Multi-state modelling of indirect chronic disease data to inform health impact models

Christopher Jackson
(MRC Biostatistics Unit, University of Cambridge)

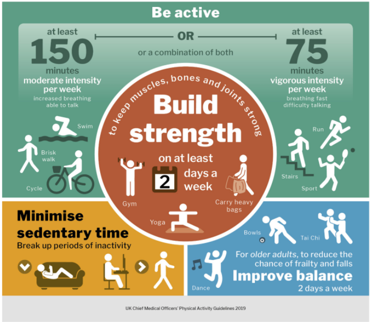
Belen Zapata-Diomedi (RMIT University, Melbourne)
James Woodcock (MRC Epidemiology Unit, University of Cambridge)

Armitage Workshop, November 10, 2021

Modelling impacts of chronic disease prevention

What if?

Everyone did the recommended amount of physical activity?



(for able-bodied adults)

Modelling impacts of chronic disease prevention

What if?

40% of car trips in a city switched to walk/bike?

Modelling impacts of chronic disease prevention

What if?

Food manufacturers achieved salt reduction targets?

Modelling impacts of chronic disease prevention

What if?

Food manufacturers achieved salt reduction targets?

Impacts on

Lives saved / healthy life expectancy (“QALY” / “DALY”)

Modelling impacts of chronic disease prevention

What if?

Food manufacturers achieved salt reduction targets?

Impacts on

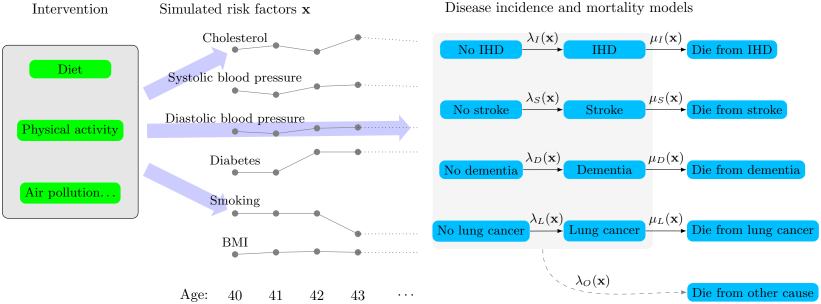
Health / social care costs, inequalities. . .

Modelling impacts of chronic disease prevention

What if?

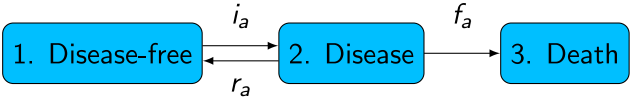
. . . Impacts on . . .

Models describe the mechanism for the impacts of prevention scenarios



Simulate outcomes under different scenarios or policy decisions

Common approach: multi-state lifetable model



1. Disease-free

i_a

r_a

2. Disease

f_a

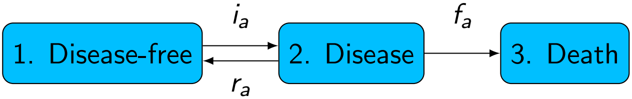
3. Death

- Represent a disease as a 3-state Markov model
- Defined by rates by age a of
  - incidence i_a , case fatality f_a , (sometimes) remission r_a
- For some population stratum (e.g. area, gender)
- Population simulated assuming multiple diseases independent

Statistical challenge: data often indirect

- Prevalence (proportion with the disease) but not incidence i_a (rate of new cases)
- Mortality (deaths among whole population) but not case fatality f_a (risk for people who have the disease)

Common approach: multi-state lifetable model



- Represent a disease as a 3-state Markov model
- Defined by rates by age a of
  - incidence i_a , case fatality f_a , (sometimes) remission r_a
- For some population stratum (e.g. area, gender)
- Population simulated assuming multiple diseases independent

Statistical challenge: data often indirect

- Prevalence (proportion with the disease) but not incidence i_a (rate of new cases)
- Mortality (deaths among whole population) but not case fatality f_a (risk for people who have the disease)

Health impact models for transport policies and scenarios

Motivating example

- ITHIM (Integrated Transport and Health Impact model) and variants (MRC Epidemiology Unit)
- Used to inform transport policy in settings around the world (Sao Paulo, San Francisco, Nashville, Accra, Delhi...)
- Version under development (METAHIT) to inform "active transport" policy (walk, bike) for the city regions of England
- Model diseases affected by physical activity, air pollution, noise exposure (also road injury)
- Need city region-specific data on disease incidence and case fatality to inform multi-state progression model

