ITHIM

September 10, 2019

Contents

1 Introduction

1.1 Summary of literature

1.2 Summary of aims

1.3 Overview

The remainder of the document presents a current working version of the ITHIM, and is a comprehensive record of all the calculations of which it is composed. The equations for the ITHIM are described in Section 2 (tabulated in Section C). The method is concretised through the example application to Accra in Section 3 and particular columns in Tables 1 and 3, with results presented in Section 14.

The ITHIM as described here is implemented in the ITHIM-R R package (`https://github.com/ITHIM/ITHIM-R`). Other cities are also included in that package (São Paulo, Delhi, Bangalore), and reference will be made to those where relevant.

2 Method: ITHIM

In the ITHIM, we combine many information inputs and compute the health burden through three pathways (depicted in Figure 1). Each pathway has its own set of definitions and outputs, and we refer to them as "modules".

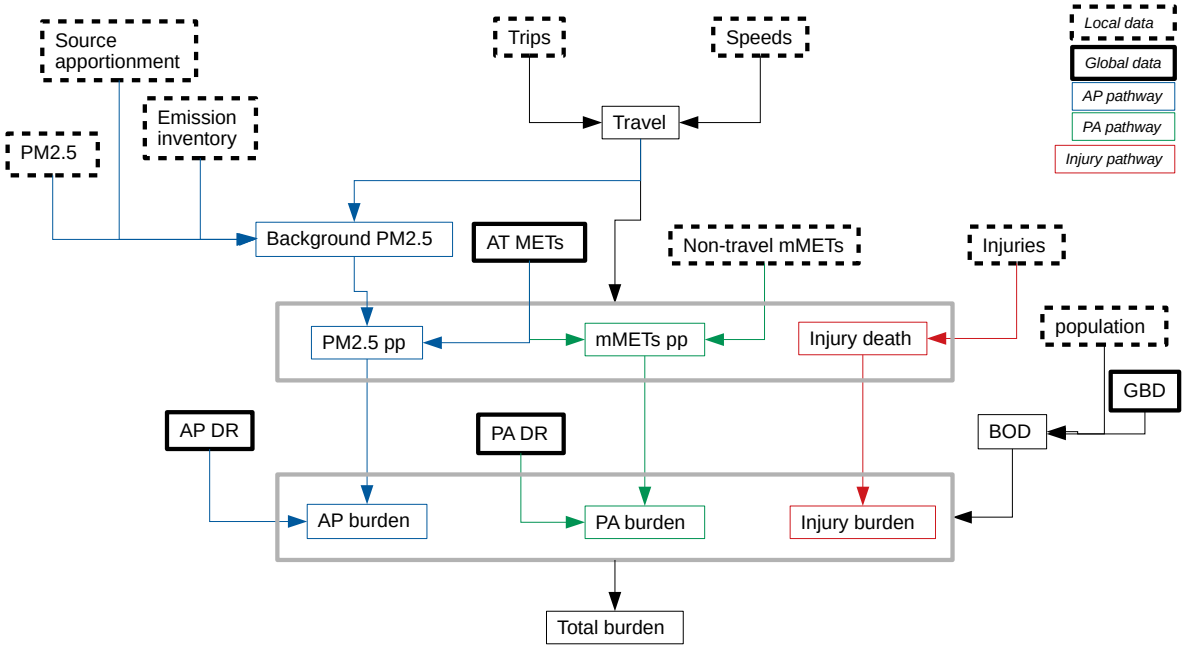

Figure 1: ITHIM workflow. Inputs to the model have a bold, black outline. The graph depicts parent–child relationships, where a child (at the head of an arrow) depends upon all its parents (at the source(s) of the arrow(s)). To aid visualisation, an arrow connected to the outside of a grey box indicates that the arrow connects to each item within the box. *AT: active travel. MET: metabolic equivalent task. pp: per person. AP: air pollution. PA: physical activity. DR: dose–response relationship. GBD: global burden of disease. BOD: (local) burden of disease.*

To describe the model, we use lower-case letters to denote indices, or dimensions, of objects. They take one of a set of possible values, detailed in Table 1. The set of input items are denoted by capital letters, and variable parameters are denoted by Greek letters. These are presented in Table 3. The equations of the ITHIM are presented below, and summarised in Table 4. These use letter modifiers to show how the inputs are modified in the course of output generation.

We calculate the ITHIM by combining the modules' outputs in terms of the health burden ($\hat{U}_{a,g,h,o,s}$). Other outputs of the model include: MMETs per person ($M_{i,s}$), PM2.5 in each scenario ($\bar{P}_s$), PM2.5 per person ($\check{W}_{i,s}$), and injuries burden ($\check{I}_{a,g,m,o,s}$).

To assess uncertainty, we sample from the ITHIM output by sampling uncertain parameters multiple times and calculating the ITHIM. We specify parametric distributions or sampling strategies for all the uncertain parameters. Evaluating the ITHIM with uncertain parameters allows assessment of their impact on the outcome (AKA sensitivity analysis). We use EVPPI to calculate the expected reduction in uncertainty in the outcome were we to learn a parameter perfectly. This means we can implement models that are basic in their parametrisation, and learn at the end for which parameters it would be worthwhile spending dedicated time learning better.

Table 1: Indices used in the ITHIM variables. The "Label" is the subscript used to refer to the index in this document. The "Name" is descriptive. The "Levels" are the levels used in the Accra model.

Label	Name	Levels
a	Age group	15–49
		50–69
g	Gender	Male
		Female
h	Disease	
i	Individual index	
j	Trip index	
m	Transport mode	walk
		cycling (cyc.)
		bus
		car
		motorbike (mot.)
		goods vehicles (GV)
		subway (sub.)
o	Outcome	Death
		YLL
s	Scenario	
w	AP RR hyperparameter	$1, 2, 3, 4$
x	PA RR hyperparameter	$1, 2, 3$
z	PA dose	$0, ..., 300$

Table 2: Common abbreviations.

Abbreviation	Meaning
AP	Air pollution
CDF	Cumulative distribution function
DR	Dose response
EVPPI	Expected value of perfect partial information
GAM	Generalised additive model
GBD	Global burden of disease
ITHIM	Integrated transport and health impact model
MMET	Marginal metabolic equivalent task
PA	Physical activity
PM2.5	Particulate matter (diameter $< 2.5\mu$m)
PT	Public transport
RR	Relative risk
VOI	Value of information
YLL	Years of life lost

Table 3: All values input to the ITHIM model.

Name	Description	Value in Accra model	*ITHIM-R reference* Label	Model/ Setting
Global values				
$U_{a,g,h,o}$	Background burden of disease	Constant	GBD_DATA	Setting
$N_{a,g}$	Population by age and gender in GBD_DATA	Constant	GBD_DATA	Setting
$\bar{N}_{a,g}$	Population by age and gender	Constant	DEMOGRAPHIC	Setting
ρ_h	Chronic disease scalar	1 or Lnorm(0,0.18)	CHRONIC_DISEASE_SCALAR	Setting
Travel values				
$Z_{m=walk}$	Walk speed	4.8	MODE_SPEEDS	Setting
$Z_{m=cyc}$	Cycle speed	14.5	MODE_SPEEDS	Setting
$Z_{m=bus}$	Bus speed	15	MODE_SPEEDS	Setting
$Z_{m=car}$	Car speed	21	MODE_SPEEDS	Setting
$Z_{m=mot.}$	Motorbike speed	25	MODE_SPEEDS	Setting
ϵ	Walk time to bus	5 or Lnorm(log(5),0.18)	BUS_WALK_TIME	Setting
$\mu_{m=HGV}$	Truck distance relative to car	0.21 or Beta(3,10)	TRUCK_TO_CAR_RATIO	Setting
$\mu_{m=bus}$	Bus driver distance relative to bus passenger distance	1/45 or Beta(20,600)	BUS_TO_PASSENGER_RATIO	Setting
T_j	Trip-level data	Constant	SYNTHETIC_TRIPS	Setting
Injury model values				
$I_{a,g,m_{\text{vic}},m_{\text{str}}}$	Fatalities table	Constant	INJURY_TABLE	Setting
σ	Injury reporting rate	1 or Beta(8,3)	INJURY_REPORTING_RATE	Setting
ω	Injury linearity	1 or Lnorm(0,log(1.2))	INJURY_LINEARITY	Model
ψ	Casualty fraction of ω	0.5 or Beta(8,8)	CASUALTY_EXPONENT_ FRACTION	Model
Pollution model values				
η	Background PM2.5 concentration	50 or Lnorm(log(50),0.18)	PM_CONC_BASE	Setting
ζ	Fraction of measured PM2.5 concentration due to road transport	0.225 or Beta(5,20)	PM_TRANS_SHARE	Setting
γ_m	Vehicle emission inventory	Constant or Dirichlet	EMISSION_INVENTORY	Setting
P	Vehicle emission confidence	1	EMISSION_INVENTORY_CONFIDENCE	Setting
Pollution & health model values				
C_1	Base-level inhalation rate	1	BASE_LEVEL_INHALATION _RATE	Model
C_2	Exposure ratio window closed	0.5	CLOSED_WINDOW_PM_ RATIO	Model
C_3	Proportion of travel with closed windows	0.5	CLOSED_WINDOW_RATIO	Setting
C_4	Parameter for on-road PM2.5 level	3.216	ROAD_RATIO_MAX	Model

C_5	Parameter for on-road PM2.5 level	0.379	ROAD_RATIO_SLOPE	Model
C_6	Exposure ratio for underground	0.8	SUBWAY_PM_RATIO	Setting
$H_{a,h,w}$	AP dose–response–curve parameters	Constant	DR_AP	Model
$\xi_{h,w}$	Quantiles for AP RR functions	0.5 or Unif(0,1)	AP_DOSE_RESPONSE_ QUANTILE	Model
Physical activity model values				
χ	Day-to-week travel scalar	7 or 7×Beta(20,3)	DAY_TO_WEEK_TRAVEL_ SCALAR	Model
Y_i	Individual non-transport MMETs	Constant	PA_SET$work_ltpa_marg_met	Setting
θ	Background PA scalar	1 or Lnorm(0,0.18)	BACKGROUND_PA_SCALAR	Setting
$\tilde{Y}$	Confidence in PA survey	1	BACKGROUND_PA_CONFIDENCE	Setting
$\tilde{\theta}$	Background PA zeros quantile	0.5 or Unif(0,1)	BACKGROUND_PA_ZEROS	Setting
$\lambda_{m=\text{cyc.}}$	Cycling MET surplus	4.63 or Lnorm(log(4.63),1)	MMET_CYCLING	Model
$\lambda_{m=\text{walk}}$	Walking MET surplus	2.53 or Lnorm(log(2.53),1)	MMET_WALKING	Model
$\lambda_{m\notin\text{walk,cyc.}}$	In-vehicle MET surplus	0		Model
Physical activity & health model values				
$G_{h,x,z}$	Look-up table for relative risk truncated normal distribution	Constant		Model
ϕ_h	Quantiles for PA RR functions	0.5 or Unif(0,1)	PA_DOSE_RESPONSE_ QUANTILE	Model

2.1 Data module

Input data for the ITHIM are listed as Roman capital letters in Table 3. They include (vehicle) mode speeds, injury/fatality records, disease dose–response relationships, measures of air pollution, surveys of travel and physical activity, population numbers by demographic group, MMETs associated with active travel, and current burden of disease (see Figure 1).

2.1.1 Synthetic

We create a "synthetic population" by matching individuals surveyed for travel to the individuals surveyed for physical activity, based on age and gender. Then each person in the synthetic population has a set of trips taken and an amount of non-travel physical activity. This population is assumed to be representative of the population of the city or setting under study.

Alternatively/additionally we could create a synthetic population as described in Appendix D.2, which allows uncertainty in the propensity to travel. This method operates with person-level, rather than trip-level, information.

2.1.2 Scenarios

From the synthetic population we create "scenarios" of different pictures of travel by changing certain trips (e.g. swapping the mode of travel for particular trips and/or particular people). For the purposes of illustration, we choose a simple scenario for Accra in which every person gains one one-kilometre walk.

2.1.3 Distances

For the ITHIM, we process the travel data into distance data of various formats. First, we augment PT trips with a walk component: we take the set $\{T_j : m(j) = \text{bus}\}$, which has J entries. We add J journeys with mode "walk" and duration ϵ to the set T. (This step would not be necessary for a travel survey which already includes all stages for each trip.)

