Use of Big Data For Actuarial Research on Longevity

Prof Elena Kulinskaya (UEA) and Mr Nigel Wright (Aviva Life Analytics)

Content

Big Data and Statistics

Big Data Research at UEA

IFoA grant on modelling longevity
- Aims
- Methodology

Case Study on Statins
- Rationale
- Study design & Methodology
- Results

Discussion

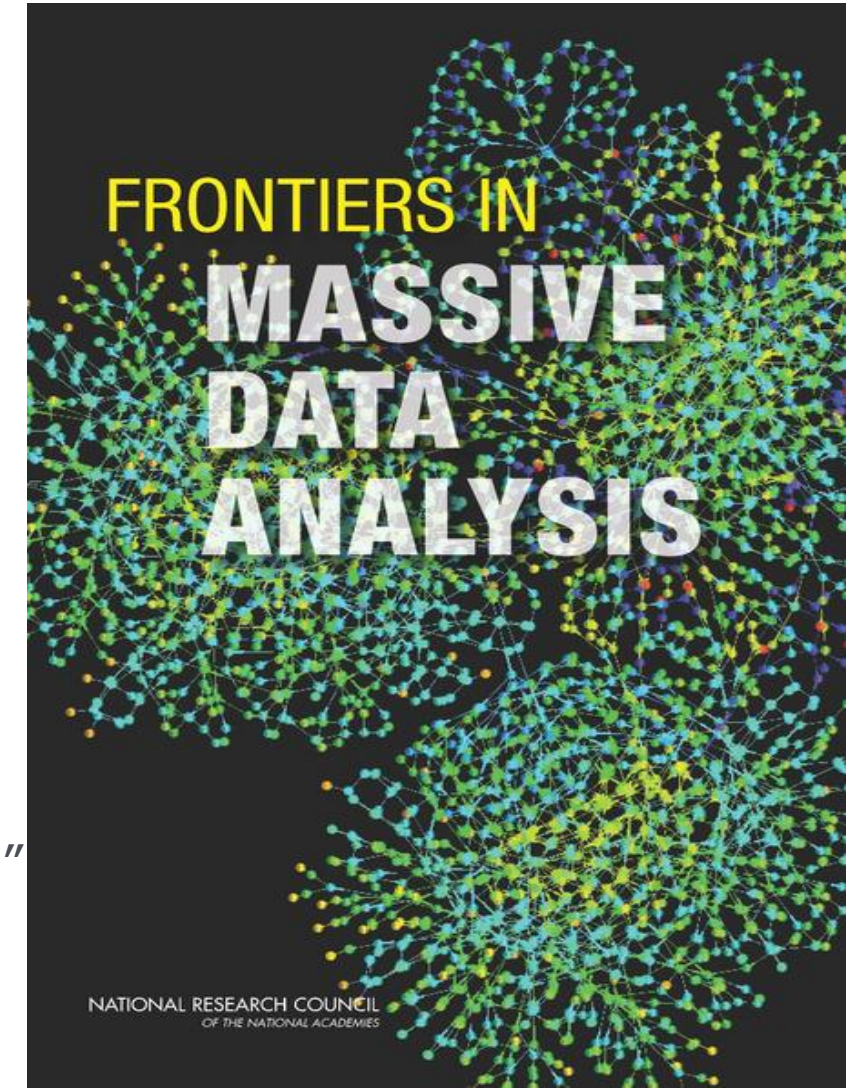
Recommendations

Data explosion

"90% of the data in the world today has been created in the last two years alone. Some estimate that data production will be 44 times greater in 2020 than it was in 2009. Others estimate an additional 2.5 quintillion bytes of data is being generated every day."

Australian Public Service Big Data Strategy, Commonwealth of Australia 2013.

"We now create as much information in two days as humans did from the beginning of history to 2003" (Eric Schmidt, Google CEO, 2010).



Big Data across the world

Access to unprecedented volumes and complexity of data collected by government, businesses, and social media provides a seemingly unlimited resource for harvesting new knowledge.

Many jurisdictions have developed reports identifying improvement of capacity to harness 'Big Data' as a major strategic priority.

Research Councils UK, 2012

Australian Government, 2013

Government of Canada, 2013

Scheveningen Memorandum, 2013 (Eurostat)

Office of Science and Technology Policy, 2012

Government of India, 2014

He *et al.,* 2014 (Commonwealth)

Fields and Disciplines

Big Data arise in such fields as genomics, public health, environmental sciences, neuroscience, and business.

Basic issues of management and storage have primarily implicated computer science, underpinning initiatives sometimes described as business analytics or data science.

Statistical science has not played a prominent role. Because practical problem solving has proceeded rapidly, the science has lagged behind and work to identify the statistical features associated with Big Data has been largely ad hoc.

The Alan Turing Institute: Cambridge, Edinburgh, Oxford, Warwick and UCL.

Statistical Problems in Big Data

False positives arising from multiple exploratory analyses

Biases due to peculiarities of units, outcomes or settings

Missing data

Inadequate linkage strategies

Data gathered at varied levels such as transaction, person, organization, community, and state

Causal Inference from (mostly) correlational data

Modelling heterogeneity

ESRC Business and Local Government Data Research Centre (BLG DRC)

- Funded under the ESRC's Big Data Network, £5m, 2014-19
- An Eastern ARC (Essex, UEA & Kent) partnership
- Exploitation of data to benefit researchers, data owners and society
- Facilitating access to data, stimulating innovative policy/practice-relevant research
- Highest ethical standards, anonymised data used non-disclosively
- Training for researchers and other users
- Business Engagement Programme

ESRC Business and Local Government Data Research Centre

University of Essex

UEA University of East Anglia

University of Kent

LSE

ESRC BLG DRC Research Programme

Regional drivers of smart economic growth (led by Essex, with input from Norwich Business School at UEA)

Support for vulnerable people (led by UEA with input also from Personal Social Services Research Unit at Kent and LSE)

Green Infrastructure (led by UEA School of Environmental Sciences)

A methodologic programme for Big Data analytics

Data can be safely accommodated at the UK Data Archive at Essex, and safe access is possible in different formats, including through the safe rooms at UEA and Essex.