Health impact models for transport policies and scenarios

Motivating example

- ITHIM (Integrated Transport and Health Impact model) and variants (MRC Epidemiology Unit)
- Used to inform transport policy in settings around the world (Sao Paulo, San Francisco, Nashville, Accra, Delhi. . .)
- Version under development (METAHIT) to inform "active transport" policy (walk, bike) for the city regions of England
- Model diseases affected by physical activity, air pollution, noise exposure (also road injury)
- Need city region-specific data on disease incidence and case fatality to inform multi-state progression model

Global Burden of Disease study (Institute of Health Metrics and Evaluation, University of Washington)

Publishes estimates of incidence, prevalence, mortality, risk factors but not case fatality

- . . . for hundreds of diseases / conditions
- . . . for countries, and regions within countries, covering the whole world. Local authority level in UK
- Synthetic / model-based estimates with credible intervals
  - ensure consistency / comparability between outcomes and settings

Also published tools for estimating the multi-state disease model with indirect data

- DisMod II (Barendregt et al 2001)
  - friendly Windows interface, widely used, poorly understood statistical basis
- DisMod-MR (Flaxman et al 2015)
  - Bayesian, code less accessible, only been used internally

Global Burden of Disease study (Institute of Health Metrics and Evaluation, University of Washington)

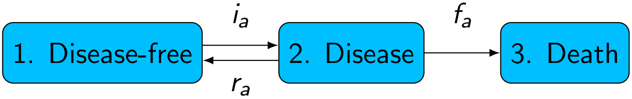
Publishes estimates of incidence, prevalence, mortality, risk factors but not case fatality

- … for hundreds of diseases / conditions
- … for countries, and regions within countries, covering the whole world. Local authority level in UK
- Synthetic / model-based estimates with credible intervals
  - ensure consistency / comparability between outcomes and settings

Also published tools for estimating the multi-state disease model with indirect data

- DisMod II (Barendregt et al 2001)
  - friendly Windows interface, widely used, poorly understood statistical basis
- DisMod-MR (Flaxman et al 2015)
  - Bayesian, code less accessible, only been used internally

This work



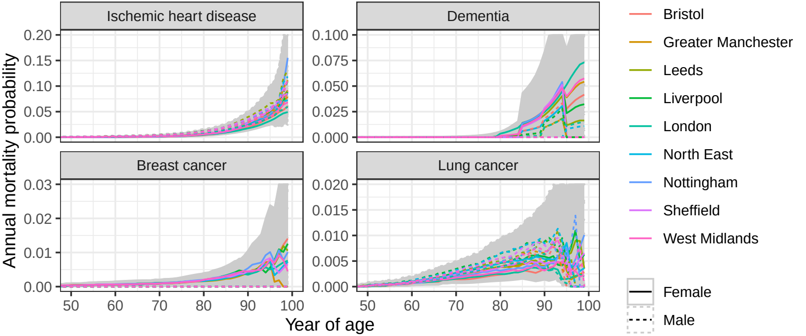
Methodological

- explained statistical basis behind DisMod inference methods
- extended the methods to make them more flexible
- provided accessible software as R package `https://chjackson.github.io/disbayes`

Application

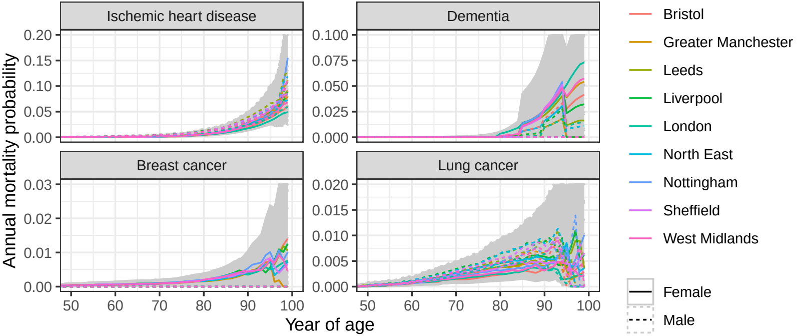
- estimate case fatality given mortality / prevalence for England city regions
- to inform transport health impact modelling for those areas

