# GBD Dietary Risks Emulator

Version 0.1

# Description of model and validation

Shaun Solomon, Estelle Paulus, Steven Lord

Environmental Change Institute, University of Oxford

Citation: Solomon, S., Paulus, E., Lord, S. (2025) GBD Dietary Risks Emulator: Description of model and validation. Documentation of the SPIQ-FS Dataset Version 0. Environmental Change Institute, University of Oxford.

Disclaimer: Views are the sole responsibility of the authors. The authors, their organisations, and funders, accept no liability for any damage resulting from the use of the results of this document or the application of the emulator model. Validation that the code reproduces the population attributable fractions of the Global Burden of Disease requires data downloadable from IHME. License terms of IHME.

# Summary


**Methodology**

This document describes the Global Burden of Disease (GBD) Dietary Risks Emulator, a stand-alone computational model written in Python that reproduces the dietary risk component of the GBD 2017 comparative risk assessment framework. The emulator uses established comparative risk assessment formulas to quantify disease burden attributable to the composition of dietary intake by linking population exposure distributions to disease outcomes via relative risk functions and population attributable fractions (PAFs). Fifteen dietary risk factors are modelled, encompassing both under- and over-consumption relative to theoretical minimum risk exposure levels, and are linked to cardiovascular diseases, diabetes and kidney diseases, and selected neoplasms.

Population dietary exposure is represented using reconstructed intake distributions parameterised by published GBD estimates of mean intake and standard deviation across country, age, and sex subpopulations. To approximate population heterogeneity, the emulator employs a weighted ensemble of candidate statistical distributions consistent with the GBD 2017 methodology. Disease burden is calculated in disability-adjusted life years (DALYs) by combining attributable fractions with years of life lost (YLLs) and years lived with disability (YLDs). Joint exposure to multiple dietary risks is handled through mediation matrices to avoid double counting of disease outcomes. In addition, the emulator includes a Monte Carlo simulation framework that propagates uncertainty in exposure, relative risks, and disease burden inputs, to emulate the GBD uncertainty estimates.

**Intended Output and Use**

The primary output of the emulator is national-level disease burden attributable to joint dietary risk exposure, expressed in DALYs, as an aggregate of disease burden by disease outcome, age, sex, and country. Beyond reproducing baseline attribution, the emulator code is designed to support counterfactual and scenario analysis. Users can assess changes in disease burden by providing shifts in dietary intake distributions, either uniformly across populations or targeted to specific subpopulations and dietary components. The code also calculates marginal attributable disease burdens, representing the rate of change in disease burden per unit change in dietary intake at current exposure. These outputs are intended to support applied research, policy analysis, and economic or environmental modelling where estimates of diet-related health impacts are required, while remaining flexible to alternative exposure or relative risk inputs supplied by users.

**Validation**

The emulator is validated against published GBD 2017 results. Individual population attributable fractions for single dietary risks are compared with values available from the GBD results tools. For national attributable disease burdens from joint dietary exposure, the emulator estimates align with published GBD values within 1-2%. Remaining discrepancies are primarily attributable to known complexities in the GBD treatment of sodium, including mediation through systolic blood pressure, and to reconstructing intake distributions. Validation demonstrates that the emulator provides a transparent stand-alone reproduction of the GBD 2017 dietary risk model suitable for analytical and applied modelling of alternative exposures. Limitations of the GBD design and inputs are detailed, and any such limitations are inherited by the emulator.

# Annex A

## Contents


1 Introduction ..... 4

2 Description of the comparative risk model ..... 6

2.1 Description of inputs ..... 8

2.1.1 Exposure from dietary intake ..... 8

2.1.2 Disease outcomes and relative risk from exposure ..... 11

2.1.3 Population disease burdens as years of life lost to mortality and morbidity ..... 15

2.2 Emulator model ..... 15

2.2.1 Population attributable fraction ..... 15

2.2.2 Attributable disease burden ..... 16

2.3 Description of outputs ..... 17

2.3.1 Aggregate attributable disease burden from current exposure to dietary risks ..... 17

2.3.2 Attributable disease burden from shifting exposure to dietary risks ..... 17

2.3.3 Marginal disease burden from shifting exposure to dietary risks ..... 18

2.4 Monte Carlo simulation of attributable disease burden due to exposure to dietary risk factors ..... 19

2.4.1 Inputs as random variables ..... 20

2.4.2 Joint sampling of exposure, response and population disease burdens ..... 20

3 Validation of the emulator ..... 22

3.1.1 Validation of population attributable fractions ..... 22

3.1.2 Validation of national attributable disease burden from joint exposure ..... 24

4 Limitations ..... 28

4.1 General limitations of the comparative risk assessment model ..... 28

4.2 Specific limitations of the Global Burden of Disease dietary risks model ..... 30

Appendix A : maximal entropy sampling given an interval and an average value ..... 33

Appendix B : recalculation of parameters for candidate distributions ..... 33

Appendix C : further graphs of validation of emulator ..... 36

Appendix D : Derivation of attributable disease burden ..... 46

Appendix E : Proportional individual risk estimates ..... 50

5 References ..... 52

# Annex A

## 1 Introduction

The food system is responsible for a significant fraction of the observed preventable, global Non-Communicable Diseases (NCDs) through dietary intake. According to the Global Burden of Disease (GBD) study in 2017 [1], dietary risks were responsible for 11 million deaths and 255 million disability-adjusted years of life lost (DALYs) among adults (15% of all DALYs among adults from preventable risks). Furthermore, stating that globally every fifth death could be potentially prevented by a change of diet, the study [1] showed that nearly all healthy foods were under consumed whereas the daily intake of all unhealthy foods was exciding the recommended threshold. Leading risk factors for mortality were found to be diets high in sodium, low in whole grains, low in fruit, low in nuts and seeds, low in vegetables, and low in omega-3 fatty acids. Dietary risk factors are associated with an increase in the risk for NCDs such as diabetes mellitus type II, or ischaemic stroke.

Public health models such as the Global Burden of Disease study (GBD) [2] use a comparative risk framework to quantify dietary risk and burden of disease by identifying dietary intake associated with negative health outcomes when over or under consumed. The Global Burden of Disease study and its data and models are developed and managed by the Institute for Health Metrics and Evaluation (IHME) at the University of Washington. The GBD studies, model and data have periodic updates, which have been every 2 years since 2017. The GBD 2017 study lists 15 different dietary risk factors, which are related to exposure to over or underconsumption from an optimal range of dietary intake. The selection of the risk factors for the GBD study was based on; the importance for either disease burden or policy, the availability of sufficient data, the strength of the epidemiological evidence and the data availability supporting the relationship per unit change of exposure, as well as the generalisability of the effects from a selection of countries to all populations.

Each risk factor is associated with some, but not necessarily all metabolic mediators of diseases, which can be summarised as related to influences on the body mass index, the total serum cholesterol, the fasting plasma glucose, as well as the systolic blood pressure. Those mediators influence the vulnerability of the human body to certain dietary related diseases as well as their likelihood. Furthermore, this risk is influenced by factors such as age, sex, or socio-economic status. Disease outcomes attributable to dietary intake fall into three groups: cardiovascular diseases, diabetes and kidney diseases and different kinds of neoplasms. Globally, 39% of all preventable cardiovascular diseases (15% of the total DALYs attributed to preventable causes) and 18% of diabetes and kidney diseases (4% of the total DALYs attributed to preventable causes) were attributed to a combination of all dietary risk factors in 2019 [3].

Exposure to over and under consumption in the GBD study is measured by distributions of per capita adult dietary intake exceeding or below, respectively, an optimal level of intake for each dietary risk category (Table 1) [2]. The distributions are estimated, and imputed from statistical modelling, from national surveys and 24 dietary recall studies by the GBD study.

*Table 1: 15 dietary intake quantities associated to human disease outcomes from over and consumption used in the 2017 Global Burden of Disease (GBD). Summary of Table on p. 3 in [2]. For risk from low intake, the optimal level of intake indicates the level below which intake is associated to preventable disease outcomes. For risk from high intake, the optimal level of intake indicates the level above which intake is associated to preventable disease outcomes. The intervals above, or below respectively, optimal intake are called the TMREL intervals. Public health models using population attributable fractions attribute the additional disease outcomes due to distributions of adult population dietary intake supported outside*

*the TMREL interval. Optimal levels of intake are uncertain, and the range from [2] given in brackets for optimal intake is used in uncertainty modelling.*

| Dietary intake quantity | Intake unit | Dietary risk factor | Optimal Level of Intake (in intake units) |
|---|---|---|---|
| **Fruit** consumption excl. fruit juice | grams per day per capita (g/capita/day) | Diet low in fruit | 250 (200-300) |
| **Vegetables** consumption excl. starchy vegetables | grams per day per capita (g/capita/day) | Diet low in vegetables | 360 (290-430) |
| **Whole grains** consumption | grams per day per capita (g/capita/day) | Diet low in wholegrains | 125 (100-150) |
| **Legumes** consumption | grams per day per capita (g/capita/day) | Diet low in legumes | 60 (50-70) |
| **Nuts and seeds** consumption | grams per day per capita (g/capita/day) | Diet low in nuts and seeds | 21 (16-25) |
| **Milk**, excl. plant milks | grams per day per capita (g/capita/day) | Diet low in milk | 435 (350-520) |
| **Fibre** | grams per day per capita (g/capita/day) | Diet low in fibre | 24 (19-28) |
| **Calcium** | grams per day per capita (g/capita/day) | Diet low in calcium | 1.25 (1.00-1.50) |
| **Red meat** consumption excl. processed | grams per day per capita (g/capita/day) | Diet high in red meat | 23 (18-27) |
| **Preserved / processed meat** consumption | grams per day per capita (g/capita/day) | Diet high in processed meat | 2 (0-4) |
| **Omega-3 fatty acids** from seafood consumption | grams per day per capita (g/capita/day) | Diet low in omega-3 fatty acids | 0.25 (0.20-0.30) |
| **Omega-6 polyunsaturated fatty acids** consumption | % of total kcal intake per capita per day | Diet low in omega-6 polyunsaturated fatty acids | 11 (9-13) |
| **Trans Fat** consumption | % of total kcal intake per capita per day | Diet high in trans fatty acids | 0.5 (0.0-1.0) |
| **Sugar-sweetened beverage** consumption incl. carbonated beverages, sodas, energy drinks, sweetened fruit drinks excl. 100% fruit and vegetable juices | grams per day per capita (g/capita/day) of beverages with ≥50 kcal per 226.8g serving | Diet high in sugar-sweetened beverages | 3 (0-5) |
| **Sodium** intake | 24h urinary sodium in grams per day per capita (g/capita/day) | Diet high in sodium | 3 (1-5) |

In many food system studies, including the authors', researchers would like to explore varying current exposure levels in the categories listed in Table 1 and estimating the change in disease

burden. From the change in disease burden, research can gain understanding of future economic or social effects from changes in food consumption.

Code for the Global Burden of Disease 2017 model was previously available in R from IHME. However, it was not standalone for only the dietary risk factors and required additional data across other categories. The GBD study implements standard calculations for comparative risk assessment in public health. For use in our own studies, we wrote a Python emulator of the dietary risk component of the comparative risk modelling from scratch, with exposure inputs approximating the published distribution format of the Global Burden of Disease Study.

The Python code is available at https://github.com/shaun-solomon/GBD-Diet-NCD-Model-Emulator and described there. The code does not need to use the GBD exposure values, or the response to exposure captured by GBD relative risk estimates from studies. User can input their own exposure and relative risk values. This document describes the comparative risk modelling performed by the code, including formulas, and the validation against the GBD 2017 dietary risk factor results using the same inputs as the GBD 2017 study.

# 2 Description of the comparative risk model

A public health comparative risk model maps from dietary intake (exposure) to disease burden attributable to exposure measured in disability-adjusted life years (DALYs), [4, 5].

Figure 1 shows the modelling process. Exposure in the form of distributions of dietary intake in populations is integrated against the relative risk of disease outcomes at given levels of intake. The integration provides a mean relative risk from current population exposure versus a population with an exposure level below the minimum threshold. The mean relative risk is used in a formula for the **population attributable fraction** (**PAF**) which, in comparative risk assessment framework, is interpreted as the fraction of a disease outcome attributable to the exposure above minimum threshold. The **attributable disease burden for that disease in that population with current exposure** is calculated when the years of life lost to mortality (**YLL**) or morbidity (**YLD**) from a disease outcome in that population are multiplied against the fraction attributable to exposure.

The three sources of data (input data) required to populate comparative risk formulas for any or each population being studied are therefore (Figure 1)

- Exposure (distribution of dietary intake is derived from mean and standard deviation of population intake)
- Effect parameters (relative risk parameters for each disease outcome including the minimum threshold (TMREL))
- Overall disease burden by disease outcome (YLLs and YLDs)

The first two types of data are required to generate PAFs. The third type of data are required to arrive at disease burden estimates in DALYs from PAFs.

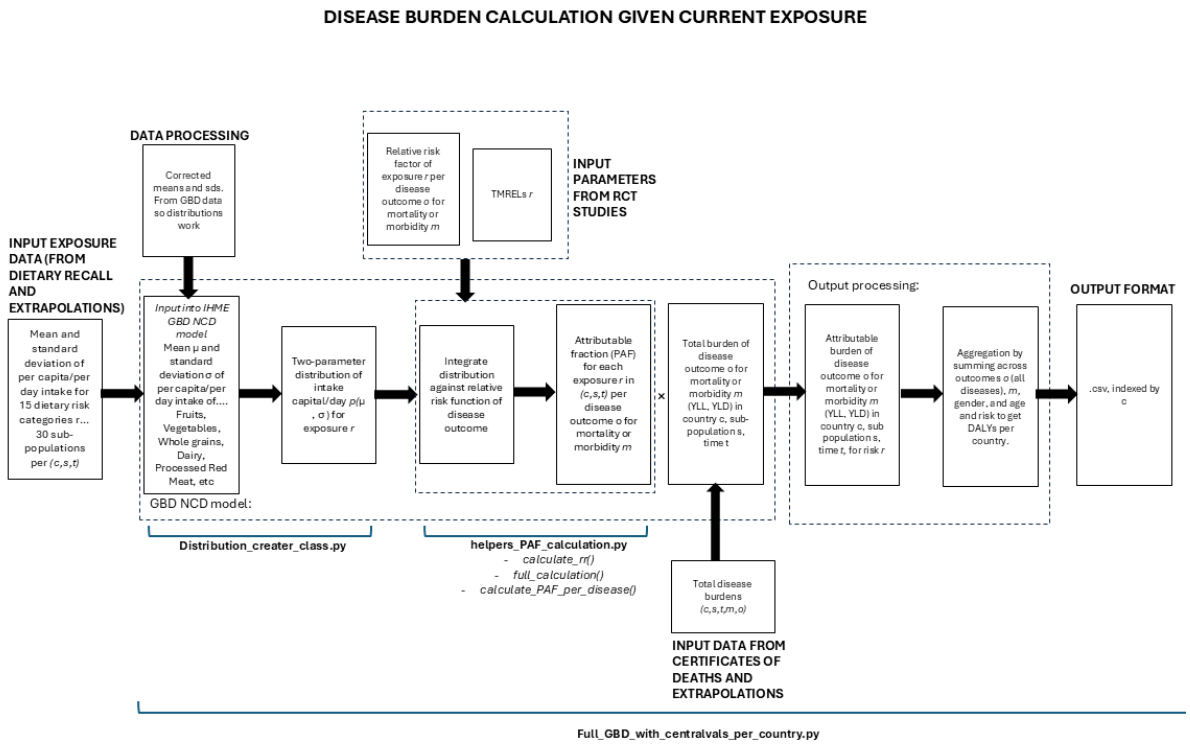



*Figure 1: Calculation of the attributable disease burden to dietary risks involves mapping 15 dietary intake exposure categories for subpopulations to the PAFs for disease outcomes, which is the fraction of the total mortality or morbidity disease outcome in the population attributable to the dietary intake exposure. The attributable disease burden from exposure to dietary intake is obtained from multiplying PAFs for disease outcomes against the total disease outcome in the population.*

One of the computational and data challenges of emulating the GBD model is that PAFs and disease burdens are calculated for exposure in national subpopulations. Every national population is broken down in thirty (30) age and sex subpopulations, and some regions such as the USA and the UK are further broken down into subnational regions. The emulator does not go to subnational regions, using only subpopulations characterised by country, age and sex. The emulator code has the flexibility to allow the user to isolate subpopulations (one country, all age and sex indices; all countries, one age and sex; all Asian countries and individuals aged 70+, etc.) and isolate disease outcomes. Aggregation is typically to national level for the subpopulations identified and across up to thirteen (13) disease outcomes. Aggregation of dimensions can also be customised in the code.

For example, to calculate the aggregate national disease burden attributable to the dietary risks listed in Table 1 for 160 countries, up to 160*30*15*13*2 = 1,872,000 individual calculations of attributable disease burden iterating over a given country, a given age and sex subpopulation, a given intake, a given disease outcome, and for morbidity and mortality, are performed. For national level estimates of attributable disease burden the individual calculations are then aggregated across subpopulations, exposures, diseases, and morbidity and mortality.

The next sections go into more detail on the inputs, the formulas for the comparative risk assessment, and the outputs. An additional output of the emulator which is not available from IHME

results is the marginal attributable disease burden given current exposure. The marginal attributable disease burden can be thought of as the rate of change of disease burden at current exposure.

## 2.1 Description of inputs

### 2.1.1 Exposure from dietary intake

The GBD 2017 study identified 15 different risk factors related to dietary intake [2] (Table 1). These risks are

- diet low in calcium (calcium),
- diet low in fibre (fiber),
- diet low in seafood omega-3 fatty acids (omega-3),
- diet low in fruits (fruits),
- diet low in whole grains (wholegrains),
- diet low in legumes (legumes),
- diet low in milk (milk),
- diet low in nuts and seeds (nuts),
- diet high in processed meat (procmeat),
- diet low in polyunsaturated fatty acids (pufa),
- diet high in red meat, diet high in sodium (sodium),
- diet high in sugar-sweetened beverages (ssbs),
- diet high in trans fatty acids (transfat),
- diet low in vegetables (vegetables).

These risks are not treated as mutually exclusive and the formulas to calculate the joint disease burden from joint exposure to dietary risks are given below.

Exposure to dietary risks is represented by the dietary intake in calcium, fibre, etc. per Column 1 and in units per Column 2 of Table 1. Dietary intake varies in populations, which is represented in the GBD and the emulator by distributions of intake. Interpreted as a relative frequency distribution, the distribution of intake can be derived from the information of the fraction of people in the population that have intake above a certain level. Interpreted as a probability distribution, the c.d.f. of the distribution of intake represents the chance that a random individual in that population has intake above a certain level.

Even in the USA, the EU, and the UK, there are few studies that have significant enough sample sizes and unbiased measurement of intake to empirically derive the intake distributions for subpopulations. The GBD therefore extrapolates distributions of intake for country, age and sex, indexed subpopulations from information that is available from thousands of data sources of dietary intake. The extrapolation method is described in Supplemental Information to the GBD 2017 study [1], but it is complicated and not always transparent.

The emulator model follows the GBD 2017 derivation of distributions of intake. Without approximating the same distributions of output, the emulator could not be validated against GBD results. Users can modify the code to use their own derived distributions of intake for populations.

#### 2.1.1.1 Parameterized family of candidate distributions of intake

The mean and standard deviation are used as parameters in an ensemble of distributions to reconstruct the distribution of intake in a subpopulation indexed by country, age group and sex [6].