Methodology research within ESRC BLGDRC

Techniques and methods on the quality, pre-processing and analysis of Big Data.

Methodologies for merging data from multiple sources and robust techniques for data quality grading and assurance

Automated data quality and cleaning procedures

Privacy preserving data mining methods, including methods for dimensionality reduction and data perturbation techniques

Techniques for extracting entities, relations between them, opinions and other elements for use to support semantic indexing and visualization and anonymisation.

Automatic methods for tracking interactions, for example, to identify service pathways in local government or business data.

Machine learning and other methods for identifying stylised facts, seasonal, spatial or other relations, patterns of behaviour at the level of the individual, group, or region from transactional data from business, local government or other organisations.

Methodology research within ESRC BLGDRC

Research on modelling, and predicting complex and adaptive socio-economic systems

Models and statistical methods for the analysis of local government health and social care data

New data mining and machine learning algorithms to identify intervention subgroups

Evidence synthesis for data at various levels of aggregation

Early warning systems for social care

Methodology of observational data analysis

Use Of Big Health And Actuarial Data For Understanding Longevity And Morbidity Risks, IFoA 2016-2020

Consortium:

University of East Anglia:
School of Computing Sciences (CMP) and Norwich Medical School (NMS).
Aviva Life Plc.

Principal Investigator Prof Elena Kulinskaya, Aviva Chair in Statistics, CMP

UEA co-investigators: Dr Beatriz de la Iglesia, Senior Lecturer, CMP;
Prof Ruth Hancock, NMS, Prof Nick Steel, NMS.

Aviva co-investigators: Mr Nigel Wright, actuary; Ms Sarah Allen, Senior Data Analyst, the Life Risk Analytics team.

Main objectives

Development of novel statistical and actuarial methods for:

modelling mortality

modelling trends in morbidity

assessing basis risk

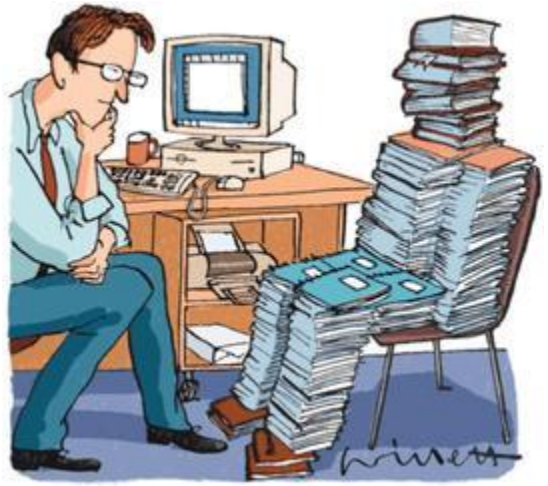
evaluating longevity improvement based on Big Health and Actuarial Data

tools to forecast longevity risk of a book

Data

The Health Improvement Network (THIN) data

- Medical records from primary care
- Representative of the UK when adjusted for deprivation



- All patients born before 1960 and followed to 01.01.2015,
  this includes 3.4 million patients

- Added various social economic status variables such as IMD and Mosaic

- The Continuing Mortality Investigation (CMI) data

Aim 1: Identification and quantification of the key factors affecting mortality/ longevity

We intend to have a target list of between 3-5 conditions or interventions.

We propose to consider statin prescription, an established longevity-improving intervention as one of the target scenarios.

Other conditions may include type 2 diabetes or heart failure.

Health interventions may include an introduction of NICE guidelines on use of particular health sustaining drugs such as calcium channel blockers, or targeted outcomes such as the blood pressure targets.

Lifestyle factors may include obesity or smoking.

Design and methods

For each of these conditions we will design a population-based prospective cohort study using an appropriate extract of the primary care data.

We intend to use a case-control design with cases matched with several controls from the same GP practice. This provides balanced and comparable cohorts of cases and controls and simplifies the study of comparatively rare conditions without loss of efficiency.

The full list of relevant confounding variables will be established from medical literature such as systematic reviews, and from expert knowledge within the team, and then the subset of these variables to be adjusted for will be found through backward elimination.

To account for the interdependence of patients from the same GP practice, multilevel modelling and multiple imputation will be used.

Aim 2. Modelling of temporal changes in the factors affecting morbidity and mortality

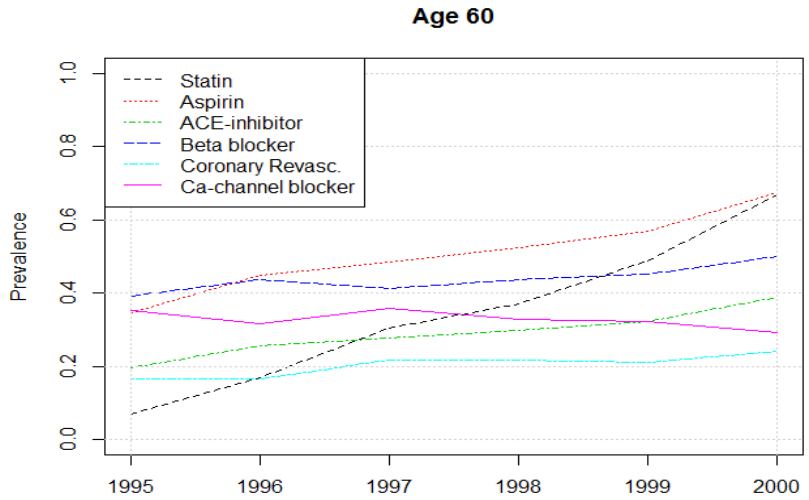
Trends in the incidence and/or prevalence of particular medical conditions and/or lifestyle factors will also be obtained from the primary care data.

