Mechanisms of Ageing and Development 199 (2021) 111570

Contents lists available at ScienceDirect

Mechanisms of Ageing and Development

journal homepage: www.elsevier.com/locate/mechagedev

Original Article

A quantile frailty index without dichotomization

Garrett Stubbings[a], Kenneth Rockwood[b], Arnold Mitnitski[b,1], Andrew Rutenberg[a,*]

[a] Department of Physics and Atmospheric Science, Dalhousie University, Halifax, Nova Scotia, Canada B3H 4R2
[b] Division of Geriatric Medicine, Dalhousie University, Halifax, Nova Scotia, Canada B3H 2E1

ARTICLE INFO

Keywords:
Frailty index (FI)
Dichotomization
Quantile
Summary measures of health
Aging

ABSTRACT

Summary measures of health quantify the aging process of individuals. They should be interpretable, associated with future adverse outcomes, and straightforward to assemble. We use the rank-ordering of risk within a population to construct a quantile frailty index (QFI) that avoids dichotomization, is convenient and interpretable, and is associated with adverse outcomes. We show that the QFI outperforms previous frailty index (FI) measures on cross-sectional laboratory data (NHANES, CSHA, and ELSA). We construct the QFI by ranking the risk of individuals with respect to a reference population. Sex-specific reference populations narrow male–female FI differences as a function of age, and improve predictive performance. With a fixed reference population of 80–85 year olds, our QFI appears similar to earlier FI measures. With an age-matched reference population for each individual, we obtain a QFI that contains very little age information and that has similar predictive performance as other age-controlled FI measures. Adding age as an auxiliary variable leads to significantly better performance. We conclude that age should be controlled for when evaluating the predictive performance of summary measures of health. This is straight-forward to do with the QFI.

1. Introduction

Population health declines with advancing age, but health trajectories vary considerably between individuals (Nicklett, 2011). There are many distinct measures used to assess aspects of health on both the individual and population level. These range from molecular details of epigenetic methylation, to laboratory blood and metabolite tests, to clinical assessment measures in the comprehensive geriatric assessment, to self-assessed functional measures such as in the activities of daily living (ADL) or independent ADL (IADL). In principle, tens of thousands of distinct measurements are accessible for any individual. Nevertheless, any one measurement varies both intrinsically and due to measurement quality control (McPherson, 2017). Furthermore, any one measurement paints an incomplete picture of individual health. To obtain a fuller picture, summary measures of health can be assembled from many disparate measurements.

Summary measures of health combine many aspects of individual health into one. They include frailty (Mitnitski et al., 2001; Fried et al., 2001; Aguayo et al., 2018), prognostic measures (Shi et al., 2020), Allostatic Load (Juster et al., 2010), epigenetic clocks (Horvath, 2013; Levine, 2020), and biological age (Li et al., 2020; Belsky et al., 2018). These metrics span the range from the tissue level of biological age, to the standard laboratory evaluations of the Allostatic Load, to the functional level of the FI or the frailty phenotype. Functional-level summary measures are strongly associated with a wide array of adverse health outcomes (Zucchelli et al., 2019).

While many summary measures of health overlap in how they are constructed or how they perform, they are generally not identical (Aguayo et al., 2018; Shi et al., 2020; Belsky et al., 2018; Levine, 2020; Li et al., 2020). This reflects multidimensional aging – including organismal scales ranging from cellular, to tissue, to functional (Ferrucci et al., 2018; Levine et al., 2018; Jazwinski and Kim, 2019). To assess multidimensional health more completely, we need to continue to both develop new summary measures of health and to improve existing ones. For example, controlling for both age and sex is important in assessing and comparing individual health. How to conveniently and effectively do this for a given summary measure of health is a persistent challenge.

Here, we focus on the frailty index (FI) (Mitnitski et al., 2001) because it is simply constructed and can be effectively adapted to a broad variety of health aspects. The FI has been defined as the proportion of measured health aspects which are considered to be in the unhealthy state. Candidate health variables considered in the FI include

[*] Corresponding author.
E-mail address: andrew.rutenberg@dal.ca (A. Rutenberg).
[1] Deceased May 26, 2021.

https://doi.org/10.1016/j.mad.2021.111570
Received 4 June 2021; Received in revised form 19 August 2021; Accepted 7 September 2021
Available online 10 September 2021
0047-6374/

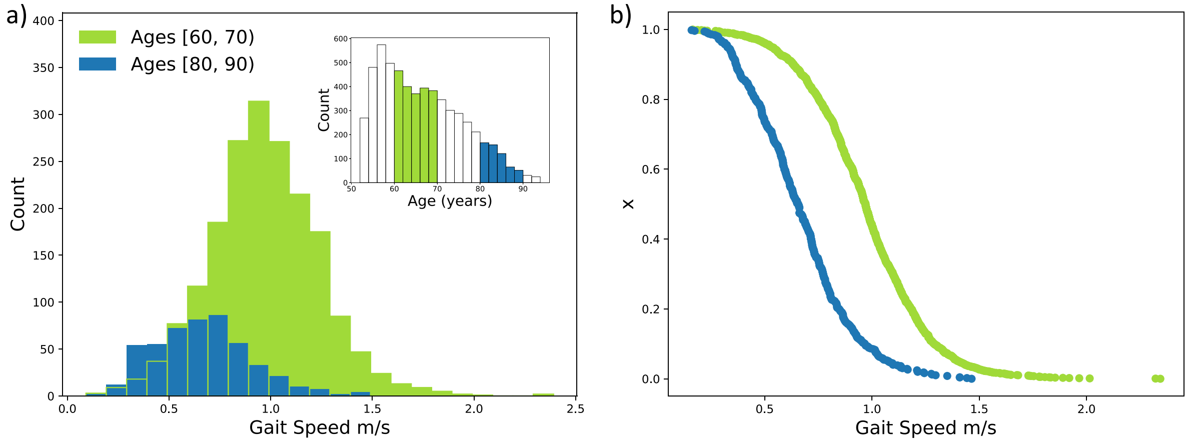

Fig. 1. Risk quantile calculation example with the ELSA dataset. We show (a) the distribution of gait speeds for example reference populations of 60–70 year-olds (green) and 80–90 year-olds (blue), with (b) the associated risk quantile *x vs.* gait speed. The inset in (b) shows the age distribution. Gait speed decreases with age, so the highest risk quantiles are associated with low gait speeds.

