Modelling cancer risk – uneven outcomes

Prof George Streftaris

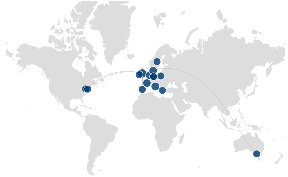
School of MACS
Maxwell Institute of Mathematical Sciences
Heriot-Watt University, Edinburgh, UK

Scor Foundation of Science, 20 September 2023

Network

Research grants from:

- SoA, Centers of Actuarial Excellence
  Predictive Modelling for Medical Morbidity Risk Related to Insurance
- SCOR Foundation for Science
  Breast cancer risk modelling



Collaborators:

- Dr A Arik (HWU)
- Prof A Cairns (HWU)
- Prof I Duncan (UCSB)
- Prof E Dodd (Southampton)
- Alex Jose (HWU)

Outline

1. Cancer rates trends over time

   – mainly all-cancer, lung, breast cancer

2. Stochastic modelling for incidence (& mortality) rates

3. Variation by region and deprivation

4. Projection into the future

5. Impact of diagnosis delays on mortality

   – also linked to delays relating to Covid-19

6. Deep learning methods for cancer rates

Cancer data

Cancer incidence and deaths data
England: Office for National Statistics (ONS)

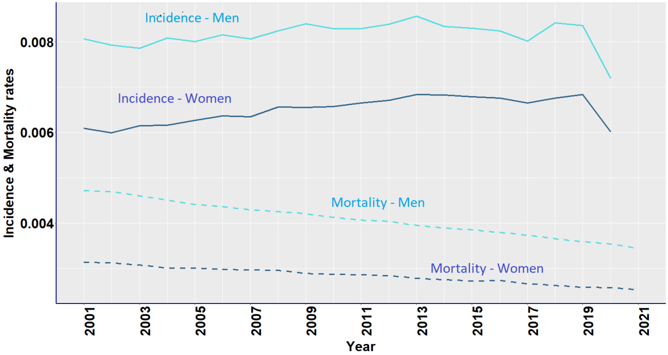
- Age groups: 0, 1-4, 5-9, ..., 95+

  Age-standardised results, based on the European Standard Population (ESP) 2013
- Gender
- Years: 2001 - 2017 (*some* up to 2021)
- Income Deprivation (ID) decile

  1: most deprived; 10: least deprived
- Regions of England: North East, North West, Yorkshire and the Humber, East Midlands, West Midlands, East, London, South East and South West

Trend over time: 2001-2021

All-cancer incidence, mortality
Age standardised rates (no modelling)



Increasing trends for incidence

Decreasing mortality trends

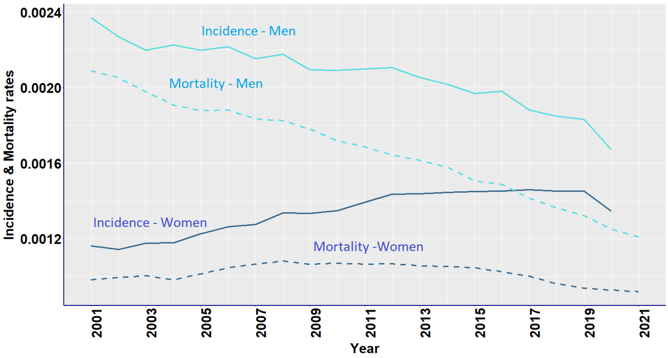
Higher rates for men

HERIOT WATT UNIVERSITY

Notable exception in trend:

Lung cancer, 2001-2021

Age standardised rates (no modelling)



Decreasing incidence for men

Increasing for women

Mortality relatively close to morbidity

Regional and/or socioeconomic differences in cancer rates?

- How big is the gap?
- Is it getting better? Worse?

We need modelling - to account for uncertainty and noise.

