RAND EUROPE

Discovery

VITALITY AGE WITH PRE-EXISTING CONDITIONS

Martin Stepanek, Cloé Gendronneau, Christian Van Stolk,
Francois Millard, Howard Bolnick

For more information on this publication, visit www.rand.org/t/RR2484z1

Published by the RAND Corporation, Santa Monica, Calif., and Cambridge, UK

© Copyright 2019 Discovery Group

RAND® is a registered trademark.

RAND Europe is a not-for-profit organisation whose mission is to help improve policy and decision making through research and analysis. RAND's publications do not necessarily reflect the opinions of its research clients and sponsors.

All rights reserved. No part of this book may be reproduced in any form by any electronic or mechanical means (including photocopying, recording, or information storage and retrieval) without permission in writing from the client.

Support RAND

Make a tax-deductible charitable contribution at

www.rand.org/giving/contribute

www.rand.org

www.rand.org/randeurope

Table of contents

Figures

Tables

Abbreviations

BD	Baseline Death
DCCPS	Division of Cancer Control and Population Sciences
GBD	Global Burden of Disease
IHD	Ischaemic Heart Disease
IHME	Institute for Health Metrics and Evaluation
IS	Ischaemic Stroke
NCI	National Cancer Institute
PAF	Population Attributable Fraction
RR	Relative Risk
SEER	Surveillance, Epidemiology, and End Results Program
SRP	Surveillance Research Program
VA.3	Vitality Age
VAPC	Vitality Age with pre-existing conditions

1. Introduction

The Vitality Age with pre-existing conditions (VAPC) is an extension to the newly-updated Vitality Age model (VA.3), allowing explicit differentiation between individuals with one or more selected pre-existing conditions that negatively affect their health and life expectancy. VA.3 is a risk-adjusted health assessment tool, giving individuals a snapshot of their overall health based on lifestyle choices – including diet, alcohol consumption and exercise habits – and clinical factors such as blood pressure, cholesterol or body mass index. This is done by calculating a bespoke life expectancy estimate using a pre-defined set of the individual's risk factor measures and comparing it to the life expectancy of an average person with the same sex, age, and country characteristics. The Vitality Age (VA) is then the actual chronological age adjusted by the difference in life expectancy; it estimates an individual's remaining years of life and produces an adjusted age estimate so that the individual's life expectancy remains unchanged.

The VA.3 concept relies on adjustment of population-level mortality rates, reflecting how one's risk factor profile compares to that of the general population. However, while the population mortality rates used in the calculator reflect a certain share of the population considered as having a variety of pre-existing conditions, such as haemophilia or cancer, the negative effects of such conditions are diluted in the aggregate statistics in proportion to the ratio of negative effects to prevalence. In other words, if a certain condition with 0.1 per cent prevalence in the population decreases an individual's life expectancy by 10 years, the population-wide aggregate effect will be negligible, although the few affected individuals are expected to have a substantially shorter life on average. The Vitality Age assessment will thus be biased for any such individuals with pre-existing conditions. The exception was made for diabetes, however, which is embedded as a possible pre-existing condition in the VA.3 model, and diabetic patients are assigned adjusted mortality rates for this condition.

In this report, we look at the feasibility of including more pre-existing conditions in the algorithm to calculate the VAPC. We have done this for a number of conditions (see Section 3) but the calculation could be extended further. It was not our intention to build a long list of pre-existing conditions into a model, but to test for a certain set of conditions whether extending the Vitality Age calculation makes sense and is feasible. Some restrictions were placed upon us in including conditions for which the necessary data was available and with reasonable prevalence. More is explained in Section 2.

Care needs to be given as to how feedback is provided to individuals with pre-existing conditions. In some cases, the pre-existing conditions may have a very significant impact on mortality. Being confronted with this through a calculation may cause distress, especially if users do not receive appropriate professional support. In the first instance, this tool is aimed at health professionals and professionals in other sectors, for instance insurance sectors.

Section 2 describes the methodology and Section 3 presents the sources of data used in the model and their processing. The questions completing the VA.3 questionnaires are laid out in Section 4. Finally, Section 5 discusses the limitations of the model.

2. Methodology

The proposed methodology for the VAPC model is largely based on the recent update of the VA.3 model undertaken by RAND Europe (Stepanek et al., 2018). In short, VA.3 works with elementary risk factor-cause of death pairs and the associated relative risks (RR) and mortality rates to compute an individual's all-cause mortality rate. That is, the algorithm looks separately at each risk factor and each cause of death in turn, determines whether and how they are connected, and the extent to which the individual's risk factor profile affects the estimated mortality rates associated with the given cause of death. This is done through the following steps.

1. First, the **population cause-specific mortality rates** (MR) for a given sex, age and country are taken as a starting point. Roughly speaking, these are applicable to an individual with the population average risk-factor profile.

2. Second, **counterfactual baseline mortality rates** (BD) for an individual with the same sex, age and country characteristics, but no history of exposure to any risk factor considered are calculated from MR using population-attributable fractions (PAFs). PAF is the proportional reduction in population mortality that would occur if exposure to a given risk factor was reduced to its theoretical minimum. For instance, lung cancer mortality lowered by the respective smoking-related PAF would represent the mortality rate in an equivalent population where no one has ever smoked. Formally, the baseline mortality rate for outcome o, age a, sex s, country c and time t is given by:

$$BD_{oasct} = MR_{oasct} \times (1 - PAF_{Joasct}), \tag{1}$$

   where MR_{oasct} is the population mortality rate, PAF_{Joasct} is the PAF for the set $J$of all risk factors j relevant for outcome o, for age a, sex s, country c and time t. PAF_{Joasct} may thus be a joint statistic if the set J contains more than one element (see Stepanek et al., 2018, for more details).

3. Third, the **cause-specific mortality rates (MR_{oasct}) are converted into probabilities of dying** within twelve months (q_{oasct}). Following Lim et al. (2015), this is done using a standard life table calculation:

$$q_{oasct} = \frac{n \times MR_{oasct}}{1 + (n - {}_n a_x) \times MR_{oasct}} \tag{2}$$

where n is length of the assumed period in years (hence $n = 1$ as we are interested in probability of dying within 12 months) and $_n\, a_x$ is the distribution of deaths in the interval [0;1], assumed to be 0.5 (a uniform distribution). The process is replicated for both the population mortality rates, resulting in population-level probability of dying q^p, and for the baseline mortality rates (BD), resulting in the baseline probability of dying q^b:

$$q^b_{oasct} = \frac{n \times BD_{oasct}}{1 + (n - {}_n a_x) \times BD_{oasct}} \tag{3}$$