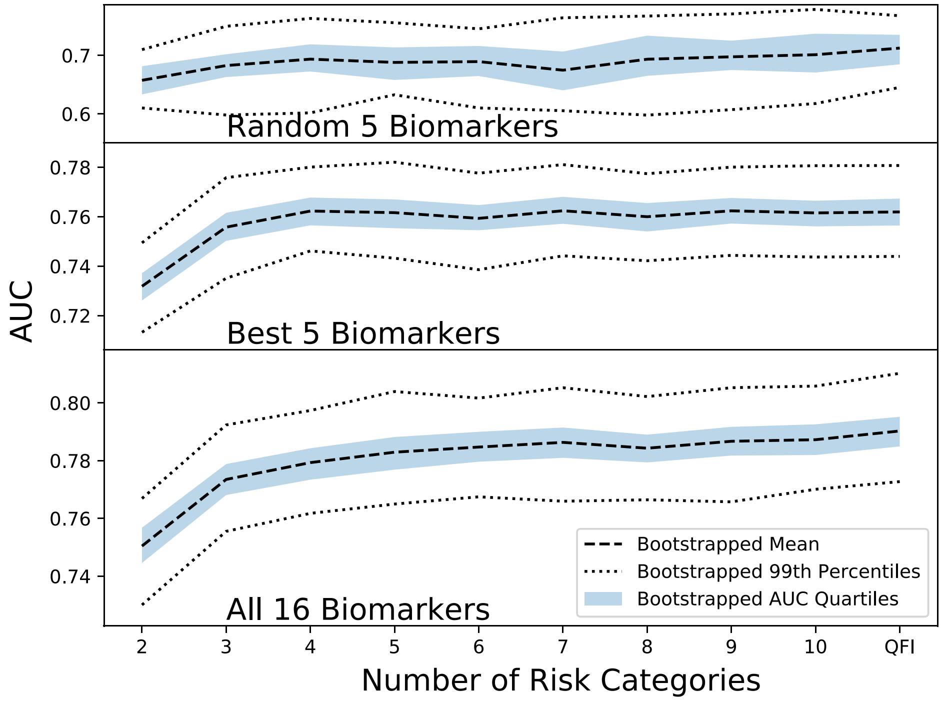

Fig. 2. The relationship between number of risk categories and the predictive value with respect to mortality within 5 years in the NHANES dataset. 2 risk categories is equivalent to dichotomization at the median, 3 risk categories equivalent to risk tertiles, and so forth. The upper plot shows the effect using 5 randomly selected biomarkers, the middle plot shows the best 5 biomarkers selected by AUC with respect to 5 year mortality, and the bottom plot shows the results using all available biomarkers. We resample the data using half the population size 400 times for each point, with the random 5 biomarkers also being re-selected 20 times. The dotted lines show the upper and lower 1st percentiles of AUC, the shaded blue region shows the upper and lower quartile range of AUC, and the dashed line shows the average AUC. Note that AUC ranges improve from the top to the bottom plots.

anything health-related that increases in prevalence with age (Searle et al., 2008). These have typically been high-level health deficits such as impairments in acts of daily living, self-rated health, and other clinically observable deficits in an FI-Clin (Rockwood et al., 2004). However, biomarkers such as the results of blood tests can also be used to create an effective FI-Lab (Blodgett et al., 2017; Mitnitski et al., 2015; Howlett et al., 2014; Stubbings et al., 2020). Both FI-Clin and FI-Lab are strongly associated with adverse health outcomes including mortality.

One challenge in calculating FI-Lab is how to properly incorporate measurements which are not already dichotomized. Typically, measurements are dichotomized based on normal reference ranges such as those found in clinician's handbooks – such as McPherson (2017) (Blodgett et al., 2017; Mitnitski et al., 2015; Howlett et al., 2014). However, diagnostic thresholds – intended to guide treatment – may not be appropriate for a summary measure of health (Stubbings et al., 2020). Furthermore, many biomarkers do not have associated diagnostic thresholds. With larger omics-style biomarker assays becoming more prevalent this absence will become increasingly pressing (Karczewski and Snyder, 2018).

There are also significant intuitive and empirical issues with dichotomization (or "binarization") of continuous variables (Cohen, 1983; Altman and Royston, 2006; Royston et al., 2006; Fedorov et al., 2009; Naggara et al., 2011; Dawson and Weiss, 2012). These are well understood for predictive measures since there are quantifiable losses in statistical power when imposing dichotomy on a continuous variable (Cohen, 1983). Individual dichotomized variables are sensitive to small variations around the cutpoint. Consider an individual measurement with a value close to the dichotomization threshold: any small variation of that measurement could result in a switch from absence to presence of deficit – the maximum penalty for a minimal variation. Frailty indices reduce these issues by averaging a large number of variables (Pe na et al., 2014; Mitnitski et al., 2015), but the scale of these effects have not

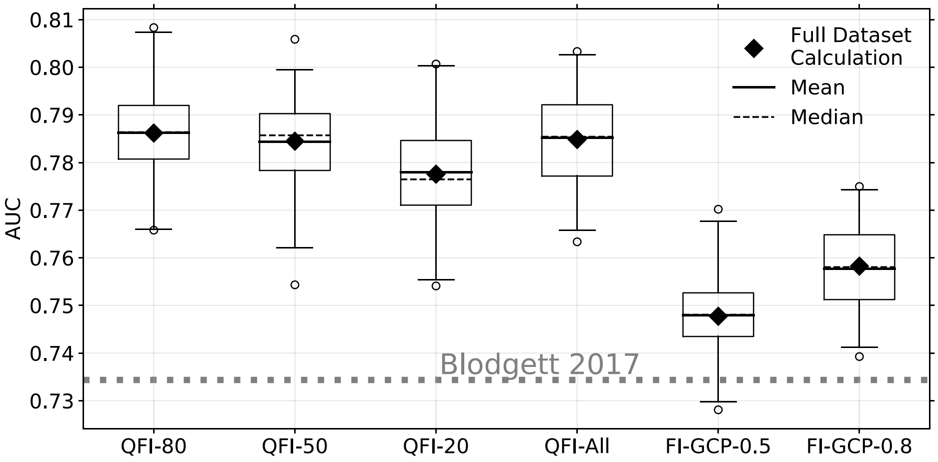

Fig. 3. The predictive value of various FI-Lab with respect to 5 year mortality in the NHANES study. From left to right we show the QFI using the 80–85 year-old reference population, the QFI with a 50–55 year-old reference, QFI with a 20–25 year-old reference, QFI using the whole NHANES population, and FI-GCP with the cutpoint at 0.5 or at 0.8 (Stubbings et al., 2020). Box and whisker plots display the data from resampling and cross-validation: the boxes represent the upper and lower quartiles, the whiskers go to the 99th and 1st percentiles, and the circles are remaining outliers. The short dashed line within each box is the median, the solid line the mean, and the diamond is the AUC for the full data set without cross validation or resampling. The horizontal grey dashed line shows the AUC of the published FI-Lab using the same data (Blodgett et al., 2017).

