University of St Andrews | FOUNDED 1413

Artificial intelligence and datamining with biological data.

Tom Kelsey
Professor of Health Data Science
School of Computer Science
University of St Andrews

ISLCCC 22
Utrecht
July 2022

Overview

- Case studies: Biological data science for late effects research
  - Original model
  - Validation
  - Application(s)
  - The future
- Common themes
- Techniques
- External validation
- Implications

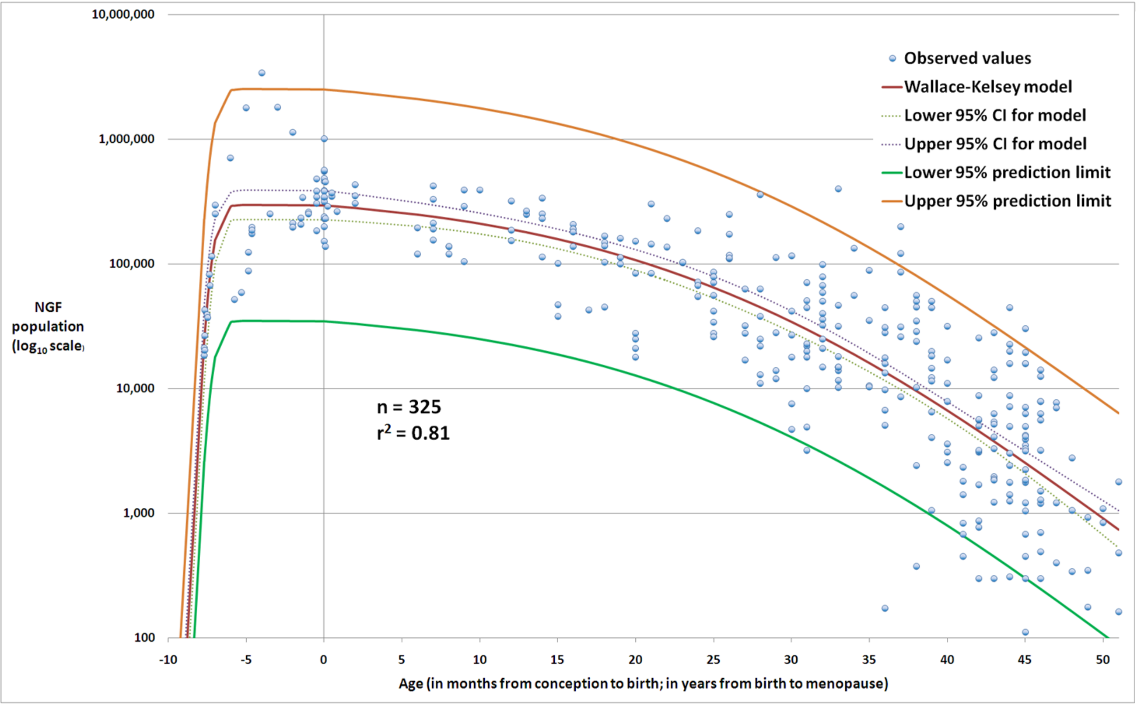
Methodology

- Data aggregation
  - Systematic search for data sources from the literature
  - Tables, charts, descriptive statistics
  - Our own data – if available
- Data selection to create data set with minimal bias
  - Exclusion & inclusion criteria (e.g. exclude infertile)
- Homogeneous data set that approximates the healthy population for a wide range of ages
- Identify model with good fit to the data and low generalisation error
  - Accurate when used to predict unseen examples