Stochastic modelling

- Stochastic modelling for cancer rates

- Transition characterised by underlying rate $\theta_{g,r,a,d,t}$
- $\theta_{g,r,a,d,t}$ depending on **g**ender, **r**egion, **a**ge, **d**eprivation, **t**ime
- Quantify uncertainty (probability intervals)

Bayesian models for incidence and mortality rates

$$
\begin{aligned}
C_{a,t,d,g,r} &\sim \text{Poisson}(\theta_{a,t,d,g,r}\ E_{a,t,d,g,r}) \\
\theta_{a,t,d,g,r} &\sim \text{Lognormal}(\mu_{a,t,d,g,r}, \sigma^2) \\
\mu_{a,t,d,g,r} &= \boldsymbol{\beta}'\boldsymbol{X} \\
\boldsymbol{\beta}\text{'s} &\sim \text{Normal}(0, 10^4) \quad [\textbf{vague priors for risk factor effects}] \\
\sigma^2 &\sim \text{Inv.Gamma}(1, 0.001)
\end{aligned}
$$

- $C_{a,t,d,g,r}$: number of cancer registrations/deaths at **age** a, in **year** t, for **gender** g, **deprivation** level d and **region** r
- $E_{a,t,d,g,r}$: mid-year population estimates
- $\theta_{a,t,d,g,r}$: incidence/mortality rates
- $\boldsymbol{X}$: vector of covariates: **age**, **year**, **deprivation**, **gender**, **region**, average age-at-diagnosis + appropriate interaction(s)
- $\boldsymbol{\beta}$: vector of coefficients

Also: change-point analysis, variable selection

HERIOT WATT UNIVERSITY

Change points

- Allow change point(s) in time trends (and age)
  - E.g. different trend after new health/screening policy introduced
  - or after a certain age

- Changepoint analysis, based on BIC, is considered for detection of changes

$$\mu_{a,t,d,g,r} = \beta_0 + \beta_1 t + \beta_2 (t - \epsilon) \mathrm{I}(t \geq \epsilon)$$

with β_2: change in trend after time point ϵ.

E.g.
$$\mu_{a,t,d,g,r} = \beta_0 + \beta_1 \text{ year} + \dots$$

may become

$$\mu_{a,t,d,g,r} = \beta_0 + \beta_1 \text{ year}_{<2006} + \beta_2 \text{ year}_{\geq 2007} + \dots$$

HERIOT WATT UNIVERSITY

Model selection

- Bayesian variable selection methodology
- Chooses the **best** model for

$$\mu_{a,t,d,g,r} = \boldsymbol{\beta}'\boldsymbol{X}$$

according to *marginal likelihood & Bayes factors*:

$$B_{jk} = \frac{Pr(D|M_j)}{Pr(D|M_k)};\ j \neq k$$

or *deviance information criterion*:

$$\text{DIC} = -2E_{\boldsymbol{\beta}|D}(\log f(D|\boldsymbol{\beta})) + 2\log f(D|\hat{\boldsymbol{\beta}}),$$

HERIOT WATT UNIVERSITY

Initial findings and main trends (Arik et al, 2020)

Variable selection:

- All-cancer and *life-style* cancers, i.e. lung and bowel cancer: all main variables (age, time, deprivation, gender, region) are important

- Breast and prostate cancer mortality: deprivation is **not** important

HERIOT WATT UNIVERSITY

Initial findings and main trends (cont.)

How do various factors affect rates? (in general ...)

- Age: higher rates at older ages
- Time:
  - higher incidence in more recent years
  - lower mortality
- Gender: higher rates for men
- Region?

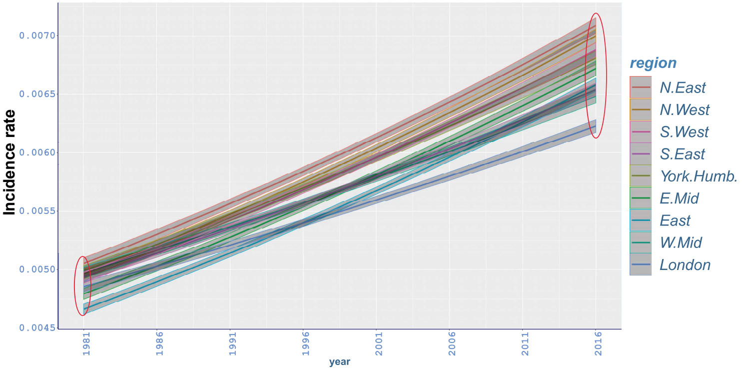
  Deprivation?