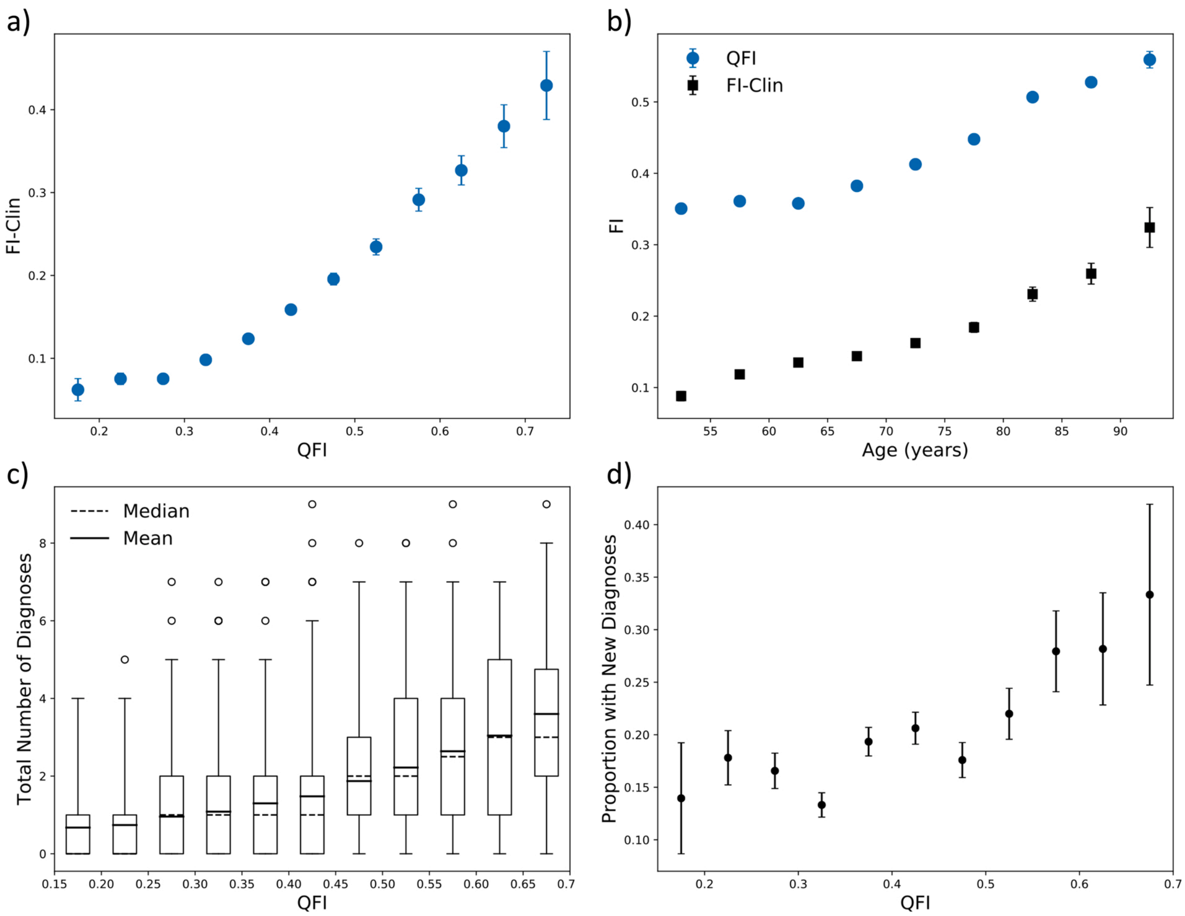

Fig. 4. QFI on wave 2 of the ELSA dataset, using a reference 80–85 year-old population. (a) The average FI-Clin binned by QFI for wave 2 of the ELSA dataset. (b) The average QFI (blue points) and FI-Clin (black squares, described in the Supplemental Information) binned by age. (c) The relationship between the QFI and the total number of existing or previous diagnoses. The boxes represent the upper and lower quartiles, the whiskers go to the 99th and 1st percentiles, and the circles are remaining outliers. The short dashed line within each box is the median and the solid line is the mean. (d) The fraction of the population with 1 or more new diagnoses in the year following the QFI evaluation.

been systematically explored within the FI literature.

To assemble an FI-Lab without dichotomization, we first need to preprocess health measurements in order to be able to combine them into a single measure. The common approach of using *Z*-scores (or standard scores), which shift measurements by their mean and then rescale by their standard deviation, does not naturally fit into the 0 (maximal health) to 1 (maximal unhealth) range of FI scores. However, ranking individuals by age-related health risk with respect to a reference population is an effective way of pre-processing an arbitrary set of biomarker measurements that naturally leads to a 0 to 1 range (Stubbings et al., 2020). Furthermore, using age-related risk ensures that scores increase with the aging trend, following the definition of a deficit

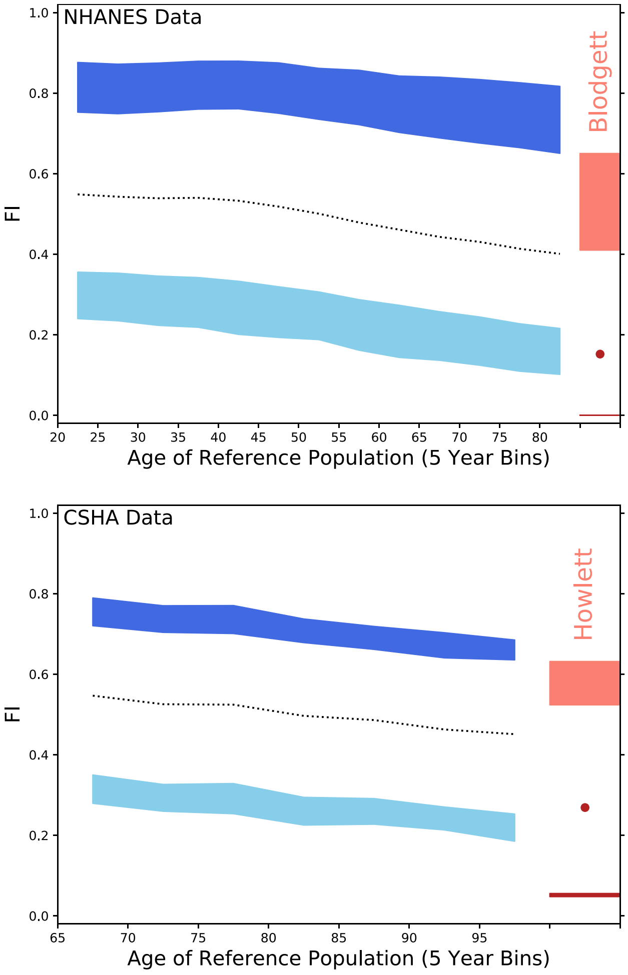

Fig. 5. The effects of changing the reference population on the distribution of QFI scores in the NHANES (top) and CSHA (bottom) studies. The upper 1% (light blue) and lower 1% (dark blue) of the QFI distributions as the age of the reference cohort changes in 5 year bins. On the right side of the plots, the red blocks show the upper 1% (light red), lower 1% (dark red), and average (red point) for the respective published FI-Lab using diagnostic thresholds (Blodgett et al., 2017; Howlett et al., 2014). We require each bin to have at least 20 individuals, removing only the 100–105 year-old bin in the CSHA study that had 2 individuals.

in Searle et al. (2008). Rank normalization is often used in e.g. pre-processing of gene expression data (Tsodikov et al., 2002), and is illustrated in Fig. 1. We found that by imposing a single global quantile cutpoint (GCP) on all of the individual rank-normalized scores, the resulting FI-GCP outperformed pre-existing FI-lab with the same data and was effective for a broad range of GCP (Stubbings et al., 2020). We show in the Methods how this quantile approach can be adapted to assemble an FI without dichotomization. Nevertheless, while quantile approaches avoid artificially grouping individual measurements, preserve aging trends, and treat all biomarkers similarly, they do require an explicit reference population.