The GBD 2017 employed individual consumption level data from nutrition and health surveys, as well as data from FAO FBS [7] and FAO SUA/USDA [8] to estimate the mean and standard deviation of intake, and parametrise the shape of the distribution of intake for subpopulations. Due to standardisation and lack of data the intake distributions for each country, age group, sex and risk factor were reconstructed as a weighted ensemble of a family of candidate distributions parameterised by mean and standard deviation. First, a weighted linear combination of candidate models (namely: exponential, gamma, Weibull, Gumbel, log-logistic, normal, log-normal, stretched beta, mirrored Gumbel, mirrored gamma, inverse gamma, and an inverse Weibull distribution) was fitted to the data and the weights were averaged over the demographic subsets to get one set of weights for each risk factor. The first two empirical moments (mean and standard deviation) were fitted to the data via the Method of Moments (MoM) and the goodness of fit estimated via a Kolmogorov-Smirnov (KS) test. Data which was not available was imputed between countries and age splitting applied using a spatial and temporal hierarchal model [2], [6].

All weights, mean with 2.5$^{th}$ and 97.5$^{th}$ percentile are included in published dataset available from the GBD. The use of the stretched beta distribution, the mirrored Gumbel, and the mirrored gamma distribution in the ensemble mean that the upper limit of the intake also became a parameter for reconstructing the intake distribution as well as upper and lower limit for each distribution. This created additional calculation steps because the upper and lower limits were not published by the GBD. Standard deviation values with 2.5$^{th}$ and 97.5$^{th}$ percentile and upper and lower values can be requested from IHME.

#### 2.1.1.2 Issues with upper and lower limits

For each risk factor $j$ and age group $a$ pair there was an upper $u_{ja}$ and a lower limit $l_{ja}$ of the intake distribution which was not specified in the GBD study but obtained directly from the GBD study authors. Those upper and lower boundaries of the intake range were needed to reconstruct the distributions in Table 2. Both upper and lower values were provided by the GBD as single number without specifying an uncertainty band.

Mean $\mu_{jcas}$ and standard deviation $\sigma_{jcas}$ of intake were available from GBD data for each risk factor $j$ and subpopulation $(c, a, s)$ of a country $c$, age group $a$ and sex $s$. The data provided also has upper $\mu^u_{jcas}, \sigma^u_{jcas}$ and lower $\mu^l_{jcas}, \sigma^l_{jcas}$ limit of a 95-uncertainty interval for mean and standard deviation of intake. For some risk factor $j$-country $c$-age $a$-sex $s$ combinations the $\mu^u_{jcas}$ value (resp. $\mu^l_{jcas}$) was larger (resp. smaller) than the upper $u_{ja}$ and lower $l_{ja}$ limit of the distribution provided, which represents inconsistency in the data provided by the GBD. In those cases, a new maximum (minimum) value was calculated via

$$u_{ja} = \mu_{jc^*as^*} + \sigma_{jc^*as^*}, \qquad \mu_{jc^*as^*} = \max_{c,s} \mu_{jcas}, \sigma_{jc^*as^*} = \max_{c,s} \sigma_{jcas}$$

$$l_{ja} = \begin{cases} \mu_{jc_*as_*} - \sigma_{jc_*as_*}, & \text{if } \mu_{jc_*as_*} - \sigma_{jc_*as_*} > 0 \\ 0, & \text{else} \end{cases}, \quad \mu_{jc_*as_*} = \min_{c,s} \mu_{jcas}, \sigma_{jc_*as_*} = \min_{c,s} \sigma_{jcas}.$$

### 2.1.1.3 Calculating the distribution

The intake distributions for the 15 different dietary related risks are calculated using the mean and standard deviation of each distribution, as well as the processed upper and lower limit. To recalculate the intake distributions used in the GBD study, the distributions were derived as weighted mixture models of the 12 distributions namely: exponential [9], gamma [10], Weibull [11], Gumbel [12], log-logistic [13], normal [14], log-normal [15], and a stretched beta-distribution[1] [16], mirrored Gumbel[2] and gamma distribution[3], as well as an inverse gamma[4] and an inverse Weibull[5] distribution [17]. The parameters for each distribution were derived for the mean and standard deviation as presented in the GBD data for each of the dietary risks, age groups and countries. See Appendix B and the difficulty with limits and the stretched beta distribution. The weights are different for each dietary risk, but do not depend on the country or age group. All weights are presented in Table 2. During sampling all distributions are truncated such that they never exceeded the prescribed lower and upper boundary.

*Table 2: Weights for mixture model for recalculating the intake distributions from mean and standard deviation per risk factor as given by the GBD.*

| **Risk factors** | **Distribution weights** | | | | | |
|---|---|---|---|---|---|---|
| | **Exponential** | **Gamma** | **Log-logistic** | **Gumbel** | **Weibull** | **Log-normal** |
| **calcium** | 0.001456 | 0.193448 | 0.001727 | 0.277735 | 0.023303 | 0.146496 |
| **fiber** | 0.002011 | 0.16453 | 0.157119 | 0.036856 | 0.099987 | 0.088336 |
| **omega-3** | 0.008863 | 0.04487 | 0.049111 | 0.155971 | 0.067336 | 0.038445 |
| **fruits** | 0.024946 | 0.007929 | 0.005742 | 0.004883 | 0.192756 | 0.003126 |
| **wholegrains** | 0.014514 | 0.079307 | 0.115052 | 0.068613 | 0.029466 | 0.043794 |
| **legumens** | 0.000943 | 0.128218 | 0.159286 | 0.000431 | 0.076946 | 0.000184 |
| **milk** | 0.004631 | 0.13236 | 0.000476 | 0.002913 | 0.007831 | 0.091233 |
| **nuts** | 0.06425 | 0.10716 | 0.053545 | 0.057451 | 0.059273 | 0.151847 |
| **procmeat** | 0.006054 | 0.086769 | 0.042766 | 0.000678 | 0.016872 | 0.072237 |
| **pufa** | 6.93E-05 | 0.029987 | 3.52E-05 | 0.223273 | 0.000581 | 0.083236 |
| **red meat** | 0.00598 | 0.002104 | 0.138592 | 3.82E-07 | 0.05576 | 8.38E-06 |
| **sodium** | 0.002311 | 0.060067 | 0.155361 | 0.036831 | 0.141137 | 0.074028 |
| **ssbs** | 0.543985 | 0.002734 | 0.000357 | 0.002663 | 0.000135 | 0.000193 |
| **transfat** | 0.001257 | 0.030775 | 0.056043 | 0.365707 | 0.158431 | 0.122688 |
| **vegetables** | 0.0077 | 0.089608 | 0.001341 | 0.007608 | 0.026765 | 0.05878 |
| | | | | | | |
| | **Normal** | **Stretched beta** | **Mirrored gamma** | **Mirrored Gumbel** | **Inverse gamma** | **Inverse Weibull** |
| **calcium** | 0.162523 | 0.065691 | 0.075873 | 0.051747 | 0 | 0 |
| **fiber** | 0.066857 | 0.174163 | 0.02075 | 0.131008 | 0.042667 | 0.015716 |
| **omega-3** | 0.064831 | 0.276366 | 0.057641 | 0.089164 | 0.147228 | 0.000173 |
| **fruits** | 0.06304 | 0.331922 | 0.110351 | 0.241891 | 0.001553 | 0.011861 |
| **wholegrains** | 0.191202 | 0.169252 | 0.063539 | 0.149856 | 0.050946 | 0.024459 |

[1] beta distribution which was stretched out from interval $[0, 1]$ to an interval $[l,\ u]$
[2] Gumbel distribution mirrored at $u/2$ such that $Y = u - X$
[3] Gamma distribution mirrored at $u/2$ such that $Y = u - X$
[4] Reciprocal of the random variable $X$ of a gamma distribution with $Y = X^{-1}$
[5] Reciprocal of the random variable $X$ of a Weibull distribution with $Y = X^{-1}$

| | | | | | | |
|---|---|---|---|---|---|---|
| **legumes** | 0.129665 | 0.256996 | 0.166263 | 0.079496 | 0.000455 | 0.001118 |
| **milk** | 0.001299 | 0.667469 | 0.008749 | 0.082666 | 0.000373 | 0 |
| **nuts** | 0.010782 | 0.241769 | 0.00363 | 0.137205 | 0.088869 | 0.02422 |
| **procmeat** | 0.124287 | 0.299958 | 0.103835 | 0.23233 | 0.013636 | 0.000578 |
| **pufa** | 0.150751 | 0.004999 | 0.373229 | 0.133839 | 0 | 0 |
| **red meat** | 0.006746 | 0.000332 | 0.629045 | 0.161413 | 1.94E-05 | 6.45E-08 |
| **Sodium** | 0.200173 | 0.095126 | 0.088303 | 0.066555 | 0.04171 | 0.038398 |
| **Ssbs** | 0.020046 | 0.153331 | 0.062552 | 0.213942 | 2.51E-05 | 0 |
| **Transfat** | 0.059806 | 0.113474 | 0.014176 | 0.077645 | 0 | 0 |
| **Vegetables** | 0.171636 | 0.279544 | 0.201525 | 0.142473 | 0.004587 | 0.008433 |

The use of complicated statistical distributions in an ensemble model is a limitation of the GBD 2017 model, risks overfitting, and makes the emulator code slower to run. Validating that the distribution code itself worked for the individual distributions was a time intensive exercise for the authors. No sensitivity analysis is provided by the GBD that simpler distributional formats might yield similar numerical results and be an equally statistically significant fit to the original data sources, nor that the shape of the distributions derived are indeed valid to use across countries and demographic subpopulations.

#### *2.1.1.4 Result*

At the end of processing input data, the code calculates and stores a numerical approximation of the two-parameter probability density function of intake

$$P_{jcast}(\mu_{jcast}, \sigma_{jcast})(x), \qquad x \geq 0$$

where $\mu_{jcast}$ is the mean intake and $\sigma_{jcast}$ is the standard deviation of intake for dietary risk factor $j$ in the subpopulation $(c, a, s)$ of a country $c$, age group $a$ and sex $s$ in a year $t$. The density function is supported on a finite range $[l_{ja}, u_{ja}]$ specified by some upper and lower limits described in the previous section. Conceptually and for formulas there is no loss in treating the probability density function generally as defined on the intake interval $[0, \infty)$.

### 2.1.2 Disease outcomes and relative risk from exposure

Disease outcomes attributable to dietary intake can be summarised into three different categories. The largest contribution are cardiovascular diseases of which intracerebral haemorrhage, ischaemic heart disease, ischaemic stroke and subarachnoid haemorrhage were included in the GBD study. The second largest group includes diabetes mellitus type II as well as kidney diseases, which were not further specified in the GBD. Furthermore, the occurrence of neoplasms can be influenced by food consumption including the occurrence of colon and rectum cancer, stomach cancer, oesophageal cancer, larynx cancer, lip and oral cavity cancer, nasopharynx cancer, other pharynx cancer, as well as tracheal, bronchus, and lung cancer.

*Table 3: Quantities associated with impact - health outcomes related to dietary intake as used by the GBD 2017 study.*

| **Disease category** | | **Diseases associated with dietary intake** | |
|---|---|---|---|
| **Index** | **Name** | **Index** | **Name** |
| B.2 | Cardiovascular diseases | B.2.2 | Ischaemic heart disease |

| | | | |
|---|---|---|---|
| | | B.2.3.1 | Ischaemic stroke |
| | | B.2.3.2 | Intracerebral haemorrhage |
| | | B.2.3.3 | Subarachnoid haemorrhage |
| B.1 | Neoplasms | B.1.1 | Lip and oral cavity cancer |
| | | B.1.2 | Nasopharynx cancer |
| | | B.1.3 | Other pharynx cancer |
| | | B.1.4 | Oesophageal cancer |
| | | B.1.5 | Stomach cancer |
| | | B.1.6 | Colon and rectum cancer |
| | | B.1.10 | Larynx cancer |
| | | B.1.11 | Tracheal, bronchus, and lung cancer |
| B.8 | Diabetes and kidney diseases | B.8.1.2 | Diabetes mellitus type 2 |

Every risk is associated with some, but not necessarily all metabolic mediators of diseases, which can be summarised as related to influences on the body mass index, the total serum cholesterol, the fasting plasma glucose, as well as the systolic blood pressure. Those mediators influence the vulnerability of the human body to the dietary related diseases as well as their likelihood. Table 4 gives an overview over which risk factors are associated with which disease outcomes. Furthermore, this risk is determined by factors such as age, sex, or socio-economic status.

*Table 4: Dietary risks and related disease outcomes as in the GBD 2017 study [2]. Some risk-disease relationships used the same relative risk function for morbidity and mortality and are listed here as 'same'. However, a few disease outcomes (for example a diet low in nuts and seeds associated with an ischemic heart disease) had individual relative risk functions for morbidity and mortality. To indicate this, they are referred to as 'differing'.*

| **Risk** | **Disease** | **Morbidity/ Mortality** |
|---|---|---|
| Diet low in fruits | Lip and oral cavity cancer | Same |
| | Nasopharynx cancer | Same |
| | Other pharynx cancer | Same |
| | Oesophageal cancer | Same |
| | Larynx cancer | Same |
| | Tracheal, bronchus, and lung cancer | Same |
| | Ischemic heart disease | Same |
| | Ischemic stroke | Same |
| | Intracerebral haemorrhage | Same |
| | Subarachnoid haemorrhage | Same |
| | Diabetes mellitus type 2 | Same |
| Diet low in vegetables | Ischemic heart disease | Same |
| | Ischemic stroke | Same |
| | Intracerebral haemorrhage | Same |
| | Subarachnoid haemorrhage | Same |

| | | |
|---|---|---|
| Diet low in legumes | Ischemic heart disease | Same |
| Diet low in whole grains | Ischemic heart disease | Same |
| | Ischemic stroke | Same |
| | Intracerebral haemorrhage | Same |
| | Subarachnoid haemorrhage | Same |
| | Diabetes mellitus type 2 | Same |
| Diet low in nuts and seeds | Ischemic heart disease | Differing |
| | Diabetes mellitus type 2 | Same |
| Diet low in milk | Colon and rectum cancer | Same |
| Diet high in red meat | Colon and rectum cancer | Same |
| | Diabetes mellitus type 2 | Same |
| | Colon and rectum cancer | Same |
| | Ischemic heart disease | Same |
| | Diabetes mellitus type 2 | Same |
| Diet low in fiber | Colon and rectum cancer | Same |
| | Ischemic heart disease | Same |
| Diet low in calcium | Colon and rectum cancer | Same |
| Diet low in seafood omega-3 fatty acids | Ischemic heart disease | Differing |
| Diet low in polyunsaturated fatty acids | Ischemic heart disease | Same |
| Diet high in trans fatty acids | Ischemic heart disease | Same |
| Diet high in sodium | Stomach cancer | Same |
| Diet high in sugar-sweetened beverages | Ischemic heart disease | Same |
| | Diabetes mellitus type 2 | Same |

The original GBD study also included four additional health outcomes related to a 'Diet high in sodium' namely ischemic heart disease, ischemic stroke, intracerebral haemorrhage, and subarachnoid haemorrhage. However, the risk estimation for those factors was done differently in the GBD study to the other risk-disease pairs listed above.

Further disease outcomes related to a 'Diet high in sodium' are presented in Table 3 of the GBD supplementary material [2]. Those diseases are hypertensive heart disease, non-rheumatic calcific aortic valve disease, other cardiomyopathy, atrial fibrillation and flutter, aortic aneurysm, peripheral vascular disease, endocarditis, other cardiovascular and circulatory disease, as well as chronic kidney disease due to diabetes mellitus type 1, diabetes mellitus type 2, hypertension, glomerulonephritis or due to other and unspecified causes. However, no direct calculation method is given for those disease outcomes in the supplementary material. Hence, these risk outcomes are excluded from the Version 0.1 of the emulator model we developed. Furthermore, as the four cardiovascular diseases mentioned above (ischemic heart disease ~7% of total DALYs and stroke ~5.3% of total DALYs) have a much higher contribution to total DALYs, than the other outcomes (largest contribution of the CVDs is hypertensive heart disease with ~0.7% of total DALYs), not including those factors will not lead to a large difference in total attributable DALYs.

# Annex A

Comparative risk assessment uses a relative risk function on intake that describe the increasing odds of disease outcomes relative to optimal intake.

For exposure of age group $a$ to intake $x \geq 0$ in dietary risk factor $j$ the relative risk of morbidity or mortality $m$ from disease $o$ is encoded by the relative risk functions $R_{joam}(x), x \geq 0$. Relative risk functions are described in the form of a parameter $rf_{joam}$ a unit parameter $\nu_j$ and the optimal intake $TMREL_j$ in the GBD 2017 study [2, 6, 18]. The relative risk function has a different shape depending on whether risk increases with over (1) or underconsumption (2):

$$R_{joam}(x) = \begin{cases} rf_{joam}^{(x-TMREL_j)/\nu_j} & x \geq TMREL_j \\ 1 & 0 < x < TMREL_j \end{cases} \quad (1)$$

$$R_{joam}(x) = \begin{cases} 1 & x \geq TMREL_j \\ rf_{joam}^{(TMREL_j-x)/\nu_j} & 0 < x < TMREL_j \end{cases} \quad (2)$$

Conceptually the parameters $\nu_j$ can be absorbed into the parameters $rf_{joam}$ (as $rf_{joam}^{1/\nu_j}$). Henceforth we consider the relative risk functions as described by $rf_{joam}$ and the optimal intake $TMREL_j$. The parameter $TMREL_j$ are available from the original GBD 2017 dietary risk study [2]. There is no loss in setting $rf_{joam} = 1$ when exposure $j$ is not associated to disease $o$ in Table 4, so that, conceptually, there are relative risk functions for all risks across all disease outcomes. The parameters $rf_{joam}$ for the GBD 2017 study are published in GBD supplemental material [18].

The form of the relative risk functions changed in GBD 2021, and the connection between risks and disease outcomes also changes in GBD iterations. From GBD 2021 the relative risk functions use a more complicated spline presentation. The effective relative risk for dietary exposure have also changed considerably over GBD iterations. This is discussed further in limitations. The emulator uses the GBD 2017 formulation due to the simpler parameterisation of the relative risk function, and to avoid the fluctuations in response to exposure over more recent iterations of the study. The emulator code can be updated to newer iterations of forming relative risk functions if the user has sufficient resources. Presently the GBD 2021 spline representation must be downloaded manually from the IHME website, and then a new code module developed to process and numerically approximate the spline function. Another complication of using the GBD 2021 formulation is the dietary risk module is less standalone, several of the dietary risks now require mediation through other metabolic risks rather than having their own relative risk function.

Disease association and relative risks was assessed by the GBD Burden of Proof (BoP) process [19, 20] which uses a standardised meta-regression framework that models study heterogeneity to produce the Burden of Proof Risk Function (BPRF), or the most conservative effect size that is still supported by the evidence. For harmful exposures, associated with overconsumption, the BPRF uses the lower bound of the uncertainty interval, and for protective exposures, associated with underconsumption, the upper bound is used. This ensures that the association is not overstated in either direction. The overall quality of the evidence is then rated from 1 to 5 stars, making it easier to compare how strong and reliable the evidence is across different risk-disease pairs.

### 2.1.3 Population disease burdens as years of life lost to mortality and morbidity

Relative risk of exposure provides a fractional attribution. To estimate an attributable disease burden in a comparative risk assessment, the fraction of the disease outcome attributed to exposure needs to be multiplied by the burden of the disease outcome. Years Lived with Disability (YLD) and the Years of Life Lost (YLL) by age, sex, country and year are measures of disease burden taken from the GBD study [6]. YLLs are estimated by the GBD and the WHO by adding up each death multiplied with the standard life expectancy at each age [21]. The causes of death were taken from the respective GBD database containing Vital Registration (VR) data, Verbal Autopsy (VA) data (both corrected for garbage coding and VR also corrected for under-registration), as well as surveys and surveillance systems for specific causes.