Data (incidence, mortality, prevalence, remission)



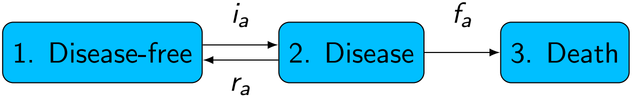
- 17 diseases, by 5-year age group, gender and local authority
- Incidence, mortality, prevalence from GBD.
- Cancer remission from 10-year survival rates published by ONS

Data (incidence, mortality, prevalence, remission)



- Data in form of estimated rates + credible intervals
- Converted to annual probabilities p, hence to implicit count of r events per year with denominator n
  - assuming CIs for p describe the Beta posterior from "data" $r \sim Bin(n, p)$. n describes the uncertainty.
- Counts for 5-year age groups smoothly disaggregated to 1-year age groups, and aggregated over city regions

Estimating multi-state transition rates from data



- Annual mortality, incidence, (remission) and prevalence as counts/denominators.
- Modelled as Binomial, with probabilities $p_a^{(mort)}, p_a^{(inc)}, p_a^{(prev)}, (p_a^{(rem)})$, for each age a
- Probabilities defined as complex functions of the parameters of interest i_a, f_a, (r_a).
  - via annual transition probability matrix P_a between 3 states
  - continuous-time Markov chain theory / analytic ODE solution
- How are rates for different ages/genders/areas related?

Relating rates from different ages, areas, genders

- Age-dependence of rates through smooth spline functions, e.g.

$$log(f_a) = \beta_0 + \beta_1 a + \sum_{k=2}^{K} \beta_k g_k(a)$$

  where $g_k()$ are basis functions (generated by `mgcv` R package)
- β_0, β_1 have vague priors, $\beta_2, \ldots, \beta_K \sim N(0, \lambda)$,
- $\lambda \sim Gamma(2, s)$ controls smoothness / deviation from linearity
- f_a assumed to be constant under a specific age if the data are insufficient (30, 50, 70 depending on the disease)
- Areas modelled independently or hierarchically (β_0 becomes random effect) with area-constant or area-dependent effect of gender

Relating rates from different ages, areas, genders

- Age-dependence of rates through smooth spline functions, e.g.

$$log(f_a) = \beta_0 + \beta_1 a + \sum_{k=2}^{K} \beta_k g_k(a)$$

  where $g_k()$ are basis functions (generated by `mgcv` R package)
- β_0, β_1 have vague priors, $\beta_2, \ldots, \beta_K \sim N(0, \lambda)$,
- $\lambda \sim Gamma(2, s)$ controls smoothness / deviation from linearity
- f_a assumed to be constant under a specific age if the data are insufficient (30, 50, 70 depending on the disease)
- Areas modelled independently or hierarchically (β_0 becomes random effect) with area-constant or area-dependent effect of gender

Relating rates from different ages, areas, genders

- Age-dependence of rates through smooth spline functions, e.g.

$$log(f_a) = \beta_0 + \beta_1 a + \sum_{k=2}^{K} \beta_k g_k(a)$$

  where $g_k()$ are basis functions (generated by `mgcv` R package)

- β_0, β_1 have vague priors, $\beta_2, \ldots, \beta_K \sim N(0, \lambda)$,
- $\lambda \sim Gamma(2, s)$ controls smoothness / deviation from linearity
- f_a assumed to be constant under a specific age if the data are insufficient (30, 50, 70 depending on the disease)
- Areas modelled independently or hierarchically (β_0 becomes random effect) with area-constant or area-dependent effect of gender

Computation and implementation

Stan `https://mc-stan.org`

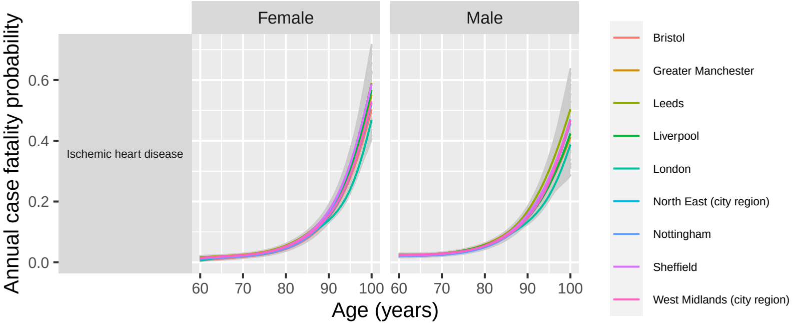
- Hamiltonian MCMC to obtain sample from posterior — for "final" results (minutes to hours)
- Optimisation to estimate the posterior mode, with normal approximation to posterior
  - instant, useful for model development