Any health assessment is implicitly with respect to one or more reference populations. For example, dichotomization of any one variable creates two reference populations – a healthy one and an unhealthy one. Repeating this for many health variables creates many small reference populations that have identical dichotomized scores across many health measures. In contrast, quantile approaches can share a reference population across many health measures. With a small number of reference populations we can more easily treat them as independent, or controllable, ingredients of our summary measure of health. In particular, we can use reference populations to control for age and sex effects.

By construction, deficits included in the FI increase in prevalence with age (Searle et al., 2008). As a result, there is often a significant correlation between included biomarkers and age. For dichotomized biomarkers, this raises the question of how to age-control thresholds. Doing this by prognosis raises the question of whether multiple age-related outcomes may lead to distinct thresholds. However age-control is done, or not done, it will affect the resulting FI-lab. We can think of age as a confounding variable in terms of assessing aging health. How much we can learn about individual health independently of an individual's age? This is broad question that has also been raised in the context of biological age (Mitnitski et al., 2017), and other summary measures of health.

Issues of dichotomization are compounded when state variables, such as sex, are considered. Summary measures of health should reflect differences in health between the sexes. In many studies men are measured as "healthier" despite having greater prevalence of negative outcomes (Gordon and Hubbard, 2018). Selecting clinical-level health deficits based on sex-dependent prevalence affects sex-dependent mortality prediction of composite measures (Kulminski et al., 2008). When biomarker measurements are used the prevalence of each deficit can be tuned by the dichotomization threshold. However, using sex-dependent diagnostic thresholds still results in large sex differences in the FI (Howlett et al., 2014). Furthermore, new biomarkers do not yet have known diagnostic relevance or sex-dependent relevance. A transparent approach may be best: treat sexes as independent populations and use identical methods for calculating FI for each sex.

While a broad reference population with natural demographics is used in e.g. the frailty phenotype (Fried et al., 2001) (gender, height, BMI) or allostatic load (Juster et al., 2010) (non-stratified), we find that three smaller reference populations are particularly useful. One is a population of older adults (80–85 year-olds). This group is more prone to adverse health and outcomes than younger adults, but is still very well represented in population studies since they are slightly below the average human lifespan in Canada. This reference population is useful since it leads to an FI that is most similar to existing FI-lab measures in appearance. The second reference population we explore is a set of age-matched populations that can be matched to each individual. We use this to critically examine the explicit and implicit role of age in the FI, particularly with respect to its association with adverse health outcomes. The third type of reference population is to use sex-specific reference cohorts in combination with either of the others. This allows investigation of sex differences in the FI with a non-parametric approach.

In this work, we show that the quantiles of age-related risk lead to a predictive and interpretable FI-lab, which we call the quantile frailty index (QFI). The QFI predicts 5 year mortality significantly better than previous dichotomized methods of creating FI-Lab. The QFI is strongly correlated with the number of accumulated diagnoses, with number of new diagnoses accumulated at a one year follow-up, and is strongly associated with independent FI-Clin for the same individuals. Furthermore, we show that changing the reference population does not significantly affect prediction, but does affect the observed distribution of the QFI. Using different reference populations, we investigate the role of age and sex in the predictive quality of the QFI.

2. Methods

2.1. Quantile frailty index (QFI)

Consider N biomarkers that are assessed for every individual in a reference population, so that the ith biomarker (y_i) has a distribution $P(y_i)$. We take the risk quantile x_i as the position of the corresponding biomarker value y_i in its cumulative distribution. For biomarkers which

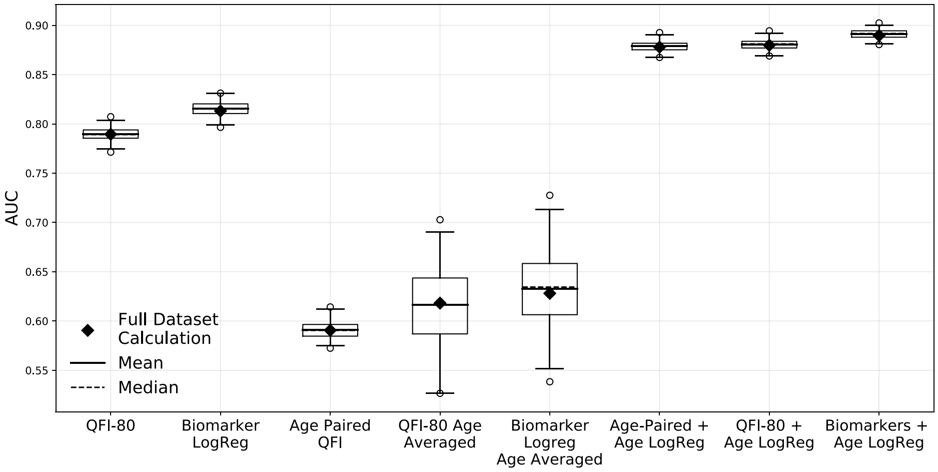

Fig. 6. Comparing the predictive value for the different methods of calculating the QFI against 5 year mortality in the NHANES dataset. From left to right we first have two "raw" points: the QFI with an 80–85 year-old reference population (QFI-80) and a logistic regression model using all of the biomarkers regressed against 5 year mortality. Then three age-controlled points: the age-paired QFI, QFI-80 with AUC averaged across performance within 5 year age bins, and a logistic regression of the biomarkers against 5 year mortality with AUC averaged across performance within 5 year age bins. Then three age-supplemented points: the age-paired QFI combined with age in a logistic regression, the QFI-80 combined with age in a logistic regression, and the raw biomarkers included with age in a logistic regression. Box and whiskers are from resampling with half the population. The boxes represent the upper and lower quartiles, the whiskers go to the 99th and 1st percentiles, and the circles are remaining outliers. The short dashed line within each box is the median, the solid line the mean, and the diamond is the AUC for the full data set without resampling. We use logistic regression to control for age in the prediction rather than for testing a logistic model, so performance is evaluated on the same individuals as the model is fit on.

increase with age (e.g. c-reactive protein), we take the quantile to increase with increases of the biomarker:

$$x_i = \int_0^{y_i} P(y_i^{'})\mathrm{d}y_i^{'}. \tag{1}$$

In the case of biomarkers that decrease with age (e.g. gait speed), we take the quantile to increase with *decreases* of the biomarker:

$$x_i = \int_{y_i}^{\infty} P(y_i^{'})\mathrm{d}y_i^{'}. \tag{2}$$

In both cases, the quantile $x_i \in [0, 1]$ increases with age on average. Alternatively, risk directions could be chosen such that increased quantile score is associated with increased mortality or other adverse health outcomes (Stubbings et al., 2020). However, using age-related risk directions satisfies the definition of a deficit defined by Searle et al. (2008). Furthermore, the differences between age-related risk and mortality-related risk are small in the datasets examined here – as also seen in (Stubbings et al., 2020). Obtaining the quantile is equivalent to performing a rank normalization of the score with respect to the population. Because many biomarkers have limited measurement precision there are frequent ties in biomarker scores. We use the minimum rank of tied scores; other methods of tie-breaking lead to similar results.