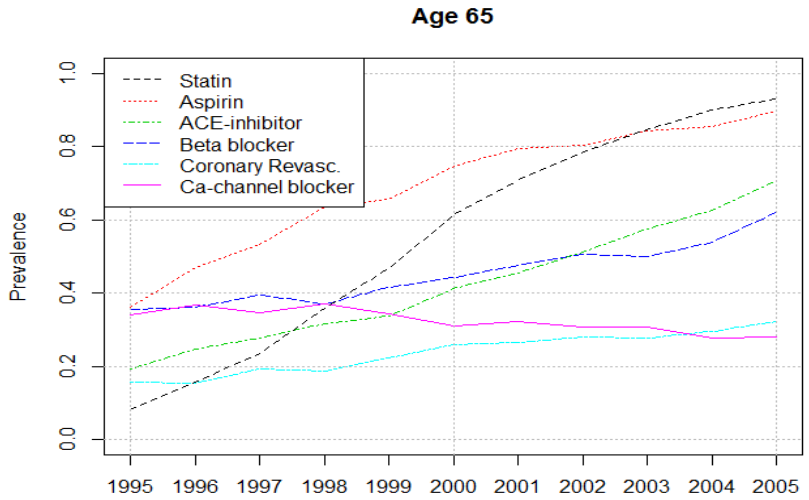
This will enable us to establish patterns due to social or geographic inequalities, such as socio-economic status (SES), age or postcode lottery.

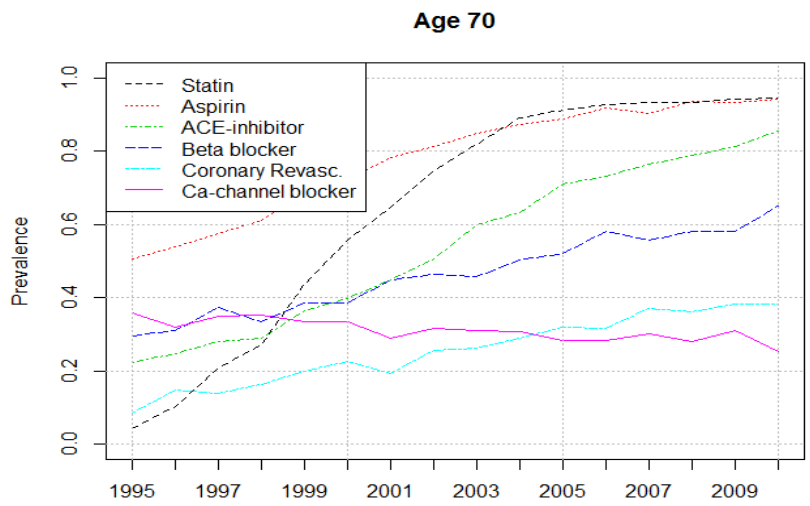
For instance, the patients in the more deprived areas may be disadvantaged in regards to the latest interventions and/or public health campaigns at least initially. This will result in widening the gap in longevity between individuals from different backgrounds.

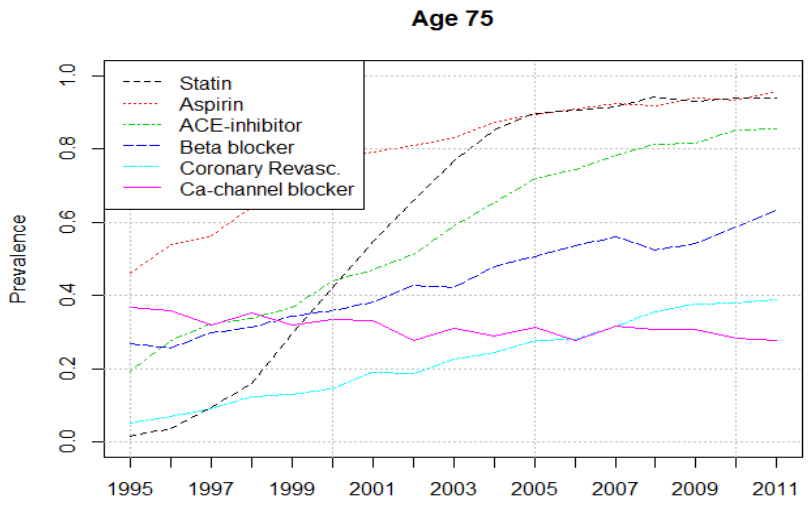
Thus to be able to ascertain an effect on longevity of a population, we need to model the incidence of a condition or an uptake of an intervention over time in parallel to modelling mortality.