Regional variation in cancer rates?

- Is there a geographical pattern?
- Does variation change over time?
- Is variation the same for different types of cancer?

Regional variation

All cancer incidence – Females, 1981-2016



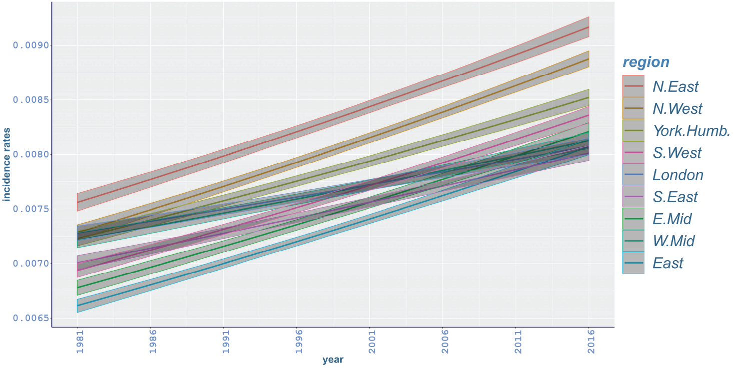
Increasing trend in all regions

Higher incidence in north

Gap widening with time

Regional variation

All cancer incidence – Males, 1981-2016



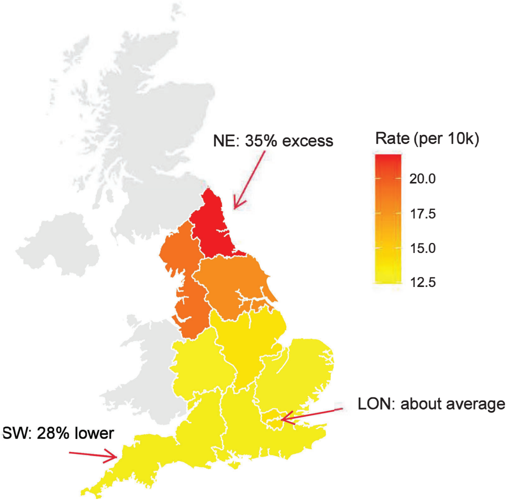
Higher incidence in north

Rates are higher than for women

Gap **not** widening for men

Regional effect

Lung cancer incidence – Females, 2017

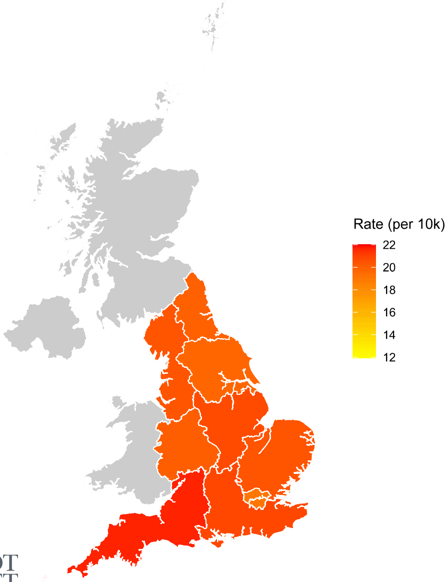


Regional effect compared to average

North v. south?

Regional effect

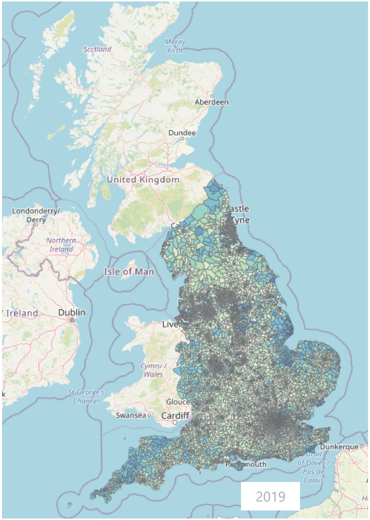
Breast cancer incidence – 2017



Not a 'life-style' cancer

Regional variation much lower

Socioeconomic inequality in cancer rates?