Our definition of the risk quantile means that x_i corresponds to the proportion of the population that has lower health-risk associated to that biomarker. So, x_i is equivalent to the "fraction unhealthier than" for a given biomarker with respect to a reference population. For example, Fig. 1 shows that having a gait speed of 1 m/s is slower than about 50% of the 60–70 year-olds, so $x_i = 0.5$ is the fraction of 60–70 year-olds that an individual with a gait speed of 1.0 m/s is unhealthier than. Calculating risk scores using this approach can be effectively done in many programming and statistics packages.

We then average the N non-dichotomized risk quantile measures for every biomarker to obtain an individual frailty index, the QFI:

$$\mathrm{QFI} = \sum_{i=1}^{N} \frac{x_i}{N} = \left\langle x_i \right\rangle, \tag{3}$$

where the angle-brackets indicate an average. We then have $QFI \in [0, 1]$.

We can quantify the advantage of having a continuous score by also examining m discrete risk categories such as dichotomization ($m = 2$) (Stubbings et al., 2020), tertiles ($m = 3$), quartiles ($m = 4$), or general m. For m risk categories, our risk scores would then be

$$d_i^{(m)} = \mathrm{floor}(x_i * m)/(m-1), \tag{4}$$

where the *floor*(*z*) function returns the greatest integer less than or equal to z. So for dichotomization ($m = 2$) scores of $x_i \in [0, 0.5)$ would give $d_i^{(2)} = 0$ while $x_i \in [0.5, 1)$ would give $d_i^{(2)} = 1$. We can then construct discrete $\mathrm{QFI}_m = \sum_{i=1}^{N} d_i^{(m)}/N$. As $m \to \infty$ we obtain $d_i \to x_i$.

The QFI can be calculated with respect to an arbitrary reference populations. We examine two age-related reference populations. The first is a fixed-age reference, which was defined as all individuals from a particular study (NHANES, CSHA, or ELSA) that were within a fixed range of ages – 80–85 year olds unless otherwise stated. We will use this 80–85 year old reference population as the default reference for the QFI – unless otherwise mentioned this is the reference population used.

A second reference population was age-matched. Here we used the same fixed-range bins for both the reference population and the individuals (so, e.g., the quantiles of 50–55 year olds were determined with respect to 50–55 year olds). For all reference populations, and unless otherwise stated, our results are for non-overlapping 5-year ranges of ages and the age of the population for plotting purposes was taken to be the middle of the range.

We also consider sex-matched reference populations. Some measurements (e.g. grip strength) vary substantially between the sexes and comparing individuals only within their group is desirable. We can combine sex and age to make very specific reference populations, for instance comparing all women in the study to a subset of 80–85 year old women, or women of similar age.

2.2. Assessment

We evaluate predictive performance of the FI using the area under the receiver operating characteristics curve (AUC) (Buitinck et al., 2013). We re-sample random halves of the population 100 times to estimate errors, while FI-GCP measures are cross validated as described in Stubbings et al. (2020). We present distributions using box and whisker plots with whiskers extending to the 99th percentiles. We exclude bins with less than 20 individuals. All analysis is available on GitHub (Stubbings, 2021); logistic regression is done using the statsmodels Python package (Seabold and Perktold, 2010).

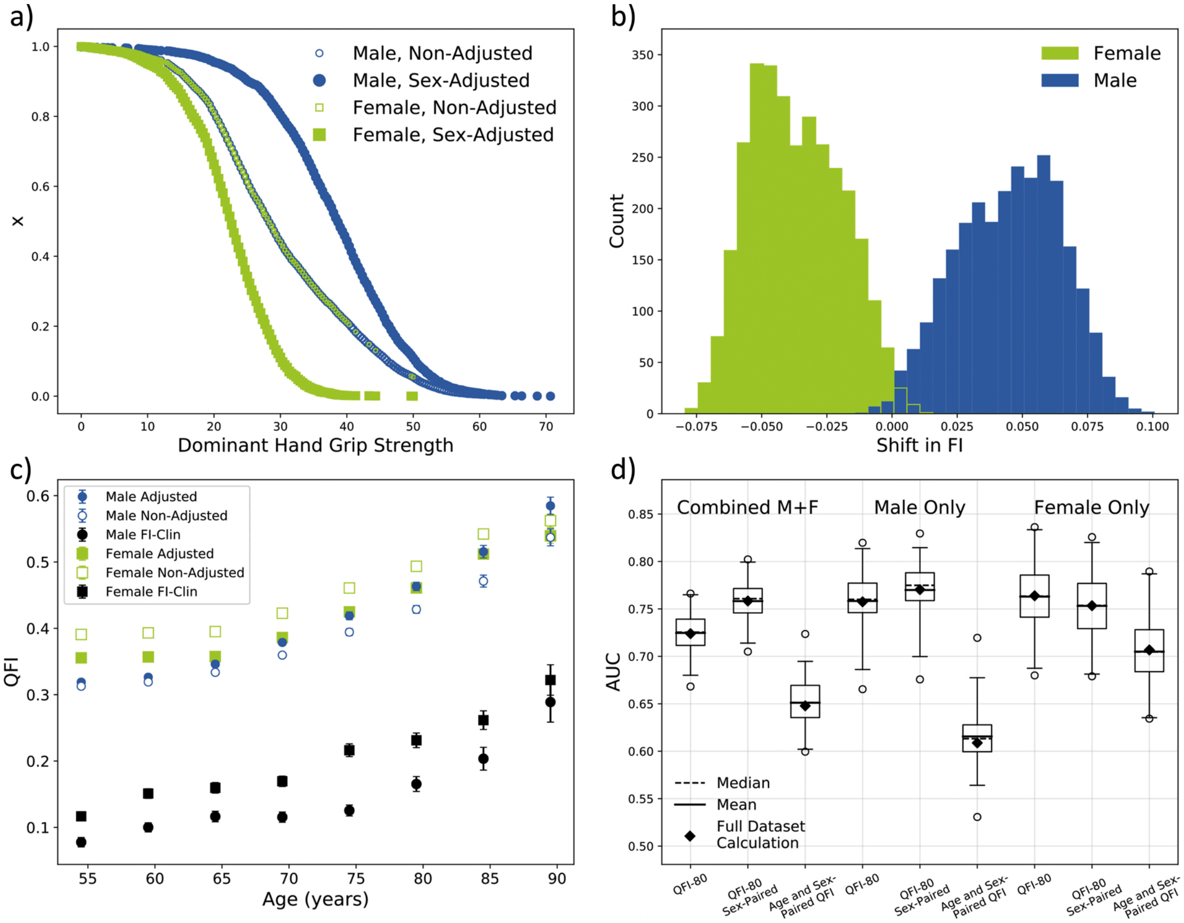

Fig. 7. The effects of using sex-specific reference populations on QFI-80 in wave 2 of the ELSA dataset. All plots show female in green and male in blue. (a) The risk quantiles for dominant hand grip strength with (filled points) and without (no-fill) using a sex-specific reference population. The non-adjusted male and female scores overlap since they are using the same reference population. (b) The difference between sex-adjusted QFI-80 and non-adjusted QFI-80 using all 80–85 year-olds as a reference population. Sex-adjusted QFI-80 uses only 80–85-year-olds of the respective sex as the reference population. (c) The average QFI for the sex-adjusted QFI (filled) and non-adjusted (no fill) binned by age in 5 year bins, the black points show the associated FI-Clin. (d) The AUC of various QFI with respect to mortality at 5 year follow up. We compare (from left to right) the QFI-80, sex-paired QFI-80, and age-and-sex-paired QFI for the combined, male, and female populations (from left to right). Box and whiskers are from resampling with half the population. The boxes represent the upper and lower quartiles, the whiskers go to the 99th and 1st percentiles, and the circles are remaining outliers. The short dashed line within each box is the median, the solid line the mean, and the diamond is the AUC for the full data set without resampling.