R package that embeds the Stan models

`https://chjackson.github.io/disbayes`

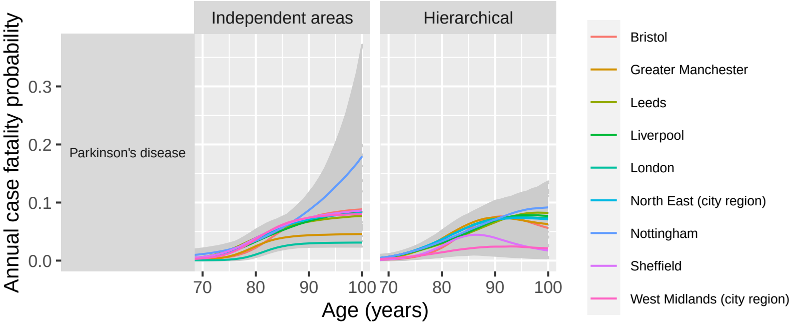
- Intended to be $\geq$ friendly / principled / flexible as previous DisMod packages

Examples of case fatality estimates under our models



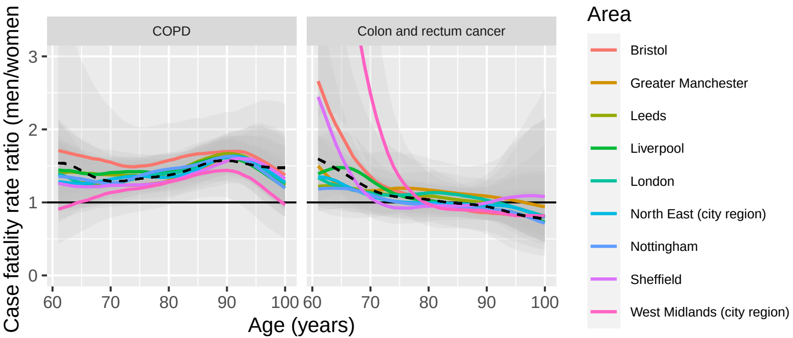
- Area/gender specific case fatality curves produced for: ischemic heart disease, stroke, lung cancer, colorectal cancer, breast cancer, dementia, COPD, diabetes, Parkinson's disease, liver cancer, non-rheumatic valvular heart disease.
- National estimates by gender for: stomach cancer, liver cancer, uterine cancer, cardiomyopathy and myocarditis, multiple myeloma

Examples of case fatality estimates under our models



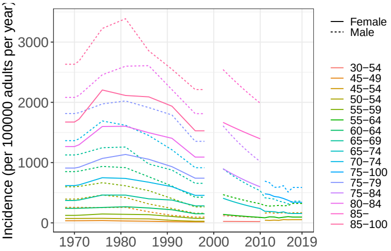
- Hierarchical models for area variations had limited utility.
- Same estimates as non-hierarchical models, except for some shrinkage at oldest ages.
- Identifiability problems for rarest diseases
- Cross-validatory comparison ("LOO-PSIS" method, Vehtari et al.) generally favoured non-hierarchical

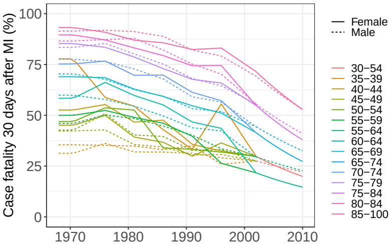
Examples of case fatality estimates under our models



- Rate ratio between men and women, as a function of age, in the hierarchical models
- Dotted lines show model where this ratio is the same for each area.
  - Cross-validatory criterion generally prefers this

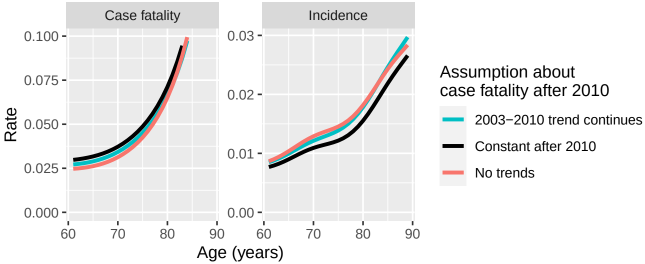
Trends through calendar time in disease risks