Based on epidemiological surveillance data, and the International Classification of Diseases (ICD-9 and ICD-10), the GBD mapped non-fatal causes, impairments and nature of diseases and estimated prevalence and incidence by using the Bayesian meta regression tool DisMod-MR 2.1 [18, 22]. The YLDs were calculated by multiplying the prevalence of a disease (number of patients) with a disability weight.

For convenience of notation, we denote YLLs and YLDs from disease outcome $o$ by

$$YL_{omcast}$$

in the subpopulation $(c, a, s)$ of a country $c$, age group $a$ and sex $s$ in a year $t$. When $m$ =mortality then YLLs are denoted and when $m$ =morbidity then YLDs are denoted.

## 2.2 Emulator model

### 2.2.1 Population attributable fraction

To derive the Population Attributable Fraction (PAF) of YLLs or YLDs resulting from the dietary intake within a given country $c$, age group $a$ and sex $s$ in the year $t$, the intake distributions are weighted by their relative risk factors and integrated for any intake $x$ between the lowest level of intake $l_{ja}$ and highest level of intake $u_{ja}$ for each risk $j$- disease $o$ pair.

$$PAF_{jomcast} = 1 - \frac{1}{\int_0^\infty R_{joam}(x)\, P_{jcast}(\mu_{jcast}, \sigma_{jcast})(x)\, dx} \qquad (3)$$

Here $P_{jcast}(\mu_{jcast}, \sigma_{jcast})$ is the probability distribution of intake for risk factor $j$ supported on the range of corresponding intake $[l_{ja}, u_{ja}]$ and parameterised by mean intake $\mu_{jcast}$ and standard deviation of intake $\sigma_{jcast}$, and $R_{joam}(x)$ is a relative risk function supported on the range $(0, \infty)$ that indicates an increased risk of morbidity or mortality disease outcome $(o, m)$ at age $a$ if intake in category $j$ is $x$ units. For a diet low in intake of a specific good, intake above the Theoretical Minimum Risk Exposure Level for adults ($TMREL_j$ - Table 1) is considered sufficient consumption and therefore assigned a relative risk factor value of 1. In analogy for any diet high in intake, any value below $TMREL_j$ is given a relative risk factor of 1. If the subpopulation dietary intake probability distribution $P_{jcast}$ is fully supported in the interval above or below TMREL, respectively, then $PAF_{jomcast}$=0. None of the disease outcomes are then attributed to dietary intake. Up to a proportion of 1, $PAF_{jomcast}$ attributes the fraction of death and disability outcomes in disease $o$ to exposure to dietary risk *j* in the subpopulation *cas* in year *t* outside the TMREL interval.

The PAFs for a given risk-disease pair are the fraction of total DALYs associated with a given disease because of the intake risk. For summing up across different dietary risk- disease pairs to get the total PAF associated with food consumption a mediation matrix is employed. The mediation factor $MF_{jio}$ is the pairwise evaluation of a dietary risk $j$ mediated through risk $i$ for disease $o$. To avoid double counting the attributed DALYs when risk factors share similar effects, the PAF for a joint exposure to the dietary risk factors $J \subset \{1, \dots, 15\}$ is derived according to Equation (4). The mediation factors are published for the 2017 study in the supplemental information of [2]. The main mediations within the set of dietary risk factors is that attributed outcomes from low calcium intake are mediated by low milk intake, and the attributed outcomes from low fibre intake are mediated by a combination of whole grains, fruit and vegetable intake.

$$PAF_{Jomcast} = 1 - \prod_{j=1}^{J} \left( 1 - PAF_{jomcast} \prod_{i=1}^{J} \left(1 - MF_{jio}\right) \right) \tag{4}$$

### 2.2.2 Attributable disease burden

Equation (3) describes the calculation of a population attributable fraction $PAF_{jomcast}$ for morbidity or mortality disease outcome $(o, m)$ due to exposure to a dietary risk $j$ by intake in the subpopulation $(c, a, s, t)$ of country $c$ at time $t$ with age range $a$ and sex $s$.

To calculate the burden of disease attributed to exposure to a dietary risk $j$ in disability adjusted life years (DALYs), the years of life lost to a disease $o$ in country $c$ in individuals of age $a$ and sex $s$, in year $t$ are multiplied by the fraction attributable to the risk factor $j$, and added to the years of life lost to disability to a disease $o$ in country $c$ in individuals of age $a$ and sex $s$, in year $t$, multiplied by the fraction attributable to the risk factor $j$.

$$DALY_{jocast} = \sum_{m} PAF_{jomcast} \cdot YL_{omcast} \tag{5}$$

Equation (4) describes the calculation of a population attributable fraction $PAF_{Jomcast}$ for morbidity or mortality disease outcome $(o, m)$ due to exposure to a joint set of dietary risks $J$ by intake in the subpopulation $(c, a, s, t)$ of country $c$ at time $t$ with age range $a$ and sex $s$.

To calculate the burden of disease attributed exposure to a set of dietary risk factors in disability adjusted life years (DALYs), the years of life lost to a disease $o$ in country $c$ in individuals of age $a$ and sex $s$, in year $t$ are multiplied by the fraction attributable to the joint risk factors $J$, and added to the years of life lost to disability to a disease $o$ in country $c$ in individuals of age $a$ and sex $s$, in year $t$, multiplied by the fraction attributable to the joint risk factors $J$.

$$DALY_{Jocast} = \sum_{m} PAF_{Jomcast} \cdot YL_{omcast} \tag{6}$$

The user should note the difference in indexing in equation (5) and (6). In equation (6) $J$ denotes a set of dietary risk factors. It is true that $DALY_{jocast}$ in equation (5) is equal to $DALY_{\{j\}ocast}$ in equation (6), however it is not true that the individual attributable disease burdens $DALY_{jocast}$ sum to $DALY_{Jocast}$. In general,

$$DALY_{Jocast} \leq \sum_{j \in J} DALY_{jocast}$$

is sub-additive. **Therefore, the user should take care interpreting and aggregating $DALY_{jocast}$ in the absence of other dietary risk exposure.** See Appendix E for a normalised version of the attributable disease burden from single intake exposure that is additive to the joint exposure.

## 2.3 Description of outputs

### 2.3.1 Aggregate attributable disease burden from current exposure to dietary risks

Aggregating across dietary risks (and other risk categories and exposure not considered in the stand-alone model such as high body mass index) is sublinear. However, the GBD model framework considers that disease burdens are exclusive for mortality and morbidity and exclusive between subpopulations (country age and sex). The emulator model default output format is national disease burden from exposure to dietary risks:

$$DALY_{Jct} = \sum_{o,m,a,s} PAF_{Jomcast} \cdot YL_{omcast}. \tag{7}$$

Though disease burdens in subpopulations from joint exposure are considered exclusive, and added, this does not imply that they should be considered independent.

### 2.3.2 Attributable disease burden from shifting exposure to dietary risks

Populations attributable fractions are dependent on current exposure. The emulator had setup code that allows the user to shift a baseline current exposure. The emulator's default shift is to move the entire distribution of intake in a subpopulation by an amount $h \geq 0$,

$$PAF_{jomcast}(h) = 1 - \frac{1}{\int_0^\infty R_{joam}(x)\, P_{jcast}(\mu_{jasct}, \sigma_{jasct})(x-h)\, dx}.$$

A sufficient but not necessary assumption for this shift is that every individual in the population increases their dietary intake associate to the risk factor $j$ by $h$. This shift is relatively easy to compute because substitution in the integral provides that

$$PAF_{jomcast}(h) = 1 - \frac{1}{\int_0^\infty R_{joam}(x+h)\, P_{jcast}(\mu_{jasct}, \sigma_{jasct})(x)\, dx}$$

and the model does not need to recompute distributions only shift relative risk. An alternative to shifting is rescaling intake. For exposure to diets low in beneficial intake, then shifting decreases the attributable fraction. For exposure to diets high in harmful intake, then shifting increases the attributable fraction.

The code has flexibility that the custom user input can designate shifts $[h_{jcast}]_{j\in J, cast\in S}$ in a set of subpopulations $S$ and exposures $J$ individually, resulting in attributable fraction to shifted joint exposure that is a vector-valued function for fixed $cast \in S$

$$PAF_{Jomcast}\left([h_{jcast}]_{j\in J}\right) = 1 - \prod_{j=1}^{J}\left(1 - PAF_{jomcast}(h_{jcast}) \prod_{i=1}^{J}(1 - MF_{jio})\right).$$

Without loss, by just setting the shift to be 0 for countries and times without a designated shift the change in the attributable disease burden due to the shift from current exposure is

$$DALY_{Jct}\left(\left[h_{jcast}\right]_{j\in J,cast\in S}\right) - DALY_{Jct} = \sum_{o,m,a,s}\left(PAF_{Jomcast}\left(\left[h_{jcast}\right]_{j\in J}\right) - PAF_{Jomcast}\right)\cdot YL_{omcast}. \quad (8)$$

If the user designated shifts are non-trivial, then the emulator will output the attributable disease burden due to current exposure and the change in the attributable disease burden due to the shift from current exposure. Note that this assumes the overall disease outcomes for the population are not changed by the shift, which is one of the caveats of comparative health risk assessment.

### 2.3.3 Marginal disease burden from shifting exposure to dietary risks

Though the user can designate the shifts for each subpopulation and exposures, studies are mostly concerned with a subset of two more specific shifts.

The first is the change of disease burden per unit shift in national exposure. For example, what is the increase in attributable disease burden if the national population intake in g/capita/day of red meat consumption increases by 1 g/capita/day and fruit and vegetable consumption decreases by 1 g/capita/day?

For the first type of study, for a vector $\eta = \left[\eta_j\right]_{j\in J}$ the user could set a shift of

$$h_{jcast}(\eta) = \eta_j$$

assuming that increase in intake occurs equally across all subpopulations and calculate the change in attributable burden per last section. If only one intake is being shifted, say red meat consumption, and the shift in national intake is not equal in subpopulations then the user can use a variant of the second type of study.

In the first type of equally across all subpopulations shift case the national attributable disease burden to joint exposure becomes a vector valued function on the space of shifts:

$$DALY_{Jct} : \mathbb{R}^{15} \to \mathbb{R}$$

and change in disease burden for small shifts in exposure is approximated by the partial derivatives

$$DALY_{Jct}(\eta) - DALY_{Jct} \cong \sum_{j\in J}\frac{\partial DALY_{Jct}}{\partial \eta_j}(0)\ \cdot\ \eta_j \quad (9)$$

The final output we designed in the emulator is the calculation of partial derivatives of the attributable disease burden to unilateral shift in intake in one of the exposures. The partial derivatives allow pre-calculation of values that can distributed to other users, so they can calculate a change in disease burden without them having to run the emulator.

The second type of study is the change in disease burden for a shift in intake that is distributed in a specified way amongst different exposures and subpopulations. For example, what is the increase in attributable disease burden for a 1% increase in energy intake in adults under 65? Or, what would be the increase in disease burden if national red meat intake decreased by 10 g/capita/day where most of decrease occurred in younger adult age groups compared to older age groups?

For the second type of study, there is a single parameter $\eta$ and the shifts $h_{jcast}(\eta)$ are a function of this parameter. The change in attributable disease burden to a small change in the parameter is approximated by

$$DALY_{Jct}(\eta) - DALY_{Jct} \cong \frac{dDALY_{Jct}}{d\eta}(0) \cdot \eta$$

where,

$$\frac{dDALY_{Jct}}{d\eta}(0) = \sum_{k,a,s} \frac{dh_{kcast}}{d\eta}(0). \frac{\partial}{\partial h_{kcast}} \sum_{o,m} \prod_{j=1}^{J} \left( 1 - PAF_{jomcast}(h_{jcast}) \prod_{i=1}^{J} (1 - MF_{jio}) \right) \cdot YL_{omcast} \Bigg|_{\{h_{jcast} = 0 \,:\, j \in J\}}$$

The factors $\frac{dh_{kcast}}{d\eta}(0)$would normally be known to users by design, they are weights that specify how the energy increase is initially distributed among exposures in subpopulations. Or, how national intake in one variable is distributed amongst age and sex stratified subpopulations.

For fixed $cast$ the partial derivatives of the vector valued function

$$h_{jcast} \mapsto \sum_{o,m} \prod_{j=1}^{J} \left( 1 - PAF_{jomcast}(h_{jcast}) \prod_{i=1}^{J} (1 - MF_{jio}) \right) \cdot YL_{omcast}$$

evaluated at 0 (current exposure) are calculated in the emulator code as part of calculating the output

$$\frac{\partial DALY_{Jct}}{\partial \eta_j}(0).$$

Therefore, a small adjustment in the emulator code to capture the partial derivatives within iteration at the subpopulation level would allow a user to conduct studies of the second type.

The partial derivates are not numerically approximated from first principles. The emulator code calculates them from the analytic formulas derived in Appendix D. The analytic formulas were separately verified however using first principles numerical approximations.

## 2.4 Monte Carlo simulation of attributable disease burden due to exposure to dietary risk factors

The emulator has a Monte Carlo simulation variant that reproduces the GBD 2017 estimation of uncertainty of the burden of disease by treating the model inputs as random variables. By taking samples of the inputs from Section 2.1, and calculating outputs given the input samples using the formulas in Section 2.2, the emulator stores the outputs as the samples of random variables of attributable disease burden (or change due to shift, or marginal attributable disease burden, etc.).

### 2.4.1 Inputs as random variables

In the Monte Carlo simulation variant of the emulator, the inputs must be provided in a file format that contains 1000 samples. The inputs files with only 1 sample are equivalent to running the emulator in "deterministic" mode.

The exposure inputs $\mu_{jcast}, \sigma_{jcast}$ , the response parameters $rf_{joam}, TMREL_j,$ and the population disease outcomes $YL_{mocact}$ are published by the GBD with a central value and a lower and upper value representing the 95% uncertainty interval. The shape of the distribution of the input parameters is not discussed or further specified in published information. Uncertainty estimates (given as 95 percentile interval) in YLL were calculated 'based on 1000 draws from the posterior distribution of cause-specific mortality by age, sex, and location for each year' [18]. Uncertainty is given as 97.5$^{th}$ and 2.5$^{th}$ percentile and were calculated from 1000 samples of the disability weight, as well as the posterior distribution [18, 22].

To check our reproduction of the GBD 2017 model for burden of disease from dietary risk factors with uncertainty against published data we reproduced 1000 input samples using maximal entropy distributed random variables on an interval given the information available. This is described in Appendix A.. A user can also provide central values, and lower and upper values of the 95% uncertainty interval and use our code to generate 1000 input samples form a maximal entropy distribution.

Due to the complication of parametric dependence of population intake distributions on maximum $u_{ja}$ and minimum $l_{ja}$ values of intake from the choice of using stretched beta, mirrored Gumbel and mirrored gamma distribution, maximum and minimum values of intake (mean intake and standard deviation of intake) were adjusted during the sampling. They were first calculated as relative values $\hat{u}_{jcast}$ for each risk factor $j$, country $c$, age $a$ and sex $s$ specific mean value from published data $\mu_{jcast}$

$$\hat{u}_{jcast} = \frac{u_{ja}}{\mu_{jcast}}.$$

When a new sample of mean intake was taken, then new minimum and maximum value for dietary intake was derived by multiplying the sample of mean intake by the factor $\hat{u}_{jcast}$ (or $\hat{l}_{jcast}$ respectively).

### 2.4.2 Joint sampling of exposure, response and population disease burdens

The modelling and the underlying data have large uncertainties. For many countries data separated by age and sex in the dietary intake categories of the Global Burden of Disease study does not exist. Data on intake is derived by a modelling step in the Global Burden of Disease study.

The complete set of random inputs

$$\mu_{jcast}, \sigma_{jcast}, rf_{joam}, TMREL_j, YL_{jmocast}, MF_{jio}$$

is in the order of greater than 2*15*30*160+2*13*15*15+15+15*2*13*160*30 = 2,021,865 elements not considering temporal correlation and mediation. Describing non-independent joint sampling of this many random variables is difficult. Even though the parameters are likely not independent, the emulator follows the GBD study and take 1000 joint samples of model inputs for Monte-Carlo simulation. The number of joint samples is limited by the speed of calculations, and it is

unlikely that the tail of the distribution of national disease burdens is well explored by the limited number of samples.

By following the GBD study the emulator is potentially underestimating the uncertainty and risk from higher disease burdens. It could be argued that the inputs for exposure

$$\mu_{jcast}, \sigma_{jcast}$$

response to exposure

$$rf_{joam}, TMREL_j, MF_{jio}$$

and population disease outcomes

$$YL_{jmocast}$$

represent independent subsets of random variables. This argument would be an extension of the basic caveats of comparative risk assessment. Large variation in exposure of a significant preventable disease burden being independent of the population total disease burden is probably the main limitation on this separation argument.

Supposing exposure inputs, response inputs, and population disease inputs do represent independent subspaces for fixed subpopulations $cast$ and disease outcome $o$, it should be noted that these inputs are used serially (multiplication of terms) in the comparative risk formulas. The serial nature in the joint population attributable fraction also makes it more likely that the random variable

$$DALY_{Jocast} = \sum_{m} PAF_{Jomcast} \cdot YL_{omcast}$$

exhibits log-normal type behaviour as a distribution due to central limit theorems [23]. Examining samples from the emulator of these components confirms distributions that are right skewed with "fatter" tails toward higher disease burdens.

The larger issue for assuming independent sampling concerns independence in inputs between subpopulations $as$ and disease outcomes $o$ rather than across exposure, response and disease inputs. The assumption of exclusivity means that the random variable of national disease burden from exposure to dietary risks

$$DALY_{Jct} = \sum_{o,a,s} DALY_{Jocast}$$

is the sum of around 300 independent random variables. Irrespective of the log-normal type behaviour of the individual terms, the summation across so many subpopulations and disease treated as independent gives the random variable of national disease burden from dietary risks a normal distribution shape. This also is confirmed by examining samples from the emulator.

The concern is that the normal shape of the national disease burden underrepresents risk by assuming independence. Correlation in the attributable disease burden across age groups and sex and disease outcomes can skew the shape of the sum as the central limit theorem does not generally apply for summing dependent random variables [24]. Correlation is a measure of when samples of one random variable in the summand are above the mean, how many samples and by

how much would samples of the other random variable in the summand expected to be above the mean. With strong correlation in summands the sum can be higher than without.

We should expect correlations in exposure measurement error across age groups and sex resulting from cultural and social factors that persist across age and sex stratified populations within countries. For example, if traditionally the national diet is based on one specific food category (for example bread or rice) with a certain preparation method. Or prevalence of ultra-processed food consumption across all age groups. Intake is systemically over or underestimated in nutrition and health questionnaires across all age groups and for both sexes [25-28]. Total YLLs and YLDs are influenced not only by dietary risk factors, but also by a wide range of concurrent external influences that affect all subpopulations. These include environmental exposures [29], levels of physical activity [30], broader lifestyle, behavioral patterns and genetics [31]. Since these factors jointly shape population health, correlations in YLL and YLD across subpopulations are expected and should be interpreted within this context. Reporting bias on death certificates can also create dependence across age groups and sexes [32]. Co-morbidities and collider bias can create correlations in YLL and YLD estimates for individual disease [33, 34].

# 3 Validation of the emulator

While it was relatively easy to download individual risk PAFs from the GBD results tool [1] to validate the emulator model against published results, comparing the disease burdens attributed to dietary risks by the emulator model and those published by the GBD is more difficult[6]. The emulator calculates disease burdens from dietary exposure for which there are direct relative risk factors. The published GBD results for sodium and thereby the disease burden from joint exposure dietary risks includes additional sodium disease pathways mediated by high systolic blood pressure for which there are no direct relative risk factors. To confirm disease burdens, and to also compare Monte Carlo simulation performed by the emulator against the GBD model simulation, requires adjusting the emulator output for the sodium pathways mediated by high systolic blood pressure.