2.3. Data

We have explored the QFI with cross-sectional data from the National Health and Nutrition Examination Study (NHANES) (Centers for Disease Control, 2014) and the Canadian Study of Health and Aging (CSHA) (Canadian Study of Health, 1994). The NHANES data set consists of the 8881 individuals from the NHANES study with data for at least 11 of the 16 available biomarkers. This sample has an age range of 20–85 years. The data used from the CSHA study has 973 individuals aged 65+ for which data is available for at least 16 of the 22 biomarkers. We use the same NHANES and CSHA data examined previously with other frailty indices (Stubbings et al., 2020; Howlett et al., 2014; Blodgett et al., 2017). We also use data from the ELSA study (Oldfield et al., 2020), described in detail in Supplemental information. We focus on data from the second and fourth waves of the ELSA study, and examine predictive value on available data in subsequent waves.

2.4. Replication

All figures are replicated in waves 2 and 4 of the ELSA dataset, as well as in the NHANES and CSHA datasets when applicable. The exceptions being the figures where diagnosis data is used, since it is not available in the NHANES and CSHA datasets. Preferentially, we show data from the NHANES study and wave 2 of the ELSA data due to their larger sample sizes and number of mortality events. Replicated figures are available in the Supplemental material (Figs. S1–S15).

3. Results and discussion

3.1. Advantages of not dichotomizing

We compared the effects of transforming biomarker measurements into categorical variables with m categories and found that using more categories gives better prediction in the NHANES data-set (Fig. 2). Prediction can be improved by using 5 risk categories rather than just 2 categories (equivalent to binarizing at the median). Using more risk categories with fewer variables can result in better prediction than dichotomization with more variables. Here, using 5 or more risk categories on the 5 highest predicting biomarkers outperforms 2 risk categories used for all 17 biomarkers. In the NHANES data-set, quintiles perform as well as any finer grouping, which suggests that variation within 20% of the population does not have a significant effect on outcomes. However, using the QFI – with as many risk categories as is possible with the available data – does not negatively affect prediction and is more convenient than restricting everything to quintiles. In the other data-sets the plateau of prediction vs number of risk categories occurs in different places, but the QFI never under-performs a coarser grouping of risk. For the remainder of the paper we will use the QFI.

In Fig. 3 we show that the QFI performs modestly better than using a global cutpoint to binarize biomarkers based on risk quantile (Stubbings et al., 2020), and is also better than binarizing biomarkers using diagnostic thresholds (Blodgett et al., 2017). Interestingly, using the full data-set as a reference population performs the same as using an 80–85 year-old reference population. We find that the choice of fixed-age reference population does not significantly affect the prediction quality of the QFI.

3.2. QFI is interpretable

The detailed characteristics of the QFI are similar to other types of FI. We show the relationship between the QFI and the FI-Clin in wave 2 of the ELSA data in Fig. 4a. The relationship between QFI and FI-Clin is close to linear at larger values. Consistent with this, the aging trend of the QFI is very similar to that of FI-Clin – as shown in Fig. 4b. Nevertheless, the average QFI is significantly larger than FI-Clin at all ages. Very few individuals exhibit a QFI below 0.3.

These detailed characteristics of the QFI are dependent on the choice of reference population, though we have seen that predictive performance is not. We know that if most of the study is younger than 80 years old that selecting an 80 year old reference will make the bulk of the study appear relatively healthy. The effect of switching the reference population to an unhealthier group is an overall lowering of QFI scores in the population. As seen in Fig. 5, selecting an older cohort as the reference population leads to a general downwards shift in the distribution, and a slight positive skew. Using an older reference makes the distribution of the QFI look much more like a typical FI. However, to achieve a QFI of 0 an individual would have to be the healthiest individual compared to the population across every single biomarker measurement. Intrinsic variability and measurement noise make this unlikely even if there is someone in perfect health.

Although the QFI looks more like a standard FI with older reference populations, we do not think that aiming to look exactly like a standard FI is necessary. The QFI has a natural interpretation as being the average relative health with respect to the reference population.

We have also used ELSA data to test the association of the QFI with the various non-mortality outcomes available. For simplicity we use a reference population ages 80–85 in all cases where the QFI is calculated. The ELSA dataset has a list of reported diagnoses recorded at every wave (see supplemental information for details). We use the wave of first reported diagnosis to relate the QFI to these diagnoses in a number of ways. Firstly we look at the total number of accumulated diagnoses as it relates to the QFI in Fig. 4c. This figure shows a strong relationship between the QFI and the total number of diagnoses before this wave of the ELSA. Fig. 4d shows the proportion of individuals with one or more new diagnoses in the wave directly following the QFI assessment (1 year later). A higher QFI is associated with an increased probability of new diagnoses in the coming year on average. The difference in expected number of diagnoses almost doubles from a QFI of 0.3 to a QFI of 0.6.

3.3. Role of age within the QFI

Since we can define health with respect to a specific age group we can also remove the confounding effects of age from the QFI. We do this by calculating the QFI with respect to a group of individuals of the same age. In this age-paired QFI we group individuals into 5 year age bins and calculate the QFI using those binned individuals as the reference population.

We compare the predictive value of the age-paired QFI to the QFI with an 80 year old reference population in Fig. 6. We find that the QFI-80 substantially outperforms the age-paired QFI. However, we find that if we calculate the AUC as the average AUC across a set of age-binned QFI measurements (see Supplemental Figs. S11–S12) the QFI-80 performs similarly to the age-paired QFI. Furthermore, we find that if we add age back in to prediction using a logistic regression of both QFI and age – then both the QFI-80 and age-paired QFI perform similarly. As a benchmark for mortality prediction with biomarker data we have included the results of logistic regression on the raw biomarker measurements. We find that raw regression of the biomarker measurements performs better than the raw QFI-80. However, the biomarker measurements perform as badly as other FI measures when age controlled and only slightly better than other FI measures when combined with age.

3.4. Sex-specific reference populations

We also consider sex-specific reference populations, where quantile scores for each sex are calculated with respect to a reference cohort of only that sex (age restricted or not). In the ELSA data there is a large sex difference in the adjusted QFI scores due to the presence of grip strength measurements. Fig. 7a and b shows the difference in quantile scores for dominant hand grip strength and the resulting shift in QFI. Fig. 7c shows that controlling for sex in the reference population has the effect of narrowing the difference between male and female across the age range for ages below 90 years. Fig. 7d shows that the AUC for 5-year mortality predictions improve when controlled for sex. The age-paired QFI also saw improved prediction when matching sex (see Fig. S10a comparison).