Prevalence of treatment by cohort's age in patients with a history of acute myocardial infarction

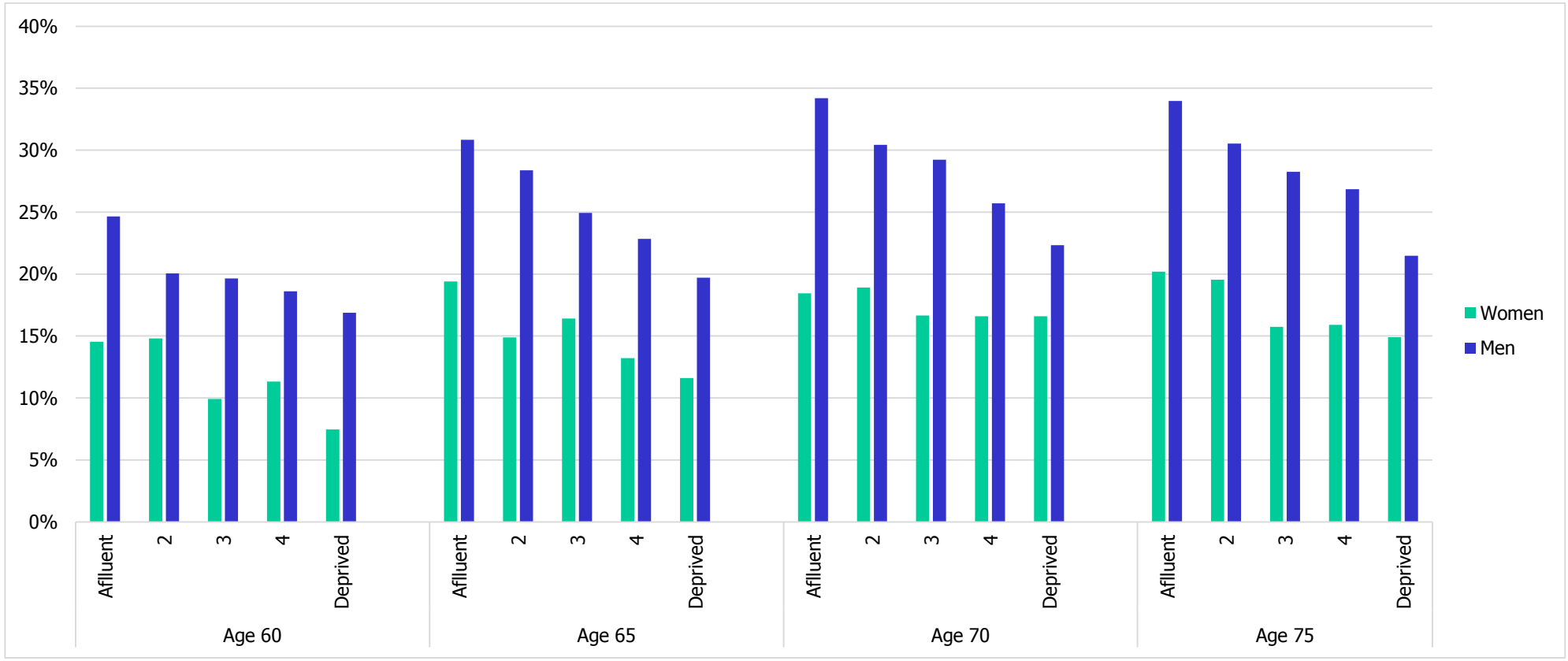








Example: Coronary Revasculisation given IHD



Aim 3. Evaluation of plausible scenarios in mortality trends due to particular medical advances or lifestyle changes on the population of insureds

As often happens with the existing portfolio of insured lives, the minute health details of a life are not available. Instead, the interest lies in the mortality trends of the whole book.

To be able to provide this information, three components are required:

- established in Aim 1 model for survival differentials associated with a particular disease or intervention;
- developed in Aim 2 model for the incidence/ prevalence of this condition or uptake of this intervention over time,
- and the sufficient knowledge of the population to which it is desired to translate trends in longevity established in general population to be able to assess the basis risk.

Example: evaluation of the contribution of the new medical guidelines

Given a model of survival benefits of lipid-lowering drugs such as statins, and a model for trends in prescription of statins over time we can evaluate the contribution of the new medical guidelines on widening statin prescription to the overall change in longevity for a population of a known composition, see the case study on statins below.

Access to the individual or high granularity level data submitted to Continuing Mortality Investigation would enable us to evaluate the composition of populations in question

Aim 4. Tools to forecast longevity risk of a book

We will develop an R package incorporating our models and providing analytical and graphical means to forecast longevity of a general UK population, and also of a population of a user defined composition under a number of scenarios for changes in disease incidence, health behaviours and treatments.

This will be an open source software available from the project website along with an accompanying manual for its use.

We also intend to develop teaching materials for the actuarial community on the modelling techniques used in the project, and the use of the developed R package. These materials will be available from the project website.

Case Study: Statins and Life Expectancy

By: Lisanne Gitsels, PhD candidate

supervisors: Prof Elena Kulinskaya, Prof Nick Steel,
Mr Nigel Wright (Aviva)