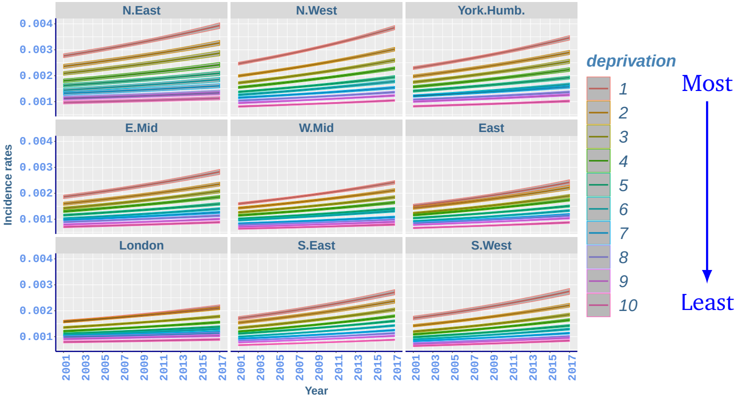


- Use Index of Income Deprivation (ID)
- Deciles: 1 (most deprived), 10 (least deprived)
- For projection (later): quintiles 1 – 5

HERIOT WATT UNIVERSITY

Deprivation inequality in cancer rates

Lung cancer incidence – Females, 2001-2017



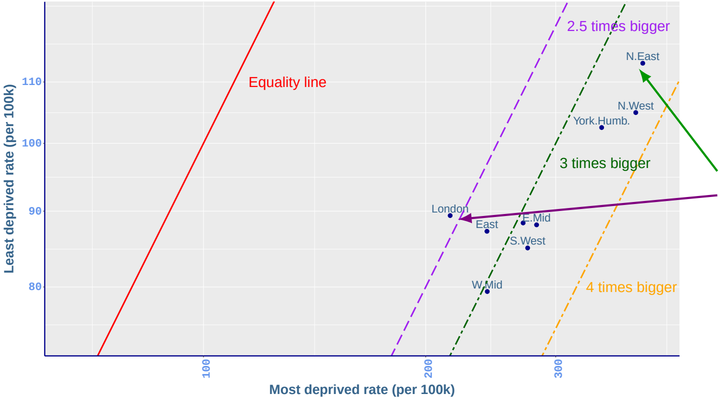
Higher rates for most deprived (1)

Variation by ID getting wider through time

Inequalities more evident in northern regions

Most v. least deprived by region

Lung cancer incidence – Females, 2017



Rates for **most deprived much higher**:

$\times$3.5 N East

$\times$2.5 London

Regional variation

Bayesian forecasting for mortality

$$C_{a,t,d,r} \sim \text{Poisson}(\theta_{a,t,d,r}\, E_{a,t,d,r})$$
$$\theta_{a,t,d,r} \sim \text{Lognormal}(\mu_{a,t,d,r}, \sigma^2)$$
$$\mu_{a,t,d,r} = \beta_0 + \beta_{1,a} + \beta_{2,t} + \beta_{3,r} + \beta_{4,d} + \beta_5 \text{AAD}_{r,d}$$
$$\sigma^2 \sim \text{Inv.Gamma}(1, 0.1)$$
$$\beta_0,\ \beta_1,\ \beta_3,\ \beta_4 \text{ and } \beta_5 \sim \text{Normal}(0, 10^4),$$