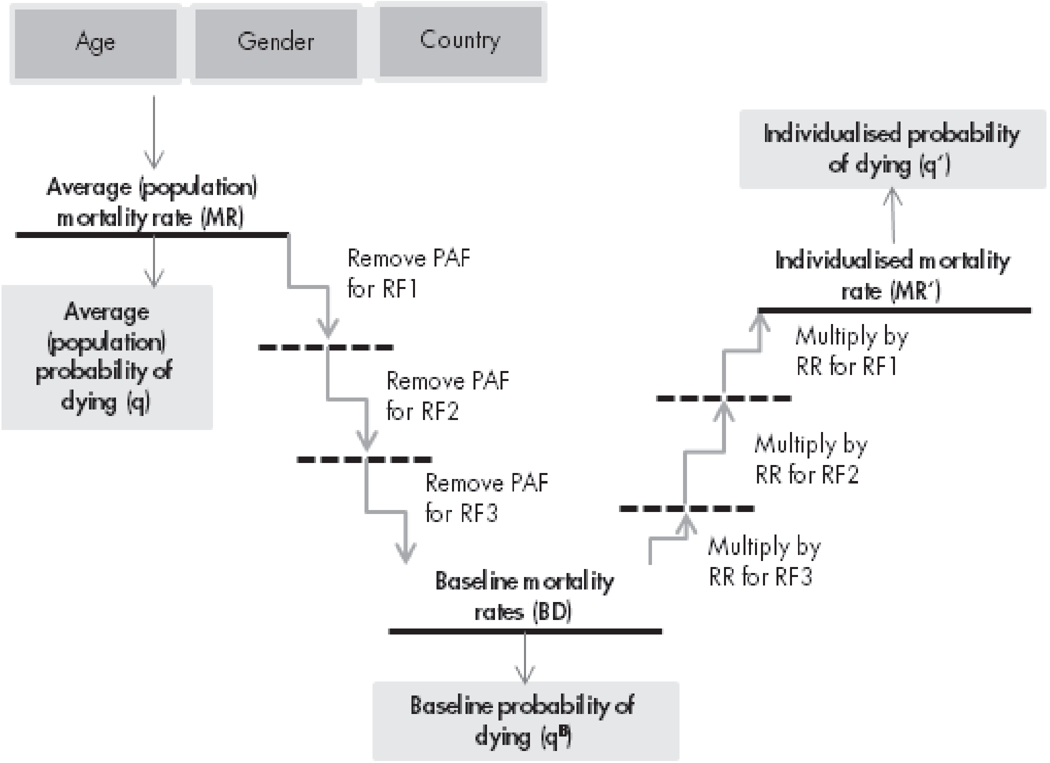
4. The baseline, cause-specific probability of dying q^b can then be **transformed into an individualised probability of dying** q' by multiplying it by the cause-specific joint relative risk (RR) reflecting an individual's risk factor profile:

$$q'_{\mathcal{P}oasct} = RR'_{\mathcal{P}oasct} \times q^b_{oasct}, \tag{4}$$

   where $RR'_{\mathcal{P}oasct}$ is the individualised RR for outcome o, age a, sex s, country c, time t, and the individual's risk factor profile $\mathcal{P} \in [0, \infty)^{\mathcal{R}}$, a vector of risk factor exposures from the risk factor space $\mathcal{R}$, which contains all the risk factors considered in the model. That is, exposure to each risk factor is characterised by a single non-negative number, the risk factor multiplier, which together form the vector $\mathcal{P}$. The risk factor multipliers are then combined with the baseline individual RR's to create a single measure $RR'_{\mathcal{P}oasct}$ that defines the ratio of mortality rate risk compared to a baseline population (see Stepanek et al., 2018, for further details).

The entire process is repeated for each year from an individual's actual age onwards until an arbitrary maximum age set to 120 in our model, providing detailed overview of the mortality risks at any stage in life. Figure 1 summarises the steps taken.

Figure 1: Determination of individualised probability of dying from a single cause of death


Age
Gender
Country
Average (population) mortality rate (MR)
Average (population) probability of dying (q)
Remove PAF for RF1
Remove PAF for RF2
Remove PAF for RF3
Baseline mortality rates (BD)
Baseline probability of dying (q^B)
Multiply by RR for RF3
Multiply by RR for RF2
Multiply by RR for RF1
Individualised mortality rate (MR')
Individualised probability of dying (q')

Source: Stepanek et al. (2018).

Table 1 shows probabilities of dying within 12 months for a hypothetical male, with an average risk profile at different ages for the selected causes of deaths further discussed in this report. Due to a limitation of the dataset used (see Section 3), the rates are only available in 5-year age groups; the model subsequently interpolates between the estimates to obtain single year-specific estimates that are used in the calculation (see Stepanek et al., 2018, for details).

Table 1: Selected sample cause-specific individualised probabilities of dying within 12 months

		Age					
Cause of death	Sex	40-44	45-49	50-54	55-59	60-64	65-69
Breast cancer	Male	0.000%	0.000%	0.000%	0.001%	0.001%	0.002%
Colon and rectum cancer	Male	0.004%	0.007%	0.014%	0.028%	0.047%	0.075%
Ischaemic heart disease	Male	0.021%	0.041%	0.074%	0.119%	0.186%	0.285%
Ischaemic stroke	Male	0.001%	0.002%	0.003%	0.005%	0.014%	0.027%
Oesophageal cancer	Male	0.003%	0.006%	0.012%	0.022%	0.036%	0.052%
Prostate cancer	Male	0.000%	0.001%	0.003%	0.010%	0.027%	0.059%

We can see that the probabilities of dying for an individual with an average risk-factor profile are rather low, well below one per cent for each selected cause of death. This is because the model implicitly works with the probability of *developing* a condition in the first place (which may be very low, especially at young ages) and only then assumes a certain probability of dying from the condition. However, for individuals who already have a certain condition, this methodology does not apply, as the probability of developing a condition is 100 per cent. Indeed, the assumption that the general population mortality rates can be used as a baseline for the calculations presented above no longer holds; instead, one needs to work with mortality rates of an equivalent population of individuals with such a condition.

In order for the VAPC model to take pre-existing conditions into account, it is possible to disregard, where appropriate, the population cause-specific mortality rates MR_{oasct} originally used (for the specific pre-existing conditions), as well as the entire process of their adjustment, and replace the resulting $q'_{\mathcal{P}oasct}$ with customised probabilities of dying obtained from an alternative data source. Specifically, a customised probability of dying for individuals with pre-existing conditions, $q''_{\mathcal{P}oasct}$, can be obtained from epidemiological studies analysing the mortality rates of individuals with given conditions, while controlling for other confounding factors.

In particular, consider an individual with an early stage pancreatic cancer. Clearly, such an individual will face a higher probability of dying from pancreatic cancer than an average person with the same characteristics. Utilising additivity of mortality rates (and subsequently using an appropriate formula, probabilities of dying), we can estimate the life expectancy of such an individual in a world in which pancreatic cancer does not exist at all (and the all-cause probability of death is thus lower than in reality). We then add the customised probability of dying due to pancreatic cancer at the end of Step 3 above to reflect that not only pancreatic cancer exists, but the individual already has it. Having made this adjustment, the VA calculation can continue as per the VA.3 methodology, as described in Figure 2 below, and set out in steps 5 and 6 hereafter.

5. The cause-specific probabilities of dying can be aggregated to an all-cause age- and sex-specific probability of dying using the standard competing risk model:

$$Q_{\mathcal{P}asct} = 1 - \prod_{o}^{\mathcal{M}}(1 - q'_{\mathcal{P}oasct}), \tag{5}$$

   where $\mathcal{M}$ refers to the set of all causes of deaths considered (i.e. those included in the Global Burden of Disease (GBD) database – see Section 3).