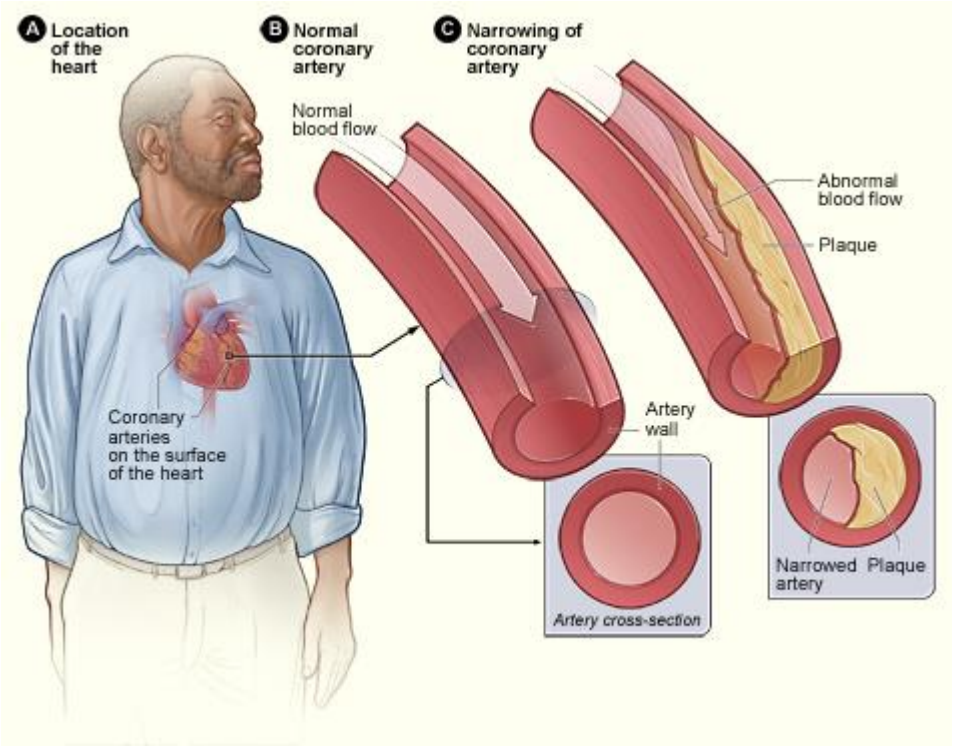
Medical dictionary

Cardiovascular disease (CVD)

- Definition: diseases of the heart and circulation
- E.g. heart attack and stroke
- Number 2 cause of death: 28%

Prevention by lipid-lowering therapy

- Definition: drugs that are used to treat high cholesterol (fatty substance)
- E.g. statins



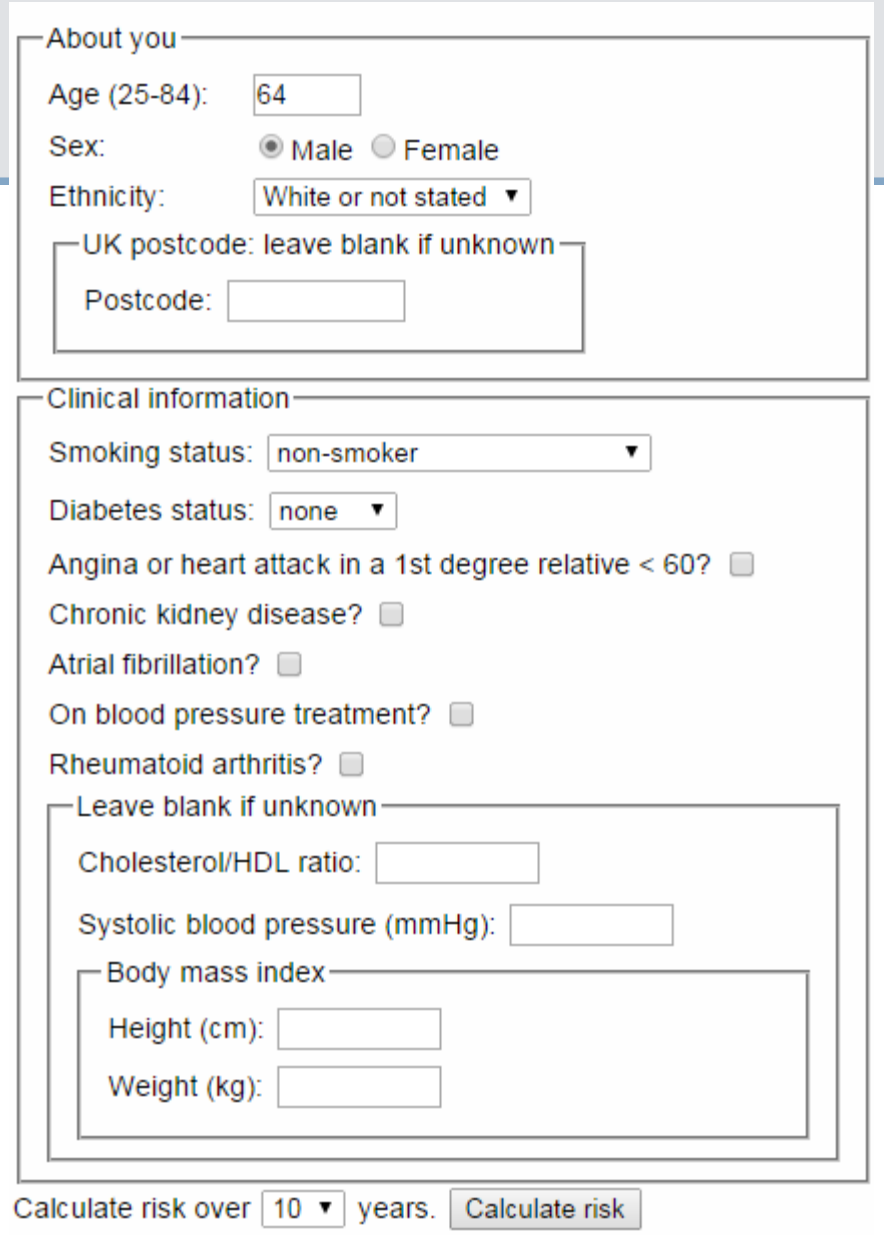
Primary prevention of CVD

Calculate 10-year risk of first cardiac event

Prescribe statins when risk is above threshold

Last year, the National Institute of Health and Clinical Excellence (NICE) lowered threshold from 20% to 10% risk

- 4.5 million UK residents eligible for statins



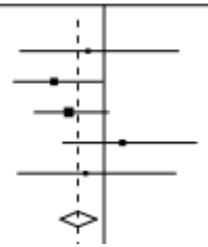
Previous research on efficacy of statins