### 3.1.1 Validation of population attributable fractions

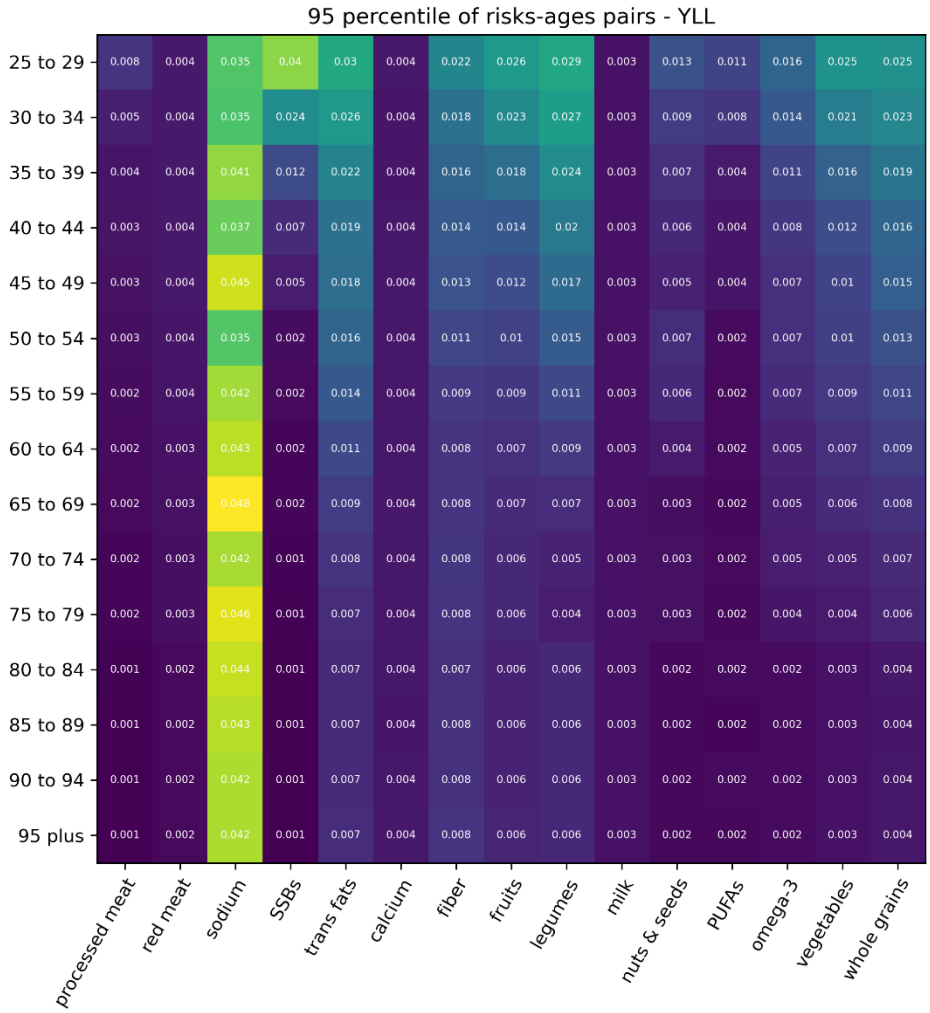
To validate our implementation of the 2017 GBD model of dietary intake to burden of disease outcomes we examined the error between the emulator model individual dietary intake exposure PAFs and the published values that were available at the GBD website [3].

For every risk factor, disease, country, mortality or morbidity, age group and sex combination of the individual dietary intake PAF value was calculated as presented in Equation (3). The differences to the PAF values as available from the GBD results [1, 3] are displayed in Figure 2 (for further figures see Appendix C). Note that the high dimensionality of the index means that only two indexing pairs were chosen for display, e.g. risk-disease pairs, and a histogram of the absolute value of the difference across the other four index dimensions was formed. Figure 2 reports the 95-percentiles of the histograms. It is interpreted, for example in Figure 2, for a diet low in fruit (risk) in ages 30-34

[6] GBD 2017 results were downloaded when still available in the IHME results tool. GBD 2017 results are currently archived and require a request to IHME. The Zenodo repository for the emulator (https://zenodo.org/records/18259172) has the data required to validate the emulator for dietary risk exposure, population attributable fractions and disease burden in DALYs.

(age group), in less than 5% of the emulated PAFs indexed further by disease, country, mortality or morbidity, and sex differ from the published PAF by greater than 0.023.

*Figure 2: Individual dietary intake PAF discrepancies for mortality. The graphic shows a fixed age group a and dietary risk j. The set of absolute PAFs differences,* $\{|\, PAF^{*}_{jomcast} - PAF_{jomcast} |\}_{o,s=0,1,c=1,\ldots,160,t=2017,m=mortality}$ *contains over 3200 values for fixed a and j (the order of the set depends on j since there are a differing number of disease outcomes for each risk factor) where* $PAF^{*}_{jomcast}$ *denotes the GBD published value. The values in the figure refer to the 95th percentile of the sets of absolute PAF differences for fixed age and dietary risk. For example, for a diet low in fruit in ages 30-34, in less than 5% of cases the emulated PAF differs from the published PAF by greater than 0.023.*




In general, the PAF estimates from the emulator model were in good agreement with the published PAF values. As the distribution is reconstructed from missing data on upper and lower limits, and both the published values and the emulator use numerical integration and numerical parameter reconstruction for the ensemble of intake distributions, we do not expect the PAFS to be in exact agreement.

However, for sodium there was a systemic discrepancy visible. The outcome of stomach cancer from diets high in sodium is persistently 0.035-0.045 higher in over 5% of cases across all age groups. In the documentation of the GBD diet model [2], the sodium risk factor and unit for the calculation were not given as available for all other risk factors. Instead, there were shifts ‘reported for diet high in sodium as the estimation is based on mediation through high systolic blood pressure’ [2].

Furthermore, the 'Diet high in sodium' was associated with four additional disease outcomes, ischemic heart disease, ischemic stroke, intracerebral haemorrhage, and subarachnoid haemorrhage which were not calculated as presented in Eq. (3), but were mediated through systolic blood pressure [6]. PAFs from those four disease outcomes associated with a 'Diet high in sodium' were not included in our reproduction of the GBD model. We could not see it described in the full mediation matrix for all exposure in the GBD 2017 supplemental information (Table 7), but there may be a missing mediation in the emulator or an adjustment for race of subpopulation groups that is not emulated. The bias toward higher sodium disease outcomes is visible in the next section and contributes toward a 0%-1% higher estimation of national disease burden from joint exposure to dietary risks.

There is also a systematic difference across age groups, likely because of the higher intake and higher standard deviation of intake in younger groups. This increases the potential differences between the distribution approximation in the emulator and the original distributions in the GBD model.

### 3.1.2 Validation of national attributable disease burden from joint exposure

The absolute error in population attributable disease fractions for single dietary risks does not translate directly into a percentage deviation of the national attributable disease burden from joint exposure to dietary risks. Some PAFs may be very small, and if the original PAF is 0.01 and the emulator is 0.02, the percentage error in attribution is large. However, if the disease burden being attributed by this PAF does not have many years of life to morbidity or mortality then the difference the PAF attribution error makes to the overall national disease burden calculation in Equation (6) will be small. What is relevant for the studies the emulator is designed for is the national attributable disease burden from joint exposure to dietary risks.

Validating the emulator calculation in Equation (6)

$$DALY_{Jocast} = \sum_{m} PAF_{Jomcast} \cdot YL_{omcast}$$

against the published values tests the estimation of joint attribution and the correct use of disease burdens for individual disease outcomes.

The results from the GBD published study include four additional disease outcomes related to a 'Diet high in sodium' (ischemic heart disease, ischemic stroke, intracerebral haemorrhage, and subarachnoid haemorrhage), which were not calculated as presented in Equation (3). Instead, the PAFs were mediated through systolic blood pressure [2]. The exact calculation method was not available through the supplementary material of the published study and those four disease outcomes for sodium mediated by exposure to high systolic blood pressure were excluded from the emulator. In order to validate the emulator DALYs output against the published study, the individual attributable fractions of diets high in sodium for ischemic heart disease, ischemic stroke, intracerebral haemorrhage, and subarachnoid haemorrhage need to be inserted into the joint exposure formula in Equation (4).

To test agreement with the published disease burdens for joint exposure to dietary risks, the emulator model calculated the individual dietary intake PAFs and then used the PAF of each of the four extra factors for sodium as available through the GBD results tool [1]. Figure 3 and Figure 4 show the relative differences of the DALYs associated with dietary risks calculated with the emulator

outputs to the values published by the GBD 2017 for the year 2017 – and the issue of validating within including the additional sodium factors. For emulator inputs, each model factor (rfs, TMRELs), as well as mean and standard deviations of exposures, and the total population disease burdens for disease outcomes in YLDs and YLLs the published mean values were used. The emulator output of attributable disease burden in DALYs with additional sodium factors is compared to the published attributable DALYs on a percentage basis.

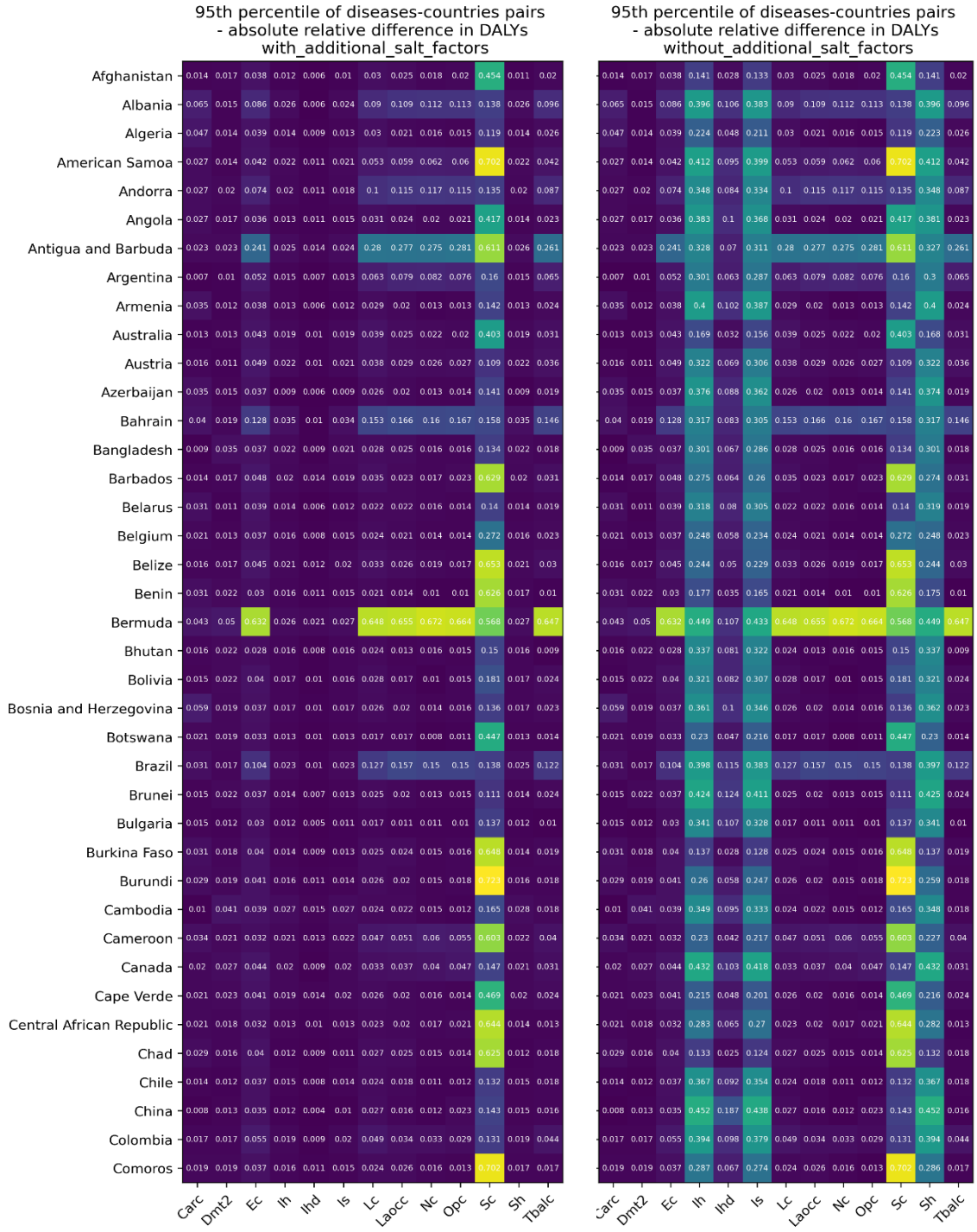
For every disease, country, age group and sex combination the absolute value of the relative difference between published values and disease burden for joint exposure to dietary risks in DALYs was calculated using the emulator as presented in Equation (6). Figure 3 shows the absolute value of the relative difference between published values and emulator values. The two indices chosen for display are country-disease pairs, which means that a histogram across the other indices contains 30 values (across age-sex subpopulations) and the 95-th percentile of the histogram is displayed. It is interpreted, for example in Figure 3, for China for Ischemic heart disease less than 5% of the emulated disease burdens across age-sex differ from the published disease burdens by more than 0.5%. Appendix C contains additional figures covering all countries examined.

The emulator output compared to the published results without the additional sodium pathways for ischemic heart disease, ischemic stroke, intracerebral haemorrhage, and subarachnoid haemorrhage mediated by high systolic blood pressure is shown in the right panel of Figure 3 for comparison. Stand-alone emulators of joint exposure to dietary risks cannot be validated against published results without broadening the scope of diets high in sodium.

The total national disease burden from joint exposure to dietary risks is calculated in the emulator by Equation (7). Figure 4 shows the percentage and absolute difference between emulator national disease burdens and published values. The national disease burden involves adding the disease burdens across disease outcomes in age and sex subpopulations. National disease burdens from joint exposure to dietary risks are the main output of the emulator, and validation shows that emulator calculations and published values agree to within ±2%.

Examining Figure 3 and Appendix C for cancer outcomes, small nations and island countries such as Antigua and Barbuda, The Bahamas, Bermuda, Virgin Islands, St Kitts, Cuba, Dominican Republic, Guam, Luxembourg, Montenegro, and African countries Rwanda, Ghana and Uganda, show outlier deviations in the 95-th percentile from other countries. On examination of absolute and not relative differences this is likely to be due to small disease burdens (< 30 DALYs) inflating percentage errors and difference in distributions of intake due to the complicated ensemble approach. The cancer outcomes displaying the deviation are associated to diets low in fruit, and the distribution for fruit intake is most heavily weighted by the stretched beta distribution, mirrored Gumbel, and mirrored Gamma. Intake distributions for diets low in fruit are therefore most heavily affected by the choice of upper and lower limits of intake and the additional processing that was required to correct the data on upper and lower limits provided by IHME. Many of the outlier countries on the cancer outcomes have high fruit intake and likely standard deviations. Middle East and North African countries where dates are highly consumed, such as Morocco, Iran, Oman, Qatar and Bahrain showed similar deviations. Brazil also shows higher errors on attributable cancer burdens.

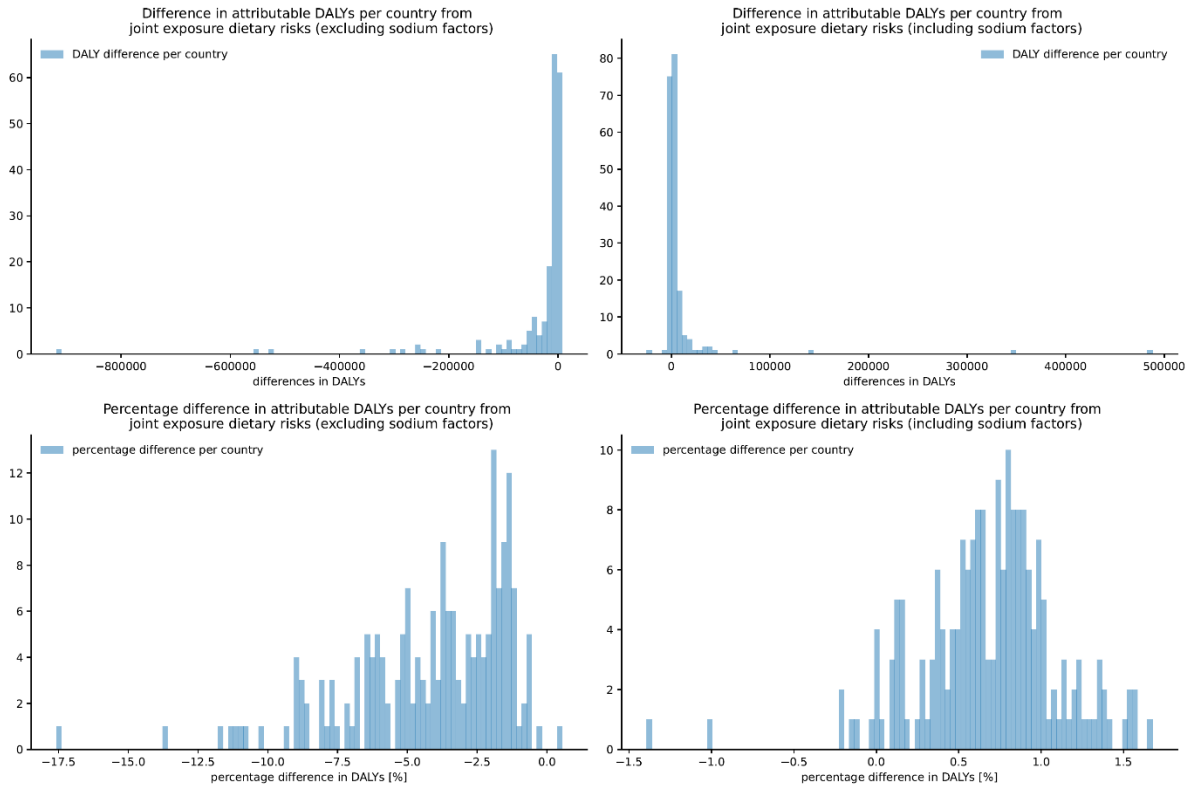
*Figure 3: Relative discrepancies from published values in DALYs disease burdens $DALY_{Jocast}$ from joint exposure to dietary risks. Shown is 95th percentile of the absolute relative difference in total DALYs per country and disease outcome (only the first 40 countries are displayed). All other countries are available in Appendix C.*




Even though discrepancies exist in younger age groups, and some cancer outcomes from fruit and sodium intake, summing across disease outcomes and subpopulations shows in Figure 4 that emulator calculations and published values on national disease burden of joint exposure to dietary

risks $DALY_{Jct}$ agree to within ±1.5%. Most countries are in the range of 0-1% including the United States, China, Russian Federation, India, and most countries in the EU, and the United Kingdom.

*Figure 4: Differences between emulated national attributable disease burden from exposure to dietary risks $DALY_{Jct}$ calculated total associated DALYs and published results in the GBD 2017 study. The left panel omits the diet high in sodium attributable fraction for disease outcomes, ischemic heart disease, ischemic stroke, intracerebral haemorrhage, and subarachnoid haemorrhage mediated through systolic blood pressure and outside the stand-alone model, and the right-panel when the diet high in sodium factors are included for comparison with published results. The histograms count the number of countries in the band of absolute (upper figures) or percentage error (lower figures) for national disease burden.*




Errors for estimated cancer outcomes due to diets high in sodium, and to a more minor degree fruit intake, create a persistent overcorrection by 0-1% (which is visible as a bias in the right panel of Figure 4). By not including high systolic blood pressure and other mediators, the GBD 2017 study considers the stand-alone emulator to be missing up to 17% (worst case China, and mostly in other East and South-East Asian countries) of the disease burden that should be attributed to diets high in sodium through a more complicated mediation calculation that is performed in the full GBD model across all risks and exposures.