Then we summarise the total distances and durations. The distance set is labelled $\hat{T}_{m,s}$, representing the total distance travelled per mode per scenario, and the duration set is constructed analogously. We make an assumption that the distance for bus drivers scales linearly with bus passengers, where the distance in the baseline is defined based on the bus driver distance relative to car.

2.2 AP module

2.2.1 Background PM2.5

We use the total distance by mode ($\hat{T}_{m,s}$) to calculate the total PM2.5 in the scenario ($\bar{P}_s$). First we calculate the emission inventory in terms of fractions. If the confidence P is 1 then we use $\dot{P}_m = \gamma_m$ directly. If $P < 1$, we sample the emission inventory fractions as follows:

$$\tilde{\gamma}_m \sim \text{Dir}(f(P, \gamma_m)),$$

$$\dot{P}_m = \tilde{\gamma}_m$$

where we choose some mapping function such as $f(P, \gamma_m) = \frac{\gamma_m}{\sum_m \gamma_m} 10^{5P}$ (Figure 5).

Then we multiply vehicle distance by vehicle emission factor (where the emission factor is the emission inventory (fraction) divided by distance at baseline).

$$\tilde{P}_{m,s} = \dot{P}_m \frac{\hat{T}_{m,s}}{\hat{T}_{m,s=\text{baseline}}}.$$

For the baseline, $\tilde{P}_{m,s}$ is equal to the emission inventory fraction. For each scenario, it is scaled. We take the sum over modes to get a scalar for emissions in scenarios.

$$\hat{P}_s = \sum_m \tilde{P}_{m,s}.$$

Then the background PM2.5 in each scenario is the sum of the transport component and the non-transport component:

$$\bar{P}_s = \eta(\zeta \hat{P}_s + 1 - \zeta).$$

2.2.2 PM2.5 per person

Individual exposures to PM2.5 are calculated using the background PM2.5 ($\bar{P}_s$) and the trip sets. There are three major components to daily exposure: one, a person's total inhalation off road; two, a person's inhalation on road in a vehicle, and, three, a person's inhalation on road while undertaking active transport. Each category has an amount of background PM2.5 and a ventilation rate which together inform overall exposure.

The ratio of exposure off road to that on road is a function of total PM2.5, defined as

$$K_s = C_4 - C_5 \log(\bar{P}_s)$$

as in Goel et al. (2015). This defines the exposure of a person in an open vehicle (i.e. pedestrian, cyclist or motorist):

$$\tilde{K}_{m\in\{\text{walk,cyc.,mot.}\},s} = K_s$$

and it is used to calculate in-vehicle exposure, assuming an exposure of C_2 with the window closed, and a proportion C_3 of vehicles having closed windows:

$$\tilde{K}_{m\notin\{\text{walk,cyc.,mot.,sub.}\},s} = C_2C_3 + K_s(1 - C_3).$$

The exposure in a subway is constant and not dependent on the road ratio K_s:

$$\tilde{K}_{m=\text{sub.},s} = C_6.$$

Ventilation rates are calculated for each mode, assuming a base-level inhalation rate of C_1:

$$V_m = C_1 + \frac{1}{2}\lambda_m,$$

where λ_m is the MMETs for mode m.[1] Then the air inhaled during travel per person is

$$\tilde{V}_{i,m,s} = \sum_{\substack{j:i(j)=i,\\ m(j)=m,\\ s(j)=s}} T_j \cdot V_{m(j)}/Z_{m(j)}$$

and the rate of inhalation during travel is

$$\bar{V}_{i,s} = \sum_m \tilde{V}_{i,m,s} \cdot \tilde{K}_{m,s}.$$

The air inhaled when not travelling is

$$\hat{V}_{i,s} = C_1 \left(24 - \sum_{\substack{j:i(j)=i,\\ s(j)=s}} T_j/Z_{m(j)} \right).$$

Together, the PM2.5 exposure is calculated as the total PM2.5 inhaled per hour as follows:

$$\check{W}_{i,s} = \frac{\bar{P}_s(\bar{V}_{i,s} + \hat{V}_{i,s})}{24}.$$

2.2.3 AP–disease dose–response relative risk

We use each person's exposure to PM2.5 ($\check{W}_{i,s}$) to calculate their relative risk of five diseases (IHD, lung cancer, COPD, stroke, LRI), using curves parametrised by four disease-specific variables (Burnett et al., 2014). Of the five diseases, two (IHD and stroke) have parameters specific to age groups starting at age 25.[2] The other three (lung cancer, LRI and COPD) have one set of parameters for all ages.

[1] Ainsworth compendium 2011 sites.google.com/site/ compendiumofphysicalactivities for walking and cycling

[2] For any person of age lower than 25, we set the relative risk to 1.

The curves are in the form of samples of the set of four parameters. We model the densities of these samples (using a quantile for parameter 3, kernel density estimation for parameter 2, and GAMs for parameters 1 and 4)[3] in order to draw either the median or random samples via their quantiles ($\xi_{h,w}$) as follows:

$$\tilde{H}_{a,h,w=3} = \text{CDF}^{-1}_{H_{a,h,w=3}}(\xi_{h,w=3}) \tag{1}$$

$$\tilde{H}_{a,h,w=2} = \text{CDF}^{-1}_{H_{a,h,w=2}|\tilde{H}_{a,h,w=3}}(\xi_{h,w=2}) \tag{2}$$

$$\tilde{H}_{a,h,w=1} = \text{CDF}^{-1}_{H_{a,h,w=1}|\tilde{H}_{a,h,w=2},\tilde{H}_{a,h,w=3}}(\xi_{h,w=1}) \tag{3}$$

$$\tilde{H}_{a,h,w=4} = \text{CDF}^{-1}_{H_{a,h,w=4}|\tilde{H}_{a,h,w=3},\tilde{H}_{a,h,w=2},\tilde{H}_{a,h,w=1}}(\xi_{h,w=4}). \tag{4}$$

From these parameters, the relative risk of mortality is defined

$$\check{H}_{h,i,s} = 1 + \tilde{H}_{a=a(i),h,w=1} \times \left\{1 - \exp\left(-\tilde{H}_{a=a(i),h,w=2}(\check{W}_{i,s} - \tilde{H}_{a=a(i),h,w=4})^{\tilde{H}_{a=a(i),h,w=3}}\right)\right\}.$$

For $h \notin \{\text{IHD, lung cancer, COPD, stroke, LRI}\}$, $\check{H}_{h,i,s} \equiv 1$.

2.3 PA module

2.3.1 Individual-level MMETs

Using trip sets and the synthetic population, we calculate total MMETs per person ($M_{i,s}$) as the sum of walking MMETs per day and cycling MMETs per day, which are scaled up to a week via the scalar χ, and work/leisure MMETs per week as follows:

$$M_{i,s} = \sum_{\substack{j:i(j)=i,\\ s(j)=s}} \chi\lambda_{m(j)}\frac{T_j}{Z_{m(j)}} + \theta Y_i. \tag{5}$$

Here, λ_m is the MMETs for mode m, T_j the distance of trip j, Z_m the speed of mode m, θ the scalar for background PA, and $\check{Y}_i$ the amount of work and leisure (background) PA of person i.

We calculate $\check{Y}_i$ according to the confidence in the PA survey, $\tilde{Y}$. First, we calculate the raw probability that a person in demographic group completes no non-travel PA:

$$\dot{Y}_{a,g} = \text{Prob}(Y_i = 0)|_{a(i)=a,g(i)=g}.$$

Then, we set the probability to use, $\bar{Y}_{a,g}$, as $\dot{Y}_{a,g}$ if our confidence $\tilde{Y}$ is 1. If $\tilde{Y} < 1$, we map the confidence to parametrise a Beta distribution somehow, e.g. (Figure 4)

$$\hat{Y} = 500^{\tilde{Y}+0.2}, \tag{6}$$

$$\beta = \hat{Y}\dot{Y}_{a,g}\left(\frac{1}{\dot{Y}_{a,g}} - 1\right), \tag{7}$$

$$\alpha = \hat{Y} - \beta, \tag{8}$$

$$\bar{Y}_{a,g} = \text{CDF}^{-1}_{\text{Beta}(\alpha,\beta)}(\tilde{\theta}). \tag{9}$$

Finally, for each person in the population, we sample non-travel MMETs as zero with probability $\bar{Y}_{a,g}$ and from the raw non-zero density of their demographic group with probability $1 - \bar{Y}_{a,g}$.

2.3.2 PA–disease dose–response relative risk

We use each person's MMETs per week ($M_{i,s}$) to calculate their relative risk of six diseases, using curves found from meta analysis. Each disease (except type 2 diabetes) has a threshold beyond which there is no further change in relative risk. (Total cancer, breast cancer, colon & rectum cancer, endometrial cancer, & coronary heart disease: 35; lung cancer: 10; stroke: 32; all cause: 16.08.) For all individual cancers, we estimate the burden in terms of ... For all other diseases, we use mortality.

[3]The four parameters refer, in numerical order, to α, β, γ and tmrel in Burnett et al. (2014).

The interpolation of the dose–response curve can be described as follows:

$$\tilde{G}_{h,x}(m) = G_{h,x,z=\lfloor m \rfloor} + (G_{h,x,z=\lceil m \rceil} - G_{h,x,z=\lfloor m \rfloor}) \frac{m - \lfloor m \rfloor}{\lceil m \rceil - \lfloor m \rfloor}$$

where we interpolate the mean, the lower bound, and the upper bound. These define the normal from which we sample, truncated at 0:

$$F_S(v_x) = \mathcal{N}\left(v_{x=2}, \frac{v_{x=3} - v_{x=1}}{1.96}\right)\Big|_{F_S(v_x)>0}.$$

Then the relative risk for each person for each disease is calculated for the quantile ϕ_h as

$$W_{h,i,s} = \mathrm{CDF}^{-1}_{F_S\left(\tilde{G}_{h,x}(M_{i,s})\right)}(\phi_h).$$

See Figure 2 for an example. For $h \notin \{$diabetes, total cancer, breast cancer, colon & rectum cancer, endometrial cancer, IHD, lung cancer, stroke, all cause$\}$, $W_{h,i,s} \equiv 1$.

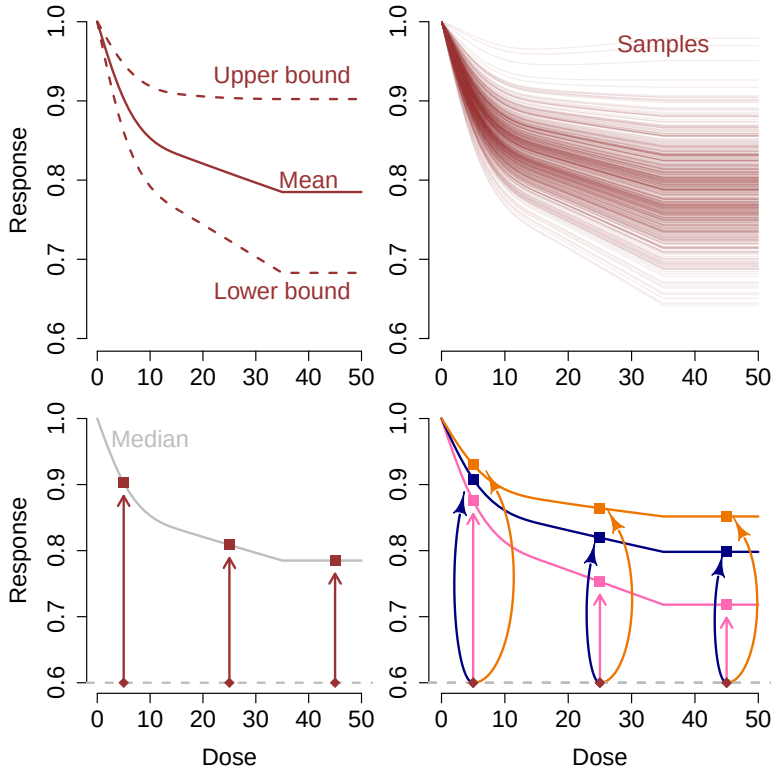

Figure 2: Example of the dose–response workflow for PA and total cancer. Top left: results of the metaanalysis: mean, lower bound and upper bound of the relationship between PA dose and relative risk of disease. Top right: examples of 500 samples from the normal distribution defined by the mean of the top-left panel and the standard deviation defined by (upper bound minus lower bound) divided by 1.96 and truncated at 0. Bottom left: relative risks of individuals with PA doses 5, 25 and 45 are found by mapping their doses onto the median curve. This is used in the "constant" ITHIM use case. Bottom right: relative risks of individuals with PA doses 5, 25 and 45 for three different samples from the distribution shown in the top-right panel. For each sample, each individual is mapped onto the response curve it defines.

