# Modelling Healthy Life Expectancy at Neighbourhood Level

Andrew J.G. Cairns

Director, Duncan Laboratory for Insurance Data Science
Heriot-Watt University, Edinburgh

Longevity 21, Rome, September 2026

DUNCAN
LABORATORY
for Insurance Data Science

HERIOT
WATT
UNIVERSITY

# Outline

- Motivation and objectives
- Data: life, health, socio-economic
- Modelling of neighbourhood mortality and health rates
- Results
- Discussion: problems with HLE

# Motivation and Objectives

- **Inequality:** *systematic, avoidable and unfair differences in mortality and health outcomes and metrics*
  Main reason:
  prevalence of *causal risk factors* varies between groups (e.g. smoking)
  + access to healthcare and related services
- **Short-term aim:**
  To develop *customised* mortality and health indices for Scotland

  Longer term: UK-wide indices (especially mortality)
- How much inequality is there in mortality and health rates at different ages?
- Which socio-economic factors are most strongly associated with mortality and health outcomes?

# Mortality-Related Data: Neighbourhoods

- Small neighbourhoods:
  "Data Zones": DZ (equivalent to LSOAs in England and Wales)
- $L = 6976$ DZs
- Average size $\approx 800$ persons

- DZs $i = 1, \ldots, L$,
  single years ($t$ =2011-2019),
  single or 5-year age groups, $x$:
  - Deaths: $D(i, t, x)$ (5-yr age groups) (by normal place of residence)
  - Exposures: $E(i, t, x)$ (single ages)

  - Many potential predictive variables, $\Rightarrow$ vector $Z(i)$

# Methodology

- National death rates (all $t$ and $x$)

$$m(t,x) = \frac{D(t,x)}{E(t,x)} \qquad \text{(single ages)}$$

- Neighbourhood death rates ($i = 1, \ldots, L$)
  - neighbourhood death rates: $m(\mathbf{i}, t, x)$
  - Question: how to model $m(i,t,x)$ relative to national $m(t,x)$?

- Predictive variables $Z(i)$ plus $m(t,x) \longrightarrow m(i,t,x)$

# Methodology: General Approach

- Over a limited age range (e.g. 70-89); and
- Over a limited range of years (e.g. 2011-2019):

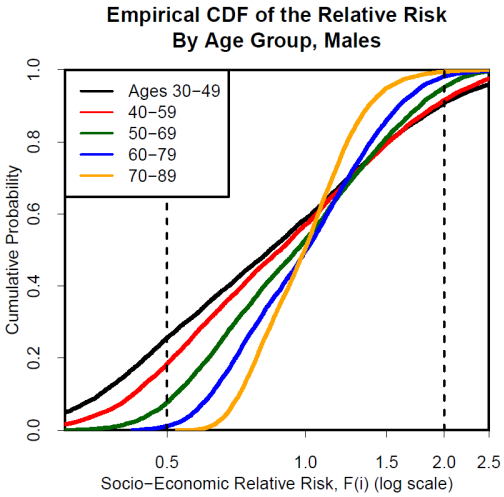
$$\text{neighbourhood } i, \qquad m(i,t,x) = m(t,x)F(i)$$

- $F(i)$ = neighbourhood-specific *relative risk* due to socio-economic and other non-spatial characteristics, $Z(i)$
- Estimate $F(i)$ using Local Linear Regression (variant of LOESS)

# Choice of Predictive Variables: what do we include in $Z(i)$

- **Scottish Index of Multiple Deprivation (SIMD):**
  - Employment deprivation
  - Education deprivation: Proportion of the population with no qualfications
  - Crime
- **Census data:**
  - Communal establishments: proportion of the population aged 65+ in a communal establishment (mostly care homes in the 65+ group)
  - Proportion married: ages 35+
- **2011-2019 population data:**
  - Old-age dependency ratio: # 65+ / # 20-64; males and females separately

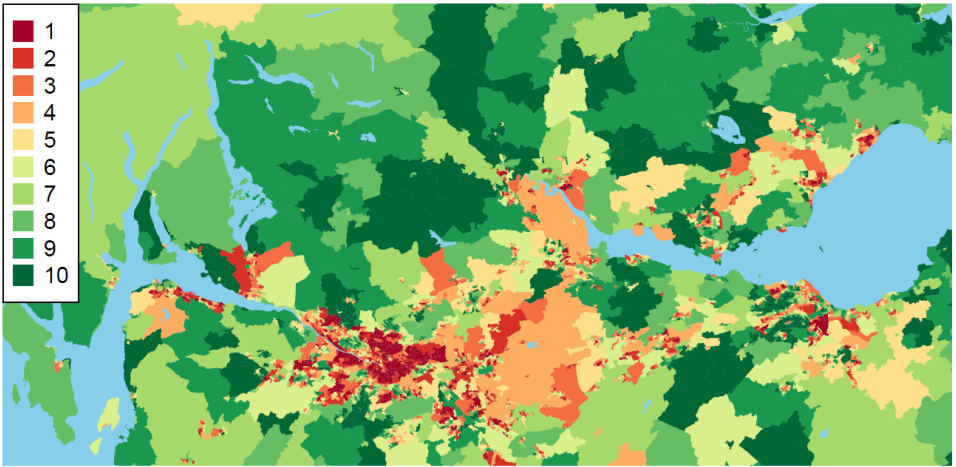
# Socio-Economic Relative Risk, $F(i)$: Significant Inequality