6. Finally, we can calculate the probability of surviving up to age i and dying exactly at age i and multiply it by i to compute $t_{i,oasct}$, using

$$t_{i,asct} = (1 - Q_a) \times (1 - Q_{a+1}) \times (1 - Q_{a+2}) \times \ldots \times (1 - Q_{i-1}) \times Q_i \times i, \tag{6}$$

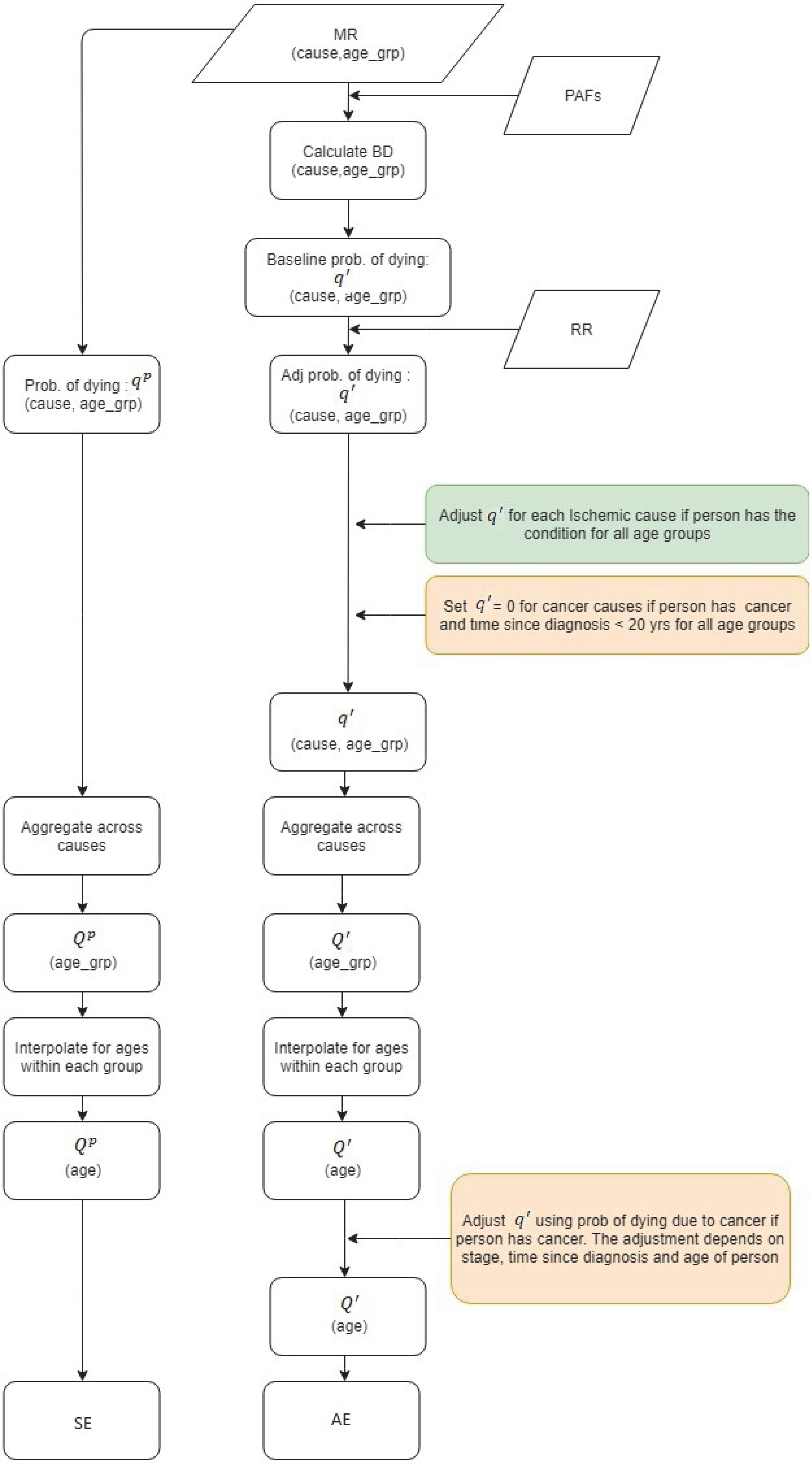
   where Q_a stands for Q_{asct} for simplicity, a denotes an individual's actual age (as of their last birthday), and $i \in \{a, \ldots, 120\}$ denotes an age between an individual's actual age a and an arbitrary maximum age, set to 120 in the model. The factors $t_{i,asct}$ and $t'_{i,asct}$ calculated using the population and individualised probabilities of dying, respectively, are used to calculate the estimated life expectancies for an average (SE_{asct}) and the assessed individual (AE_{asct}) (see Stepanek et al., 2018, for details), which in turn form the Vitality Age estimate:

$$VA = a + SE_{asct} - AE_{asct}, \tag{7}$$

Note that replacing only the estimated individualised probability of dying from a given cause with a customised value disregards the fact that having a pre-existing condition may influence the probability of dying from other causes as well (e.g. due to severity of treatment or generally poorer health), potentially underestimating the actual value. In the model, we mitigate this issue by considering deaths from any cause in the affected population (rather than the disease itself) and comparing them to the equivalent all-cause deaths in the general population. This **relative survival rate** can be calculated by dividing the proportion of patients with a disease who are still alive at the end of the period of time (e.g. one year) by the proportion of people with similar characteristics in the general population who are alive at the end of the same time period. This way, the customised probability of dying q'' combines the effect of a pre-existing condition on probabilities of dying due to other causes of death as well.

The main issue with this relatively simple adjustment is the lack of appropriate data in the database used in the VA.3 model (see below). The data, therefore, need to be supplied from other sources, specifically from epidemiological studies. In addition, because every pre-existing condition needs to be treated separately and the framework of effects can be very complex – depending on type and stage of the condition – only a subset of all existing long-term conditions can be considered.

Figure 2: Flow chart of the Vitality Age with pre-existing conditions algorithm


MR
(cause,age_grp)
PAFs
Calculate BD
(cause,age_grp)
Baseline prob. of dying: q'
(cause, age_grp)
RR
Prob. of dying : q^p
(cause, age_grp)
Adj prob. of dying : q'
(cause, age_grp)
Adjust q' for each Ischemic cause if person has the condition for all age groups
Set q'= 0 for cancer causes if person has cancer and time since diagnosis < 20 yrs for all age groups
q'
(cause, age_grp)
Aggregate across causes
Aggregate across causes
Q^p
(age_grp)
Q'
(age_grp)
Interpolate for ages within each group
Interpolate for ages within each group
Q^p
(age)
Q'
(age)
Adjust q' using prob of dying due to cancer if person has cancer. The adjustment depends on stage, time since diagnosis and age of person
Q'
(age)
SE
AE

3. Data

Analogously to VA.3, the VAPC calculator works with a dataset obtained from the Global Burden of Disease (GBD) database.[1] Led by the Institute for Health Metrics and Evaluation (IHME) at the University of Washington, the GBD database is a result of a project involving over 1,000 researchers from more than 100 countries, including 26 low- and middle-income countries. The researchers collect all appropriate published epidemiological studies and data sources, such as hospital data, disease registry data or censuses, and combine the information using a Bayesian meta-regression framework. The GBD database provides a consistent and comparative quantification of more than 300 diseases and injuries in nearly 200 countries, by age and sex, from 1990 to the present day, and is updated regularly. The GBD data are widely used in scientific literature and have the advantage of not being associated with a single cohort of individuals or methodology, but rather are a product of a multitude of studies from countries all over the world. The database contains data on MRs, PAFs and RRs, as well as supporting documentation on how the risk factor multipliers may be calculated. The data are available in 5-year age groups (up to a '95+ years' category) for both sexes.