2.4 Injury module

2.4.1 Processing

For the injury module, we require distance per travel mode per demographic group per scenario. It requires matching of mode names in the travel survey to mode names in the injury records, for example "taxi", "shared auto", "shared

taxi" and "car" combine to form "car". Some work is required to separate drivers from passengers; we currently assume all travellers are drivers, with the exception of "bus".[4]

2.4.2 Modelling

We model injuries via regression by predicting the number of fatalities of each demographic group on each mode, which we calculate as a sum over all the ways in which they might have been injured (i.e., all the modes with which they might have collided).

The predictive covariates include the distances travelled by the parties and their demographic details. These requirements lead to a natural separation of the (training) dataset into two groups: a set for which we have distance data for the other party, and a set for which we do not have distance data for the other party. The former equates to a "who hit whom" ("whw") matrix (albeit in a higher dimension), and will account for changes in injuries resulting of a change in strike-mode travel. The latter corresponds to causes of injury that will not change across scenarios, including "no other vehicle" and modes of transport that we do not consider to change, which might include trucks and buses if they are not somehow explicitly included in the trip set. We label this group "noov": no or other vehicle.

We model the number of injuries as a Poisson-distributed variable with an offset depending on the distance(s) and the reporting rate. We write the model as follows:

$$\tilde{I}_{a,g,m_{\text{cas}},m_{\text{str}},s,y} \sim \text{Poisson}(\bar{I}_{a,g,m_{\text{cas}},m_{\text{str}},s}); \tag{10}$$

$$\begin{aligned}\log\left(\bar{I}_{a,g,m_{\text{cas}},m_{\text{str}},s} | m_{\text{str}} \in \{m_{\text{whw}}\}\right) =& \kappa_0 + \sum_{i=1}^{n} \kappa_i X_i - \log(\sigma) + \\ & \log(\bar{T}_{a,g,m_{\text{cas}},s}) - E_{m_{\text{cas}}} \log(\hat{T}_{m_{\text{cas}},s}) + (1 - E_{m_{\text{str}}}) \log(\hat{T}_{m_{\text{str}},s});\end{aligned} \tag{11}$$

$$\log\left(\bar{I}_{a,g,m_{\text{cas}},m_{\text{str}},s} | m_{\text{str}} \in \{m_{\text{noov}}\}\right) = \kappa'_0 + \sum_{i=1}^{n} \kappa'_i X'_i - \log(\sigma) + \log(\bar{T}_{a,g,m_{\text{cas}},s}). \tag{12}$$

We choose this form of equation to enable (a) linearity in injuries with respect to total travel combined across modes and (b) linearity in injuries as subdivided by travellers of each mode.[5]

In general, one makes the best model that one can with the data available. We use, for Accra, the covariates `casualty mode*strike mode, casualty age,` and `casualty gender`, to form the model matrix X. We also have another covariate, `year`, but our travel data are for one year only, so we model instead many observations of the same quantities.

We fit the coefficients κ of the model using data for ten years, assuming the same travel each year, which corresponds to the scenario $s =$ baseline. To incorporate the injury reporting rate (σ), we set it as an offset, whose value is 1 in fitting the model, and as specified in making predictions.

2.4.3 Prediction

We predict the number of injuries in each scenario based on the training model built from the baseline scenario. For Accra, we predict for the year 2016, using scenario-specific travel data. The number of fatalities is taken as the sum of fatalities over the strike modes:

$$\check{I}_{a,g,m,o=\text{death},s} = \sum_{m_{\text{str}}} \tilde{I}_{a,g,m_{\text{cas}},m_{\text{str}},s}.$$

We extrapolate injury deaths to injury YLLs via the ratio in the GBD:

$$R_{a,g} = U_{a,g,h=\text{road injuries},o=\text{YLL}} / U_{a,g,h=\text{road injuries},o=\text{death}},$$

$$\check{I}_{a,g,m,o=\text{YLL},s} = \check{I}_{a,g,m,o=\text{death},s} \cdot R_{a,g}.$$

[4] This is a problem for the examples of Delhi and Bangalore, whose trip sets have car travellers under the age at which driving is permitted and who therefore should not contribute to striking distance.

[5] Note that the assymmetry in the offset term for the whw expression arises due to the absence of demographic information for striking vehicles. Were we to have that, the offset would read $\log(\bar{T}_{a,g,m_{\text{cas}},s}) - E_{m_{\text{cas}}} \log(\hat{T}_{m_{\text{cas}},s}) + \log(\bar{T}_{a,g,m_{\text{str}},s}) - E_{m_{\text{str}}} \log(\hat{T}_{m_{\text{str}},s})$. For our case, the latter two terms simplify to $\log(\hat{T}_{m_{\text{str}},s}) - E_{m_{\text{str}}} \log(\hat{T}_{m_{\text{str}},s}) = (1 - E_{m_{\text{str}}}) \log(\hat{T}_{m_{\text{str}},s})$.

2.5 Health module

2.5.1 PA and AP relative risks combined

We combine the relative risks of disease as a function of AP and disease as a function of PA through multiplication:

$$\tilde{W}_{h,i,s} = W_{h,i,s} \cdot \check{H}_{h,i,s}.$$

Not all diseases have a dose–response relationship for both AP and PA. Just stroke, lung cancer, and IHD have both. For the other diseases, only one of $W_{h,i,s}$ and $\check{H}_{h,i,s}$ is different from one.

2.5.2 Total disease burden

We calculate the total health burden relative to the reference scenario (which, for our Accra example, is the baseline) using the injury and health outputs combined with GBD data, via population attributable fractions (PAF). The PAF is defined:

$$\hat{W}_{a,g,h,s} = \sum_{\substack{i:a(i)=a,\\ g(i)=g}} \tilde{W}_{h,i,s}.$$

We then calculate the PAF relative to the reference scenario ("ref") as:

$$\bar{W}_{a,g,h,s} = \frac{\hat{W}_{a,g,h,s=\text{ref}} - \hat{W}_{a,g,h,s}}{\hat{W}_{a,g,h,s=\text{ref}}}.$$

We estimate the background burden of disease using Global Burden of Disease data and scaling based on the ratio of populations between country and city. For country populations $N_{a,g}$, city populations $\bar{N}_{a,g}$ and country burden $U_{a,g,h,o}$, we estimate city burden as

$$\bar{U}_{a,g,h,o} = U_{a,g,h,o} \frac{\bar{N}_{a,g}}{N_{a,g}}.$$

If we are scaling the background burden of non-communicable diseases ("Neoplasms", "Ischemic heart disease", "Tracheal, bronchus, and lung cancer", "Breast cancer", "Colon and rectum cancer", "Uterine cancer"), we do so here. (Otherwise set $\rho \equiv 1$.)

$$\tilde{U}_{a,g,h,o} = \rho_h \cdot \bar{U}_{a,g,h,o}.$$

We combine the burden with the PAF:

$$\hat{U}_{a,g,h,o,s} = \bar{W}_{a,g,h,s} \cdot \tilde{U}_{a,g,h,o}.$$

2.5.3 Injury burden

And, finally, for the injuries, we sum over modes to compute the burden,

$$\bar{U}_{a,g,o,s} = \sum_m \check{I}_{a,g,m,o,s},$$

and subtract from the values for the reference scenario:

$$\hat{U}_{a,g,h=\text{road injury},o,s} = \bar{U}_{a,g,o,s=\text{ref}} - \bar{U}_{a,g,o,s}.$$

3 Example: Accra

Here we show the ITHIM as evaluated for the city of Accra, with a simple scenario of every person in the synthetic population gaining a single, 1 km walk. We define XX uncertain univariate parameters (Section 3.1), two confidences, and two uncertain parameter sets (Section 3.3).

3.1 Parametric distributions for uncertain variables

The XX uncertain univariate parameters are shown in Figure 3. Most of the parameters are self explanatory. Cycling and walking MMETs are the number of MMETs per hour when undertaking cycling and walking, and determine also the ventilation rates. Motorcycle distance is the total distance travelled by motorcycles relative to the total distance travelled by cars in the baseline scenario. Non-travel PA, injury reporting rate and NCD burden all act as scalars for the corresponding datasets. Note that the non-travel PA scalar does not affect the ≈40% of the population whose non-travel PA is 0.

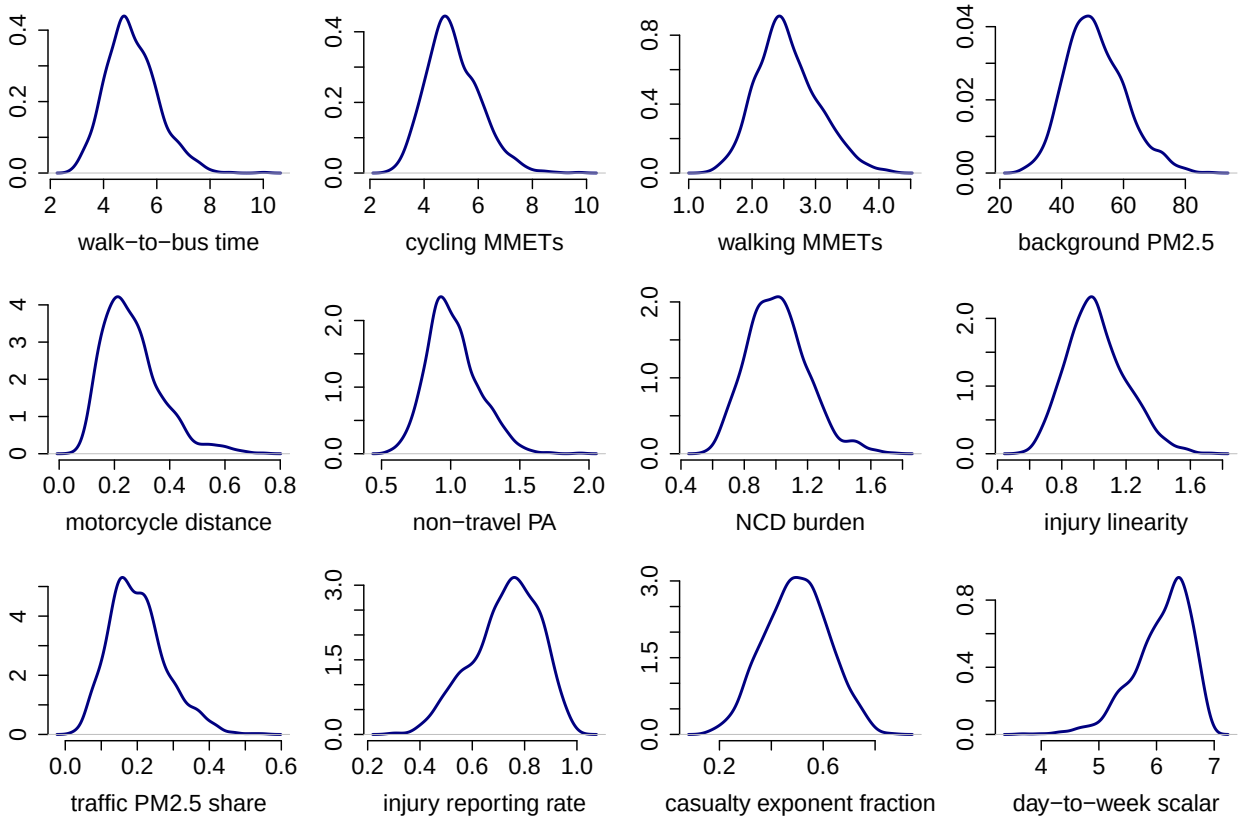

Figure 3: Distributions of univariate uncertain parameters (1024 samples).

3.2 Confidences

For physical activity propensity and emission inventories, we use "confidences", values between 0 and 1 that represent how confident we are about the data source. We map these values to Beta (Figure 4) and Dirichlet (Figure 5) distributions, respectively.