- High levels of inequality
  [Higher levels of inequality than England.]
- Inequality diminishes with age
- Age 40-59 main predictors:
  unemployment, married, education
- Age 70-89 main predictors:
  care homes, education, dependency ratio

- Females slightly less spread out
  High correlation between male and female relative risks

# Geographic Spread of Inequality: Males 60-79; Central Scotland

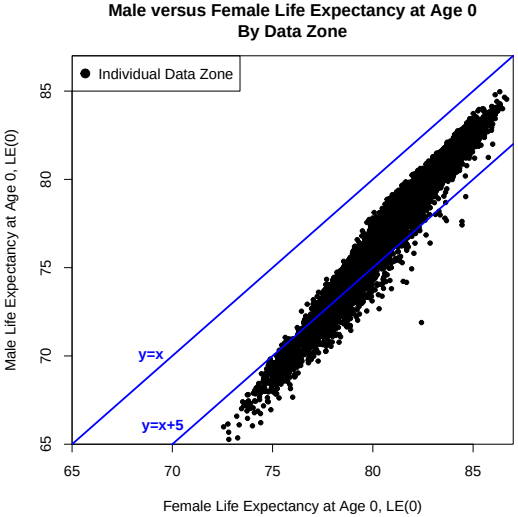



Decile 1: high mortality, Decile 10: low mortality

# Life Expectancies

- End users understand life expectancies better than relative risks
- Here: Period life expectancies
  (i.e. based on mortality rates for a given group of years: 2017-2019)
- Challenges and assumptions:
  - LE calculations assume that there is no migration between areas as people get older
  - Relative risks (RR) by neighbourhood $i$
    Modelling $\Rightarrow$ age groups 30-49; ...; 70-89
    Need $RR(i, x)$ by single ages 0 to 110
    Interpolate RR between 40, 50, 60, 70 and 80
    RR tapers to 1 at age 100
    Extrapolate back to age 0
- Combine $m(i, x) = RR(i, x)m_{NAT}(x) \Rightarrow LE(0)$ and $LE(x)$ for higher ages

# Life Expectancies at Age 0, LE(0): Males versus Females




- Axes: range = 22 on both axes
- Males: 64.1 to 83.1, StDev= 3.7
- Females: 69.8 to 86.0, StDev= 2.5
- Takeaway:

**Most deprived areas:**
female-male gap $\approx$ 8 years

**Least deprived areas:**
female-male gap $\approx$ 2 years

# Healthy Life Expectancy (HLE)

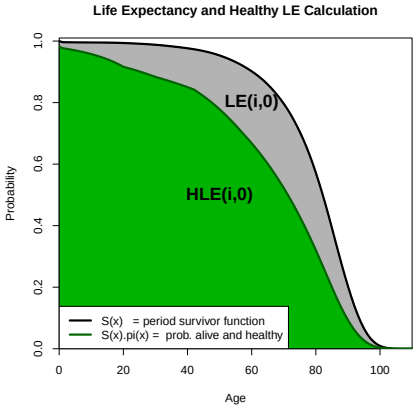
- Good health $\leftrightarrow$ poor health $\leftrightarrow$ dead
- LE measures average years spent in both good and poor health
- **How many years are spent in good health?**
- HLE calculation relies on **self-assessment** of "general health"
- Census 2022 and Health Survey:
  *How is your health in general?*
  Possible answers: Very good; good; fair; bad; very bad

- $\pi(x)$ = proportion at exact age $x$ who report that they are in **good or very good health**

- $m(y)$ = death rate at age $y$
  $S(x) = \exp[-\sum_{y=0}^{x-1} m(y)]$ =*period* survival function age 0 to age $x$

# Sullivan Method for HLE

$$HLE(0) \equiv \frac{1}{2}\pi(0)S(0) + \sum_{x=1}^{\infty} \pi(x)S(x)$$




- *Period* HLE
- $HLE(0)$ = national HLE
- $HLE(i,0)$ = data zone $i$ HLE using $\pi(i,x)$ and $S(i,x)$
- As with $LE(i,0)$, $HLE(i,0)$ assumes no migration between subgroups
- Different from LE: HLE is based on subjective responses