Meta-analysis of clinical trials by Cholesterol Treatment Trialists' (CTT) Collaborators

- There is an overall survival benefit by statins
- Survival benefit for individual risk groups remain uncertain

Webfigure 9: Effects on any deaths per 1.0 mmol/L reduction in LDL cholesterol at different levels of risk, by history of vascular disease and overall

5−year MVE risk at baseline	Deaths (% per annum) Statin/more	Control/less	RR (CI) per 1.0 mmol/L reduction in LDL cholesterol	Trend test
Participants without vascular disease				
< 5%	164 (0.38)	177 (0.41)	0.94 (0.71 − 1.26)	
≥ 5%,<10%	372 (0.77)	446 (0.93)	0.83 (0.69 − 0.99)	
≥ 10%,<20%	703 (1.99)	778 (2.19)	0.88 (0.76 − 1.02)	χ^2_1=1.57
≥ 20%,<30%	363 (5.13)	339 (4.73)	1.06 (0.86 − 1.32)	(p=0.2)
≥ 30%	192 (10.76)	192 (11.44)	0.94 (0.70 − 1.25)	
Subtotal	**1794 (1.33)**	**1932 (1.42)**	**0.91 (0.85 − 0.97) p= 0.007**	



Limitations of studies

Selection bias

- Exclusion of e.g. patients with comorbidities or on multiple drugs

Follow-up not sufficiently long

No adjustment for confounders

- Confounder = factor related to both exposure and outcome

Research question

What is the survival benefit of statins prescription
as primary prevention of cardiovascular disease
for different risk groups at various ages
in the general population?

Design and Data Selection

Population-based prospective cohort study

Restrictions data:

- Medical records from 1987 to 2011 of people born between 1920 and 1940

Target ages:

- 60, 65, 70, and 75

Exclusion:

- Patients with a history of cardiovascular disease

Missing data

Incomplete records in: BMI, smoking status, and risk of cardiac event

Multiple imputation

- Joint modelling
  - » Linear regression for BMI and risk of cardiac event
  - » Ordered probit regression for smoking status
- Multilevel on GP practice
- MCMC (Monte Carlo Markov Chain) 500 iterations resulting in 10 imputed datasets
- REALCOM-Imputation software

Model specification

Cox's proportional hazard regression estimates the hazard λ_{ij} for patient *i* from GP practice *j*: $\lambda_{ij} = \lambda_0(t)\, Z_j\, e^{\beta X_{ij}}$

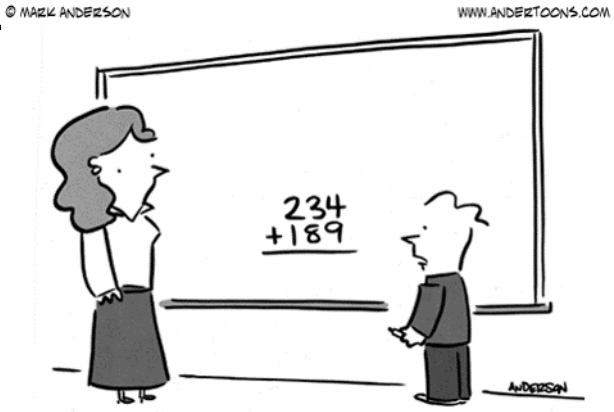
where λ_0 = baseline hazard (function of time),
Z_j = shared frailty term on GP practice,
β = coefficients (constant),
and X_{ij} = exposures, e.g. statins (constant).

Models specified:

- Ages: 60, 65, 70, and 75
- Risk groups:
  - » Low <10% risk of cardiac event
  - » Moderate 10-19% risk of cardiac event
  - » High ≥20% risk of cardiac event