Additionally to the GBD data used in VA.3, we conducted a Google search to identify epidemiological studies providing empirical estimates of mortality rates and/or probabilities of dying for a set of selected pre-existing conditions. Given time and resource constraints, only five conditions resulting in six causes of death (in addition to diabetes, already embedded in VA.3) were included in the study; additional conditions can be added to the model by following the methodology described in this report. However, note that one must be able to break down any additional conditions to a mutually exclusive set of stages or categories that would be identifiable by the assessed individuals (e.g. a localised vs regional stage of a cancer – see below) in order to properly differentiate between the respective probabilities of death.

This section presents the selection process for the five conditions and data used to estimate the VAPC model.

3.1. Selection of pre-existing conditions

Sex, age and location are not sufficient to predict the mortality rate of a specific condition. Often, the stage or type of disease, the treatment used and the time since diagnosis are important determinants of the probability of dying, particularly for aggressive diseases such as cancers. This data, however, is generally not readily available for more than a few diseases from a single database and must be extracted for each

[1] http://www.healthdata.org/gbd/data (As of 4 February 2019).

condition considered from different sources. Therefore, the number of conditions taken into consideration is constrained. Four criteria were used to select the set of conditions to be included in the model.

The first criterion is that the condition must cause a significant excess mortality rate in the affected population. The main point is that this excess mortality has to affect the mortality risk of an individual in a material way. If not, it may not make sense to include pre-existing conditions in the Vitality Age. This criterion ensures that it is necessary to adapt the algorithm for persons affected by such diseases. For instance, individuals with cancer, although there may not be many of them, face, on average, substantially different probability of dying than those who do not have cancer. The mortality rate in the affected population was estimated using the GBD data, dividing mortality rate of a given condition (for the general population) by its prevalence in the general population.[2]

The second criterion is the prevalence of the condition in the general population, which directly relates to the number in the population who will benefit from the algorithm update. For instance, certain genetic disorders may have a severe impact on the affected individuals (and thus satisfy the first criterion), but may be very rare. Data on prevalence is also available in the GBD, as indicated above. In some ways, including rare pre-existing conditions could be useful to assess the mortality risk of a specific population, and could be built into additional versions of this model. However, in the first instance, we were interested in more prevalent conditions.

Building on the second criterion, the ramifications of certain diseases were also of interest. This does not only relate to the costs of certain conditions on the health system, but also the delivery of wider support to specific groups who have certain conditions in our societies, be it support (e.g. healthcare), pension and financial advice, and even insurance. Although this was not the most important criterion, we also note the complexity of providing adequate and tailored support and services to a group at higher mortality risk.

Finally, the last criterion is the availability of the data for a given condition and is, in fact, necessary in order to be able to include the condition in the algorithm. As this can only be determined through a thorough literature review, the first two criteria were used to narrow down the list of conditions under consideration during the literature review.

Based on these four criteria, the following conditions were chosen for inclusion in the model[3]:

- Breast cancer
- Colon and Rectum cancer
- Oesophageal cancer
- Prostate cancer
- Ischaemic heart diseases and stroke (ischaemia).

[2] The estimates were done across all ages 18+. The UK was used as a country of reference throughout the project.

[3] Out of a set of 12 conditions: Ischemic heart disease; Ischemic stroke; Haemorrhagic stroke; Chronic obstructive pulmonary disease; Breast cancer; Colon and rectum cancer; Pancreatic cancer; Oesophageal cancer; Prostate cancer; Liver cirrhosis; Dementia.

3.2. Cancers

As we discuss in the following subsections, the methodological adjustments in the VAPC model are very similar for all cancers considered, yet substantially different from ischaemic heart disease and ischaemic stroke. We, therefore, present them in two separate sections, accordingly.

3.2.1. Sources of data

A short review of the literature allowed the main factors affecting the prognosis of the cancers under study to be identified. The most important are: sex, age of patient at diagnosis and time since diagnosis, stage of cancer, treatment method and lifestyle (substance or food intake as well as the level of physical activity) (National Cancer Institute, 2018; Eker et al., 2015; Zare-Bandamiri et al., 2016; Giovannucci, 2007).[4] Although mortality data is not available for each combination of these factors, the Surveillance, Epidemiology, and End Results Program (SEER)[5] of the National Cancer Institute in the United States provides information on the first four (sex, age, stage and treatment method). SEER is supported by the Surveillance Research Program (SRP) in the National Cancer Institute's Division of Cancer Control and Population Sciences (DCCPS). The SRP provides national leadership in the science of cancer surveillance, as well as analytical tools and methodological expertise in collecting, analysing, interpreting and disseminating reliable population-based statistics. Although the data is collected in the United States, and this model is mainly aimed at the United Kingdom (UK) population, we consider the two populations and countries sufficiently similar in cancer treatment and survival rates,[6] allowing us to overcome the data limitation problem by using the SEER database.

SEER offers access to several databases, containing statistics for different timespan and geographic coverage in the United States. The available statistics from SEER include incidence, mortality, survival, stage, prevalence, and lifetime risk. For the VAPC model we use the SEER 18 (2007 –2013) database to obtain data on relative survival by stage, which provides a detailed breakdown of survival rates by age, and the SEER 9 (2000–2013) database, which covers longer periods of time since diagnosis.

As discussed in Section 2, relative all-cause mortality rates (or relative survival rates), as opposed to cause-specific rates, must be used in the model to reflect that the pre-existing conditions may have indirect effects on probability of dying from other causes (e.g. through a weakened immune system). The SEER databases contain relative all-cause survival rates, calculated using expected life tables as substitutes for a cohort of cancer-free individuals (see Cho et al., 2011, for details). In essence, the estimates compare survival of people with cancer to similar cancer-free individuals over a period of time, estimating whether the disease shortens life.[7] In the following subsection we describe the two datasets in detail, including their

[4] https://www.cancer.gov/about-cancer/diagnosis-staging/prognosis (As of 4 February 2019).

[5] The online database is accessible at http://seer.cancer.gov (As of 4 February 2019).

[6] We accept that there may be differences between geographies, but our assumption is that the UK and the United States will show similar trends on all major cancer-related indicators. As we are testing the feasibility of this approach, we are not too worried about not having UK data. Such data can always be added when it becomes available.

[7] See also https://www.cancer.gov/publications/dictionaries/cancer-terms/def/relative-survival-rate for more information on the relative survival rates (As of 4 February 2019).

format and where data can be downloaded from, and explain how all data were merged together, creating a consistent dataset used in the calculation.

3.2.2. The SEER 9 dataset – Relative survival rates by stage of cancer

This first dataset provides detailed relative cumulative survival rate estimates for each of the four types of cancer considered, broken down by stage of cancer, sex, age category at diagnosis and time since diagnosis (up to ten years after diagnosis in yearly intervals) (SEER, 2016). Table 2 describes the categories for each of the variables (note that an estimate is available for each combination of the categories, i.e. $4 \times 5 \times 2 \times 3 \times 11 = 1{,}320$ estimates in total).