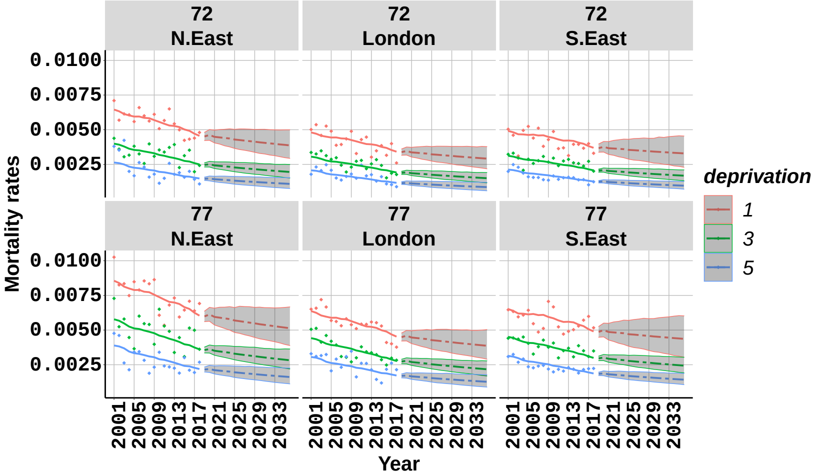
Add random walk with drift for 'period' effect:

$$\beta_{2,t} = \text{drift} + \beta_{2,t-1} + \epsilon_t$$
$$\text{drift} \sim \text{Normal}(0, \sigma^2_{\text{drift}})$$
$$\epsilon_t \sim \text{Normal}(0, \sigma^2_{\beta_2})$$
$$\sigma^2_{\beta_2} \sim \text{Inv.Gamma}(1, 0.001),$$

for $t = 2001, 2002, \ldots, 2018$, where $\hat{\sigma}^2_{\text{drift}} = \frac{\hat{\sigma}^2_{\beta_2}}{2018-2001}$.

Projected mortality – Lung cancer, 2001 - 2035

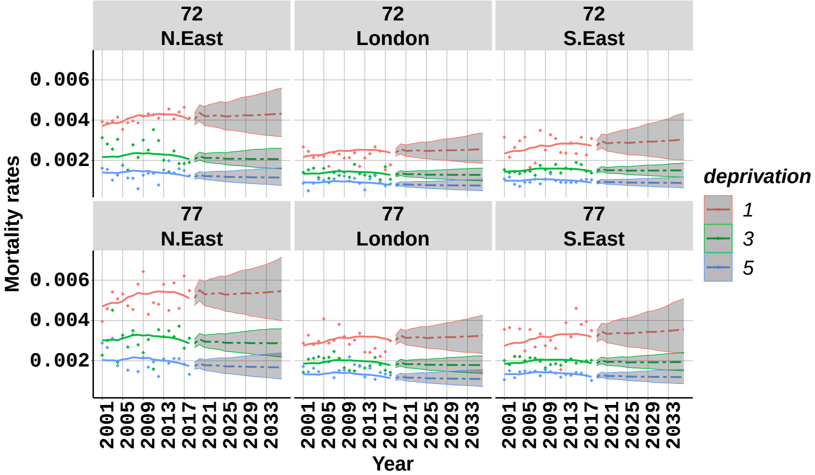
Men 72, 77 yo, deprivation *quintiles*



- Projected rates for most & least deprived NOT overlapping

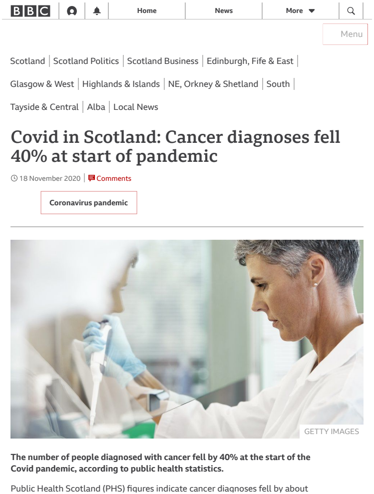
Projected mortality – Lung cancer, 2001 - 2035

Women 72, 77 yo, deprivation *quintiles*



- Mortality for women *NOT* decreasing
- Still rates for most deprived not catching up

Impact of diagnosis delays on mortality



BBC | Home | News | More

Menu

Scotland | Scotland Politics | Scotland Business | Edinburgh, Fife & East | Glasgow & West | Highlands & Islands | NE, Orkney & Shetland | South | Tayside & Central | Alba | Local News

Covid in Scotland: Cancer diagnoses fell 40% at start of pandemic

18 November 2020 | Comments

Coronavirus pandemic

GETTY IMAGES

The number of people diagnosed with cancer fell by 40% at the start of the Covid pandemic, according to public health statistics.