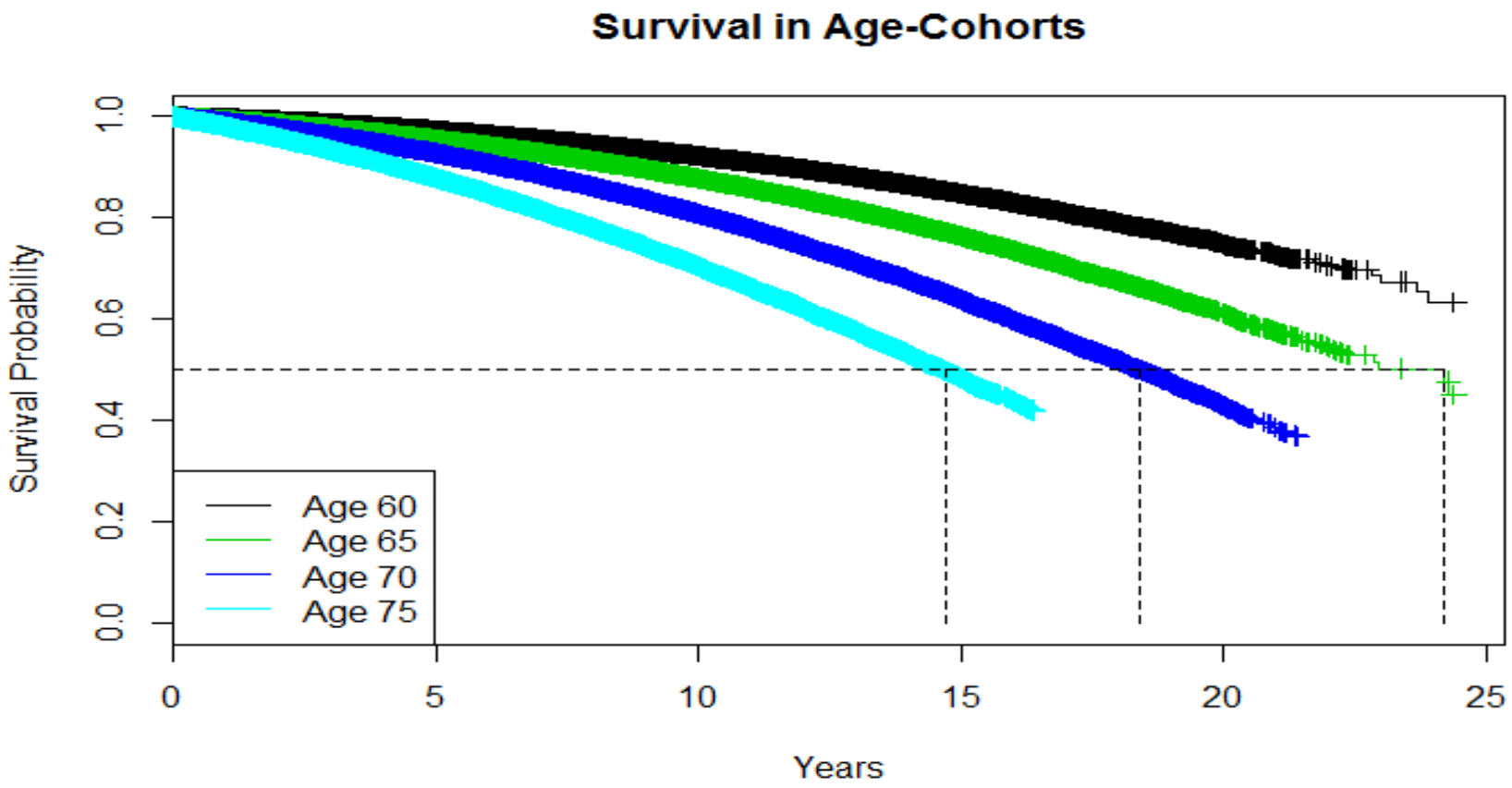
Cohorts' characteristics

Cohort	Number of patients	Number of deaths	Average follow-up time	Maximum follow-up time
Age 60	118,700	15,296 (12.8%)	12 years	24 years
Age 65	199,574	28,848 (14.5%)	10 years	24 years
Age 70	247,149	40,699 (16.5%)	7 years	21 years
Age 75	194,085	37,356 (19.2%)	6 years	16 years



"Does this count as big data?"