3.3 Dose–response relationships

For the dose–response relationships between physical activity (PA) and disease and air pollution (AP) and disease, we assume that there is uncertainty, but no variability, in the relationship. This means that we sample a relationship from the distribution of relationships, and apply that relationship to all individuals precisely. This means that, given fixed doses, responses between individuals will be perfectly correlated.

We achieve this by use of the probability integral transform: we sample a random variable uniformly distributed on the space (0,1) and map it, via a cumulative distribution function, to the distribution describing the dose–response relationship. See Figure 6 for an illustration.

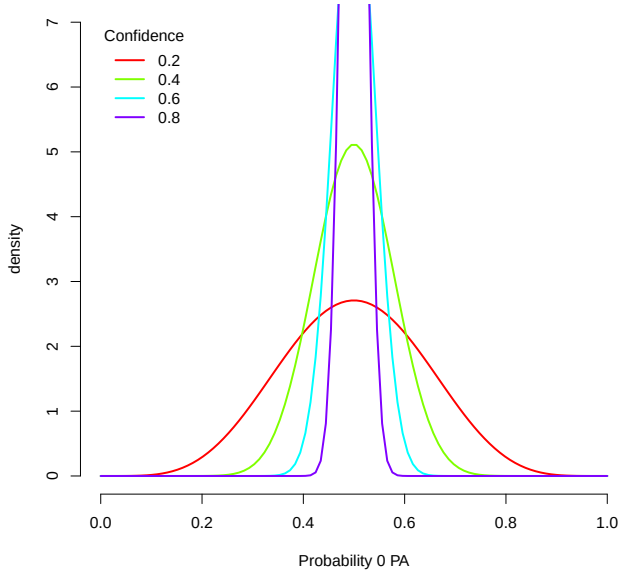

Figure 4: Distributions of a Beta-distributed propensity to non-travel PA with varying confidence values in the raw input data. The sampled value gives the fraction of the population that engages in non-travel PA. [Re-do with Accra parameters.]

3.3.1 Physical activity

Each disease's PA dose–response relationship is defined by a truncated normal. For each dose, there is a mean value, an upper bound, and a lower bound. For each person's dose, we get the response by mapping the uniform random variable onto the truncated normal defined by the mean and bounds for that dosage.

3.3.2 Air pollution

For the AP relationship, there are four parameters per disease. We sample the first from an empirical distribution using the probability integral transform. We sample the second via the same method, conditioned on the value of the first, constructing their joint density with e.g. kde2d. The third parameter is sampled conditioned on the first and second, constructing their joint density using a GAM. The final parameter is sampled conditioned on the first, second, and third, constructing their joint density using a GAM.

As before, there is perfect correlation between individuals, i.e. if person A's dose is greater than person B's, then person A's response is strictly greater than person B's response.

The empirical distributions come from Burnett et al. (2014). There are four parameters per disease: IHD, lung cancer, COPD, and stroke. In addition, for stroke and IHD, there is a set of four parameters for each age group from 25 to 95 in five-year increments. In addition to our assumption that there is perfect correlation between individuals for diseases, we assume perfect correlation between ages for diseases. I.e., our four quantiles per disease will be applied to all age groups.

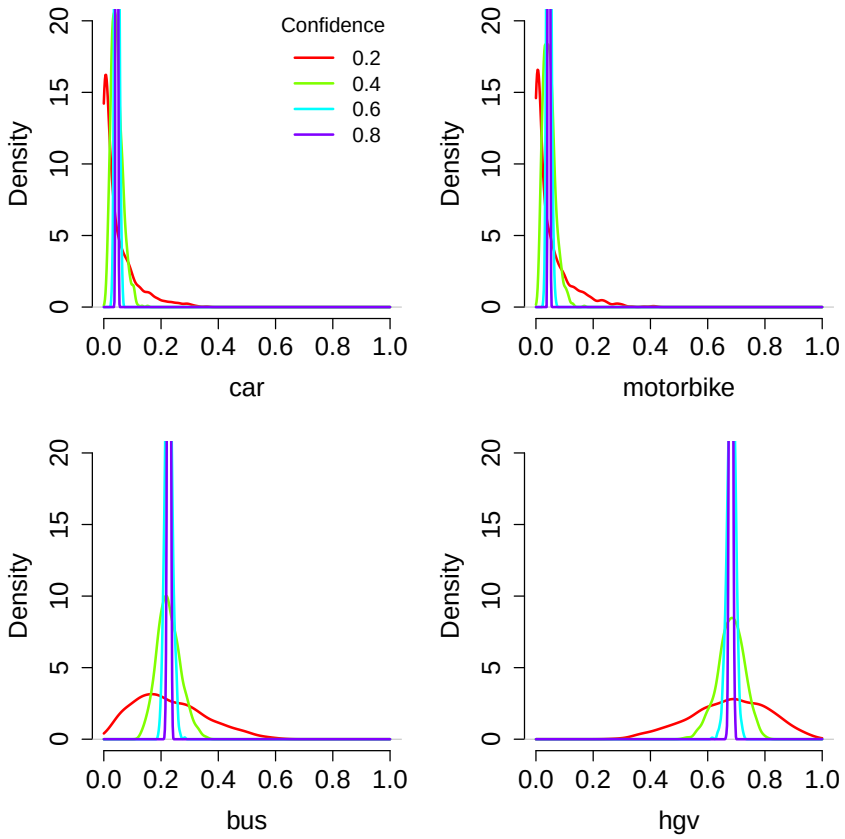

Figure 5: Distributions of Dirichlet-distributed emission inventories for four modes and varying confidence values. Raw input values were 4, 4, 20, and 60. [Re-do with Accra parameters and inventory.]

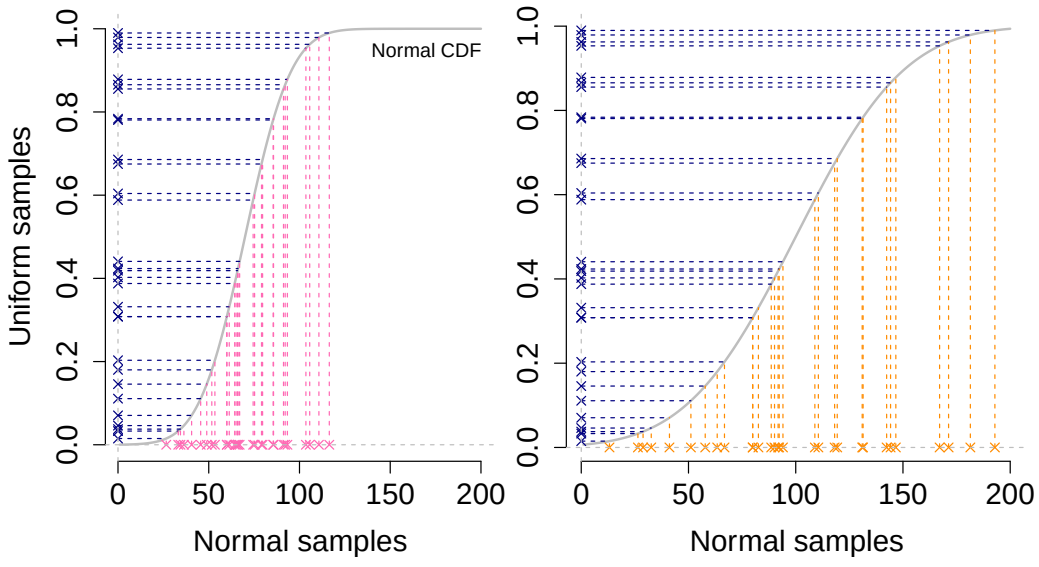

Figure 6: We use the same uniform samples to generate random variables from different distributions. It means that the samples are perfectly correlated.

4 Results

4.1 Health burdens in scenarios

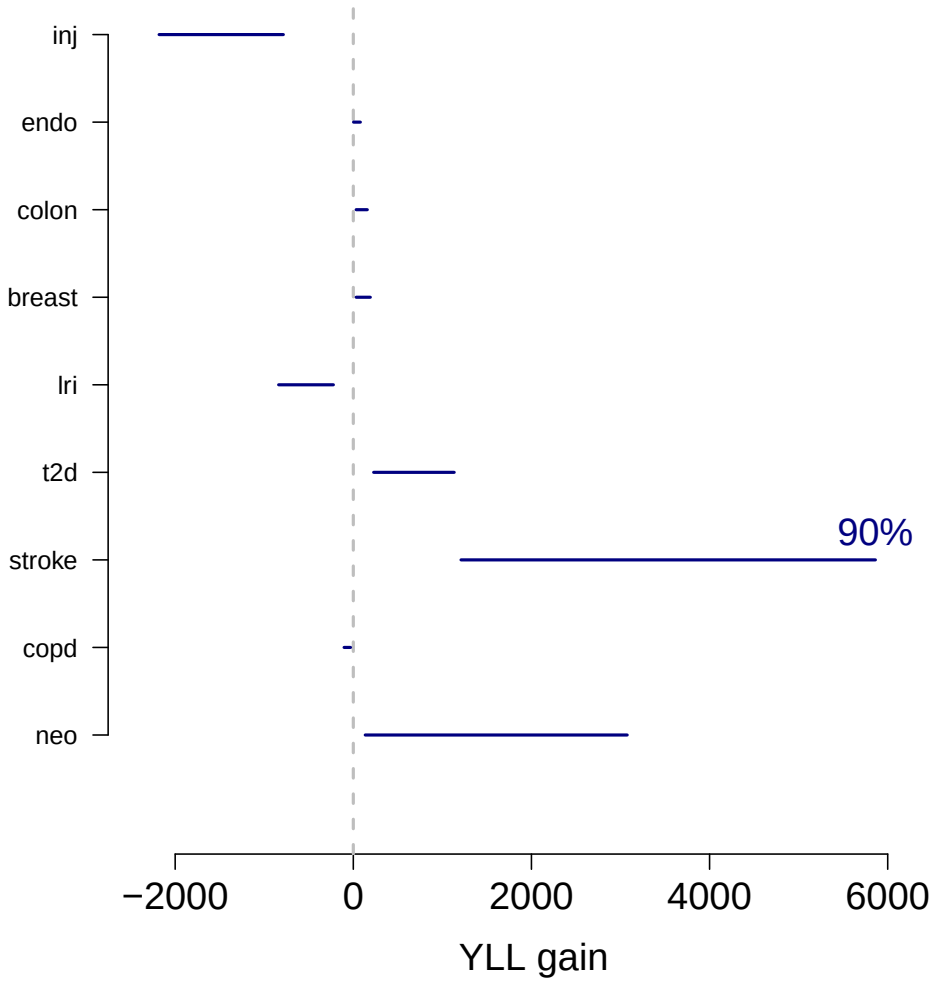

Figure 7: YLLs in the scenario minus YLLs in the baseline for Accra.