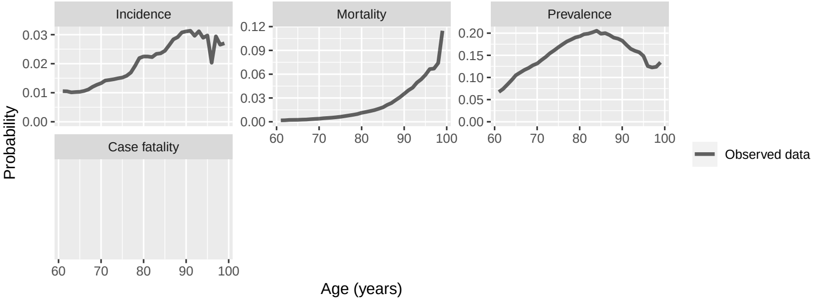
- Incidence and case fatality from ischemic heart disease declined in last 50 years (evidence from a variety of publications)
- Trends are age-dependent (previous "DisMod" software didn't account for this)
- Can adjust for this in the (non-hierarchical) model, assuming rate in a previous year is a fixed multiplier of the current rate (age-specific, from smoothing/interpolating published data)
- Different inferred rates under different assumptions

Trends through calendar time in disease risks



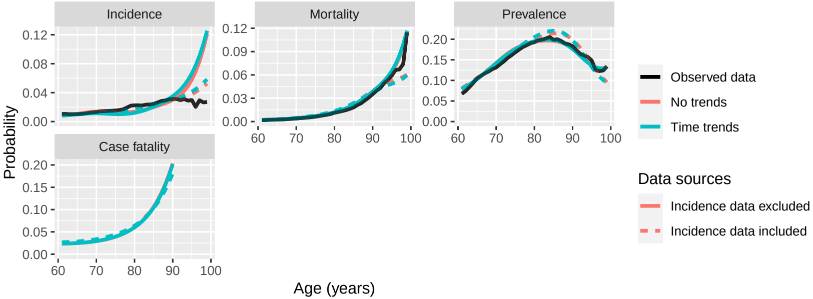
- Incidence and case fatality from ischemic heart disease declined in last 50 years (evidence from a variety of publications)
- Trends are age-dependent (previous "DisMod" software didn't account for this)
- Can adjust for this in the (non-hierarchical) model, assuming rate in a previous year is a fixed multiplier of the current rate (age-specific, from smoothing/interpolating published data)
- Different inferred rates under different assumptions

Consistency between estimates and different data sources



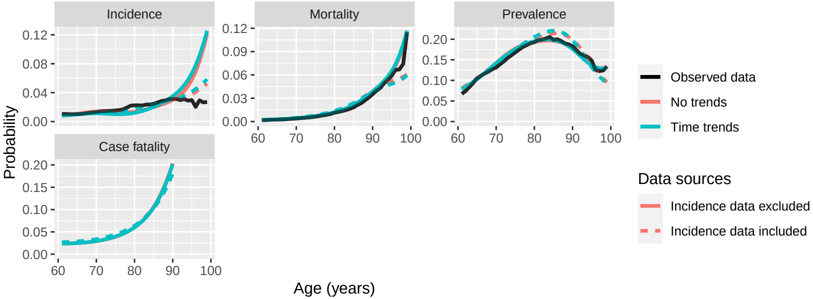
- The model synthesises observed data on current prevalence, mortality and incidence
  - to produce estimates of (unobserved) case fatality
  - estimates of prevalence, mortality and incidence also produced that are coherent with all data sources
- Check fit of model-based estimates to the direct data on incidence, prevalence and mortality
- Fit of the model estimates to the mortality and prevalence