Table 2: Relative survival rate by stage – variables

Variable	Type of cancer	Stage[8]	Sex	Age	Years after diagnosis
Category	Breast	All stages	Male	All Ages	Diagnosis
	Colon and Rectum	Localised	Female	Ages <65	1 year
	Oesophageal	Regional		Ages 65+	2 years
	Prostate	Distant			[…]
		Unstaged			9 years
					10 years

The relative survival rates are based on the November 2016 data submission from the population-based SEER 9 registries.[9] Table 3 shows a snapshot of the dataset.

Table 3: Relative survival rate by stage – snapshot of the data

Type of cancer	Sex	Stage	Age	Survival time	Relative survival rate (%)
Breast	Female	Unstaged	All Ages	10 years	44.77
Breast	Male	Localised	All Ages	10 years	95.92

We can see from Table 2 that the age categorisation, as opposed to time after diagnosis differentiation, is very broad, giving estimates only for the <65 and 65+ age categories. At the same time, the SEER 18 dataset (see Table 5 below) shows that the relative survival rates vary significantly within a single cancer-

[8] Summary Stage (as used in the SEER datasets presented in this report) categorises how far a cancer has spread from its point of origin. For more information on the methods and definitions used by SEER to assign stages: https://seer.cancer.gov/tools/ssm/ (As of 4 February 2019).

[9] SEER 9 areas [http://seer.cancer.gov/registries/terms.html] (San Francisco, Connecticut, Detroit, Hawaii, Iowa, New Mexico, Seattle, Utah, and Atlanta) (As of 4 February 2019).

stage category by age.[10] We therefore use the SEER 9 dataset to understand how the survival rates vary by time since diagnosis and, assuming that the shape of the survival curves are independent of age, fit those to the age-specific estimates obtained from the SEER 18.

3.2.3. The SEER 18 dataset – 5-year relative survival by stage of cancer

The second dataset contains the same set of relative survival rate estimates, except only five years after diagnosis as compared to the cumulative yearly estimates available in the SEER 9 dataset. On the other hand, the SEER 18 dataset offers better age differentiation, providing estimates for the <20, 20-49, 50-64, 65-74 and 75+ age groups. Table 4 shows the detailed breakdown of the available variables and categories.

Table 4: 5-year relative survival – variables

Variable	Type of cancer	Stage	Sex	Age	Years after diagnosis
Category	Breast	All stages	Male	All ages	5 years
	Colon and Rectum	Localized	Female	<20	
	Oesophageal	Regional		20-49	
	Prostate	Distant		50-64	
		Unstaged		65-74	
				75+	

The relative survival rates are based on the November 2016 data submission from the population-based SEER 18 registries.[11] Table 5 shows a snapshot of the dataset.

Table 5: 5-year relative survival – snapshot

Cancer	Sex	Age	Stage at Diagnosis	Relative Survival Rate
Esophagus	Male	20–49	Regional	30.3%
Esophagus	Male	50–64	Regional	24.3%
Esophagus	Male	65–74	Regional	24.7%
Esophagus	Male	75+	Regional	13.5%

[10] For example, according to the SEER 18 dataset, the 5-year survival rate for women with distant breast cancer aged 20 to 49 is about 37 per cent, while it is about 27 per cent for women aged 50 to 65.

[11] SEER 18 areas [http://seer.cancer.gov/registries/terms.html] (San Francisco, Connecticut, Detroit, Hawaii, Iowa, New Mexico, Seattle, Utah, Atlanta, San Jose-Monterey, Los Angeles, Alaska Native Registry, Rural Georgia, California excluding SF/SJM/LA, Kentucky, Louisiana, New Jersey and Georgia excluding ATL/RG) (As of 4 February 2019).

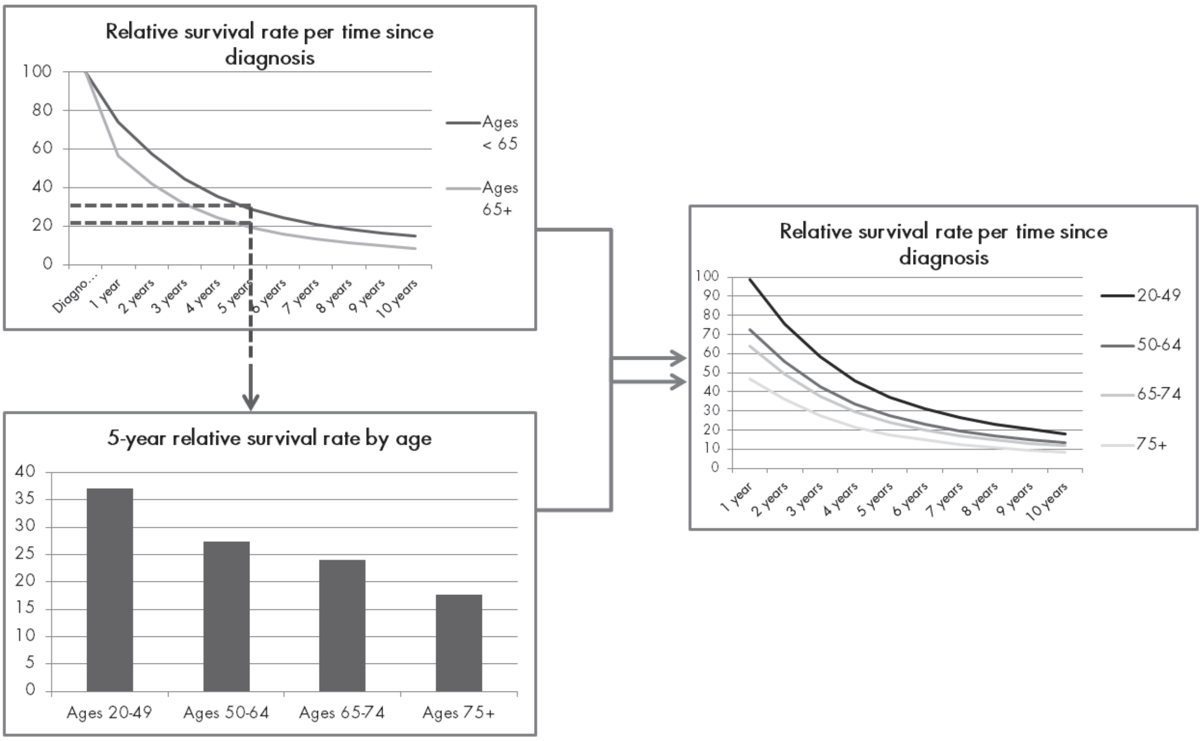
3.2.4. Combining the datasets

Since the two datasets contain the same variables and categories, they are based on the same set of population records, and the only differences are the age and time after diagnosis categorisation, we can use their combination to obtain a more precise differentiation across all of the variables. Specifically, we assume that the shape of the relative survival curve (i.e. the way the relative survival rates at time t increase or decrease compared to those at time $t - 1$ for the same type of individuals) is independent of age. While this may not be true in reality, the approach is arguably more accurate than the alternatives: assuming flat survival rates following diagnosis and/or working with only two age categories. More specifically, we assume that the '<65 years' survival curves would apply to all age groups in that range while the '65+ years' curves would apply to the higher ages.