# Data for estimating HLE at the data zone level

- Five health states:
  very good; good; fair; bad; very bad

- Health Survey:
  too small to estimate health rates accurately at the data zone level

- Census 2022:
  Pro: reasonably full coverage, so data-zone analysis is theoretically possible

  Con: 2022 $\Rightarrow$ ??? more people reporting poor health post-Covid
  Con: no annual updating
  Con: males and females combined (due to be separated)

# Data for estimating HLE at the data zone level (cont.)

- Census outputs: headcounts by data zone subdivided into

  Males and females combined
  5 health states $\times$ 6 age groups (A)
  0-15; 16-24; 25-34; 35-49; 50-64; 65+

- $GH(i, A)$ = number in Good or VG health in age group A in data zone $i$
- $P(i, A)$ = total persons in all states of health

- $\pi(i, A) = GH(i, A)/P(i, A)$ empirical proportion in good health, age group A

# Estimating HLE at the data zone level

- Model:
  Use the empirical $\pi(j, A)$ to estimate

$$\hat{\pi}(i, A) \equiv f(A, Z_H(i))$$

  where $Z_H(i)$ = vector of socioeconomic predictive variables
- Use local linear regression (LLR) again to estimate the $\hat{\pi}(i, A)$
- Age group $\hat{\pi}(i, A) \longrightarrow$ piecewise logit-linear $\hat{\pi}(i, x)$ for single ages $x$
- Combine with data-zone period survival curve $\hat{S}(i, x)$

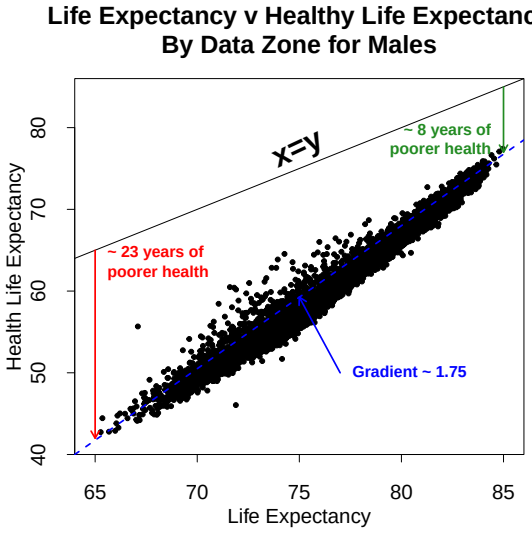
$$HLE(i, 0) \equiv \frac{1}{2}\hat{\pi}(i, 0) + \sum_{x=1}^{\infty} \hat{\pi}(i, x)\hat{S}(i, x)$$

# Predictive Variables for Health versus Mortality

Current variable selections:

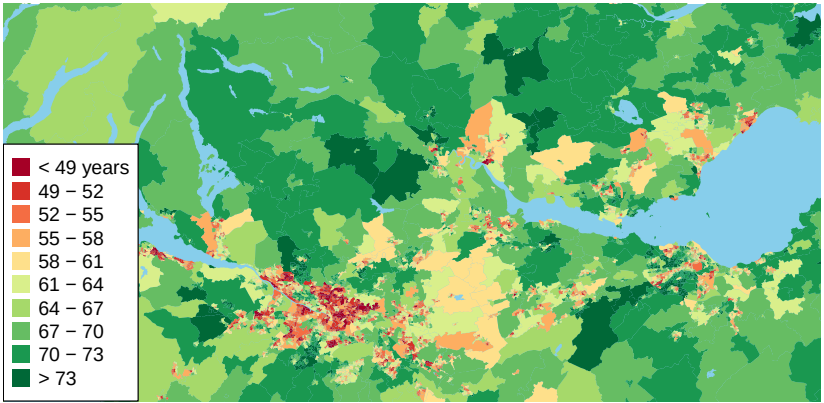
| Mortality | Health |
|---|---|
| Communal Establishment 65+ | Communal Establishment 65+ |
| Employment Deprivation | Employment Deprivation |
| Education: Low (SIMD) | Education: Low (Census) |
| Married | Married |
| Old-age Dependency Ratio | Old-age Dependency Ratio |
| Crime | Education: High (Census) |
| | UK born |