4.2 Value of information

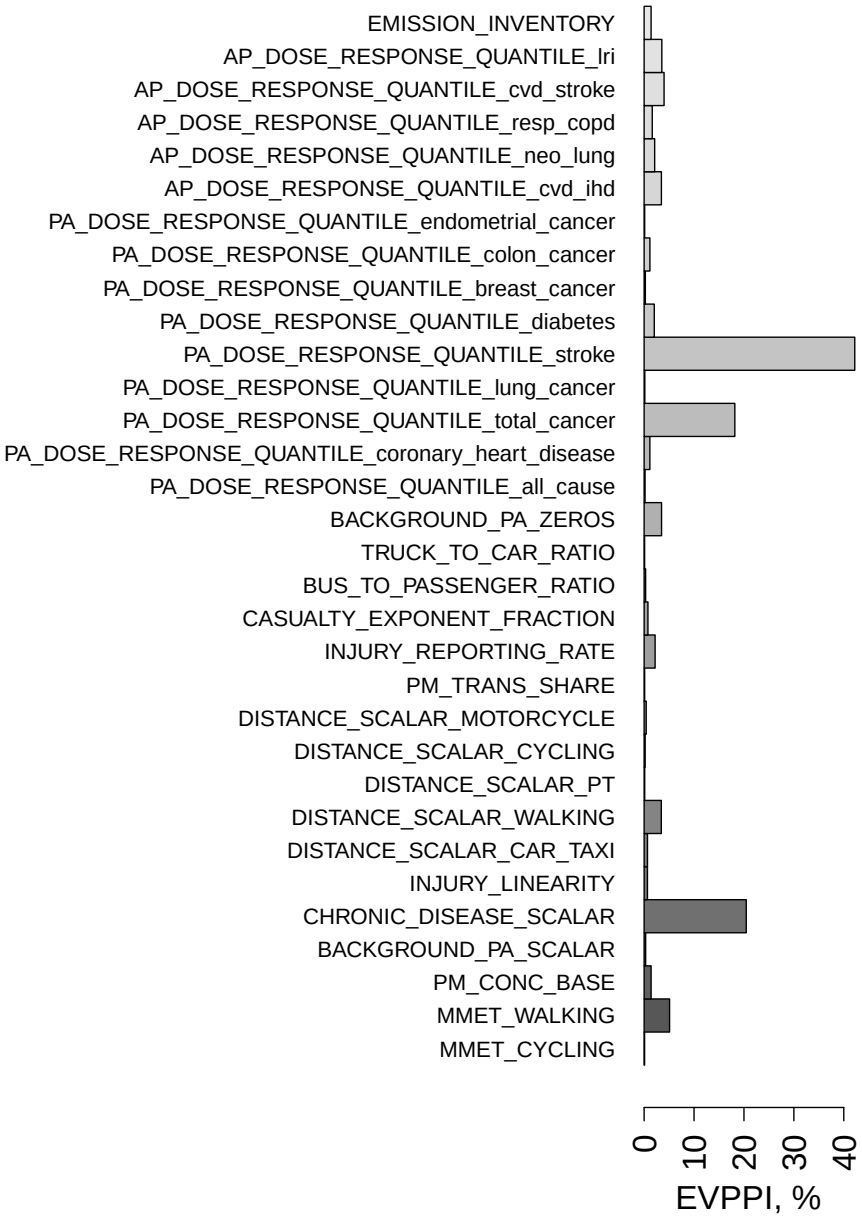

Figure 8: EVPPI for Accra's "walking" scenario for all causes of YLLs excluding "all cause" and "neoplasm".

A ITHIM with parameters

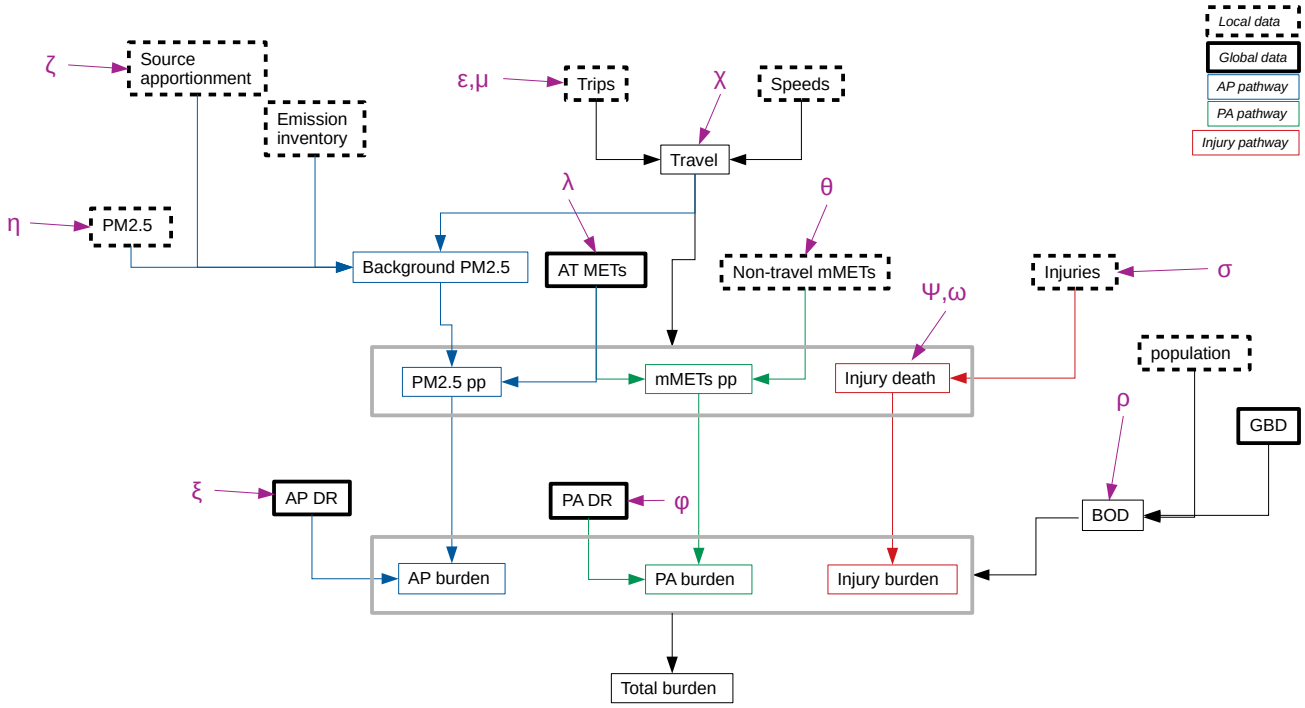

Figure 9: ITHIM-R workflow, with variable parameters in purple. Fixed inputs to the model have a bold, black outline, which is solid for inputs we consider "global" and dashed for inputs we consider "local". There are four global inputs, which will be embedded in ITHIM-R. There are seven local inputs, which users should provide. The graph depicts parent–child relationships, where a child (at the head of an arrow) depends upon all its parents (at the source(s) of the arrow(s)). To aid visualisation, an arrow connected to the outside of a grey box indicates that the arrow connects to each item within the box. *AT: active travel. MET: metabolic equivalent task. pp: per person. AP: air pollution. PA: physical activity. DR: dose–response relationship. GBD: global burden of disease.* NB: ρ only impacts the NCD burden of GBD, not all of it. Missing parameter: motorcycle distance. λ is a 2D parameter. ξ is 4D.

B Dose–response relationships

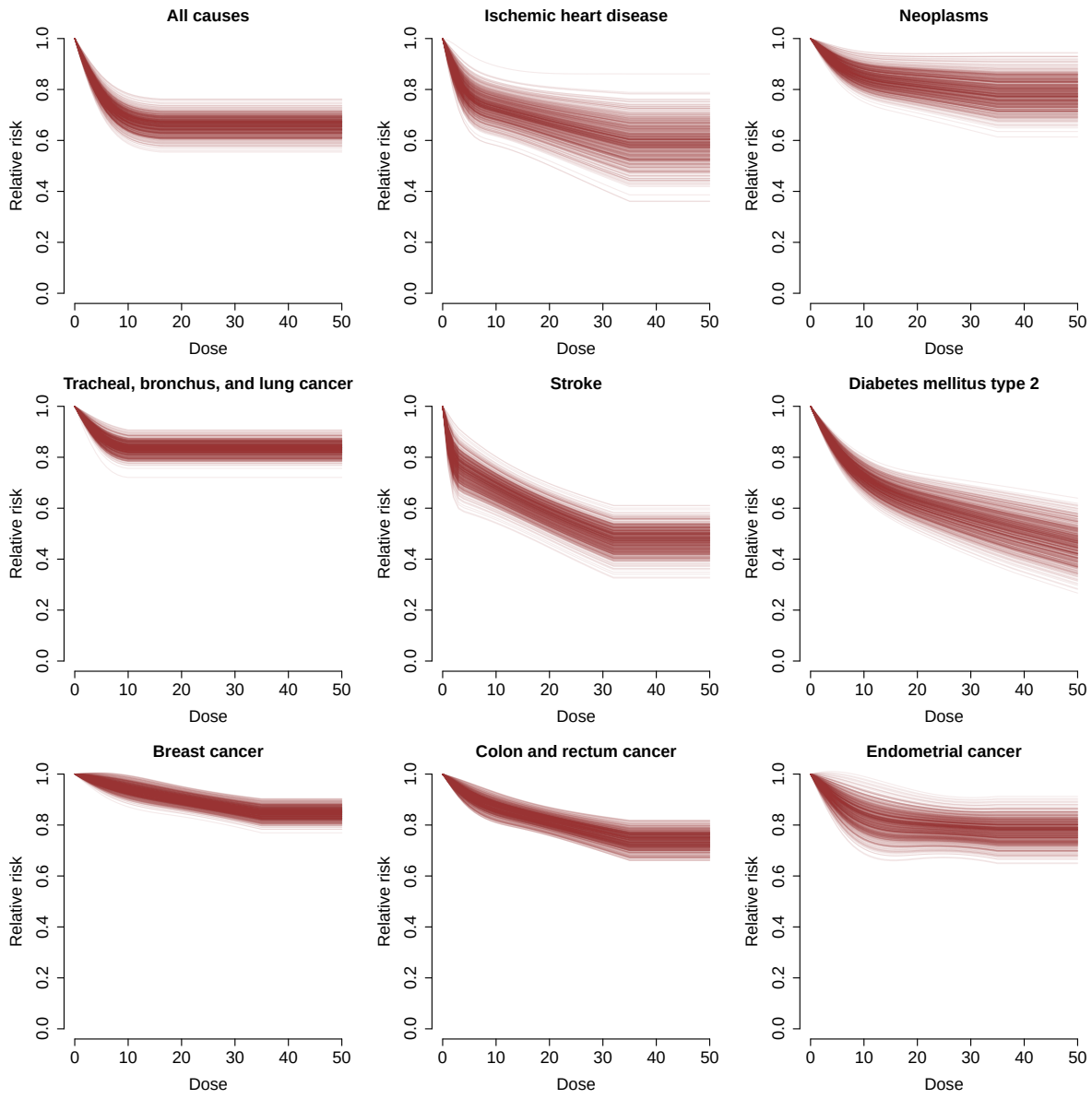

Figure 10: 1000 samples from PA dose–response curves

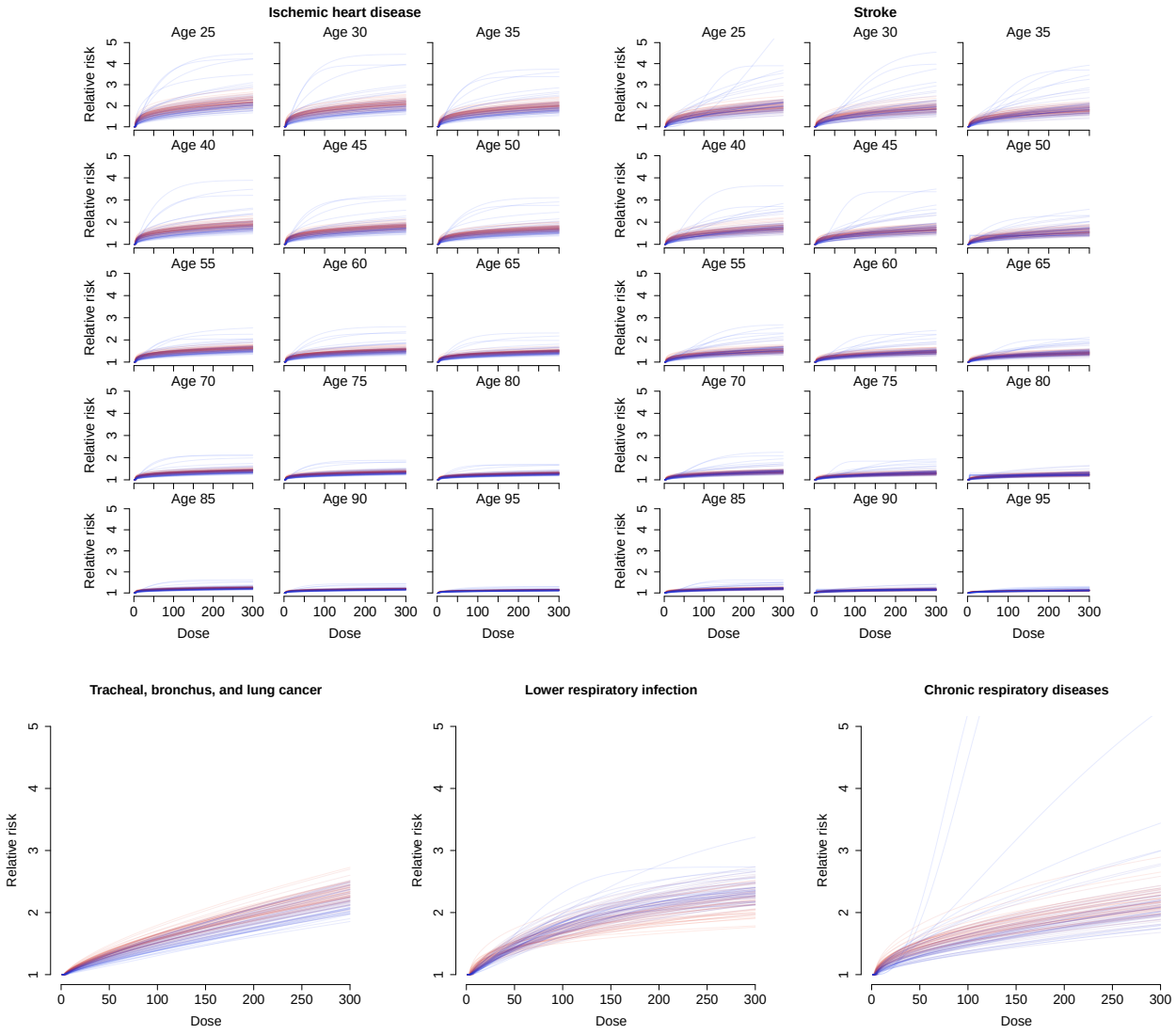

Figure 11: 100 samples from AP dose–response curves

C Tabulated ITHIM equations

Table 4: All calculations in the ITHIM.