We then use the age-agnostic survival rate curves from the SEER 9 dataset in combination with the more detailed age-specific 5-year survival rate estimates for the respective sex and cancer type and stage, in a way that the relative survival rates 5 years after diagnosis remain unchanged (i.e. as obtained from the SEER 18 dataset), and the survival rates before and after 5 years after diagnosis are obtained by fitting the age-agnostic survival curves to the 5-year estimates.

For instance, the SEER 9 dataset may suggest that the relative cumulative survival probability for women aged <65 with distant breast cancer 3 years after diagnosis is 45 per cent and 5 years after diagnosis 29 per cent, while the SEER 18 dataset may suggest that, in fact, the cumulative 5-year relative survival probability for women aged 20–49 is 37 per cent, whereas for women aged 50–64 it is 27 per cent. Using the SEER 9 estimates we may therefore underestimate the relative survival probability of women aged 20–49, while using the SEER 18 data we may underestimate it for anyone with less than 5 years after diagnosis. Using our approach, we assume that for all women with this type of cancer aged <65, the relative survival rate 3 years after diagnosis would be 0.45/0.29 = 1.55 times higher than the 5-year relative survival rate, i.e. 1.55 × 0.37 = 57.4 per cent relative survival rate for women aged 20–49 when combined with the SEER 18 estimates, as opposed to the initial flat estimate of 37 per cent. The process is illustrated in Figure 3.

Figure 3: Adjusting the distribution of relative survival rates per time since diagnosis by age



Note: example for distant breast cancer in women.

3.2.5. Survival rates vs mortality rates

Although the two states, 'being alive' and 'being dead', are two sides of the same coin, in cancer statistics, the terms 'survival' and 'mortality' can be seen as two sides of two different coins – as the applicable populations may not be corresponding (Mariotto et al., 2014). In particular, the GBD mortality rate estimates are computed as the number of deaths attributable to cancer among the entire population. In contrast, SEER estimates represent the number of living patients among people with cancer. While the former estimate is the probability that a person will get cancer and die in a given period of time (from any cause, as we are working with the relative mortality rates), the latter is the probability that a person with cancer will survive over a period of time. For the purpose of the VAPC model, mortality rates are used to compute the probability of dying from cancer for someone without cancer. However, if an individual of interest has been diagnosed with cancer, it is appropriate to work with the SEER estimates obtained from a population of individuals with cancers, yet again transformed into (relative) probability of dying for the particular individual.

Note that the survival rates cannot be simply converted into probability of dying, but must first be transformed into mortality rates – taking into account the shrinking population alive at any given time used in the denominator:

$$MR_t = \begin{cases} \dfrac{100 - SR_t}{100} \; if\ t = 1 \\ \dfrac{SR_{t-1} - SR_t}{SR_{t-1}} \; if\ t \in [2\,;10] \end{cases}$$

where MR_t is the mortality rate for people with cancer from $t-1$ to t years after diagnosis and SR_t is the relative survival rate t years after diagnosis. Since the available SEER estimates are only available up to 10 years after diagnosis, we assume that the relative survival curves (and hence the relative mortality curves) for years 11 to 20 after diagnosis look the same way as from 9 to 10 years after diagnosis.[12] Here we cannot exclude the possibility that there may be something specific about the mortality rate at years 9–10, but it is an assumption that we make. The computation for years 11 to 20 after diagnosis is as follows:

$$MR_t = \begin{cases} \dfrac{(MR_{t-1})^2}{MR_{t-2}} \; if \; MR_{t-1} < MR_{t-2} \\ MR_{t-1} \; otherwise \end{cases}$$

Using this formula allows to gradually reduce the mortality rates from cancer until 20 years after diagnosis – when the mortality rates for cancer patients are similar to the mortality rates of people without cancer. Consequently, we assume that anyone who survives 20 and more years after diagnosis has the same probability of dying due to the cancer as anyone else in the population. The mortality rates are then converted into probabilities of dying using formula (2) in Section 2.

3.2.6. Resulting probabilities of dying

By combining the two datasets described above, the final dataset contains relative survival rate estimates for the four types of cancers, per sex, age, stage of cancer and time since diagnosis from one to ten years, as described in the Table 6 below.

Table 6: Condition specific relative survival rate - variables

Variable	Cancer	Sex	Age at diagnosis	Stage	Time since diagnosis
Category	Breast	Male	All ages	All stages	1 year
	Colon and Rectum	Female	< 20	Localized	2 years
	Oesophageal		20 – 49	Regional	3 years
	Prostate		50 - 64	Distant	…
			65 – 74	Unstaged	9 years
			75 +		10 years

Ultimately, the probability of dying estimates for people with one of the specified cancers can be summarised by the snapshot in Table 7.

Table 7: Condition specific probability of dying – snapshot

Cancer	Sex	Stage	Age at diagnosis	Time since diagnosis	Probability of dying

[12] In some instances, the mortality rate between years nine and ten after diagnosis is increasing. This might be due to the SEER data tracking slightly different cohorts and does not seem realistic. In these instances, mortality rates are assumed to be constant at the level of ten years after diagnosis. See Section 3.2.7 for further details.

Breast Cancer	Female	Distant	20–49	1	0.01258
Breast Cancer	Female	Distant	20–49	2	0.208446
Breast Cancer	Female	Distant	20–49	3	0.207472

3.2.7. Methodological and data considerations

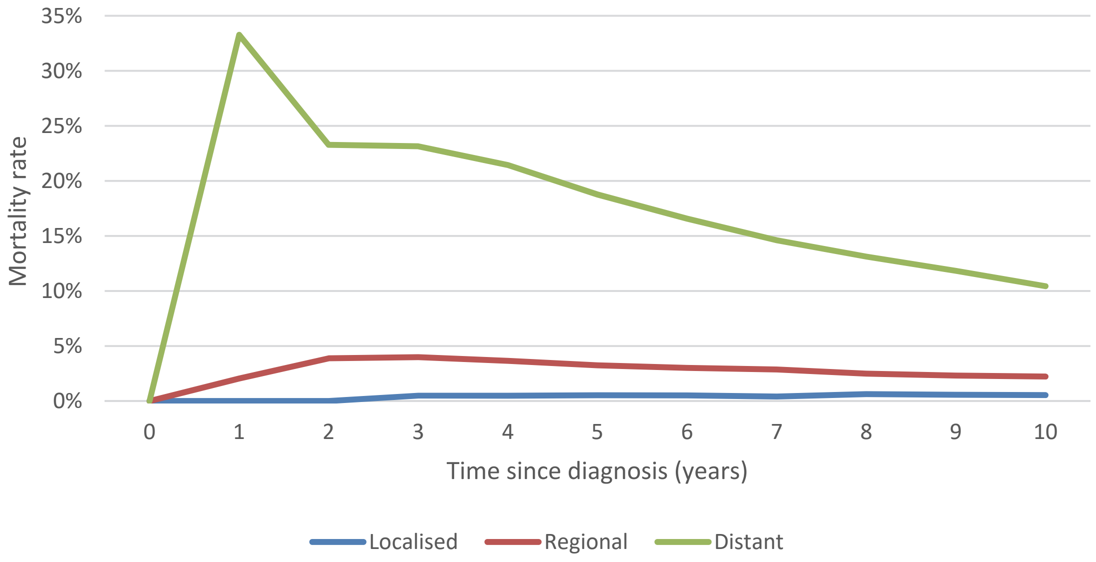
Although the conditions were chosen for their high prevalence, some of the estimates are not available due to a limited number of observations. In particular, young people (below 20 years old) are rather unlikely to have cancer, and the respective relative survival rate estimates are thus missing from the SEER database. Similarly, data for unstaged breast cancer in men is missing for all age categories. Table 8 summarises all the situations in which data is missing. In the case of unstaged breast cancer in men, the statistics for all stages of breast cancer are assigned. Cancers in young people, however, might have different effects compared to older populations. It is therefore treated as a rare disease and is not addressed in the VAPC.