Public Health Scotland (PHS) figures indicate cancer diagnoses fell by about

- Estimate average age-at-diagnosis (AAD) with incidence data
- Include AAD as risk factor in mortality model
  e.g.
  $$\mu_{a,t,d,r} = \beta_0 + \beta_{1,a} + \beta_{2,t} + \beta_{3,r} + \beta_{4,d} + \beta_5 \mathrm{AAD}_{r,d}$$
- Estimate impact on mortality

Projected mortality – Lung cancer, 2001 - 2035

Quantify Covid-19 impact on future mortality

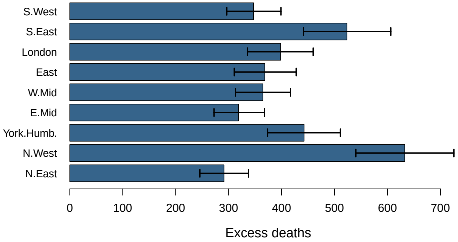
- Assume increase in AAD: e.g. 1 month, 3 months etc.
  - Use ONS region future population estimates
  - Assume future deprivation structure unchanged

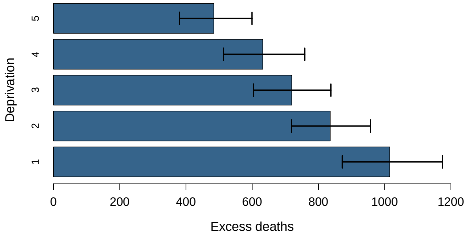
- Fit Bayesian forecasting model:
  - under no change in AAD (baseline scenario)
  - under 1-month AAD increase (Covid scenario)
  - estimate excess deaths

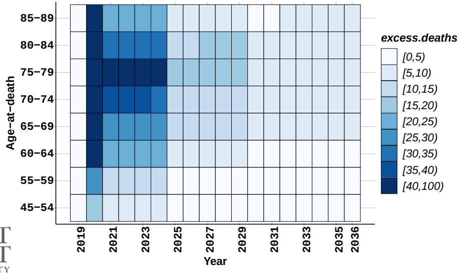
HERIOT WATT UNIVERSITY

Projected mortality – Lung cancer, women, 2001 - 2035

Excess mortality due to 1-month increase in AAD







Total excess deaths: 3,687

Cancer admissions data (US, 2016-2019)

- Source: Merative (formerly IBM Watson Health)
- **Response**: number of hospital (or similar) admissions
- **Explanatory** variables:

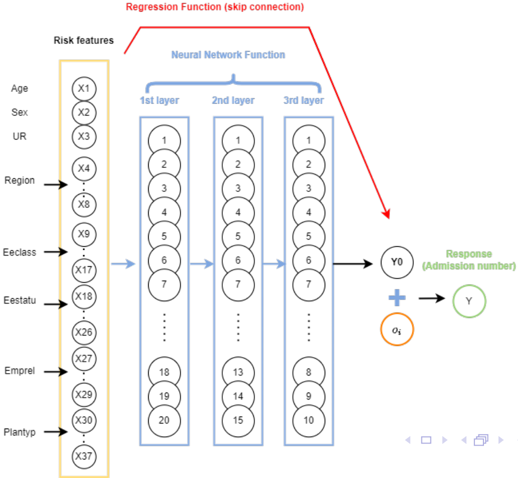
Variable	Description	Type
PLANTYP	Type of plan individual is part of	Factor w/8 levels
AGE	Age of the individual	num 30-65
REGION	Geographical region of residence	Factor w/5 levels
EGEOLOC	Geographic location based on postal code	Factor w/53 levels
UR	Urban/rural ndicator	Factor w/2 levels
EECLASS	Employee classification	Factor w/9 levels
EESTATUS	Status of employment	Factor w/9 levels
EMPREL	Relation to the primary beneficiary	Factor w/3 levels
SEX	Gender of patient	Factor w/2 levels