Consistency between estimates and different data sources



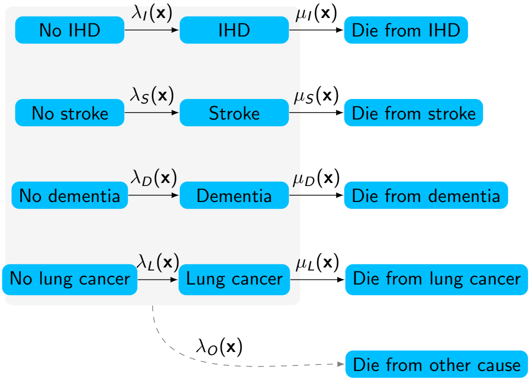
- The model synthesises observed data on current prevalence, mortality and incidence
  - to produce estimates of (unobserved) case fatality
  - estimates of prevalence, mortality and incidence also produced that are coherent with all data sources
- Check fit of model-based estimates to the direct data on incidence, prevalence and mortality
- Fit of the model estimates to the mortality and prevalence

Consistency between estimates and different data sources



- Fit of the model estimates to the mortality and prevalence data is better if the incidence data are excluded from the evidence synthesis
- Current incidence data (new cases) in conflict with current prevalence (old cases), even if we adjust for time trends
  - Conflicts remaining between data sources, including the time trend data
  - Case fatality estimates not greatly affected though

Combining multiple diseases: "multistate lifetable"



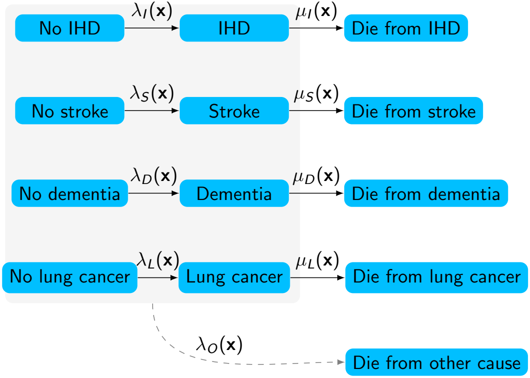
A common approach

- Parallel multistate models
- Assumes multiple diseases independent, neglecting multimorbidity effects
- Exposure (physical activity, air pollution...) may modify incidence and case fatality

Simulate at aggregate level

- proportion of population with each disease at each time
- accumulate health-adjusted life expectancy / costs to compare policies / scenarios

Combining multiple diseases: “multistate lifetable”



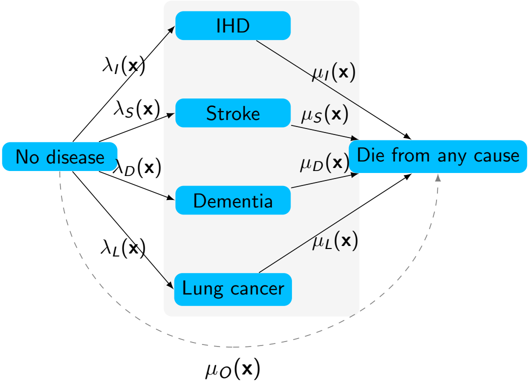
A common approach

Requires case fatality and effect of exposures on case fatality, needing data which attributes death to a specific cause

Simulate at aggregate level

- proportion of population with each disease at each time
- accumulate health-adjusted life expectancy / costs to compare policies / scenarios

Combining multiple diseases: competing risks framework?



Simulate at individual-level ("microsimulation") rather than aggregate level

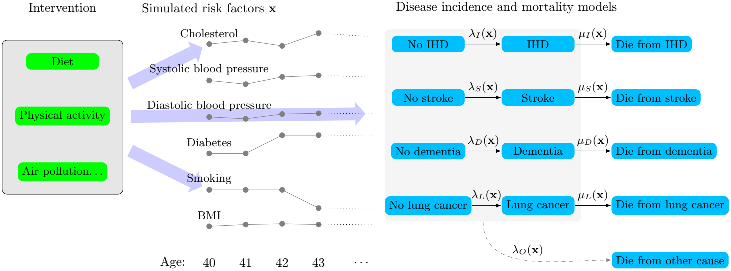
Alternative framework

- People can transition to only one disease state
- First disease that they get determines their outcome

Needs all-cause mortality rate μ for people with each disease
(and exposure effects on this)

- Easier to measure than cause-specific mortality?
- Multimorbidity effects included in each μ

Summary



Challenges of disease burden modelling to inform health impacts

- Long timespans, multiple diseases
- Disparate data sources, covering multiple populations

Further work on different aspects of this impact modelling picture

Paper soon, see `https://chjackson.github.io/disbayes` for software