There is a myriad of other possible deviations from published results such as inputs to the published results by IHME differing from the input data available for download from IHME. Including updates to mean dietary intakes, and errors in standard deviations data provided. For example, data downloaded by authors from the GBD data website on mean dietary intake for SSBs in the first instance was incorrect and subsequently corrected by a request to IHME.

# 4 Limitations

Long term studies on the influence of dietary intake and risk disease outcome on this level of specification and global extent do not exist to validate the GBD results. The emulator model repeats the comparative risk assessment methodology for joint exposure. The emulator validation is performed against the published results, it is not a statement about the validity of the GBD input data on exposure, responses or population disease burdens. Limitations in input data from the GBD study if it is used alongside the emulator and mediation are discussed below.

The limitations of the emulator on this basis are the reconstruction of distributions of intake based on an ensemble model, particularly the emulation of the beta and mirrored distributions which are sensitive to upper bounds of intake. By making the emulator stand-alone on dietary risks, the user can implement and input their own population distribution estimates. The emulator, for stand-alone reasons, does not incorporate mediated effects of diets high in sodium. The emulator also uses mediation to remove or reduce double counting in the joint dietary risk exposure. The default published mediation among the single intake categories is limited to removing fibre attribution in the presence of fruit, vegetable, and whole grain intake, and milk attribution in the presence of calcium. The user can implement their own mediation matrices in the emulator.

The emulator is based on the GBD 2017 model, for reasons mentioned in limitations below. Part of this is because of the additional complexity in specifying relative risk introduced after 2019, and part of it is choice because of the instability of some of the relative risk factors for dietary risks in subsequent GBD model. Updating the emulator to use more complicated relative risk formulations makes it less flexible for other uses. However, including additional mediation from systolic blood pressure and fasting plasma glucose is to be included in an update to the emulator.

## 4.1 General limitations of the comparative risk assessment model

The Global Burden of Disease model for multiple exposures that result in preventable mortality and morbidity is a comparative risk assessment developed originally at Harvard and WHO [4, 35]. Low fruit and vegetables consumption was a risk factor in the first WHO burden of disease study published in 2004 [35]. Comparative risk assessment of multiple dietary risks featured in the model before 2011 [36]. Other comparative risk models of joint dietary exposure use the same components of exposure, response, population attributable fractions and population disease burdens but not the same values, [37, 38]. General limitations of the comparative risk assessment model are discussed in [39].

Attributable burden is relative to the baseline exposure specified by the TMRELs. Linking attributable burden above a baseline to a “preventable” disease burden implies that the TMREL level of exposure is attainable and represents lower risk than other exposures [39]. Sometimes the range of lowest risk exposure is clearer for certain outcomes, such as zero smoking and lung cancer, but is less clear for diets [20]. What exposure level could theoretically be attained to minimize risk, such as zero smoking, and what reduction in exposure can be economically or socially achieved are different. Comparative risk assessment is best used to identify risks and policy assessment alongside costs of treatment.

The separation of exposure, response, and population disease burden decouples the total disease burden from the attribution. A reduction in exposure linked to cardiovascular disease should change

both the total disease in the population as well as the reduction in the percentage attribution of a disease burden. Projecting forward reductions in disease burden over time for different exposure trajectories requires a separate projection of population disease burden. The response modelled by relative risk is usually homogenous. For dietary risks some relative risks are specified by age group, but there is no variation across national or subnational regions and no variation in response within income or racial groups. That a population has a lower disease burden to the same amount of exposure for reasons of availability of medical treatment, technology, adaptation to specific foods, exercise, etc. is reflected generally in the total population disease rates. Therefore, there is no specific dietary aspect to difference in response per unit of exposure between, say whole grain intake in Ethiopia or the UK, except the current exposure level and the population disease rates. Population comparative risk models should not be used to examine demographic subpopulations unless a user has specific relative risk studies for a subpopulation.

A comparative risk model has no information of when a disease burden will be incurred from current exposure. It estimates the disease burden incurred currently from past exposure assuming stability in past exposure at the same levels as current exposure. This can complicate economic analysis, since present value calculations of economic costs of current exposure require knowledge of disease processes over time.

Comparative risk assessment uses mediation and multiplication in joint exposure to estimate interactions between exposure pathways. Dietary risks have a relatively simple mediation format. Lower exposure to dietary risks does not imply that the population disease burden will be lowered by the amount attributed to dietary risks. Air pollution or other exposures may shorten the life or quality of life of individuals in the population even though the risk from dietary sources has been lowered. As a stand-alone model the emulator can only calculate the attribution to dietary risks. Comparative risk assessment is sensitive to the number of risks included in joint exposure unless mediation is accurate. Mediation, since it infers causality in pathways from exposure to disease outcome, is highly uncertain even though IHME attempts to derive the mediation factors from studies and evidence.

We make a final comment on socio-economic distribution of exposure and disease burden. Often the highest exposures contain a disproportionate level of the associated risk. An advantage of using distributions of intake to represent exposure for dietary risks and not just mean intake is that the contribution of the population consuming well above average exposure (in the case of diets high in) or well below (in the case of diets low in) can be formulated. However, accuracy of the representation of tails for distributions of dietary intake is problematic. First, even for countries with gold standard dietary recall, the distribution of intake is relatively unknown. In the UK, 1000 individuals or less across all age groups are sampled as part of the UK National Diet and Nutrition Survey (NDNS). The sample size has been determined for cost reasons as statistically sufficient to assess mean intake values. It is highly insufficient to understand the distribution of intake within fifteen age brackets. The ensemble model of distributions is a modelled extrapolation of intake across countries and subpopulations that is challenging to validate, and is a complication, potentially an unnecessary one, in the GBD model. When bad diet quality is a more prevalent phenomena for low income and food-insecure families [40-42], accurate representation of exposure levels and the shape and tails of intake distributions are challenging beyond just low sampling numbers.

The use of DALYs to measure disease burden has limitations noted in literature including implicit representation of disability and value [5, 43]. Disability adjusted life years are not a relative disutility

measure. Except through potential difference in exposure, there is no representation in the comparative assessment model of additional burdens for low income and food-insecure families.

The following table captures briefly the above points and other limitations from [4] and [39]:

*Table 5: List of limitations of comparative risk assessment compiled from reference [39].*

| Limitation | Implication for interpreting GBD risk-attributable burden |
|---|---|
| Normative choice of counterfactual (TMREL) | Results depend heavily on assumed theoretical minimum exposure; alternative plausible counterfactuals would yield different attributable burdens. |
| Incomplete and heterogeneous exposure data | Modelled exposure distributions may misrepresent true population exposure, directly biasing PAFs and attributable DALYs. |
| Exposure measurement error | Simplified proxies (e.g. surveys, ambient averages) can systematically bias estimates for some risks. |
| Extrapolation of response outside of observed exposure data | Exposure–response functions may be applied outside the range of empirical data, risking over- or under-estimation of burden. |
| Uncertainty in exposure–response relationships | Relative-risk functions combine heterogeneous studies and are a major source of uncertainty in burden estimates. |
| Dependence on established causal evidence | Risks or outcomes lacking strong causal evidence are excluded, leading to underestimation of total preventable burden. |
| Difficulty handling correlated and overlapping risks | Non-additivity of PAFs and residual overlap can still misallocate burden across risks despite mediation adjustments. |
| Simplified mediation modelling | Stylised causal pathways may not capture real-world heterogeneity in how risks interact. |
| Limited treatment of time lags and cumulative exposure | Current exposure is often linked to current outcomes, misaligning burden for diseases with long latency or gradual exposure. |
| Attributable burden does not equate to achievable prevention | Estimates hypothetical avoidable burden, not the impact of feasible or cost-effective interventions. |
| Structural and model uncertainty | Uncertainty from model structure, assumptions, and counterfactual choice is difficult to quantify fully. |
| Uneven comparability across risks and locations | Differences in data quality and definitions limit robustness of cross-risk and cross-country comparisons. |

## 4.2 Specific limitations of the Global Burden of Disease dietary risks model

For disease burden estimates from dietary risks using GBD input data, there are several specific limitations in the GBD dietary risks studies. Overlap between categories, such as sodium and processed meats or wholegrains and fibre, are imputed by simple mediation factors; moreover, relative risk factors and risk-outcome pairs change between studies [44, 45]. The GBD uses historical consumption and outcomes in age groups from control studies as an indicator of the burden of disease. To reflect the outcomes from consumption now, it assumes that the future dietary patterns and the preventative factors and technology remain similar in historical patterns. The GBD extrapolates dietary intake and its distribution in populations for national and sub-national population using statistical methods, and there can be significant differences in intake estimates between the GBD and other global sources such as the Global Dietary Database (GDD) [46]. There is significant epistemological uncertainty in the distributions of dietary intake derived for exposure.

**Variation across GBD studies**

The GBD 2017 is emulated despite updates in 2019 and 2021 to the GBD dietary risk model. Between the GBD 2019 [22] and 2021 studies [47], the DALY estimates changed for most of the 15

categories, as the model updated its burden of proof, risk-outcome pairs and relative risk functions [19]. The emulator error of ±1.5% is dwarfed by the changes between GBD model updates. For the 2017 GBD study, dietary risks were responsible for 255 million DALYs globally in 2017. According to the GBD 2019 study, they were responsible for around 190 million DALYs in 2017. For the GBD 2021 study dietary risks were responsible for 167 million DALYs in 2017. Disease burden from unprocessed red meat consumption experienced a sharp increase between the 2017 and 2019 rounds of the GBD study [45], while, for example for Brazil, there was a reduction by nearly 50 percent between 2019 and 2021 [20]. In the GBD 2021 study, both diets low in milk and diets high in red meat intake have U-shaped relative risk curves [20], meaning that both overconsumption of milk and underconsumption of red meat are associated with dietary risks, with this reflected in the modelling. Due to categorical quirks in the labelling of dietary risks in the GBD 2021, this sometimes results in a negative disease burden. For example, if the protective effect of consuming red meat outweighs its potential adverse health impacts, a diet high in red meat could be associated with a negative national disease burden in the GBD 2021 study. Overall, the relative risk factors for dietary risks are not stabilised, so the GBD 2017 model was preferred on these grounds and the fact that the representation of relative risk is more accessible in the GBD 2017 version.

**Joint exposure**

The burden of preventable morbidity and mortality for the 15 individual dietary risks in the GBD study are not mutually exclusive and not additive. The joint burden of disease for dietary risks quantified by DALYs is less than the sum of the DALYs estimated for the 15 individual categories of dietary risks. The diets low in milk and fibre are mediated in the GBD formulation of joint dietary risks by diets low in calcium and diets low in fruits, vegetables and wholegrains. But some of the nutrients and additives included in the 13 categories such as sodium also have a complicated relationship in some countries with the category of processed meats and with fruits and vegetables through the mediation of high systolic blood pressure.

The global burden of disease does have another model for nutritional deficiency risks. Globally, protein-energy malnutrition and dietary iron deficiency are the largest disease outcomes in that model [48]. For adults, it is unclear without more sophisticated public health modelling how to resolve the potential double count between nutrient deficiencies in hidden hunger, like calcium and iron deficiency, and the overlap between risk from diets low in fruits, vegetables, legumes, nuts and seeds, and diets high in red meat. Iron deficiency is probably the most important separate nutritional deficiency for future efforts to capture.

Joint exposure in the GBD study can capture coinciding high exposure across the dietary intakes by age group. In some countries, but not in the emulator, subnational regions are investigated. In general, coinciding high exposure in lower income groups cannot be captured. Several studies have shown inverse correlations between poverty and diet quality [40-42] In the US food insecurity was associated with higher consumption of processed foods and palatable meals, such as high-fat dairy products or salty snacks [40]. Reasons are the cost of food, as in general high-energy-diets are cheaper than nutrition-dense diets, and low-cost diets were found to be the least healthy [41]**.** Poverty and food insecurity is therefore often associated with a focus on diets which are cheap, energy rich and shelf-stable with a low consumption of goods with high water content, such as fruit and vegetables [41, 49]. Furthermore, food access was identified as an influence, as low-income neighbourhoods were found to have a lower density of supermarkets and grocery stores. As easy access and availability of fresh food products is associated with a higher intake of fruit and

vegetables [50], in a poorer area is correlated with lower quality of food choices [49, 51]. Additionally, a lack of nutrition knowledge [52], less time and for food shopping and cooking [53], as well as a lack of cooking equipment are potential factors which lead to bad diet quality among low-income and poor individuals [41].

**Overlap with undernourishment and obesity**

The intake distributions of the GBD dietary risk exposures are normalised to sufficient calorie intake (around 2100 calories) using FAO estimates and conventions. The GBD dietary risks model is largely meant to identify the risk from dietary composition and be exclusive to insufficient calorie intake resulting in undernourishment according to the FAO definition or excess calorie intake result in high body mass index (BMI) [54]. Theoretically, the studies used to inform the relative risks in the GBD model dietary risks model have also controlled for energy intake. However, this is difficult to ascertain and difficult to achieve. The dietary risks model does not include protein-energy malnutrition as a disease outcome, which is the primary outcome of chronic deficit in energy intake, however it is unclear if dietary composition is decoupled from energy intake.

Factors such as poverty or limited access to food resources, which lead to undernutrition and chronic hunger also lead to an imbalanced diet. Especially dietary risks associated with no or low consumption of certain food products such as vegetables, fruits or omega-3 fatty acids are likely to be present in undernourished individuals. Food-insecurity is often associated with bad diet quality [40]. A study on indigenous woman in Peninsular Malaysia found higher fat and sodium intake for children in the child-hunger group, than for food secure or insecure individuals [55]. In sub-Saharan Africa, [56] identified a strong correlation between the nutritional inadequacy and prevalence of undernutrition on national level.

Excessively high-caloric diets and the consumption of foods that impair metabolic functioning have contributed to the human burden of disease through high BMI [57]. Attribution of the burden of disease from high BMI to dietary intake alone is debated in literature due to cofactors such as physical activity [58-62]. There are likely correlations in joint exposure. A rapid nutrient transition and increased availability of cheap ultra-processed foods are enhancing the prevalence of overweight and obesity [63]. High-consumption of energy dense foods which lead to obesity and overweight can be indicators of diets high in sodium, trans-fatty acids, and sugar-sweetened beverages. Unlike protein-energy malnutrition, the disease outcomes associated with dietary composition, such as diabetes mellitus type 2, ischaemic heart disease, ischaemic stroke, intracerebral haemorrhage and subarachnoid haemorrhage are also shared by high BMI [2, 64]. However, more research is needed on causality and shared biological pathways. Is the process leading to the development of a neoplasm from having high body fat enhanced or mitigated by consuming a certain level of vitamins, minerals and micronutrients through fruit or vegetable intake? Shared disease outcomes are covered by the mediation matrix for all metabolic risks available through the GBD. The mediation between dietary composition factors and high BMI is relatively low in the GBD studies – limited to correction for double counting from sugar-sweetened beverage intake. The emulator does not consider exposure to high BMI as an additional risk category outside of dietary composition.

## Appendix A: maximal entropy sampling given an interval and an average value

The GBD model samples uncertainty in model parameters to report uncertainty in attribution of burden of disease. The inputs $\mu_{jcast}, \sigma_{jcast}$ , and the model parameters $rf_{joam}, TMREL_j$, and $YL_{mocast}$ are published in the GBD study with a central value (mean) and lower and upper values representing the 95% uncertainty interval.

The shape of the distribution of those factors within the uncertainty interval is not specified in published information. We therefore used a maximum entropy distribution to draw a new sample for each parameter, using the published lower and upper value as the lower and upper limits of known values. The published value of the inputs $\mu_{jcast}, \sigma_{jcast}$ , and the model parameters $rf_{joam}, TMREL_j$, and $YL_{mocast}$ were assumed to be the mean value.

A random variable supported in the interval $[a, b]$ with mean $\mu$ has a unique maximal entropy distribution. A sample $x$ of the maximal entropy distributed random variable in the interval $[a, b]$ with mean $\mu$ can be obtained by taking a random number $u \in [0,1]$ and using the formula

$$x = a + \frac{\log\left((e^{\alpha\,(b-a)}-1)u+1\right)}{a}.$$

Here the parameter $\alpha$ needs to be solved by numerical methods. The parameter α is related to the mean value μ and limits $[a, b]$ by the equation

$$\mu = \frac{(b\, e^{\alpha b} - a\, e^{\alpha a})}{(e^{\alpha b} - e^{\alpha a})} - \frac{1}{\alpha}.$$

The emulator uses optimisation routine in Python to solve for α by numerical methods.

## Appendix B: recalculation of parameters for candidate distributions

For creating a distribution from mean μ and standard derivation $\sigma$ the characteristic parameters of each distribution needed to be derived from the mean and standard deviation as well as upper $u$ and lower $l$ limit of the distribution.

For most of the candidate models: exponential, gamma, log-normal, normal, Gumbel, mirrored gamma, mirrored Gumbel and inverse gamma this could be done analytically. However, for stretched beta, log-logistic, Weibull and inverse Weibull distributions the characteristic parameters had to be calculated numerically, as there was no analytical solution to the problem.

For all parameters which were derived through optimisation in the emulator, the final values were checked by identifying whether the resulting distribution had the correct first moment and second central moment.

**<u>Beta distribution</u>**

For the beta distribution stretched on $[l, u]$ an analytical solution was used wherever possible, so three different cases were considered. One of the main issues with the ensemble approach of the GBD is that, given some mean $\mu$ and variance $V = \sigma^2$, there may not exist a stretched beta

distribution with that first moment and second central on the interval $[l, u]$. In initial testing of the emulator with the means, standard deviations and the lower and upper values provided by IHME, the beta distribution code failed because there could be no beta distribution for the given lower and upper values that had the given mean and standard deviation. It is unknown how IHME dealt with restricted mean and standard deviation combinations for the beta distribution. After processing the standard deviations provided by IHME the representation of the beta distribution still failed. This required extending the support of the interval $[l, u]$ to the smallest interval for which a beta distribution exists with the stated mean and variance. Having to change the beta distribution in the ensemble in this way could potentially have a large impact on the validation of the emulator. We sensitivity tested the validation by using the beta adjustment below and reweighting the ensemble so that beta distributions had a weight of zero (eliminating them).

Given mean $\mu$ and variance $V = \sigma^2$ , then there exists a beta distribution with this mean and variance if

$$V < (\mu - l)(u - \mu)$$

The stretched beta probability distribution on $[l, u]$ with mean $\mu$ and variance $V$ is then given by

$$f(x, \mu, V) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} y^{a-1}(1-y)^{b-1}, \qquad y = \frac{(x - l)}{(u - l)},$$

where

$$\mu_{\boldsymbol{\beta}} = \frac{(\mu - l)}{(u - l)}, \quad V_{\beta} = \frac{V}{(u - l)^2}$$

$$a = \mu_{\boldsymbol{\beta}}{}^2 \frac{(1 - \mu_{\boldsymbol{\beta}})}{V_{\boldsymbol{\beta}}} - \mu_{\boldsymbol{\beta}}, \quad b = \mu_{\boldsymbol{\beta}} \frac{(1 - \mu_{\boldsymbol{\beta}})^2}{V_{\boldsymbol{\beta}}} - 1 + \mu_{\boldsymbol{\beta}}.$$

If

$$V > (\mu - l)(u - \mu)$$

then we adjust the interval so that it can support a beta distribution with the stated mean and variance.