# Life Expectancy versus Healthy Life Expectancy from Age 0

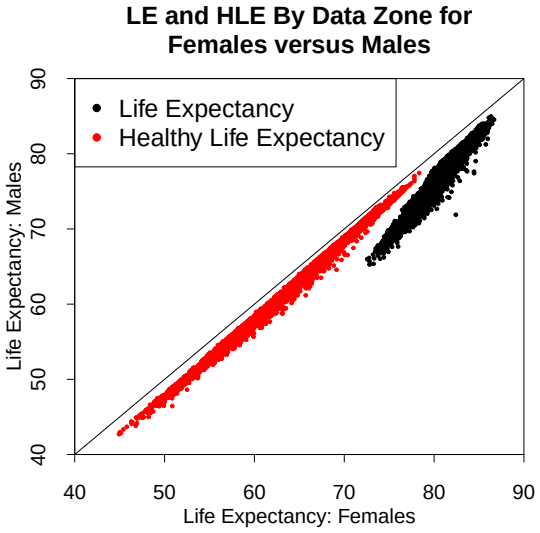



- High correlation between LE and HLE
  (Strong correlation between $S(i,x)$ and $\pi(i,x)$)
- Highest LE $\Rightarrow$ 8 years in "poorer" health
  (females 9)
- Lowest LE $\Rightarrow$ 23 years in "poorer" health
- One year reduction in LE $\Rightarrow$ 1.75 reduction in HLE
  (females 2.35)

# HLE(0) for Males, Central Scotland



# (Healthy) Life Expectancy: Females vs Males




- **Life Expectancy:**
  Gradient $> 1$
  $\Rightarrow$
  Males much more variable than females
- **Healthy Life Expectancy:**
  Gradient $\approx 1$
  BUT: conclusions limited by the data:
  Persons only; not M and F

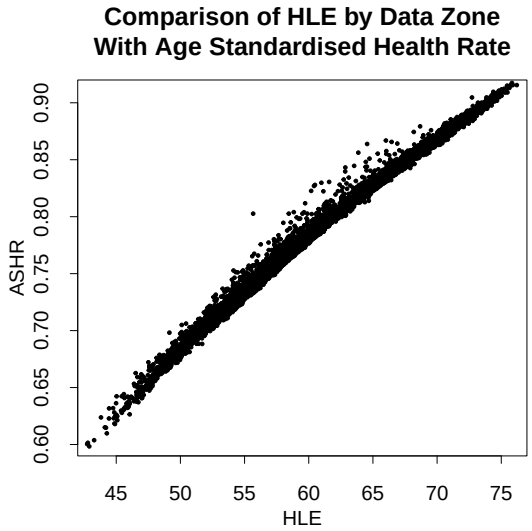
# Alternatives to Healthy Life Expectancy

$$\text{Healthy Life Expectancy:} \quad HLE(i,0) = \sum_{x=0}^{\infty} \hat{\pi}(i,x)\hat{S}(i,x) - \frac{1}{2}\hat{\pi}(i,0)$$

$$\text{Age Standardised Health Rate:} \quad ASHR(i) = \frac{\sum_{x=0}^{\infty} \hat{\pi}(i,x)S(x)}{\sum_{x=0}^{\infty} S(x)}$$

where $S(x)$ = national survival curve (or other standard population)

- HLE has an implicit assumption:
  individuals stay in the same data zone for life
- ASHR does not require this assumption
  it just gives a measure of the average health across all ages
- Formulas are very similar:
  - HLE uses custom weights
  - ASHR uses standardised weights




- Very high correlation between HLE and ASHR
- Less than 100% correlation: differences in the profile of $S(i,x)$ versus national $S(x)$ period survival curves

# Problems with HLE

- Health is on a sliding scale
  not 0/1 like mortality
- Health is also multidimensional
- Health status is subjective $\Rightarrow$
  - Perception of what is *good/very good* health varies with age
    Mental benchmark for *good health* falls with age
  - Possible bias by socio-economic or cultural group e.g.
    more deprived $\Rightarrow$ ? relatively pessimistic $\Rightarrow$ "poorer" health
    less deprived $\Rightarrow$ ? relatively optimistic $\Rightarrow$ "better" health
  - Census: head of household might be responding of behalf of others
- Interpreting trends is hard!
  E.g. mental health more likely now to be taken into account when self-assessing health status

# Problems with HLE (cont.)

- Non-response bias in the health survey and Census 2022
  Less likely to respond: lower SE group; lower education; poor health
  and, non-response bias is getting worse
- Our data: Census 2022 $\Rightarrow$
  potential Covid impact on health rates (real and subjective)

Possible improvements:

- ONS are looking at the use of administrative health data to replace survey data
  - More objective and better coverage
  - But relies on equal access to and willingness to access healthcare
  - How do you calculate an **objective score** for each individual?
  - Score $<$ **threshold** $\Rightarrow$ good or very good health
    **Calibrate this threshold** to approximate old HLE's

# Thank You!

E: A.J.G.Cairns@hw.ac.uk
W: www.macs.hw.ac.uk/~andrewc/ARCresources

THe Index of Scottish Life Expectancy – THISLE

DUNCAN LABORATORY
for Insurance Data Science

HERIOT WATT UNIVERSITY