Name	Description	Label
Scenario generation		
...		
Travel calculations		
$\{T_{j^+} \leftarrow \epsilon; m(j^+) \leftarrow \text{walk}\}\, \forall\, T_j : m(j) = \text{bus}$	New walk trips to bus	walk_to_bus_and_combine_scen
$\bar{T}_{a,g,m,s} = \sum_{j:a(j)=a,g(j)=g,m(j)=m,s(j)=s} Q_{i(j)} T_j$	(Weighted) total travel per age, gender, mode and scenario	
$\check{T}_{a,g,m,s} = \frac{\bar{T}_{a,g,m,s}}{\sum_{a,g} \bar{T}_{a,g,m,s}}$	Relative travel per mode and scenario by age and gender	inj_distances[[1]]
$\hat{T}_{m,s} = \sum_{a,g} \bar{T}_{a,g,m,s}$	Total travel per mode and scenario	dist
$\tilde{T}_{m,s} = \frac{\hat{T}_{m,s}}{\hat{T}_{m,s=\text{baseline}}}$	Total travel per mode and scenario relative to baseline, with car modes summed, and Tuktuk = 1	inj_distances[[2]]
Injury calculations		
$E_{m_{\text{cas}}} = \psi\omega$	Casualty exponent	CAS_EXPONENT
$E_{m_{\text{str}}} = \omega - E_{m_{\text{cas}}}$	Striker exponent	STR_EXPONENT
$\tilde{I}_{a,g,m_{\text{vic}},m_{\text{str}},s} = \mathcal{S}\left(\sigma \cdot I_{a,g,m_{\text{vic}},m_{\text{str}}}, \bar{T}_{a,g,m,s}, E_{m_{\text{cas}}}, E_{m_{\text{str}}}\right)$	Fatalities in scenarios by strike mode	injuries
$\check{I}_{a,g,m,o=\text{death},s} = \sum_{m_{\text{str}}} \tilde{I}_{a,g,m_{\text{vic}},m_{\text{str}},s}$	Fatalities in scenarios	victim_deaths
$R_{a,g} = U_{a,g,h=\text{road injuries},o=\text{YLL}} / U_{a,g,h=\text{road injuries},o=\text{death}}$	Road YLL-to-death ratio	GBD_INJ_YLL
$\check{I}_{a,g,m,o=\text{YLL},s} = \check{I}_{a,g,m,o=\text{death},s} \cdot R_{a,g}$	YLLs scaled from fatalities	deaths_yll_injuries
Pollution calculations		
$\dot{P}_m = \begin{cases} \gamma_m & P = 1 \\ \tilde{\gamma}_m, \tilde{\gamma}_m \sim \text{Dir}(f(P, \gamma_m)) & P < 1 \end{cases}$	Emissions by mode	trans_emissions
$\tilde{P}_{m,s} = \dot{P}_m \hat{T}_{m,s} / \hat{T}_{m,s=\text{baseline}}$	Emissions by mode in scenario relative to baseline	trans_emissions
$\hat{P}_s = \sum_m \tilde{P}_{m,s}$	Emissions scalar for scenarios	baseline_sum
$\bar{P}_s = \eta(\zeta \hat{P}_s + 1 - \zeta)$	Background PM2.5 concentration in scenarios	scenario_pm
$K_s = C_4 - C_5 \log(\bar{P}_s)$	On-road PM2.5 exposure ratio	on_road_off_road_ratio
$\tilde{K}_{m \in \{\text{walk,cyc.,mot.}\},s} = K_s$	On-road PM2.5 exposure ratio for open vehicle	on_road_off_road_ratio

$\tilde{K}_{m\in\{\text{sub.}\},s} = C_6$	Subway PM2.5 exposure ratio	subway_ratio
$\tilde{K}_{m\notin\{\text{walk,cyc.,mot.,sub.}\},s} = C_2C_3 + K_s(1 - C_3)$	In-vehicle PM2.5 exposure ratio	in_vehicle_ratio
$V_m = C_1 + \frac{1}{2}\lambda_m$	Ventilation rates	vent_rates
$\tilde{V}_{i,m,s} = \sum_{j:i(j)=i,m(j)=m,s(j)=s} T_j \cdot V_{m(j)}/Z_{m(j)}$	In-travel air inhaled	
$\bar{V}_{i,s} = \sum_m \tilde{V}_{i,m,s} \cdot \tilde{K}_{m,s}$	In-travel PM2.5 inhalation rate	
$\hat{V}_{i,s} = C_1 \left(24 - \sum_{j:i(j)=i,s(j)=s} T_j/Z_{m(j)}\right)$	Non-travel air inhaled	
$\check{W}_{i,s} = \bar{P}_s(\bar{V}_{i,s} + \hat{V}_{i,s})/24$	Total PM2.5 inhaled per hour	
$\tilde{H}_{a,h,w=3} = \text{CDF}^{-1}_{H_{a,h,w=3}}(\xi_{h,w=3})^\dagger$	AP dose–response–curve parameter	DR_AP_LIST
$\tilde{H}_{a,h,w=3} = \text{CDF}^{-1}_{H_{a,h,w=3}\mid\tilde{H}_{a,h,w=3}}(\xi_{h,w=3})^\dagger$	AP dose–response–curve parameter	DR_AP_LIST
$\tilde{H}_{a,h,w=3} = \text{CDF}^{-1}_{H_{a,h,w=3}\mid\tilde{H}_{a,h,w=3},\tilde{H}_{a,h,w=3}}(\xi_{h,w=3})^\dagger$	AP dose–response–curve parameter	DR_AP_LIST
$\tilde{H}_{a,h,w=3} = \text{CDF}^{-1}_{H_{a,h,w=3}\mid\tilde{H}_{a,h,w=3},\tilde{H}_{a,h,w=3},\tilde{H}_{a,h,w=3}}(\xi_{h,w=3})^\dagger$	AP dose–response–curve parameter	DR_AP_LIST
$\check{H}_{h,i,s} = 1 + \tilde{H}_{a=a(i),h,w=3}\times$ $\left\{1 - \exp\left(-\tilde{H}_{a=a(i),h,w=3}(\check{W}_{i,s} - \tilde{H}_{a=a(i),h,w=3})^{\tilde{H}_{a=a(i),h,w=3}}\right)\right\}$	RR of disease given pollution	RR_AP_calculations
Physical activity calculations		
$\dot{Y}_{a,g} = \text{Prob}(Y_i = 0)\vert_{a(i)=a,g(i)=g}$	Raw probability non-travel MMETs = 0 for demographic group a, g	raw_zero
$\bar{Y}_{a,g} = \begin{cases} \dot{Y}_{a,g} & \tilde{Y} = 1 \\ \text{CDF}^{-1}_{\text{Beta}(f(\tilde{Y},\dot{Y}_{a,g}))}(\tilde{\theta}) & \tilde{Y} < 1 \end{cases}$	Probability non-travel MMETs = 0 for demographic group a, g	zeros
$\check{Y}_i = \begin{cases} 0 & \bar{Y}_{a(i),g(i)} \\ \text{Sample}(\{Y_{i'}\}\vert_{a(i')=a(i),g(i')=g(i),Y_{i'}>0}) & 1 - \bar{Y}_{a(i),g(i)} \end{cases}$	Non-travel MMETs	SYNTHETIC_POPULATION$ work_ltpa_marg_met
$M_{i,s} = \sum_{j:i(j)=i,s(j)=s}(\chi T_j \cdot \lambda_{m(j)}/Z_{m(j)}) + \theta\check{Y}_i$	Total MMETs per person	mmets
$\tilde{G}_{h,x}(m) = G_{h,x,z=\lfloor m\rfloor} + (G_{h,x,z=\lceil m\rceil} - G_{h,x,z=\lfloor m\rfloor})\frac{m-\lfloor m\rfloor}{\lceil m\rceil-\lfloor m\rfloor}$	Interpolated dose–response parameters	rr, lb, ub
$F_S(s_x) = \mathcal{N}(s_{x=2}, (s_{x=3} - s_{x=1})/1.96)\vert_{0<F_S(s_x)}$	Truncated normal function	
$W_{h,i,s} = \text{CDF}^{-1}_{F_S\left(\tilde{G}_{h,x}(M_{i,s})\right)}(\phi_h)^\dagger$	Relative risk of disease given PA	RR_PA_calculations
Burden-of-disease calculations		
$\tilde{W}_{h,i,s} = W_{h,i,s} \cdot \check{H}_{h,i,s}$	Combined PA and AP relative risks	RR_PA_AP_calculations
$\hat{W}_{a,g,h,s} = \sum_{i:a(i)=a,g(i)=g} \tilde{W}_{h,i,s}$	Population-attributable fractions	pif_temp
$\bar{W}_{a,g,h,s} = (\hat{W}_{a,g,h,s=\text{scen1}} - \hat{W}_{a,g,h,s})/\hat{W}_{a,g,h,s=\text{scen1}}$	Population-attributable fractions relative to scenario 1	pif_scen

$\bar{U}_{a,g,h,o} = U_{a,g,h,o}\frac{\bar{N}_{a,g}}{N_{a,g}}$	GBD scaled to local population	DISEASE_BURDEN
$\tilde{U}_{a,g,h,o} = \rho_h \cdot \bar{U}_{a,g,h,o}$	Scaled background disease for Neoplasms, IHD, Lung cancer	gbd_data_scaled
$\hat{U}_{a,g,h,o,s} = \bar{W}_{a,g,h,s} \cdot \tilde{U}_{a,g,h,o}$	Combined health and PIF	yll_dfs, death_dfs
$\hat{U}_{a,g,h=\text{road injury},o,s} = \sum_m \check{I}_{a,g,m,o,s}$	Injury health burden	inj

† CDF: cumulative distribution function. We use generative distributions to generate random numbers so that samples are correlated. For $t = \text{CDF}_F^{-1}(q)$, t is the q-th quantile of function F with CDF CDF_F. (I'd like to find an improved notation for this.)

D Uncertainty in travel data

We model uncertainty in travel via the definition of "travel attributes" (in the terminology of Mohammadian et al. (2010)). Specifically, we define the propensity to travel in a day, and the distance travelled in a day, given that travelled occurred. Each travel attribute is defined per demographic group, per mode.

In contrast to other travel-related synthetic populations, we have four demographic groups (rather than clusters) which consist of individuals (rather than households). A crucial difference between our objectives and the existing work on simulating populations in travel-demand models (that I've found so far) is that we don't need trip-level data for our calculations. Simulating trip-level data is computationally expensive and high-dimensional in terms of uncertain parameters. For our downstream computations, we need only summary data: travel per person per mode. Beyond scenario generation, there is no need for us to consider trips.

E.g. Saadi et al. (2016) focuses on (a) synthetic population via matching covariates and (b) the sequence of trips taken by each person. (a) isn't a big thing for Accra; we're only using two covariates. And for (b), I think we can focus on people as it's a lot simpler. For our purposes, the things in this work that might be useful to us include matching principles for (a), when we revisit it, and distributions (in particular joint distributions) for travel. More useful to us would be distributions for total travel, but I haven't found any in this literature (e.g. zero-inflated compound Poisson distribution).