If

$$V \geq (\mu - l)(u - \mu) \text{ and } V < \mu\,(u + l - \mu)$$

then set

$$\partial = -\frac{(u - l)}{2} + \sqrt{\left(\frac{(u - l)}{2}\right)^2 + V - (\mu - l)(u - \mu)} + \varepsilon$$

where $\varepsilon = 10^{-12}$ and

$$\tilde{u} = u + \partial, \tilde{l} = l - \partial$$

If

$$V \geq (\mu - l)(u - \mu) \text{ and } V \geq \mu\,(u + l - \mu)$$

then set

$$\tilde{u} = \frac{V}{\mu} + \mu + \varepsilon, \tilde{l} = 0.$$

The stretched beta probability distribution on $[\tilde{l}, \tilde{u}]$ with mean $\mu$ and variance $V$ is then given by

$$f(x, \mu, V) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} y^{a-1}(1-y)^{b-1}, \qquad y = \frac{(x - \tilde{l})}{(\tilde{u} - \tilde{l})},$$

where

$$\mu_{\boldsymbol{\beta}} = \frac{(\mu - \tilde{l})}{(\tilde{u} - \tilde{l})}, \; V_{\beta} = \frac{V}{(\tilde{u} - \tilde{l})^2}$$

$$a = \mu_{\boldsymbol{\beta}}{}^2 \frac{(1 - \mu_{\boldsymbol{\beta}})}{V_{\boldsymbol{\beta}}} - \mu_{\boldsymbol{\beta}}, \; b = \mu_{\boldsymbol{\beta}} \frac{(1 - \mu_{\boldsymbol{\beta}})^2}{V_{\boldsymbol{\beta}}} - 1 + \mu_{\boldsymbol{\beta}}.$$

**<u>Log-logistic</u>**

For the log-logistic distribution with mean $\mu > 0$ and variance $V = \sigma^2$, the parameter

$$\beta(u, V) = \min_x \left( \operatorname{sinc} x^{-1} - \frac{\left(1 + \frac{V}{\mu^2}\right)}{2} \right)^2$$

is estimated using optimisation. Set

$$\alpha(\mu, V) = \mu \operatorname{sinc} \beta^{-1}.$$

Then the log-logistic distribution with mean $\mu > 0$ and variance $V = \sigma^2$ is given by

$$f(y, \mu, V) = \frac{\beta\, y^{-\beta-1} \left(1 + y^{-\beta}\right)^{-2}}{\alpha}, \qquad y = \frac{x}{\alpha}$$

**<u>Weibull</u>**

For the Weibull distribution with mean $\mu$ and variance $V = \sigma^2$, then the parameter

$$k(\mu, V) = \min_x \left( \frac{\Gamma(1 + 2\,x^{-1})}{\Gamma(1 + x^{-1})^2} - 1 - \left(\frac{V}{\mu}\right)^2 \right)^2$$

is determined through optimisation. Set

$$\lambda(\mu, V) = \frac{\mu}{\Gamma(1 + k^{-1})}.$$

Then the Weibull distribution with mean $\mu$ and variance $V = \sigma^2$ is given by

$$f(x,\mu,V) = \frac{k\, y^{k-1}\, e^{-y^k}}{\lambda}, \qquad y = \frac{x}{\lambda}.$$

**Inverse Weibull**

For the inverse Weibull distribution with mean $\mu$and variance $V = \sigma^2$, then the parameter

$$k(\mu,V) = \min_x \left( \frac{\Gamma(1-\ 2\,x^{-1})}{\Gamma(1-\ x^{-1})^2} - \ 1\ - \left(\frac{V}{\mu}\right)^2 \right)^2$$

is determined through optimisation. Set

$$\lambda(\mu,V) = \frac{\mu}{\Gamma(1-\ k^{-1})}$$

Then the inverse Weibull distribution with mean $\mu$ and variance $V = \sigma^2$ is given by

$$f(x,\mu,V) = \frac{k\, y^{-k-1}\, e^{-y^{-k}}}{\lambda}, \qquad y = \frac{x}{\lambda}.$$

# Appendix C: further graphs of validation of emulator

## C.1 Attributable population fraction from single exposure

Figure 2 showed the differences between emulator and published population attributable fractions for mortality from single dietary exposures: for every risk factor, disease, country, age group and sex. The high dimensionality of the index means that only two indexing pairs were chosen for display, e.g. risk-disease pairs, and a histogram of the absolute value of the difference across the other four index dimensions was formed. Figure 2 showed the breakdown by age-risk pairing. reports the 95-percentiles of the histograms.

Figure 5 and Figure 6 show the mortality breakdown by age-disease and disease-risk pairing respectively. Figure 7, Figure 8 and Figure 9 show the morbidity breakdown by age-risk, age-disease and disease-risk pairing respectively.

*Figure 5: Individual dietary intake PAF discrepancies for mortality. The values in the figure refer to the 95th percentile of the sets of absolute PAF differences for fixed age and disease outcome.*

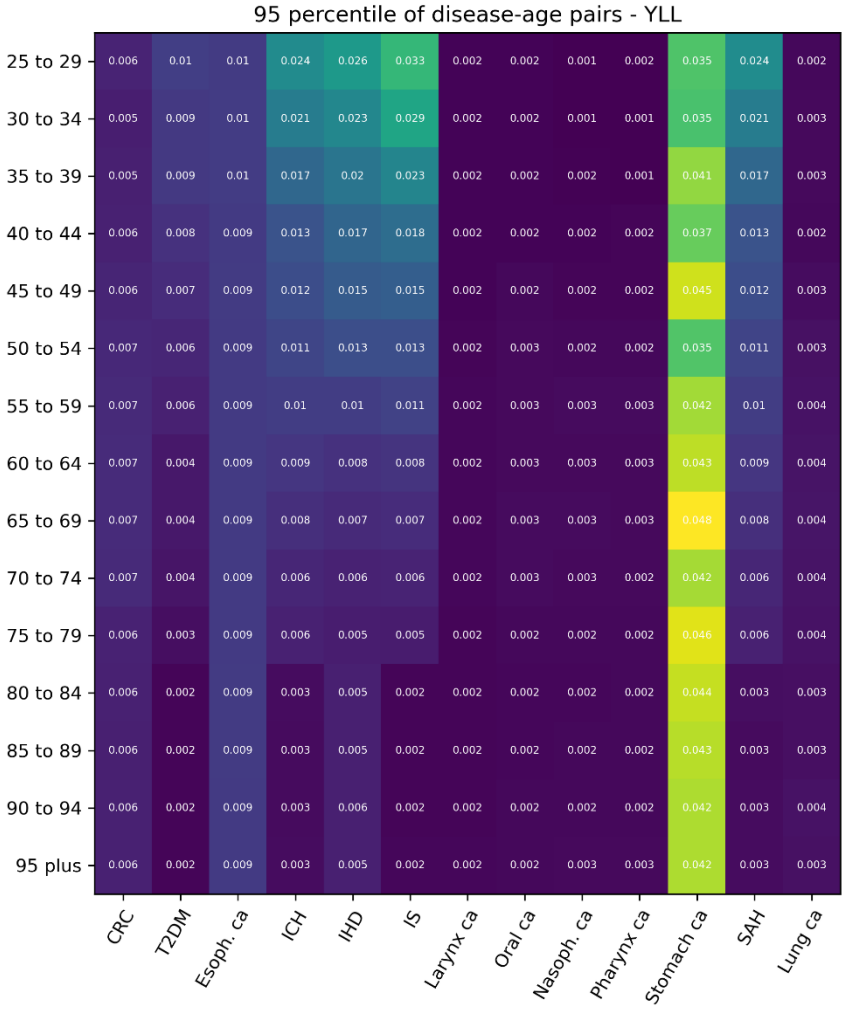


*Figure 6: Individual dietary intake PAF discrepancies for mortality. The values in the figure refer to the 95th percentile of the sets of absolute PAF differences for fixed risk and disease outcome.*

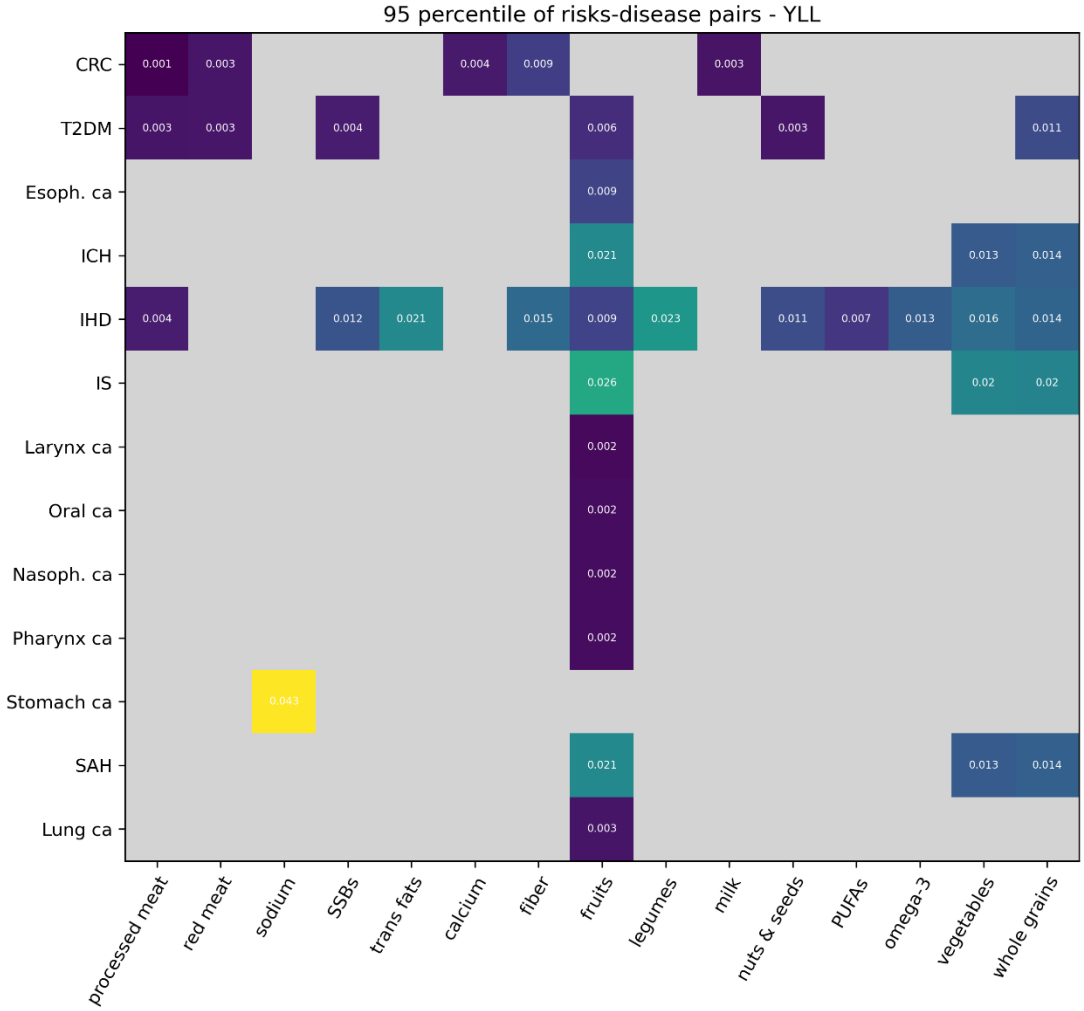


*Figure 7: Individual dietary intake PAF discrepancies for morbidity. The values in the figure refer to the 95th percentile of the sets of absolute PAF differences for fixed age and risk outcome.*

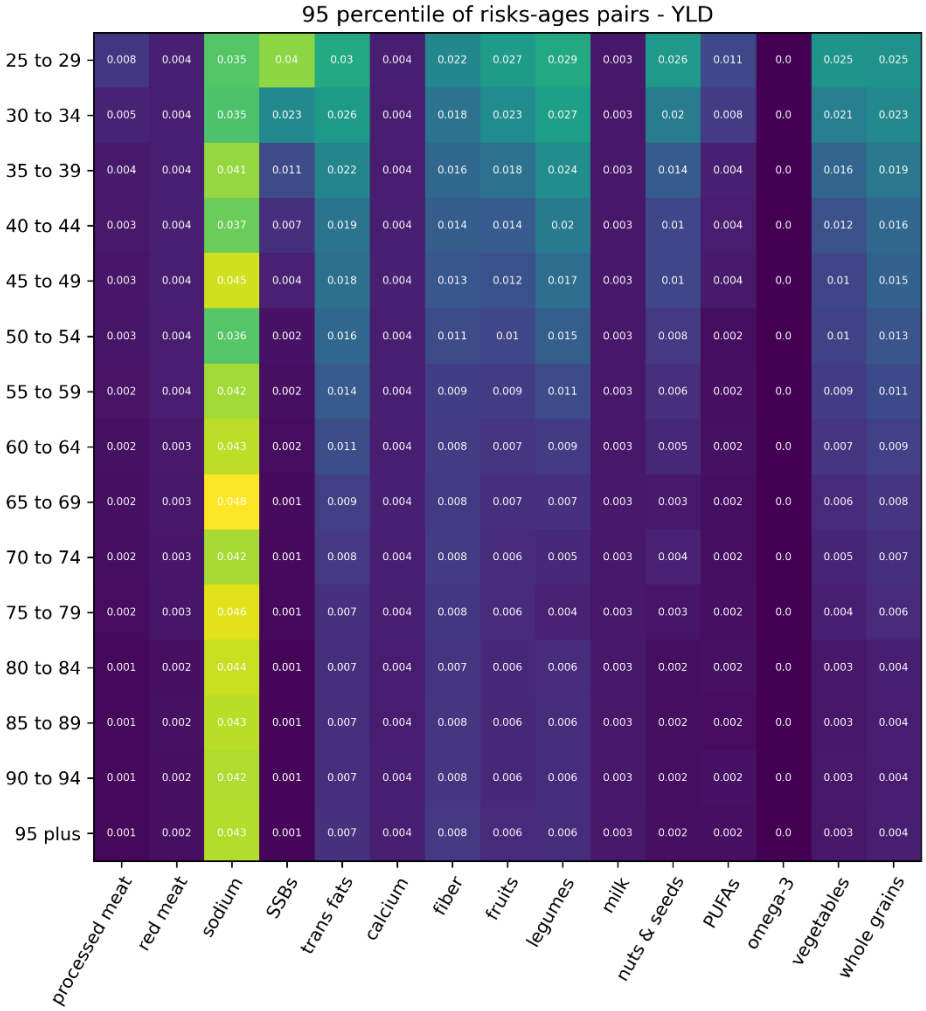


*Figure 8: Individual dietary intake PAF discrepancies for morbidity. The values in the figure refer to the 95th percentile of the sets of absolute PAF differences for fixed age and disease outcome.*

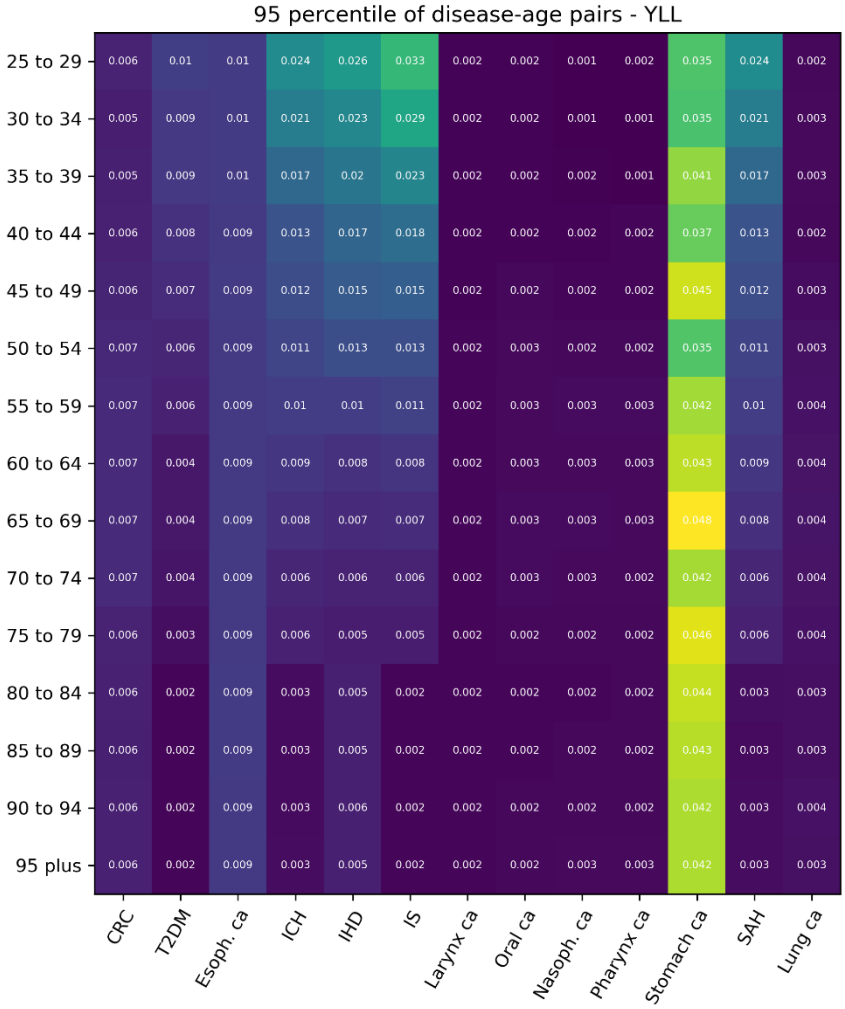
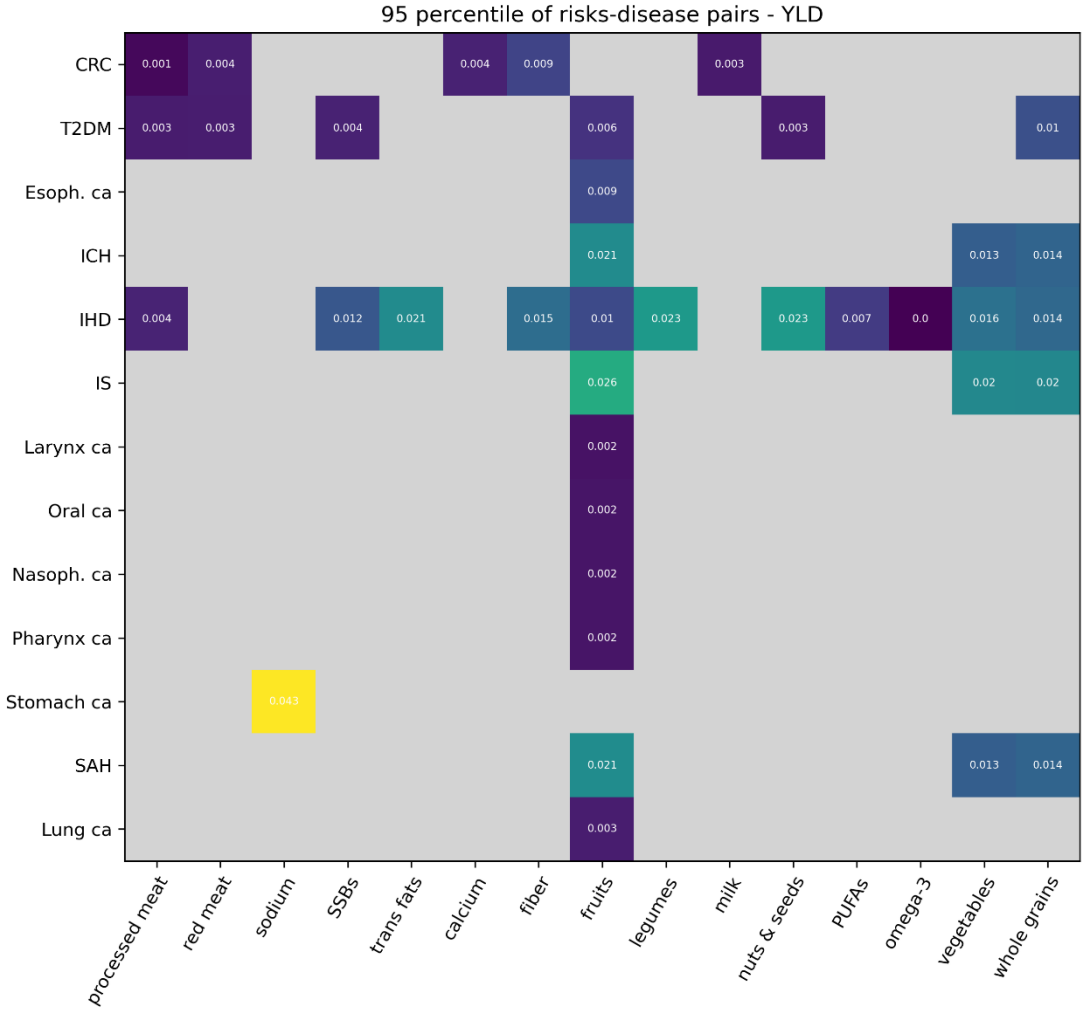