Table 8: Summary of missing data per disease and stage for each subpopulation

	Male					**Female**				
	<20	**20–49**	**50–64**	**65–74**	**75+**	**<20**	**20–49**	**50–64**	**65–74**	**75+**
Breast Cancer										
All Stages	X									
Distant	X					X				
Localized	X									
Regional	X					X				
Unstaged	X	X	X	X	X	X				
Colon and Rectum Cancer										
Distant	X					X				
Unstaged	X					X				
Esophageal cancer										
All Stages	X					X				
Distant	X					X				
Localized	X					X				
Regional	X					X				
Unstaged	X					X				
Prostate cancer										
All Stages	X									
Distant	X									
Localized	X									
Regional	X									
Unstaged	X									

The second issue to consider is unavailability of data beyond ten years after diagnosis – a result of difficulty in tracking individuals for an extended period of time, lower number of individuals to track, and also that probability of dying is considerably lower at that point than shortly after diagnosis. Indeed, the SEER 9 data show that there is generally a consistent, decreasing trend in the probability of dying with time since diagnosis, particularly for the more aggressive types of cancer. Given the consistent trends in each of the time-series (see, for example, breast cancer), it is reasonable to expect that the probability of dying would decrease further. We assume that this would be at the same rate as in the last observed year; this is an arbitrary assumption, but using any form of decreasing function would yield virtually the same results in terms of the VA estimate. In the few cases where the probability of dying slightly increased from nine to ten years after diagnosis, we assume that the rate would remain flat after that, at the level of year ten. Arguably, this is due to a relatively small number of individuals being observed rather than a particular change in the way the disease affects an individual at that time, which is also supported by estimates for other age or stage categories, which again show decreasing trends.

Figure 4: Breast cancer mortality rates by stage and time since diagnosis; all ages; all races



Source: SEER 9 dataset

3.3. Ischaemia

As opposed to cancers, ischaemic heart disease (IHD) and ischaemic stroke (IS) are both a result of a common cause: ischaemia. Ischaemia is a restriction in blood supply to a body part (such as brain or heart), causing a shortage of oxygen with resultant damage to or dysfunction of tissue. Cardiac ischaemia most frequently results from a long-term accumulation of cholesterol-rich plaques in the coronary arteries. According to the Global Burden of Disease (2004), ischaemic heart diseases are among the leading causes of death in the world, especially in high-income countries.

Ischaemia does not have stages nor can it be fully healed by a treatment or a surgical operation; the treatment aims at improving blood flow using various medications. Ischaemia can be prevented by having a healthy lifestyle and staying on top of other conditions such as diabetes, hyperlipidaemia and hypertension.

3.3.1. Data

The short literature review identified several factors affecting the likelihood of death from ischaemia, such as age, sex, lifestyle, or time since the last heart attack or stroke (Pietrangelo 2018; Delgado, 2018; Raby 1990). Since all of the factors (except for time since last heart attack or stroke) are covered in the GBD database, we use it as the main data source for the exercise. Importantly, the GBD database covers not only mortality rates, PAF's and relative risks associated with the appropriate risk factors, but also prevalence of ischaemic heart disease and ischaemic stroke (and therefore ischaemia in general) in the general population. We can utilise the prevalence estimates to limit the otherwise general population to only a population with ischaemia and use the mortality rates (i.e. the number of deaths in a given location in a year) to calculate probability of dying if one has ischaemia.

Specifically, since mortality rates obtained from the GBD database[13] are simply the number of deaths in a year divided by the total number of people alive at the beginning of the year, we may directly divide the mortality rates by the prevalence estimate to obtain an approximation of the mortality rate in the reduced population, which is directly applicable to anyone with ischaemia as a pre-existing condition. Further, assuming that the RR estimates defining changes in the probability of dying to IHD or IS due to the various risk factors do not differ for individuals with and without ischaemia, we can follow the entire MR adjustment process described in Section 2. This is essentially equivalent to the way diabetes is treated in the standard VA.3 model.

The mortality rate data for both IHD and IS were downloaded from the GBD database in form of deaths per 100,000 person-years in the UK, for each 5-year age group and sex.[14,15] Table 9 below shows a snapshot of the population mortality rates data.

Table 9: Snapshot of population mortality rates data

Location	Cause of death	Sex	Age	Year	Metric	Mean
United Kingdom	Ischaemic Heart Disease	Male	50–54	2016	Deaths per 100,000	73.77
United Kingdom	Ischaemic Heart Disease	Female	50–54	2016	Deaths per 100,000	16.77

Source: GBD 2016 database.

Analogously, prevalence of both IHD and IS were downloaded from the GBD database in form of prevalence per 100,000 person-years in the UK, for each 5-year age group and sex. Table 10 below shows a snapshot of the prevalence data.

[13] ghdx.healthdata.org/gbd-results-tool (As of 4 February 2019).

[14] The full list of age categories is: 0–6 days, 7–27 days, 28–364 days, 1–4 years, 5–9 years, 10–14 years, 15–19 years, 20–24 years, 25–29 years, 30–34 years, 35–39 years, 40–44 years, 45–49 years, 50–54 years, 55–59 years, 60–64 years, 65–69 years, 70–74 years, 75–79 years, 80–84 years, 85–89 years, 90–94 years, and 95+ years.

[15] The 2016 data for the UK can be downloaded using the following permalink: http://ghdx.healthdata.org/gbd-results-tool?params=gbd-api-2016-permalink/f844acd3149f44c33bee12fc1759c31f (As of 4 February 2019).

Table 10: Snapshot of prevalence data

Location	Cause of death	Sex	Age	Year	Metric	Mean
United Kingdom	Ischaemic Stroke	Male	50–54	2016	Prevalence per 100,000	828.00
United Kingdom	Ischaemic Stroke	Female	50–54	2016	Prevalence per 100,000	828.24

Source: GBD 2016 database.

The adjusted mortality rates among individuals with ischaemia are then computed by dividing the population mortality rate by the population prevalence for each category. It is then transformed into probability of dying using formula 2. Table 11 shows a snapshot of the resulting probability of dying.