In contrast, my main questions are:

- How to account for multi-mode trips.
- What to do about zero travel for e.g. older cyclists. One possibility is to smooth over all modes and/or demographic groups.
- In the ITHIM-R programme, are bus and truck drivers participants?
- How should we include correlations between modes? This could also be aided by a smooth model across modes.
- Whether there is a better method to get distance travelled per person than by resampling raw data.

D.1 Method

D.1.1 Summarise the travel survey

Our first step is to summarise the demographic information and the travel survey. In Table 5 are the demographic data: the populations of the city, and the number of people in the survey. In Table 6 are the summary statistics from the travel survey which we will use to generate synthetic travel. From the raw travel data (actually the adjusted trip set, which is the raw travel plus motorcycle, bus and truck trips), we calculate the probability to travel, $B_{a,g,m}$, as the number of people who travelled by that mode divided by the number of people in the demographic group. It is to these values that we assign uncertainty through the variable PROPENSITY_TO_TRAVEL.

Table 5: Population by demographic group.

Age	Gender	Surveyed	Accra population
15–49	Male	279	5,847,716
50–69	Male	56	1,016,476
15–49	Female	328	6,360,535
50–69	Female	69	1,110,037

Table 6: Travel summary table.

Age	Gender	Mode	Probability in raw trip set	Probability in baseline trip set
15-49	Male	Bus	0.4014	0.3862
50-69	Male	Bus	0.4286	0.3636
15-49	Female	Bus	0.3506	0.3495
50-69	Female	Bus	0.4203	0.4203

15-49	Male	Taxi	0.0681	0.0655
50-69	Male	Taxi	0.0357	0.0303
15-49	Female	Taxi	0.0671	0.0669
50-69	Female	Taxi	0.1159	0.1159
15-49	Male	Walking	0.6559	0.631
50-69	Male	Walking	0.4643	0.3939
15-49	Female	Walking	0.5427	0.541
50-69	Female	Walking	0.4203	0.4203
15-49	Male	Private Car	0.1254	0.1207
50-69	Male	Private Car	0.1964	0.1667
15-49	Female	Private Car	0.0549	0.0547
50-69	Female	Private Car	0.058	0.058
15-49	Male	Bicycle	0.0179	0.0172
50-69	Male	Bicycle	0	0
15-49	Female	Bicycle	0.003	0.003
50-69	Female	Bicycle	0	0
15-49	Male	Short Walking	0	0.3862
50-69	Male	Short Walking	0	0.3636
15-49	Female	Short Walking	0	0.3495
50-69	Female	Short Walking	0	0.4203
15-49	Male	Motorcycle	0	0.0621
50-69	Male	Motorcycle	0	0.1515
15-49	Female	Motorcycle	0	0.0061
50-69	Female	Motorcycle	0	0.0145

D.1.2 Adjust for trip weight

We define a trip-mode weight per mode, L_m, so that a single trip by mode m in the baseline scenario represents L_m trips in reality. If in a scenario the mode of the trip is changed ($m' \to m$), it retains the weight of its original mode m'. We use the trip-mode weight to (a) sample trips for scenario generation and (b) scale the probabilities to travel. Note that we don't use it as trip weights are ordinarily used: to scale trip distance/duration. The result is a scaled set of probability parameters:

$$\hat{B}_{a,g,m} = L_m B_{a,g,m}.$$

(L_m should actually be the average trip-mode weight for all trips in the scenario. For baseline, it will be the same. It can/should differ for scenarios. Need to work out notation.)

D.1.3 Smooth probabilities

We estimate smoothed probabilities for any mode that has zero or one probability, using the rest of the probabilities for reference, replacing $\hat{B}_{a,g,m}$. The reference to other modes is slightly problematic if e.g. in the scenario all travel increases. It means a 0 in the baseline is not the same as a 0 in the scenario. This could be fixed with more rigid coding.

D.1.4 Resample probabilities

We resample the adjusted probability parameters of Table 6. We describe each parameter with a Beta distribution as follows:

$$\tilde{B}_{a,g,m} \sim \text{Beta}(\alpha_{a,g,m}, \beta_{a,g,m}); \tag{13}$$

$$\alpha_{a,g,m} + \beta_{a,g,m} = 200; \tag{14}$$

$$\hat{B}_{a,g,m} = \frac{1}{1 + \frac{\beta_{a,g,m}}{\alpha_{a,g,m}}}. \tag{15}$$

Here, $\tilde{B}_{a,g,m}$ is the new, variable parameter that we resample for each simulation. We sample via a quantile, $\nu_m \sim \mathcal{U}(0,1)$, which we use to calculate $\tilde{B}_{a,g,m}$. We set $\alpha_{a,g,m}$ and $\beta_{a,g,m}$ to sum to 200, making the distributions

concentrated, and we constrain their relative values to enforce the distribution's mean to equal $\hat{B}_{a,g,m}$ (the rightmost column in Table 6).

D.1.5 Inform individual travel

Each individual is assigned their own propensity to travel per mode, $\delta_{i,m}$, which is a uniformly distributed random variable that is fixed and not resampled. Whether or not individual i travels by mode m depends on $\tilde{B}_{a(i),g(i),m}$, which does vary: specifically, individual i travels by mode m if $\delta_{i,m} < \tilde{B}_{a(i),g(i),m}$. (To mitigate the effects of these fixed random variables, a large population is required, e.g. more than 5000 individuals.)

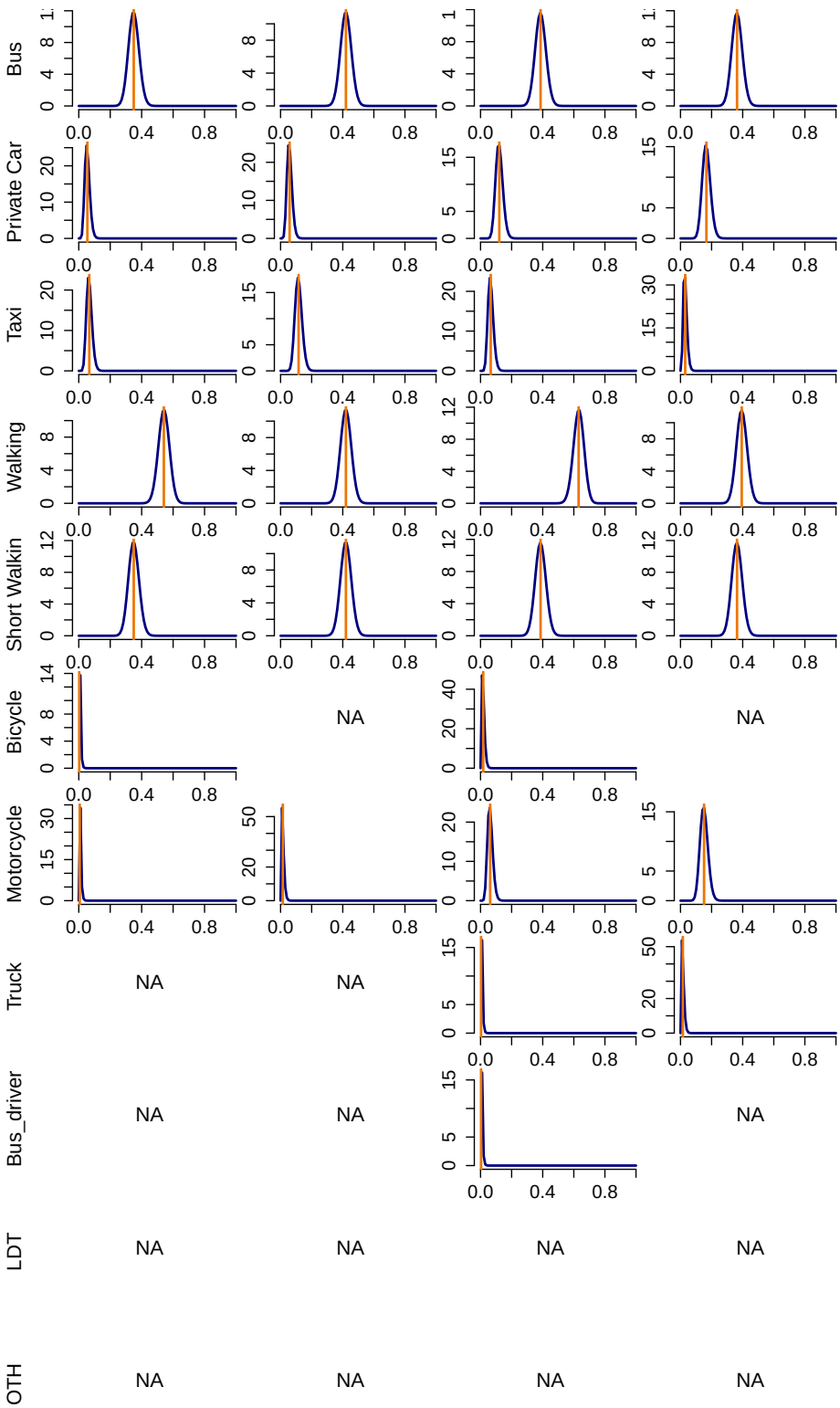

Figure 12: Distributions for propensities to travel. The propensity is the probability that a member of a demographic group will travel by a particular mode on a particular day. In orange are the raw probabilities based on the synthetic trip set, here shown for the baseline scenario. Beta distributions are defined via a pointiness parameter, set to 200, such that $\alpha + \beta = 200$.

D.2 Simulating a synthetic population

1. Choose a population size of individuals N. Label each individual $i = 1, ..., N$.
2. Allocate each individual to a demographic group according to the proportions in the Accra population (Table 5). Then each demographic group has number $\tilde{N}_{a,g}$ and $\sum_{a,g} \tilde{N}_{a,g} = N$.
3. Generate one standard uniform variable per person per mode, $\delta_{i,m} \sim \mathcal{U}(0,1)$. $\delta_{i,m}$ is individual i's propensity to travel by mode m.
4. $\tilde{B}_{a,g,m,s}$ is the propensity of group $\{a,g\}$ to travel by mode m in scenario s. It is the proportion of raw person–duration values that are zero. It is parametrised by $\alpha_{a,g,m,s}$ and $\beta_{a,g,m,s}$ (Equation 13).
5. Let $\mathcal{F}_{a,g,m,s}$ be the raw density for raw person–duration values for demographic group $\{a,g\}$ and mode m in scenario s: $\mathcal{F}_{a,g,m,s} = \left\{\sum_{j:i(j)=i,m(j)=m,s(j)=s} T_j\right\}_{i:a(i)=a,g(i)=g}$ i.e. we isolate all individuals i in group $\{a,g\}$ and sum their trip durations by mode m in scenario s.
6. Let $\tilde{\mathcal{F}}_{a,g,m,s}$ be that density corrected to have proportion $\tilde{B}_{a,g,m,s}$ zeros.
7. Define each individual's travel duration as $D_{i,m,s} = 7 \cdot \text{CDF}^{-1}_{\tilde{\mathcal{F}}_{a(i),g(i),m,s}}(\delta_{i,m})$ (see Figure 13 for resulting distributions).

For sampling, we loop over steps 6–7, where each sample has unique $\tilde{B}_{a,g,m,s}$ values. In total, the number of uncertain variables is one per mode.