*Figure 9: Individual dietary intake PAF discrepancies for morbidity. The values in the figure refer to the 95th percentile of the sets of absolute PAF differences for fixed risk and disease outcome.*

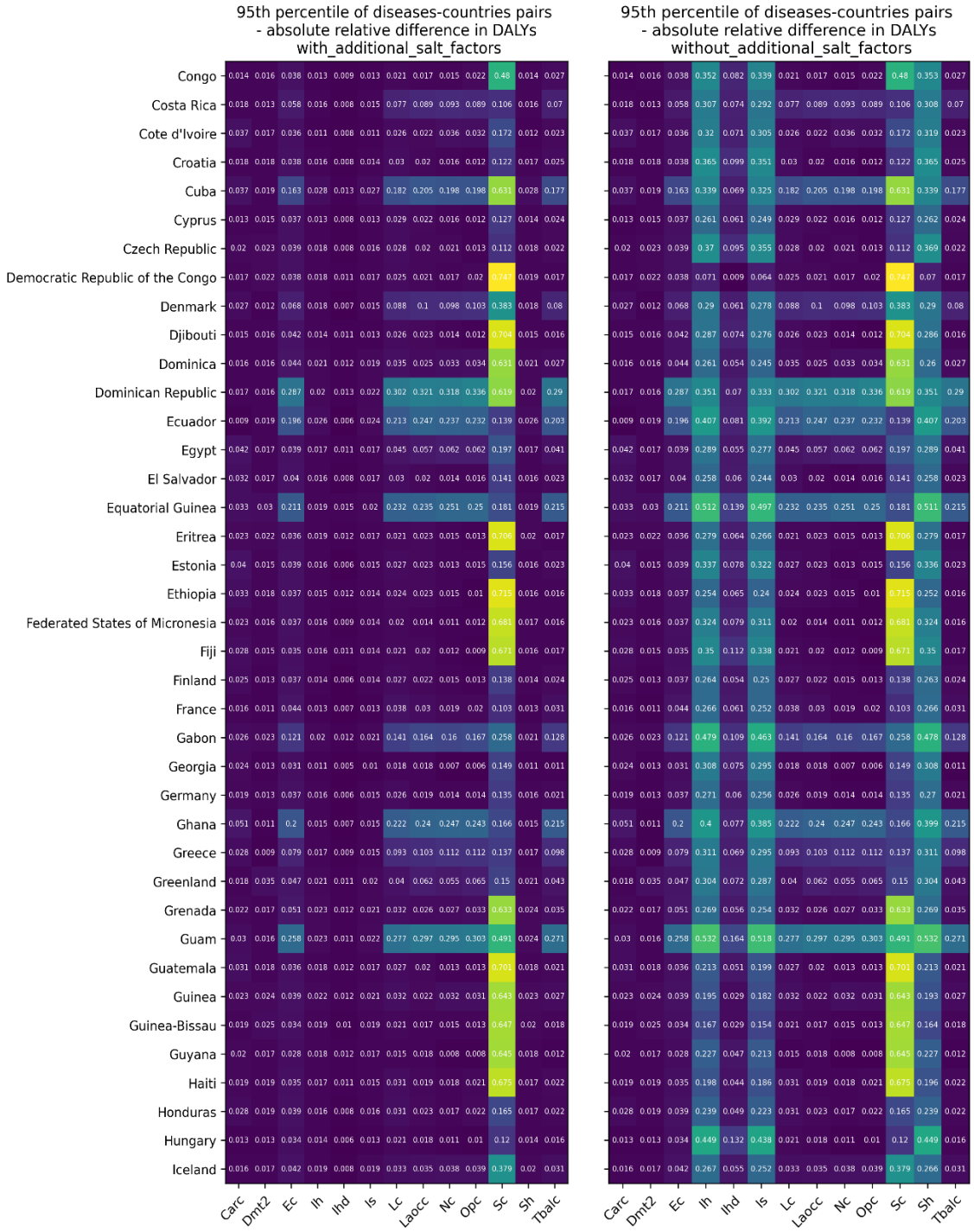



## C.2 Attributable disease burden from joint exposure

Figure 3 showed the absolute value of the relative difference between published values and emulator values for the first 40 countries listed alphabetically. The two indices chosen for display are country-disease pairs, which means that a histogram across the other indices contains 30 values (across age-sex subpopulations) and the 95-th percentile of the histogram is displayed. Figure 10, Figure 11, Figure 12, and Figure 13 display the absolute value of the relative difference between published values and emulator values for the remaining countries listed alphabetically.

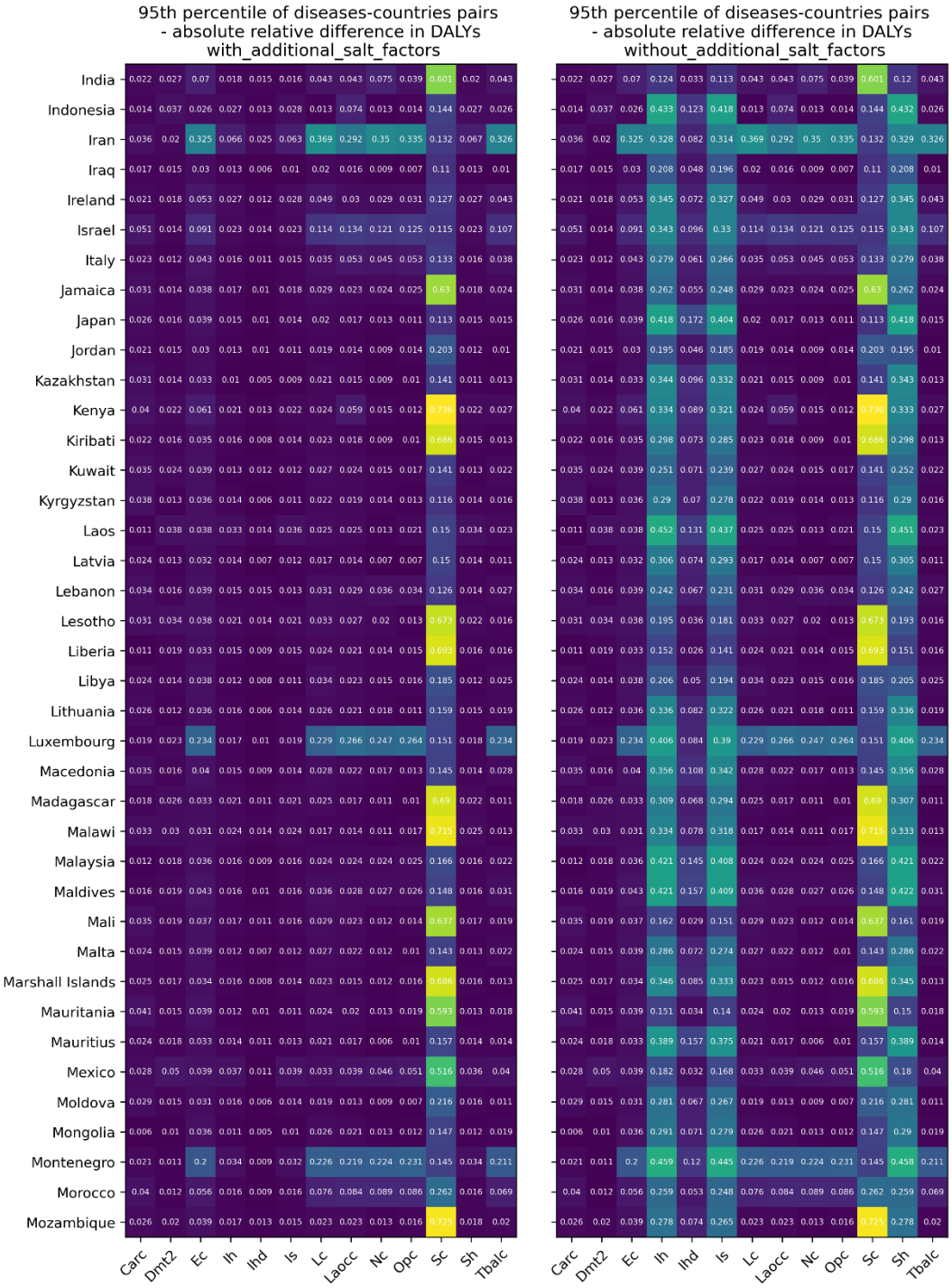
# Annex A

*Figure 10: 95th percentile of absolute value of relative difference in total DALYs per country and disease outcome. Countries: Congo to Iceland.*



# Annex A

*Figure 11: 95th percentile of absolute value of the relative difference in total DALYs per country and disease outcome. Countries: India to Mozambique.*

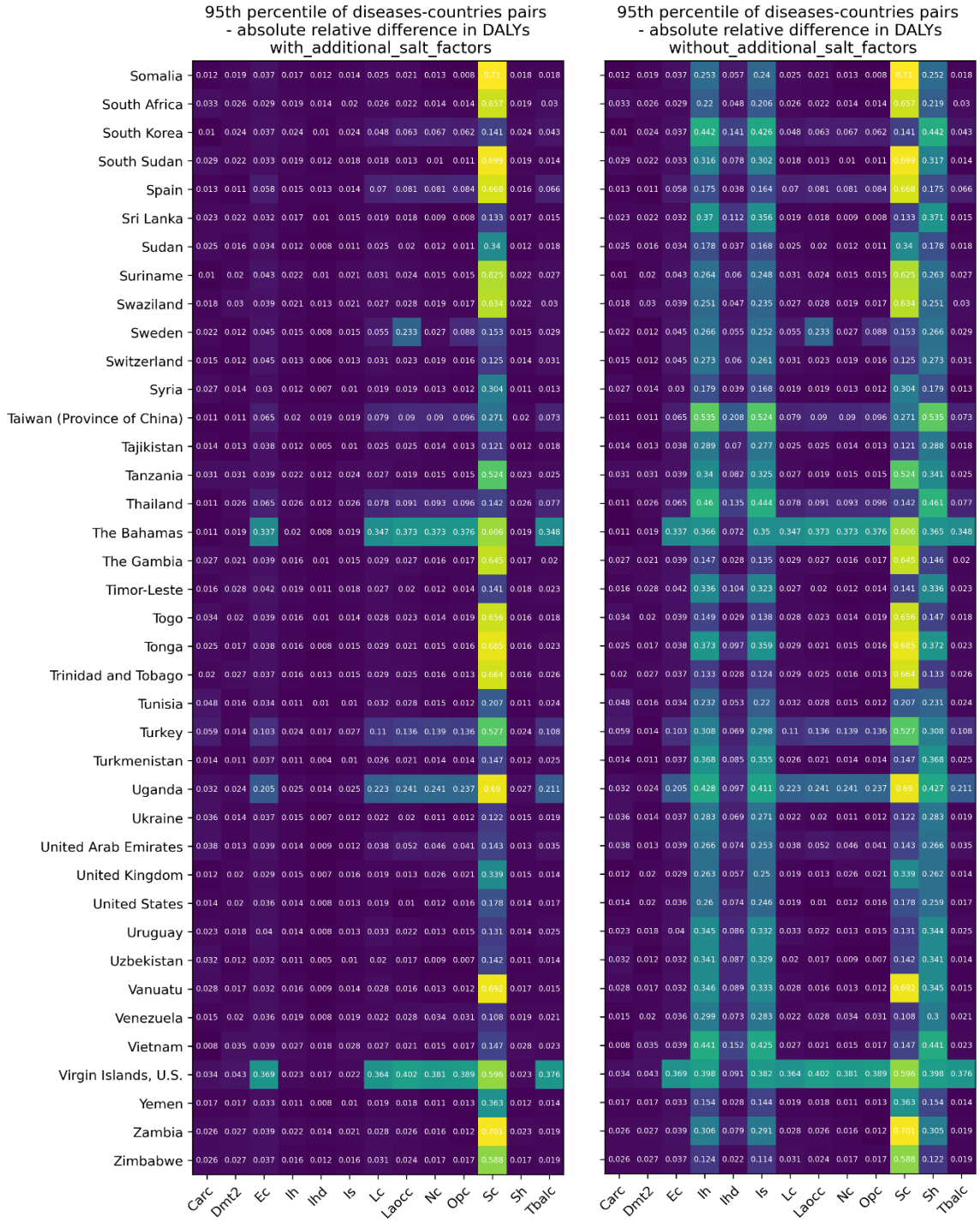


# Annex A

*Figure 12: 95th percentile of absolute value of the relative difference in total DALYs per country and disease outcome. Countries: Myanmar to Solomon Islands.*

# Annex A

*Figure 13: 95th percentile of the relative difference in total DALYs per country and disease outcome. Countries: Somalia to Zimbabwe.*



# Appendix D: Derivation of attributable disease burden

Let $c$ index countries, *a* index age, *s* index sex, *o* index disease outcome and *j* index dietary risk factor, and $m$ index mortality or morbidity. We make the simplification that the year is fixed and not indexed. The argument below is extended simply to index different years of exposure, response, and population disease outcomes.

Using the notation in Section 2.2.2, we shifted intake for a fixed $c$ index country, *a* index age, *s* index sex, *o* index disease outcome and *j* index dietary risk factor, and $m$ index mortality or morbidity

$$PAF_{jomcas}(h) = 1 - \frac{1}{\int_0^\infty R_{joam}(x)\, P_{jcas}(\mu_{jasc}, \sigma_{jasc})(x-h)\, dx}.$$

which leads to an easier calculation

$$PAF_{jomcas}(h) = 1 - \frac{1}{\int_0^\infty R_{joam}(x+h)\, P_{jcas}(\mu_{jasc}, \sigma_{jasc})(x)\, dx}.$$

The formula applies if one assumes a uniform change in intake *j* for every individual in the population.

Other changes can be represented, such as scaling intake by a proportion $k$. The following formula applies if every individual in the population is assumed to reduce their intake by $100 * k$ percent

$$\overline{PAF}_{jomcas}(h) = 1 - \frac{1}{\int_0^\infty R_{joam}(x)\, P_{jcas}(\mu_{jasc}, \sigma_{jasc})((1-k)^{-1}x)\, dx}.$$

which leads to a calculation that, again, only shifts the relative risk function

$$\overline{PAF}_{jomcas}(h) = 1 - \frac{(1-k)}{\int_0^\infty R_{joam}((1-k)x)\, P_{jcas}(\mu_{jasc}, \sigma_{jasc})(x)\, dx}.$$

In both cases, shift and scaling, the original exposure distributions can be used in the integral and the emulator is faster. The emulator can be adjusted by a user to take in custom counterfactual distributions, but currently only shifting of baseline exposure is implemented in the code.

In this section we derive an analytical formula, implemented in the code, for the shifted and marginally shifted disease burden.

## D.1 Attributable disease burden from joint exposure shifted

From Section 2.2.2,

$$\begin{aligned} DALY_{Jc}\left(\left[h_{jcas}\right]_{j\in J, cas\in S}\right) - DALY_{Jc} \\ = \sum_{o,m,a,s} \left(PAF_{Jomcas}\left(\left[h_{jcas}\right]_{j\in J}\right) - PAF_{Jomcas}\right) \cdot YL_{omcas} \end{aligned} \tag{10}$$

where custom user input can designate shifts $\left[h_{jcas}\right]_{j\in J, cas\in S}$ in a set of subpopulations $S$ and exposures $J$ individually, resulting in attributable fraction to shifted joint exposure that is a vector-

valued function for fixed $c$ index country, $a$ index age, $s$ index sex, $o$ index disease outcome and $m$ index mortality or morbidity

$$PAF_{Jomcas}\left(\left[h_{jcas}\right]_{j\in J}\right) = 1 - \prod_{j=1}^{J}\left(1 - PAF_{jomcas}(h_{jcas})\prod_{i=1}^{J}(1 - MF_{jio})\right).$$

From Section 2.2.2, if the shift is the same across all subpopulations, then

$$h_{jcas} = \eta_j$$

equally across all subpopulations shift case and the national attributable disease burden to joint exposure becomes a vector valued function on the space of shifts:

$$DALY_{Jc} : \mathbb{R}^{15} \to \mathbb{R}$$

and change in disease burden for small shifts in exposure is approximated by the partial derivatives

$$DALY_{Jc}(\eta) - DALY_{Jc} \cong \sum_{j\in J}\frac{\partial DALY_{Jc}}{\partial \eta_j}(0) \cdot \eta_j$$

where

$$\frac{\partial DALY_{Jc}}{\partial \eta_j}(0) = \sum_{o,m,a,s}\left(\frac{\partial PAF_{Jomcas}}{\partial h_{jcas}}(0)\right) \cdot YL_{omcast}. \tag{11}$$

To calculate (10) and (11) we derive analytic formulas for

$$PAF_{Jomcas}\left(\left[h_{jcas}\right]_{j\in J}\right) - PAF_{Jomcas}$$

and

$$\frac{\partial PAF_{Jomcas}}{\partial h_{jcas}}(0).$$

## D.2 Expanding out the joint product into single intake exposures

For brevity, set

$$M_{jo} = \prod_{i=1}^{J}(1 - MF_{jio}).$$

Then the difference in population attributable fractions is given by the difference of products,

$$PAF_{Jomcas}\left(\left[h_{jcas}\right]_{j\in J}\right) - PAF_{Jomcas}(0) =$$
$$= \prod_{j=1}^{J}\left(1 - M_{jo} \cdot PAF_{jomcas}\right) - \prod_{j=1}^{J}\left(1 - M_{jo} \cdot PAF_{jomcas}(h_{jcas})\right)$$

$$= \prod_{j=1}^{J} \left( 1 - M_{jo} \cdot PAF_{jomcas} \right) - \prod_{j=1}^{J} \left( 1 - M_{jo} \cdot PAF_{jomcas} - M_{jo} \cdot \left( PAF_{jomcas}(h_{jcas}) - PAF_{jomcas} \right) \right) \quad (12)$$

The attributable fractions $PAF_{jomcas}$ at current exposure are previously calculated by the code. To calculate the difference of attributable fractions in joint exposure, the formula requires only the calculation of the difference for the individual dietary exposures.