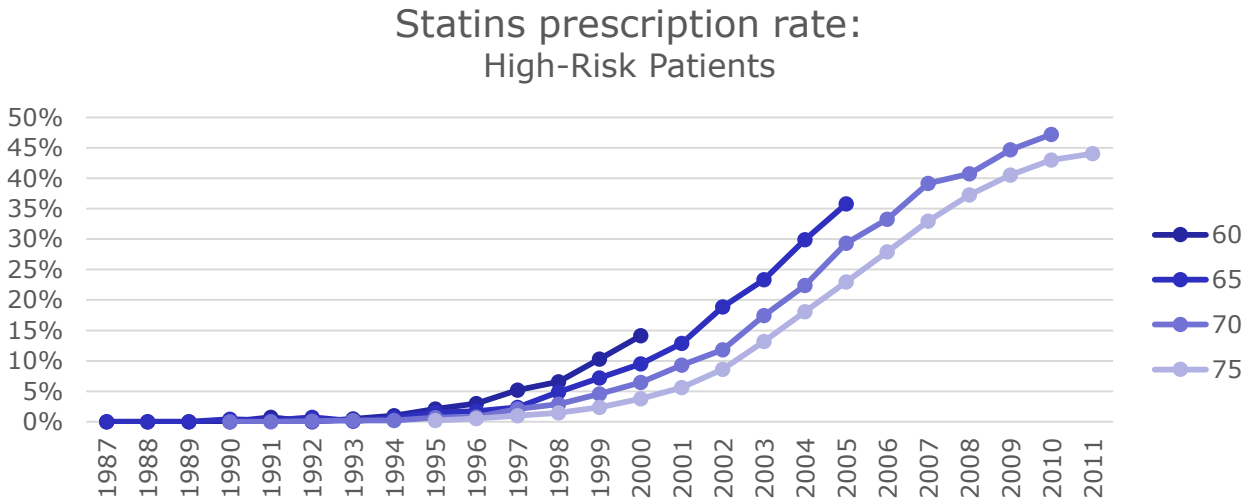
Kaplan-Meier plots

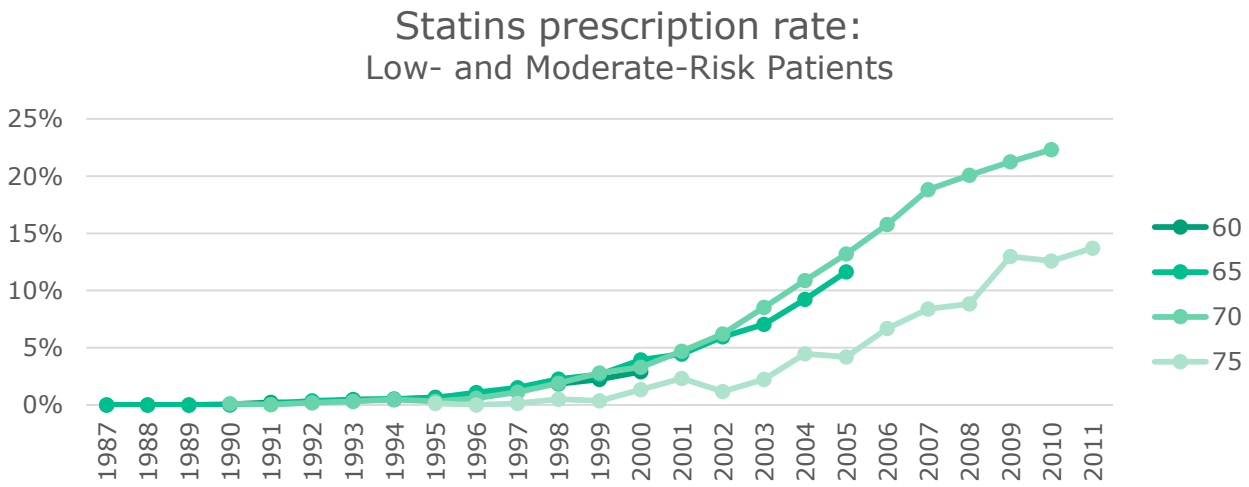


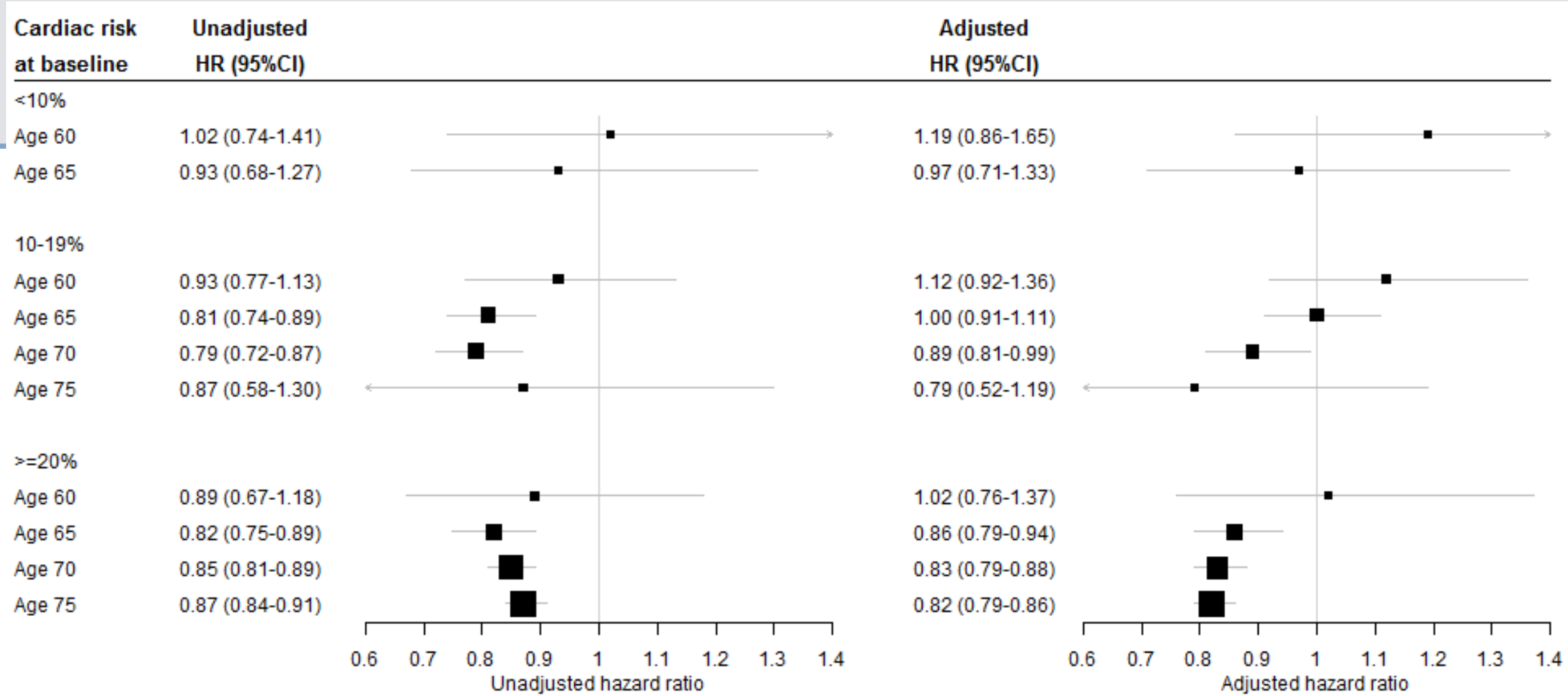
Distribution men and women across risk groups

Cohort	Cardio risk	Women % (Statins %)	Men % (Statins %)
Age 60	Low	83 (1.2)	16 (0.4)
	Moderate	16 (3.7)	78 (1.3)
	High	1 (11.9)	6 (5.2)
Age 65	Low	40 (2.2)	.
	Moderate	55 (7.4)	72 (3.2)
	High	5 (26.9)	28 (12.4)
Age 70	Moderate	80 (9.5)	17 (5.4)
	High	20 (28.2)	83 (17.4)
Age 75	Moderate	15 (4.6)	.
	High	85 (19.6)	100 (19.1)

Uptake of statins by risk group







The adjusted HRs take into account gender, year of birth, postcode, diabetes, high cholesterol level, blood pressure regulating drugs, BMI, smoking status, and GP practice.

Impact

Medicine and Public Health:

Research on the prescription of statins in the general population

- Age discrimination

Risk and age (!) dependent guidelines on prescription of statins

Insurance and Government:

Pricing and reserving for longevity risk (annuities, pension liabilities, etc.) and morbidity and mortality risk

Predicting volumes of coverage of medical procedures

Predicting changes in population life expectancy.

Personal:

Calculate your average life expectancy (and confidence limits)

Decide how you should structure your retirement funds

See if any lifestyle changes can be made (e.g. stop smoking)

By age 70 you would potentially increase the life expectancy by being on statins