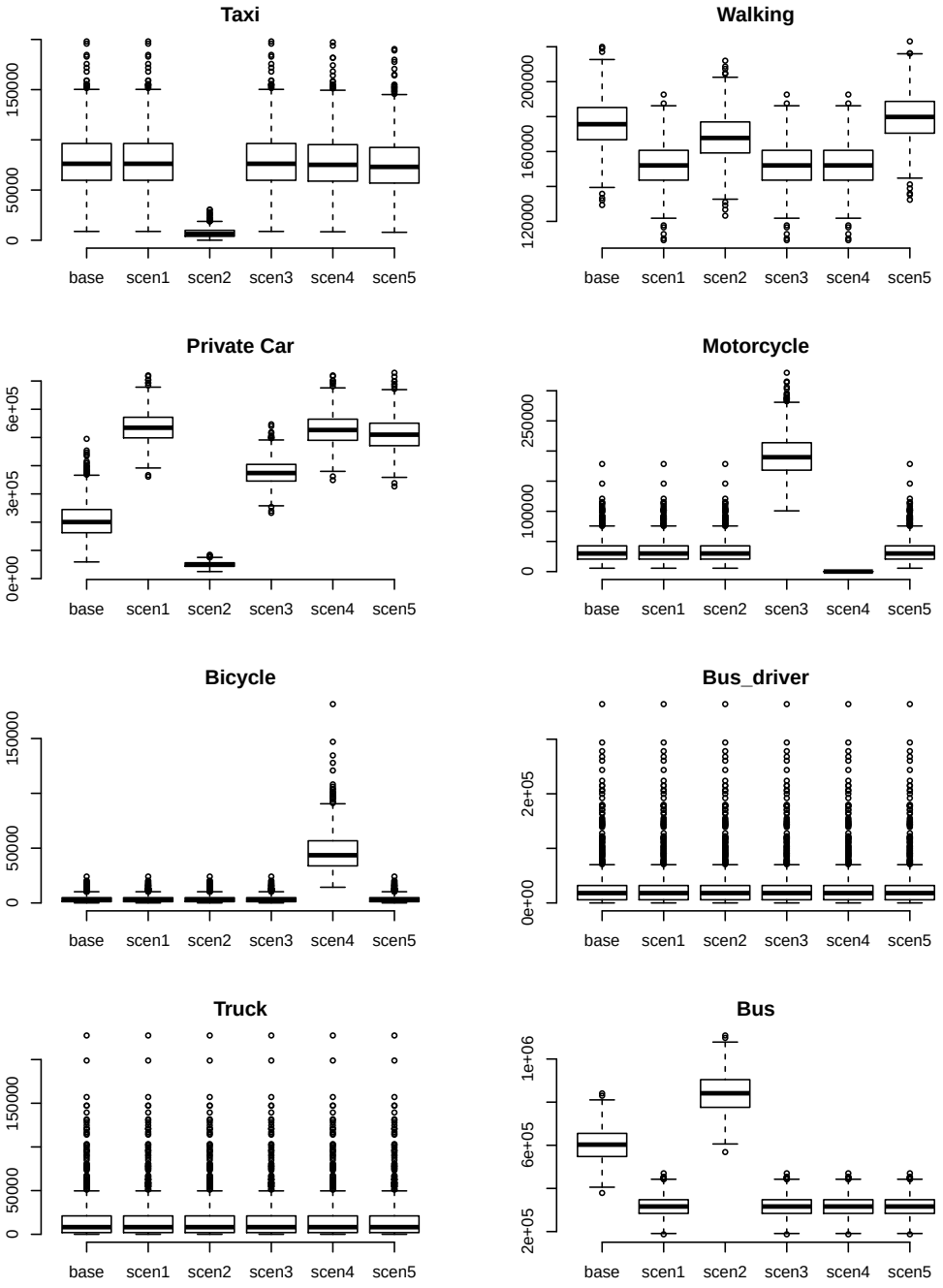

Figure 13: Samples of total distances travelled by mode and scenario.

D.3 Results

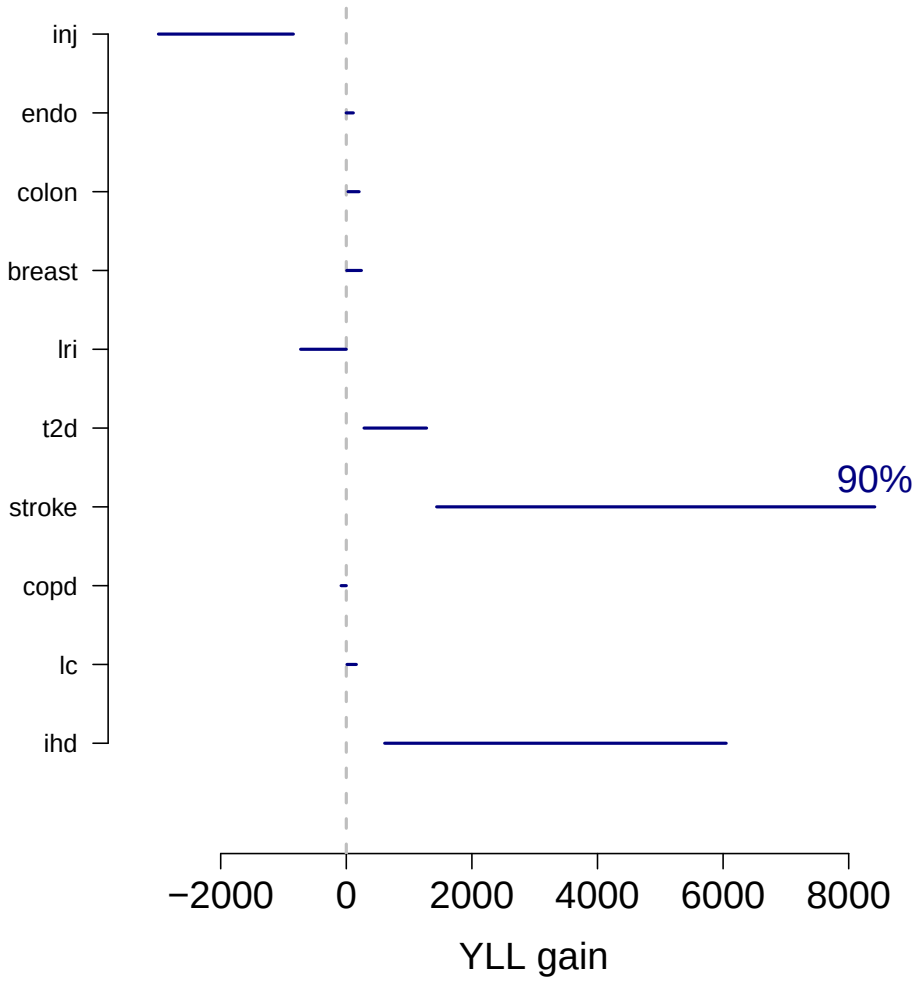

Figure 14: YLLs in the scenario minus YLLs in the baseline for Accra, with motorcycle probability scalar 2 in Accra.

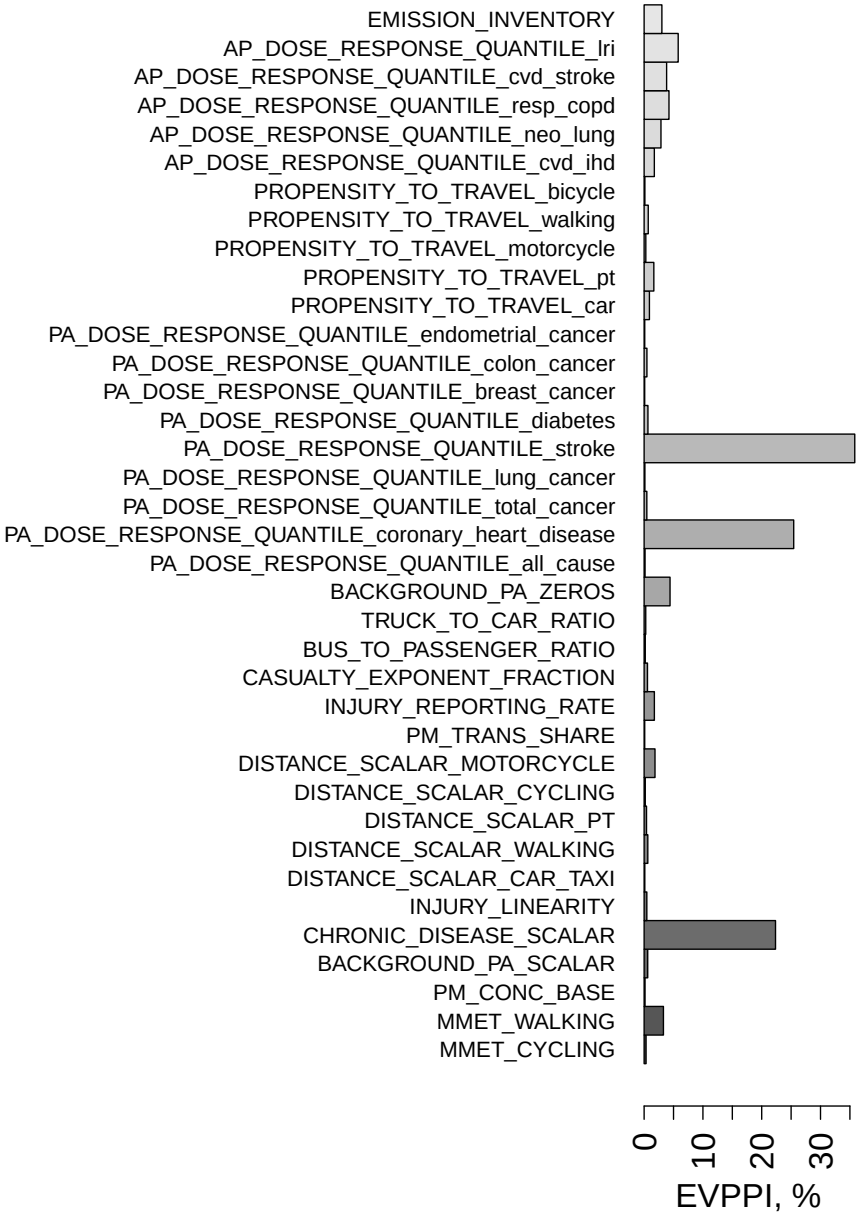

Figure 15: EVPPI for Accra's "walking" scenario for all causes of YLLs excluding "all cause" and "neoplasm", with propensity to motorcycle = 2.

E Glossary by letter

Letter	Meaning	Letter	Meaning	Letter	Meaning
α		a	age	A	
β		b		B	raw travel probabilities
γ	emission uncertainty	c		C	ventilation constants
δ	propensity to travel	d	demographic group	D	
ϵ	walk-to-bus time	e		E	injury distance exponent
ζ	traffic PM2.5 fraction	f		F	PA DR function
η	PM2.5 concentration	g	gender	G	PA DR curves
θ	PA uncertainty	h	disease	H	AP DR curves
ι		i	individual	I	injuries
		j	trip	J	
κ	injury regression parameter	k		K	AP exposure
λ	travel MMETs	l		L	
μ	mode distances	m	mode	M	MMETs per person
ν		n		N	population
ξ	AP DR quantile	o	outcome	O	
π	travel repetitiveness	p		P	vehicle emissions
		q		Q	person weights
ρ	chronic disease scalar	r		R	death-to-injury ratio
σ	injury reporting rate	s	scenario	S	
τ		t		T	trips
υ		u		U	burden of disease
ϕ	PA DR quantile	v		V	AP per person
		w	AP DR parameter	W	relative risks
χ	day-to-week travel scalar	x	PA DR parameter	X	injury model matrix
ψ	casualty fraction of ω	y	year	Y	individual MMETs
ω	injury linearity	z	PA dose	Z	speeds

References

Burnett, R. T., Arden Pope, C., Ezzati, M., Olives, C., Lim, S. S., Mehta, S., Shin, H. H., Singh, G., Hubbell, B., Brauer, M., Ross Anderson, H., Smith, K. R., Balmes, J. R., Bruce, N. G., Kan, H., Laden, F., Prüss-Ustün, A., Turner, M. C., Gapstur, S. M., Diver, W. R. and Cohen, A. (2014), 'An integrated risk function for estimating the global burden of disease attributable to ambient fine particulate matter exposure', *Environmental Health Perspectives* **122**(4), 397–403.

Goel, R., Gani, S., Guttikunda, S. K., Wilson, D. and Tiwari, G. (2015), 'On-road PM2.5 pollution exposure in multiple transport microenvironments in Delhi', *Atmospheric Environment* **123**, 129–138.
URL: *http://dx.doi.org/10.1016/j.atmosenv.2015.10.037*

Mohammadian, A. K., Javanmardi, M. and Zhang, Y. (2010), 'Synthetic household travel survey data simulation', *Transportation Research Part C* **18**, 869–878.
URL: *http://dx.doi.org/10.1016/j.trc.2010.02.007*

Saadi, I., Mustafa, A., Teller, J. and Cools, M. (2016), 'Forecasting travel behavior using Markov Chains-based approaches', **69**, 402–417.