- $425,202$ records

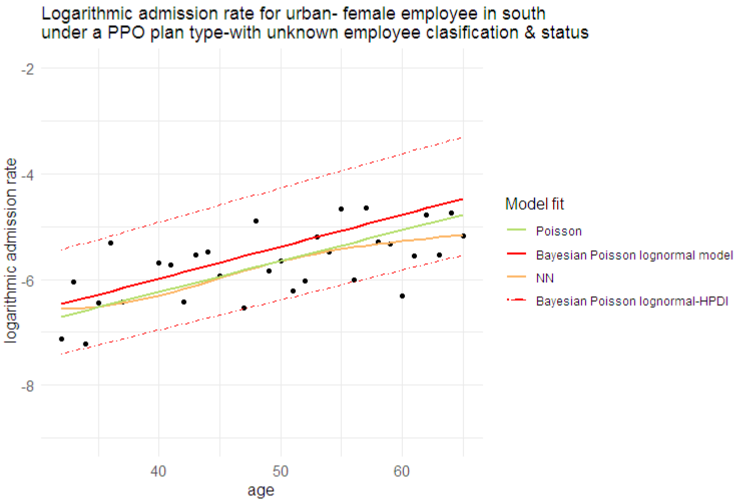
Artificial Neural Network (ANN) models

Replace predictor of GLM with **ANN predictor** – Poisson likelihood:

$$\mu^{CNNPoisR}(\boldsymbol{x}_i) = E_i \exp\left(\underbrace{\langle \boldsymbol{\beta}, \boldsymbol{x}_i \rangle} + \underbrace{\left\langle \boldsymbol{w}^{(d+1)}, \left(\boldsymbol{z}^{(d)} \circ \cdots \circ \boldsymbol{z}^{(1)} \right)(\boldsymbol{x}_i) \right\rangle} \right)$$



Modelling results – learning data



NN fit more flexible

Predictive performance: GLM v Bayes v ANN

Table: Average loss over 10-fold validation

Model	Learning loss	Testing loss	Portfolio average
Observed			0.0027
GLM	16.747	16.849	0.0030
Bayes	16.771	16.785	0.0030
NN_{Pois} (20,15,10)	16.378	**16.652**	0.0027
$CANN_{Pois}$ (20,15,10)	16.475	16.830	0.0027

- 90-10 training-testing split
- NN approach: better predictive performance over testing data
- Followed by Bayesian model

Summary

1. Regional and socioeconomic gap for cancer rates is widening in the UK

   ... but not for all cancer types

2. Covid-related delays in diagnoses can lead to significant increase in cancer deaths

   – also region dependent

3. Projection for lung cancer mortality shows persistent deprivation gap

   – and significant excess deaths due to covid-like disruptions

4. ANNs can provide enhanced rate predictions

   – but we need to address interpretability

5. Can public health interventions at regional and deprivation level contribute to lower cancer incidence and deaths?

More details in:

- Arık, A., Cairns, A., Dodd, E., Macdonald, A.S., Streftaris, G. (2023) The effect of the COVID-19 health disruptions on breast cancer mortality for older women: A semi-Markov modelling approach, *arXiv:2303.16573*.
- Yiu, M.T.L., Kleinow, T., Streftaris, G. (2023) Cause-of-death contributions to declining life expectancies using cause-specific mortality reversion scenarios, *to appear, North American Actuarial Journal*.
- Kwok, W. M., Dass, S. C., & Streftaris, G. (2023). Deep Learning Aided Laplace Based Bayesian Inference for Epidemiological Systems, *Computing and Statistics*.
- Jose, A., MacDonald, A. S., Tzougas, G., & Streftaris, G. (2022). A Combined Neural Network Approach for the Prediction of Admission Rates Related to Respiratory Diseases. *Risks*.
- Arık, A., Dodd, E., Cairns, A., Streftaris, G. (2021) Socioeconomic disparities in cancer incidence and mortality in England and the impact of age-at-diagnosis on cancer mortality, *PLOS ONE*.
- Arık, A., Dodd, E., Streftaris, G. (2020) Cancer morbidity trends and regional differences in England - a Bayesian Analysis, *PLOS ONE*.