Table 11: Snapshot of the adjusted probability of death data

Location	Cause of death	Sex	Age	Year	Metric	Mean
United Kingdom	Ischaemic Stroke	Male	50–54	2016	Probability of dying	0.00348
United Kingdom	Ischaemic H. D.	Female	50–54	2016	Probability of dying	0.00953

4. Limitations

The VAPC model as described in this report contributes to addressing one of the main limitations of the VA.3 model. VAPC takes into account six possible existing long-term conditions that have substantial impact on the life expectancy of assessed individuals. However, there are ethical considerations and potential methodological limitations of the proposed approach, some of which could be addressed by potential future updates. These are presented in this section.

Ethical considerations

VAPC extends the VA.3 model to individuals with certain pre-existing conditions. However, such a population might be negatively impacted by the estimation of their prognosis. It would also be problematic for those with pre-existing conditions to engage with a calculator of this nature without adequate professional support. Moreover, and as explained hereafter, the VAPC has a number of limitations that might make it biased and incorrect for certain individuals. Steps must be taken when feeding back the resulting VAPC estimate to affected individuals to ensure that they are aware of these limitations and to inform them that, given their condition(s), a more accurate estimation of their prognosis can be provided by a specialist.

Quality of the data

First, similarly to the VA.3 model, the VAPC model is fully reliant on the existence and accuracy of the existing empirical estimates. The VA.3 model and the VAPC data on ischaemia builds on the GBD database, which uses Bayesian statistics to estimate data points that are completely missing or do not have a sufficient empirical base. However, data on cancers is based on the SEER database, which relies on US registries and does not offer estimates beyond the United States or impute missing data or data based on a low number of estimates. The VA.3 model parameters are therefore more comprehensive than in the VAPC model, but are estimated using a less consistent approach.

In addition, although the SEER data seem to be of a good quality overall (guaranteed by regular audits and quality improvement methods[16][17][18]), it is based on the US population which is different from that of the UK. Moreover, disparities in the health system and other geographical characteristics are likely to affect the results.

[16] https://seer.cancer.gov/about/uses.html (As of 4 February 2019).

[17] https://seer.cancer.gov/about/factsheets/SEER_QI_Fact_Sheet.pdf (As of 4 February 2019).

[18] https://seer.cancer.gov/qi/tools/ (As of 4 February 2019).

Factors affecting prognoses

The second limitation relates to complexity of the selected conditions, and particularly the potential differences in their effects on individuals' mortality rates. For example, the probability of dying due to a cancer is conditional on the age and general health of the patient, existence of other conditions, lifestyle or treatment used, but also principally on the stage at which the cancer was identified and the particular subtype of cancer. Similarly, the probability of dying from an ischaemic heart attack or stroke for individuals previously affected is correlated with the time since the last heart attack or stroke. Even though a large proportion of such variance should be captured by the breakdown into stages for cancers and other lifestyle factors for ischaemia, it might not be sufficient to be able to accurately predict the probability of death.

Relative survival rates

Third, although using relative survival rates seems the most appropriate in our situation, it might lead to some double counting and overestimation of the mortality rates (and therefore the Vitality Age) of cancer patients. As explained in Section 3, the proportion of cancer patients who survived a given time period reflects surviving *all* causes of death, rather than cancer only. In an ideal world, this should be divided by the proportion of the same persons who would have survived if they did not have cancer. However, such data is not available and is instead replaced by the proportion of survivors in a similar (age and sex) group of people without cancer. The assumption being made is that cancer occurs randomly, independently of other characteristics, which is likely to not be true. There are a number of risk factors influencing the probability of having a cancer, and some of these risk factors also increase the probability of dying from some other causes. It is likely that the groups used in the numerator and in the denominator of the relative survival rate calculation have different risk-factors profiles. Therefore, the group in the numerator is more likely to die from other causes, so the probability of dying from cancer computed using the relative survival rates is likely to count deaths from other causes as well, potentially leading to double counting.

Other pre-existing conditions

The final limitation relates to the sole focus on a selected list of pre-existing conditions – any other pre-existing conditions would not be captured in the model, although the model can be readily extended again in the future, conditional on existence of suitable parameters. As such, the model tests the feasibility of including pre-existing conditions and the value added of this approach.

References

Cho, Hyunsoon, et al. 'Estimating relative survival for cancer patients from the SEER Program using expected rates based on Ederer I versus Ederer II method.' Surveillance Research Program NCI, editor (2011).

Eker, B., Ozaslan, E., Karaca, H., Berk, V., Bozkurt, O., Inanc, M., ... & Ozkan, M. 2015. Factors affecting prognosis in metastatic colorectal cancer patients. Asian Pac J Cancer Prev, 16(7), 3015–3021.

Giovannucci, E., Liu, Y., Platz, E. A., Stampfer, M. J., & Willett, W. C. 2007. Risk factors for prostate cancer incidence and progression in the health professionals follow-up study. International journal of cancer, 121(7), 1571–1578.

Delgado, A. 2018. Stroke Recovery: What to Expect. (Medically reviewed by Seunggu Han, MD on June 7, 2018). https://www.healthline.com/health/stroke/recovery (As of 4 January 2019).

Pietrangelo, A. 2018. Ischemic Cardiomyopathy: Symptoms, Causes and Treatment. (Medically reviewed by Stacy Sampson, DO on January 26, 2018). https://www.healthline.com/health/ischemic-cardiomyopathy (As of 4 January 2019).

Lim, S., E. Carnahan, E. Nelson, C. Gillespie, A. Mokdad, C. Murray & E. Fisher. 2015. 'Validation of a new predictive risk model: measuring the impact of the major modifiable risks of death for patients and populations.' *Population Health Metrics* 13(1):27.

Mariotto, Angela B., Anne-Michelle Noone, Nadia Howlader, Hyunsoon Cho, Gretchen E. Keel, Jessica Garshell, Steven Woloshin, and Lisa M. Schwartz. 'Cancer survival: an overview of measures, uses, and interpretation.' Journal of the National Cancer Institute Monographs 2014, no. 49 (2014): 145–186.

National Cancer Institute. 2018. Understanding Cancer Prognosis. Reviewed. Reviewed August 29, 2018. https://www.cancer.gov/about-cancer/diagnosis-staging/prognosis (As of 4 January 2019).

Raby, K. E., Goldman, L., Cook, E. F., Rumerman, J., Barry, J., Creager, M. A., & Selwyn, A. P. 1990. Long-term prognosis of myocardial ischemia detected by Holter monitoring in peripheral vascular disease. American Journal of Cardiology, 66(19), 1309–1313.

Stepanek, Bolnick, Millard, van Stolk, Garrod, Saunders, van Belle. 'Vitality Age v.3.' 2018. *To be published.*

Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) SEER*Stat Database: Incidence - SEER 18 Regs Research Data + Hurricane Katrina Impacted Louisiana Cases, Nov

2016 Sub (2000-2014) , National Cancer Institute, DCCPS, Surveillance Research Program, Surveillance Systems Branch, released April 2017, based on the November 2016 submission.

World Health Organization. 2004. 'Global Burden of Diseases, 2004 update.' http://www.who.int/healthinfo/global_burden_disease/GBD_report_2004update_full.pdf (As of 4 January 2019).

Zare-Bandamiri, M., Khanjani, N., Jahani, Y., & Mohammadianpanah, M. 2016. Factors affecting survival in patients with colorectal cancer in Shiraz, Iran. Asian Pac J Cancer Prev, 17(1), 159–63.