4. Discussion and summary

We have shown that the dichotomization of continuous biomarker data into binary health deficits negatively affects the predictive value of the resulting FI-Lab (Fig. 3). Using categorical variables for deficit scores increases predictive value of the resulting quantile frailty index (QFI) when compared to dichotomization: increasing the number of risk categories improves the predictive value of the FI (Fig. 2). These results are replicated in the CSHA, NHANES and ELSA datasets.

The QFI allows us to easily explore the average relative risk with respect to the reference population across many biomarkers. For example, an individual with a QFI score of 0.3 is healthier than 70% of the reference population. Selecting an appropriate reference population can enhance the interpretability of the QFI: a QFI score of 0.6 with respect to a reference population of 80 year olds means that the individual is in worse health than 60% of 80 year olds. By using a common reference population, the relative health of individuals in different populations or subpopulations can be assessed. This could be useful when study populations are heterogeneous. For example, with mixtures of community dwelling and institutionalized individuals.

The QFI is intended to be convenient, interpretable, predictive, and scalable to large new data sets. We have used increasing age prevalence of biomarkers as a readily available risk metric that can be used to rank variables (Searle et al., 2008). Any other risk metric can be used instead, if it is available. Similarly, we have used reference populations extracted from and applied within specific studies. Using different studies to provide reference populations, or different reference populations for different biomarkers, could be easily implemented with our approach. For any such innovation, it will be important to examine predictive performance and robustness.

We have critically addressed the implicit inclusion of age in the QFI through the age-correlation of included health attributes. By using age-paired reference populations, or by considering predictive value of narrow ranges of age, we see that the predictive value of the QFI is strongly degraded (Fig. 6). Conversely, by combining the QFI with age explicitly within a logistic regression we see that predictive power is greatly enhanced. Including age explicitly in this way leads to approximately the same predictive power whether we use an age-paired QFI, a reference population on 80-year olds, or raw biomarker values. This indicates that the mortality-associated age-independent health information contained in the biomarkers comprising the QFI is retained in the QFI.

In clinical practice, any summary health measure for an individual will be available together with age – so both should be used for prognosis. A single summary health measure may also be desirable. By including age explicitly in assessing predictive power of the QFI, we can assess how much a single summary measure of health could be improved by constructing it with more age-associated components – either implicitly or explicitly. For the QFI this requires a fixed-age reference population.

If we prefer a summary measure of health that excludes age, then we need to show that the age-averaged predictive quality agrees with the overall quality. For the QFI, we can construct this age-excluded measure using age-paired reference populations.

If we want to compare the predictive power of two summary measures of health through, e.g., the AUC of an ROC, it is clear that any differences in the implicit inclusion of age will dominate the comparison. Age should be either explicitly added or explicitly controlled for in such a comparison – ideally both.

A similar discussion of age-dominated composite measures exists in the biological age literature (Levine, 2020). When chronological age was controlled for, early epigenetic clocks lost many of their significant associations with health outcomes (Ryan et al., 2019). Later epigenetic clocks addressed this issue by including biomarkers associated with adverse health outcomes independently of age (Levine et al., 2018).

The QFI can also non-parametrically control for sex. Using sex-dependent reference populations ensures that the male and female individuals are treated the same. In Fig. 7c we found a crossing of male and female QFI as a function of age, so that males have a higher average QFI at later ages (above 85 years). This does not exhibit a mortality-morbidity paradox, since male mortality is somewhat higher than female at higher ages. Accordingly, we see slightly improved AUC for the sex-adjusted QFI. While there are real biological differences between male and female aging populations (Gordon and Hubbard, 2018), our finding raises the intriguing possibility that the mortality–morbidity paradox could be significantly reduced with proper control populations. This is worth further study with different aging populations using agnostic approaches to controlling for sex, such as the QFI.

Individual health is high-dimensional. There are a vast number of individual characteristics of good or poor health. In contrast, populations are often described by only a few characteristics such as just age and sex. Nevertheless, it is important to condition individual health on comparable populations. For binarized health variables this can be done with population-dependent cutpoints (Blodgett et al., 2017; Howlett et al., 2014; McPherson, 2017). For the QFI, we can explicitly choose the reference population. In this paper we have explored the role of age and sex, but any demographic differences in health and aging can be addressed with our approach. Furthermore, biases in the reference population due to selection or composition effects can be interrogated directly, since the reference population is explicitly defined.

We have mostly used a fixed-age reference population of 80–85 year olds. This leads to a natural interpretability of the resulting QFI. We do not suggest that this is the only reference population that should be used, but it is well represented in many studies and the resulting QFI is highly interpretable. Even in cases where there is not a wide range of ages to choose from the QFI will still be effective as a predictive measure.

By varying the fixed-age reference, we see in Fig. 5 that commonly reported maxima and minima of the FI (0.7 and 0.0, respectively) (Searle et al., 2008; Mitnitski et al., 2015) appear to be approached as we increase the age of the reference populations towards super-centenarians. While that would gives an appealing interpretation of the QFI as your health quantile with respect to the "very old", we do not yet have a large enough sample of the very old to explore that limit.

We have developed a summary health measure, the quantile frailty index (QFI), from continuous biomarker or health measurements without any dichotomization. The QFI is both predictive of mortality, and interpretable as a frailty index. Different reference populations can be easily used to construct the QFI. We have investigated the role of age in the QFI, and demonstrate that the QFI effectively includes the non-age related aspects of considered biomarkers. The QFI can control for other important population state variables with appropriate reference populations.

Acknowledgements

ADR thanks the Natural Sciences and Engineering Research Council (NSERC) for an operating Grant (RGPIN 2019-05888). KR has operational funding from the Canadian Institutes of Health Research (PJT-156114) and personal support from the Dalhousie Medical Research Foundation as the Kathryn Allen Weldon Professor of Alzheimer Research.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.mad.2021.111570.

References

Aguayo, G.A., Vaillant, M.T., Donneau, A.F., Schritz, A., Stranges, S., Malisoux, L., Chioti, A., Guillaume, M., Muller, M., Witte, D.R., 2018. Comparative analysis of the association between 35 frailty scores and cardiovascular events, cancer, and total mortality in an elderly general population in England: an observational study. PLoS Med. 15 (3), e1002, 543.

Altman, D.G., Royston, P., 2006. The cost of dichotomising continuous variables. BMJ (Clin. Res. Ed.) 332 (7549), 1080.

Belsky, D.W., Moffitt, T.E., Cohen, A.A., Corcoran, D.L., Levine, M.E., Prinz, J.A., Schaefer, J., Sugden, K., Williams, B., Poulton, R., Caspi, A., 2018. Eleven telomere, epigenetic clock, and biomarker-composite quantifications of biological aging: do they measure the same thing? Am. J. Epidemiol. 187 (6), 1220–1230.

Blodgett, J.M., Theou, O., Howlett, S.E., Rockwood, K., 2017. A frailty index from common clinical and laboratory tests predicts increased risk of death across the life course. GeroScience 39 (4), 447–455.

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., VanderPlas, J., Joly, A., Holt, B., Varoquaux, G., 2013. API design for machine learning software: experiences from the scikit-learn project. ECML PKDD Workshop: Languages for Data Mining and Machine Learning 108–122.

