwig, S. L. Volchenboum, K. Chou, M. Pearson, S. Madabushi, N. H. Shah, A. J. Butte, M. Howell, C. Cui, G. Corrado, J. Dean, Scalable and accurate deep learning for electronic health records. arXiv:1801.07860 [cs.CY] (11 May 2018).

19. M. D. Aldridge, A. S. Kelley, The myth regarding the high cost of end-of-life care. *Am. J. Public Health* **105**, 2411–2415 (2015). doi:10.2105/AJPH.2015.302889 Medline

20. A. S. Kelley, K. E. Covinsky, R. J. Gorges, K. McKendrick, E. Bollens-Lund, R. S. Morrison, C. S. Ritchie, Identifying older adults with serious illness: A critical step toward improving the value of health care. *Health Serv. Res.* **52**, 113–131 (2017). doi:10.1111/1475-6773.12479 Medline

21. A. S. Kelley, E. Bollens-Lund, K. E. Covinsky, J. S. Skinner, R. S. Morrison, Prospective identification of patients at risk for unwarranted variation in treatment. *J. Palliat. Med.* **21**, 44–54 (2018). doi:10.1089/jpm.2017.0063 Medline

22. Research Data Assistance Center, How to identify hospital claims for emergency room visits in the Medicare claims data; https://www.resdac.org/resconnect/articles/144.

23. Y. P. Tabak, X. Sun, C. M. Nunez, R. S. Johannes, Using electronic health record data to develop inpatient mortality predictive model: Acute Laboratory Risk of Mortality Score (ALaRMS). *J. Am. Med. Inform. Assoc.* **21**, 455–463 (2014). doi:10.1136/amiajnl-2013-001790 Medline

24. A. Silva, P. Cortez, M. F. Santos, L. Gomes, J. Neves, Mortality assessment in intensive care units via adverse events using artificial neural networks. *Artif. Intell. Med.* **36**, 223–234 (2006). doi:10.1016/j.artmed.2005.07.006 Medline

25. G. C. Pope, J. Kautter, R. P. Ellis, A. S. Ash, J. Z. Ayanian, L. I. Lezzoni, M. J. Ingber, J. M. Levy, J. Robst, Risk adjustment of Medicare capitation payments using the CMS-HCC model. *Health Care Financ. Rev.* **25**, 119–141 (2004). Medline

26. A. D. Pozzolo, O. Caelen, R. A. Johnson, G. Bontempi, in *2015 IEEE Symposium Series on Computational Intelligence* (Institute of Electrical and Electronics Engineers, 2015), pp. 159–166.

27. A. Finkelstein, M. Gentzkow, P. Hull, H. Williams, Adjusting risk adjustment—Accounting for variation in diagnostic intensity. *N. Engl. J. Med.* **376**, 608–610 (2017). doi:10.1056/NEJMp1613238 Medline

29. A. Avati, K. Jung, S. Harman, L. Downing, A. Ng, N. H. Shah, Improving palliative care with deep learning. arXiv:1711.06402 [cs.CY] (17 November 2017).

29. A. Finkelstein, M. Gentzkow, H. Williams, Sources of geographic variation in health care: Evidence from patient migration. *Q. J. Econ.* **131**, 1681–1726 (2016). doi:10.1093/qje/qjw023 Medline

30. Y. Song, J. Skinner, J. Bynum, J. Sutherland, J. E. Wennberg, E. S. Fisher, Regional variations in diagnostic practices. *N. Engl. J. Med.* **363**, 45–53 (2010). doi:10.1056/NEJMsa0910881 Medline

31. R. Pirracchio, M. L. Petersen, M. Carone, M. R. Rigon, S. Chevret, M. J. van der Laan, Mortality prediction in intensive care units with the Super ICU Learner Algorithm (SICULA): A population-based study. *Lancet Respir. Med.* **3**, 42–52 (2015). doi:10.1016/S2213-2600(14)70239-5 Medline

32. P. Genevès, T. Calmant, N. Layaïda, M. Lepelley, S. Artemova, J.-L. Bosson, Predicting at-risk patient profiles from big prescription data. hal-01517087v4 [Preprint] (2017); https://hal.inria.fr/hal-01517087v4.

33. A. Awad, M. Bader-El-Den, J. McNicholas, J. Briggs, Early hospital mortality prediction of intensive care unit patients using an ensemble learning approach. *Int. J. Med. Inform.* **108**, 185–195 (2017). doi:10.1016/j.ijmedinf.2017.10.002 Medline

34. Agency for Healthcare Research and Quality (AHRQ), Healthcare Cost and Utilization Project (HCUP), Clinical Classifications Software (CCS) for ICD-9-CM (AHRQ, Rockville, MD, 2017).

35. The Dartmouth Atlas of Health Care, ZIP code crosswalks; www.dartmouthatlas.org/tools/downloads.aspx?tab=39.

36. The Dartmouth Atlas of Health Care, Appendix on the geography of health care in the United States; www.dartmouthatlas.org/downloads/methods/geogappdx.pdf.

37. Chronic Conditions Data Warehouse, Condition categories; www.ccwdata.org/web/guest/condition-categories.

38. The Dartmouth Atlas of Health Care, Research methods; www.dartmouthatlas.org/downloads/methods/research_methods.pdf.

39. E. B. French, J. McCauley, M. Aragon, P. Bakx, M. Chalkley, S. H. Chen, B. J. Christensen, H. Chuang, A. Côté-Sergent, M. De Nardi, E. Fan, D. Échevin, P.-Y. Geoffard, C. Gastaldi-Ménager, M. Gørtz, Y. Ibuka, J. B. Jones, M. Kallestrup-Lamb, M. Karlsson, T. J. Klein, G. de Lagasnerie, P.-C. Michaud, O. O'Donnell, N. Rice, J. S. Skinner, E. van Doorslaer, N. R. Ziebarth, E. Kelly, End-of-life medical spending in last twelve months of life is lower than previously reported. *Health Aff. (Millwood)* **36**, 1211–1217 (2017). doi:10.1377/hlthaff.2017.0174 Medline