Some of the indices and products are obscuring difference of products formula. Set

$$b_j = M_{jo} \cdot PAF_{jomcas}$$

and

$$c_j(h_j) = M_{jo} \cdot \left( PAF_{jomcas}(h_j) - PAF_{jomcas} \right).$$

Then $c_j(0) = 0$. If we write

$$f\left( [h_j]_{j \in J} \right) = -\prod_{j=1}^{J} \left( 1 - b_j - c_j(h_j) \right)$$

Then, for the relative change

$$PAF_{Jomcas}\left( [h_j]_{j \in J} \right) - PAF_{Jomcas} = f\left( [h_j]_{j \in J} \right) - f(0).$$

For the instantaneous change

$$\frac{\partial PAF_{Jomcas}}{\partial h_j}(0) = \frac{\partial f}{\partial h_j}(0).$$

Using the product rule on $f$, the instantaneous change is easy to calculate

$$\frac{\partial f}{\partial h_j} = \sum_j \frac{\partial c_j}{\partial h_j}(0) \prod_{k \neq j}^{J} \left( 1 - b_k \right).$$

Hence

$$\frac{\partial PAF_{Jomcas}}{\partial h_{jcas}}(0) = \sum_j M_{jo} \cdot \frac{\partial PAF_{jomcas}}{\partial h_{jcas}}(0) \cdot \prod_{k \neq j}^{J} \left( 1 - M_{ko} \cdot PAF_{komcas} \right) \quad (13)$$

The attributable fractions $PAF_{komcas}$ at current exposure are previously calculated by the code. To calculate the rate of change of attributable fractions in joint exposure, the formula requires only the calculation of the rate of change for the individual dietary exposures.

## D.3 Single intake exposures

Using the formula for individual exposure shifted PAFs,

$$PAF_{jomcas}(h_{jcas}) = 1 - \left(\int_0^\infty R_{joam}(x + h_{jcas})\rho_{jcas}(x)dx\right)^{-1}.$$

Then

$$PAF_{jomcas}(h_{jcas}) - PAF_{jomcas} = \left(\int_0^\infty R_{joam}(x)\rho_{jcas}(x)dx\right)^{-1} - \left(\int_0^\infty R_{joam}(x + h_{jcas})\rho_{jcas}(x)dx\right)^{-1}.$$

The emulator code can calculate this difference efficiently, by encoding a shift parameter into the function call to calculate the integral and calling the same function twice. The original PAF and the PAF difference are substituted in equation (12). Equation (12) is substituted into (10) to calculate the shift in the attributable disease burden.

For the rate of change in attributable disease burden if the dietary intake is shifted

$$\frac{\partial PAF_{jomcas}}{\partial h_{jcas}}(0) = \left(\int_0^\infty R_{joam}(x)\rho_{jcas}(x)dx\right)^{-2} \cdot \int_0^\infty \frac{\partial R_{joam}}{\partial h_{jcas}}(x + h_{jcas})\bigg|_{h_{jcas}=0} \rho_{jcas}(x)dx$$
$$= \left(1 - PAF_{jomcas}\right)^2 \cdot \int_0^\infty \frac{\partial R_{joam}}{\partial h_{jcas}}(x + h_{jcas})\bigg|_{h_{jcas}=0} \rho_{jcas}(x)dx.$$

The emulator code can calculate this difference efficiently, since the original PAF is previously calculated and the rate of change in relative risk can be calculated.

When $j$ is a dietary intake that has increased risk above the threshold value $TMREL_j$, recall that the relative risk functions are of the form

$$R_{joam}(x) = \begin{cases} rf_{joam}^{(x-TMREL_j)/\nu_j} & x \geq TMREL_j \\ 1 & 0 \leq x < TMREL_j \end{cases},$$

and

$$\frac{\partial R_{joam}}{\partial h_{jcas}}(x + h_{jcas})\bigg|_{h_{jcas}=0} = \begin{cases} \frac{\ln(rf_{joam})}{\nu_j} rf_{joam}^{(x-TMREL_j)/\nu_j} & x \geq TMREL_j \\ 0 & 0 \leq x < TMREL_j \end{cases}.$$

So

$$\int_0^\infty \frac{\partial R_{joam}}{\partial h_{jcas}}(x + h_{jcas})\bigg|_{h_{jcas}=0} \rho_{jcas}(x)dx = \frac{\ln(rf_{joam})}{\nu_j} \cdot \int_{TMREL_j}^\infty R_{joam}(x)\rho_{jcas}(x)dx$$

So only the current exposure distributions $\rho_{jcas}$ and the functions $R_{joam}$ are required in the code to calculate the rate of change.

When $j$ is a dietary intake that has increased risk below the threshold value $TMREL_j$, then relative risk functions are of the form

$$R_{joam}(x) = \begin{cases} rf_{joam}^{(TMREL_j - x)/\nu_j} & 0 \le x < TMREL_j \\ 1 & x \ge TMREL_j \end{cases},$$

and

$$\left.\frac{\partial R_{joam}}{\partial h_{jcas}}(x + h_{jcas})\right|_{h_{jcas}=0} = \begin{cases} -\frac{\ln(rf_{joam})}{\nu_j} rf_{joam}^{(TMREL_j - x)/\nu_j} & 0 \le x < TMREL_j \\ 0 & x \ge TMREL_j \end{cases}.$$

So

$$\int_0^\infty \left.\frac{\partial R_{joam}}{\partial h_{jcas}}(x + h_{jcas})\right|_{h_{jcas}=0} \rho_{jcas}(x)dx = -\frac{\ln(rf_{joam})}{\nu_j} \cdot \int_0^{TMREL_j} R_{joam}(x)\rho_{jcas}(x)dx.$$

Again, only the current exposure distributions $\rho_{jcas}$ and the functions $R_{joam}$ are required in the code to calculate the rate of change.

The calculation of $\frac{\partial PAF_{jomcas}}{\partial h_{jcas}}(0)$ is substituted in equation (13). Equation (13) is substituted into (11) to calculate the rate of change in the attributable disease burden when the shift is the same across all national subpopulations.

## Appendix E: Proportional individual risk estimates

The calculation of attributable national burden of disease from joint exposure to dietary risks is

$$DALY_{Jct} = \sum_{o,m,a,s} PAF_{Jomcast} \cdot YL_{omcast}.$$

The attributable national burden of disease from exposure to a single dietary risk is

$$DALY_{jct} = \sum_{o,m,a,s} PAF_{jomcast} \cdot YL_{omcast}$$

Or corrected for mediation with other dietary risk exposures

$$DALY'_{jct} = \sum_{o,m,a,s} PAF_{jomcast} \cdot YL_{omcast} \cdot \prod_{i=1}^{J}(1 - MF_{jiom}).$$

The attributable national burden of disease from single dietary risk exposure, whether mediated or not, are not additive to the attributable national burden of disease from joint exposure, i.e.

$$DALY_{Jc} < \sum_{j=1}^{J} DALY'_{jc} < \sum_{j=1}^{J} DALY_{jc}$$

This comes from equation (4) since disease burden from preceding risks are removed for the attribution of subsequent risks in the joint PAF:

$$PAF_{Jomcast} = 1 - \prod_{j=1}^{J}\left(1 - PAF_{jomcast}\prod_{i=1}^{J}\left(1 - MF_{jiom}\right)\right).$$

To produce individual attributed DALYs such that the individual attributed disease burden adds to the joint attributable disease burden requires proportional attribution. Set

$$PAF''_{jomcast} = PAF_{jomcast} \cdot \prod_{i=1}^{J}\left(1 - MF_{jiom}\right) \cdot \frac{PAF_{Jomcast}}{\sum_{k=1}^{J} PAF_{komcast} \prod_{n=1}^{J}(1 - MF_{knom})}$$
$$= PAF_{jomcast} \cdot \left(\prod_{i=1}^{J}\left(1 - MF_{jiom}\right) - O_{Jomcast}\right)$$

When $PAF_{Jomcast}$ is expanded out then $O_{Jomcast}$ becomes a correction term that is second order in $PAF_{komcast}$, $k \in J$. The correction term equally distributes among the individual risk factors the deficit from removing the effect of preceding risks in the joint calculation. Setting

$$DALY''_{jc}(\rho) = \sum_{o,m,a,s} PAF''_{jomcast} \cdot YL_{omcast}$$

as proportional disease burdens from single intake exposure, by design they are additive to the joint exposure

$$DALY_{Jc} = \sum_{j=1}^{J} DALY''_{jc}.$$

# 5 References


1. Global Burden of Disease Collaborative Network, *Global Burden of Disease Study 2017 (GBD 2017) Results*. 2018, Seattle, WA: Institute for Health Metrics and Evaluation (IHME), University of Washington.
2. Stanaway, J.D., et al., *Global, regional, and national comparative risk assessment of 84 behavioural, environmental and occupational, and metabolic risks or clusters of risks for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017.* The Lancet, 2018. **392**(10159): p. 1923–1994.
3. IHME, *GBD Compare Data Visualization*. 2020, Seattle, WA: IHME, University of Washington, 2020: Institute for Health Metrics and Evaluation. p. https://vizhub.healthdata.org/gbd–compare/.
4. Murray, C.J.L., et al., *Comparative quantification of health risks: Conceptual framework and methodological issues.* Population Health Metrics, 2003. **1**(1): p. 1.
5. Anand, S. and K. Hanson, *Disability-adjusted life years: a critical review.* Journal of Health Economics, 1997. **16**(6): p. 685–702.
6. Afshin, A., et al., *Health effects of dietary risks in 195 countries, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017.* The Lancet, 2019. **393**(10184): p. 1958–1972.
7. FAO, *Food Balance Sheets*. 2022, Food and Agriculture Organization of the United Nations: https://www.fao.org/faostat/en/#data/FBS.
8. Schmidhuber, J., et al., *The Global Nutrient Database: availability of macronutrients and micronutrients in 195 countries from 1980 to 2013.* The Lancet Planetary Health, 2018. **2**(8): p. e353–e368.
9. Balakrishnan, K., *Exponential distribution: theory, methods and applications*. 2019: Routledge.
10. Weisstein, E.W. *Gamma Distribution*. Available from: https://mathworld.wolfram.com/GammaDistribution.html.
11. Papoulis, A. and S.U. Pillai, *Probability, random variables, and stochastic processes*. 2002: Tata McGraw-Hill Education.
12. Gumbel, E.J. *Les valeurs extrêmes des distributions statistiques*. in *Annales de l'institut Henri Poincaré*. 1935.
13. Ekawati, D., W. Warsono, and D. Kurniasari, *On the Moments, Cumulants, and Characteristic Function of the Log-Logistic Distribution.* IPTEK The Journal for Technology and Science, 2015. **25**(3).
14. Weisstein, E.W. *Normal Distribution*. 2022 [cited 2022 June 28, 2022]; Available from: https://mathworld.wolfram.com/NormalDistribution.html.
15. Weisstein, E.W. *Log Normal Distribution*. [cited 2022 June 28, 2022]; Available from: https://mathworld.wolfram.com/LogNormalDistribution.html.
16. Johnson, N.L., S. Kotz, and N. Balakrishnan, *Continuous univariate distributions, volume 2*. Vol. 289. 1995: John wiley & sons.
17. Khan, M.S., G. Pasha, and A.H. Pasha, *Theoretical analysis of inverse Weibull distribution.* WSEAS Transactions on Mathematics, 2008. **7**(2): p. 30–38.
18. Murray, C.J., et al., *Global burden of 87 risk factors in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019.* The Lancet, 2020. **396**(10258): p. 1223–1249.
19. Zheng, P., et al., *The Burden of Proof studies: assessing the evidence of risk.* Nature Medicine, 2022. **28**(10): p. 2038–2044.

20. Lescinsky, H., et al., *Health effects associated with consumption of unprocessed red meat: a Burden of Proof study.* Nature Medicine, 2022. **28**(10): p. 2075–2082.
21. Naghavi, M., et al., *Global burden of bacterial antimicrobial resistance 1990-2021: a systematic analysis with forecasts to 2050.* The Lancet, 2024. **404**(10459): p. 1199–1226.
22. Vos, T., et al., *Global burden of 369 diseases and injuries in 204 countries and territories, 1990-2013;2019: a systematic analysis for the Global Burden of Disease Study 2019.* The Lancet, 2020. **396**(10258): p. 1204–1222.
23. Fischer, H., *A History of the Central Limit Theorem: From Classical to Modern Probability Theory*. 2011, New York: Springer.
24. Heinrich, L., *A Method for the Derivation of Limit Theorems for Sums of Weakly Dependent Random Variables:A Survey.* Optimization, 1987. **18**(5): p. 715–735.
25. Black, A.E. and T.J. Cole, *Biased Over- Or Under-Reporting is Characteristic of Individuals Whether Over Time or by Different Assessment Methods.* Journal of the American Dietetic Association, 2001. **101**(1): p. 70–80.
26. Murakami, K. and M.B.E. Livingstone, *Prevalence and characteristics of misreporting of energy intake in US adults: NHANES 2003–2012.* British Journal of Nutrition, 2015. **114**(8): p. 1294–1303.
27. Westerterp, K.R. and A.H.C. Goris, *Validity of the assessment of dietary intake: problems of misreporting.* Current Opinion in Clinical Nutrition & Metabolic Care, 2002. **5**(5).
28. Yamamoto, N., et al., *Bias in nutrition-health associations is not eliminated by excluding extreme reporters in empirical or simulation studies.* eLife, 2023. **12**: p. e83616.
29. Rajagopalan, S., S.G. Al-Kindi, and R.D. Brook, *Air pollution and cardiovascular disease: JACC state-of-the-art review.* Journal of the American College of Cardiology, 2018. **72**(17): p. 2054–2070.
30. Cuenca-García, M., et al., *Combined influence of healthy diet and active lifestyle on cardiovascular disease risk factors in adolescents.* Scandinavian journal of medicine & science in sports, 2014. **24**(3): p. 553–562.
31. Manolio, T.A., et al., *Finding the missing heritability of complex diseases.* Nature, 2009. **461**(7265): p. 747–753.
32. Laanani, M., et al., *Collider and reporting biases involved in the analyses of cause of death associations in death certificates: an illustration with cancer and suicide.* Population Health Metrics, 2023. **21**(1): p. 21.
33. O'Neil, A., et al., *The health-related quality of life burden of co-morbid cardiovascular disease and major depressive disorder in Australia: findings from a population-based, cross-sectional study.* Quality of Life Research, 2013. **22**(1): p. 37–44.
34. Gadermann, A.M., et al., *COMORBIDITY AND DISEASE BURDEN IN THE NATIONAL COMORBIDITY SURVEY REPLICATION (NCS-R).* Depression and Anxiety, 2012. **29**(9): p. 797–806.
35. Ezzati, M., et al., *Comparative quantification of health risks: Global and regional burden of disease attributable to selected major risk factors*. 2004, Geneva, Switzerland: World Health Organization.
36. Danaei, G., et al., *The Preventable Causes of Death in the United States: Comparative Risk Assessment of Dietary, Lifestyle, and Metabolic Risk Factors.* PLOS Medicine, 2009. **6**(4): p. e1000058.
37. Springmann, M., et al., *Global and regional health effects of future food production under climate change: a modelling study.* The Lancet, 2016.
38. Willett, W., et al., *Food in the Anthropocene: the EAT–Lancet Commission on healthy diets from sustainable food systems.* The Lancet, 2019. **393**(10170): p. 447–492.

39. Plass, D., et al., *Estimating risk factor attributable burden – challenges and potential solutions when using the comparative risk assessment methodology.* Archives of Public Health, 2022. **80**(1): p. 148.
40. Leung, C.W., et al., *Food insecurity is inversely associated with diet quality of lower-income adults.* Journal of the Academy of Nutrition and Dietetics, 2014. **114**(12): p. 1943–1953. e2.
41. Darmon, N. and A. Drewnowski, *Does social class predict diet quality?* Am J Clin Nutr, 2008. **87**(5): p. 1107–17.
42. Champagne, C.M., et al., *Poverty and food intake in rural America: diet quality is lower in food insecure adults in the Mississippi Delta.* Journal of the American Dietetic Association, 2007. **107**(11): p. 1886–1894.
43. Arnesen, T. and E. Nord, *The value of DALY life: problems with ethics and validity of disability adjusted life years.* BMJ (Clinical research ed.), 1999. **319**(7222): p. 1423–1425.
44. Beal, T., L.M. Neufeld, and S.S. Morris, *Uncertainties in the GBD 2017 estimates on diet and health.* The Lancet, 2019. **394**(10211): p. 1801–1802.
45. Stanton, A.V., *Unacceptable use of substandard metrics in policy decisions which mandate large reductions in animal-source foods.* npj Science of Food, 2024. **8**(1): p. 10.
46. Beal, T., et al., *Differences in modelled estimates of global dietary intake.* The Lancet, 2021. **397**(10286): p. 1708–1709.
47. Ferrari, A.J., et al., *Global incidence, prevalence, years lived with disability (YLDs), disability-adjusted life-years (DALYs), and healthy life expectancy (HALE) for 371 diseases and injuries in 204 countries and territories and 811 subnational locations, 1990-2021: a systematic analysis for the Global Burden of Disease Study 2021.* The Lancet, 2024. **403**(10440): p. 2133–2161.
48. Lord, S., *Refining national true cost accounting for agrifood systems – Considerations for moving beyond The State of Food and Agriculture 2023 and 2024. Background paper to the FAO flagship The State of Food and Agriculture 2024 – Value-driven transformation of agrifood systems*, in *Technical note*. 2024, FAO: Rome.
49. Drewnowski, A. and N. Darmon, *Food choices and diet costs: an economic analysis.* The Journal of nutrition, 2005. **135**(4): p. 900–904.
50. Morland, K., S. Wing, and A.D. Roux, *The contextual effect of the local food environment on residents' diets: the atherosclerosis risk in communities study.* American journal of public health, 2002. **92**(11): p. 1761–1768.
51. Diez-Roux, A.V., et al., *Neighbourhood differences in diet: the Atherosclerosis Risk in Communities (ARIC) Study.* Journal of Epidemiology & Community Health, 1999. **53**(1): p. 55–63.
52. Variyam, J.N., J. Blaylock, and D. Smallwood, *A probit latent variable model of nutrition information and dietary fiber intake.* American Journal of Agricultural Economics, 1996. **78**(3): p. 628–639.
53. Dubowitz, T., et al., *Lifecourse, immigrant status and acculturation in food purchasing and preparation among low-income mothers.* Public health nutrition, 2007. **10**(4): p. 396–404.
54. FAO, et al., *The State of Food Security and Nutrition in the World 2020. Transforming food systems for affordable healthy diets.* 2020, Food and Agriculture Organization of the United Nations: Rome.
55. Pei, C.S., G. Appannah, and N. Sulaiman, *Household food insecurity, diet quality, and weight status among indigenous women (Mah Meri) in Peninsular Malaysia.* Nutrition Research and Practice, 2018. **12**(2): p. 135–142.
56. Liu, Z., et al., *Global burden of childhood nutritional deficiencies, 1990–2019.* Public Health, 2024. **235**: p. 26–32.

57. Dai, H., et al., *The global burden of disease attributable to high body mass index in 195 countries and territories, 1990–2017: An analysis of the Global Burden of Disease Study.* PLOS Medicine, 2020. **17**(7): p. e1003198.
58. McCarthy, M., *The economics of obesity.* The Lancet, 2004. **364**(9452): p. 2169–2170.
59. Jan, S. and G.H. Mooney, *Childhood obesity, values and the market.* International Journal of Pediatric Obesity, 2006. **1**(3): p. 131–132.
60. Moodie, R., et al., *Childhood obesity – a sign of commercial success, but a market failure.* International Journal of Pediatric Obesity, 2006. **1**(3): p. 133–138.
61. Harker, D., M. Harker, and S. Svensen, *Attributing Blame.* Journal of Food Products Marketing, 2007. **13**(2): p. 33–46.
62. Karnani, A., B. McFerran, and A. Mukhopadhyay, *The Obesity Crisis as Market Failure: An Analysis of Systemic Causes and Corrective Mechanisms.* Journal of the Association for Consumer Research, 2016. **1**(3): p. 445–470.
63. Popkin, B.M., C. Corvalan, and L.M. Grummer-Strawn, *Dynamics of the double burden of malnutrition and the changing nutrition reality.* The Lancet, 2020. **395**(10217): p. 65–74.
64. Guh, D.P., et al., *The incidence of co-morbidities related to obesity and overweight: a systematic review and meta-analysis.* BMC public health, 2009. **9**(1): p. 88.