Canadian Study of Health and Aging Working Group, 1994. Canadian study of health and aging: study methods and prevalence of dementia. Can. Med. Assoc. J. 150 (6), 899.

Centers for Disease Control and Prevention National Center for Health Statistics (Updated 2014). National health and nutrition examination survey data. http://www.cdc.gov/nchs/nhanes.htm.

Cohen, J., 1983. The cost of dichotomization. Appl. Psychol. Meas. 7 (3), 249–253.

Dawson, N.V., Weiss, R., 2012. Dichotomizing continuous variables in statistical analysis: a practice to avoid. Med. Decis. Making 32 (2), 225–226.

Fedorov, V., Mannino, F., Zhang, R., 2009. Consequences of dichotomization. Pharm. Stat. 8 (1), 50–61.

Ferrucci, L., Levine, M.E., Kuo, P.L., Simonsick, E.M., 2018. Time and the metrics of aging. Circ. Res. 123 (7), 740–744.

Fried, L.P., Tangen, C.M., Walston, J., Newman, A.B., Hirsch, C., Gottdiener, J., Seeman, T., Tracy, R., Kop, W.J., Burke, G., McBurnie, M.A., 2001. Frailty in older adults: evidence for a phenotype. J. Gerontol. Ser. A: Biol. Sci. Med. Sci. 56 (3), M146–M157.

Gordon, E., Hubbard, R., 2018. Physiological basis for sex differences in frailty. Curr. Opin. Physiol. 6, 10–15.

Horvath, S., 2013. DNA methylation age of human tissues and cell types. Genome Biol. 14, R115.

Howlett, S.E., Rockwood, M., Mitnitski, A., Rockwood, K., 2014. Standard laboratory tests to identify older adults at increased risk of death. BMC Med. 12 (1).

Jazwinski, S.M., Kim, S., 2019. Examination of the dimensions of biological age. Front. Genet. 10, 263.

Juster, R.P., McEwen, B.S., Lupien, S.J., 2010. Allostatic load biomarkers of chronic stress and impact on health and cognition. Neurosci. Biobehav. Rev. 35 (1), 2–16.

Karczewski, K.J., Snyder, M.P., 2018. Integrative omics for health and disease. Nat. Rev. Genet. 19 (5), 299–310.

Kulminski, A.M., Culminskaya, I.V., Ukraintseva, S.V., Arbeev, K.G., Land, K.C., Yashin, A.I., 2008. Sex-specific health deterioration and mortality: the morbidity–mortality paradox over age and time. Exp. Gerontol. 43 (12), 1052–1057.

Levine, M.E., 2020. Assessment of epigenetic clocks as biomarkers of aging in basic and population research. J. Gerontol.: Ser. A 75 (3), 463–465.

Levine, M.E., Lu, A.T., Quach, A., Chen, B.H., Assimes, T.L., Bandinelli, S., Hou, L., Baccarelli, A.A., Stewart, J.D., Li, Y., Whitsel, E.A., Wilson, J.G., Reiner, A.P., Aviv, A., Lohman, K., Liu, Y., Ferrucci, L., Horvath, S., 2018. An epigenetic biomarker of aging for lifespan and healthspan. Aging 10 (4), 573–591.

Li, X., Ploner, A., Wang, Y., Magnusson, P.K., Reynolds, C., Finkel, D., Pedersen, N.L., Jylhävä, J., Hägg, S., 2020. Longitudinal trajectories, correlations and mortality associations of nine biological ages across 20-years follow-up. eLife 9, 132.

McPherson, R., 2017. Henry's Clinical Diagnosis and Management by Laboratory Methods. Elsevier, St. Louis, Mo.

Mitnitski, A., Collerton, J., Martin-Ruiz, C., Jagger, C., von Zglinicki, T., Rockwood, K., Kirkwood, T.B.L., 2015. Age-related frailty and its association with biological markers of ageing. BMC Med. 13 (1).

Mitnitski, A., Howlett, S.E., Rockwood, K., 2017. Heterogeneity of human aging and its assessment. J. Gerontol. Ser. A 72 (7), 877–884.

Mitnitski, A.B., Mogilner, A.J., Rockwood, K., 2001. Accumulation of deficits as a proxy measure of aging. Sci. World 1, 323–336.

Naggara, O., Raymond, J., Guilbert, F., Roy, D., Weill, A., Altman, D.G., 2011. Analysis by categorizing or dichotomizing continuous variables is inadvisable: an example from the natural history of unruptured aneurysms. Am. J. Neuroradiol. 32 (3), 437–440.

Nicklett, E.J., 2011. Socioeconomic status and race/ethnicity independently predict health decline among older diabetics. BMC Public Health 11 (1).

Oldfield, Z., Rogers, N., Phelps, A., Blake, M., Steptoe, A., Oskala, A., Marmot, M., Clemens, S., Nazroo, J., Banks, J., 2020. English Longitudinal Study of Ageing: Waves 0–9, 1998–2019.

Peña, F.G., Theou, O., Wallace, L., Brothers, T.D., Gill, T.M., Gahbauer, E.A., Kirkland, S., Mitnitski, A., Rockwood, K., 2014. Comparison of alternate scoring of variables on the performance of the frailty index. BMC Geriatr. 14 (1), 25.

Rockwood, K., Mogilner, A., Mitnitski, A., 2004. Changes with age in the distribution of a frailty index. Mech. Ageing Dev. 125 (7), 517–519.

Royston, P., Altman, D.G., Sauerbrei, W., 2006. Dichotomizing continuous predictors in multiple regression: a bad idea. Stat. Med. 25 (1), 127–141.

Ryan, J., Wrigglesworth, J., Loong, J., Fransquet, P.D., Woods, R.L., 2019. A systematic review and meta-analysis of environmental, lifestyle, and health factors associated with DNA methylation age. J. Gerontol.: Ser. A 75 (3), 481–494.

Seabold, S., Perktold, J., 2010. Statsmodels: econometric and statistical modeling with Python. 9th Python in Science Conference.

Searle, S.D., Mitnitski, A., Gahbauer, E.A., Gill, T.M., Rockwood, K., 2008. A standard procedure for creating a frailty index. BMC Geriatr. 8 (1).

Shi, S.M., McCarthy, E.P., Mitchell, S.L., Kim, D.H., 2020. Predicting mortality and adverse outcomes: comparing the frailty index to general prognostic indices. J. Gen. Internal Med. 60 (10), 1–7.

Stubbings, G., 2021. Quantile Frailty Index https://github.com/GarrettStubbings/QuantileFrailtyIndex.

Stubbings, G., Farrell, S., Mitnitski, A., Rockwood, K., Rutenberg, A., 2020. Informative frailty indices from binarized biomarkers. Biogerontology 21 (3), 345–355.

Tsodikov, A., Szabo, A., Jones, D., 2002. Adjustments and measures of differential expression for microarray data. Bioinformatics 18 (2), 251–260.

Zucchelli, A., Vetrano, D.L., Grande, G., Calderón-Larrañaga, A., Fratiglioni, L., Marengoni, A., Rizzuto, D., 2019. Comparing the prognostic value of geriatric health indicators: a population-based study. BMC Med. 17 (1).