University of St Andrews

Observational data model - external validation

Number of potential eggs (NGFs) in the human ovary



Predictions compared to later observations from a population-based cohort

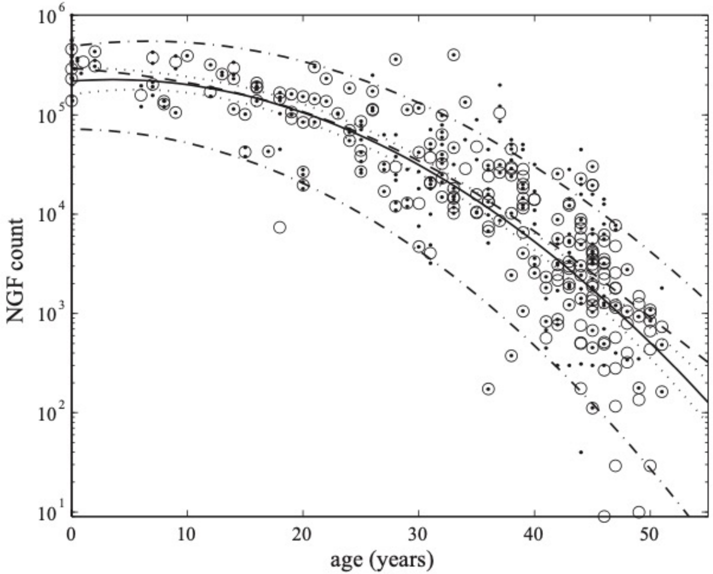


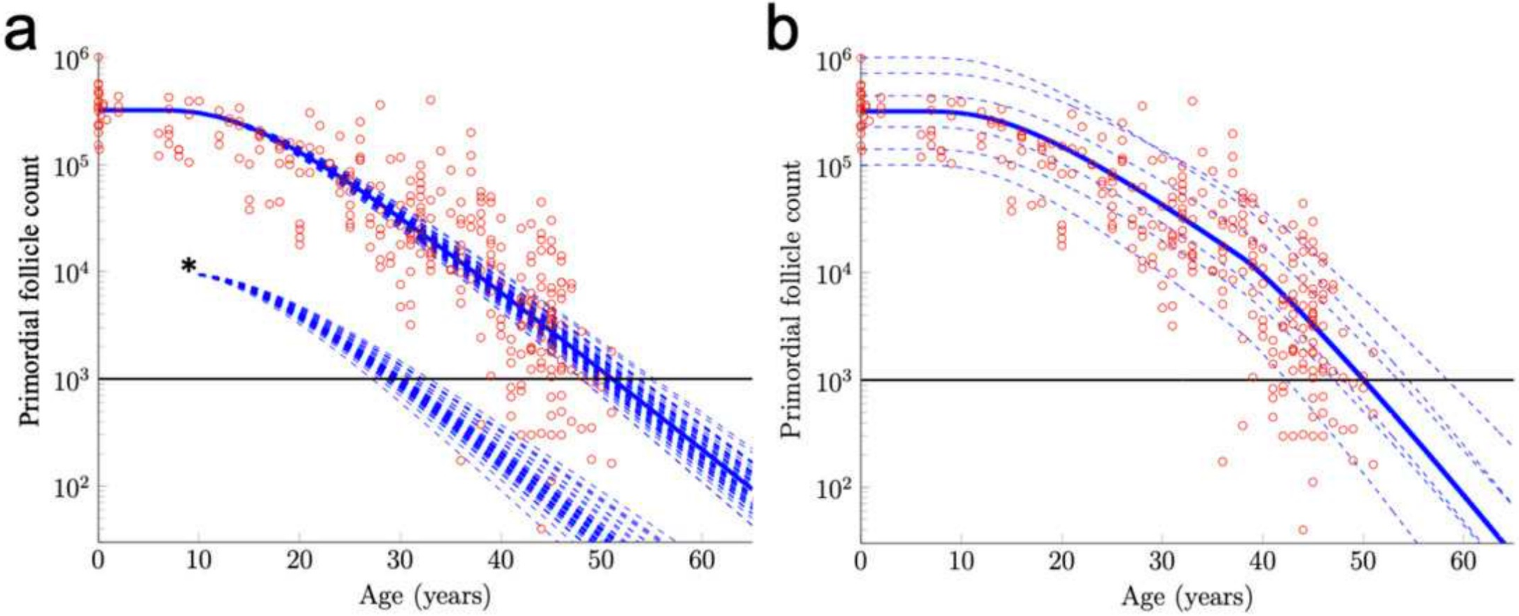
Figure 1. NGF counts (circles) from the 2015 data and (dots) from the 2010 data. The quadratic regression (solid line) fitted to the 2015 data with 95% confidence intervals (dotted lines) and 90% prediction intervals (dash-dotted lines). The Wallace-Kelsey model as fitted to the 2010 data is also depicted (dashed line).

Wallace & Kelsey, PLoS ONE 2010; Depmann *et al.*, JCEM 2015

University of St Andrews

Random walk approach

Number of potential eggs (NGFs) in the human ovary



J. Johnson, JW. Emerson, SD. Lawley; https://doi.org/10.1101/2022.05.31.494088

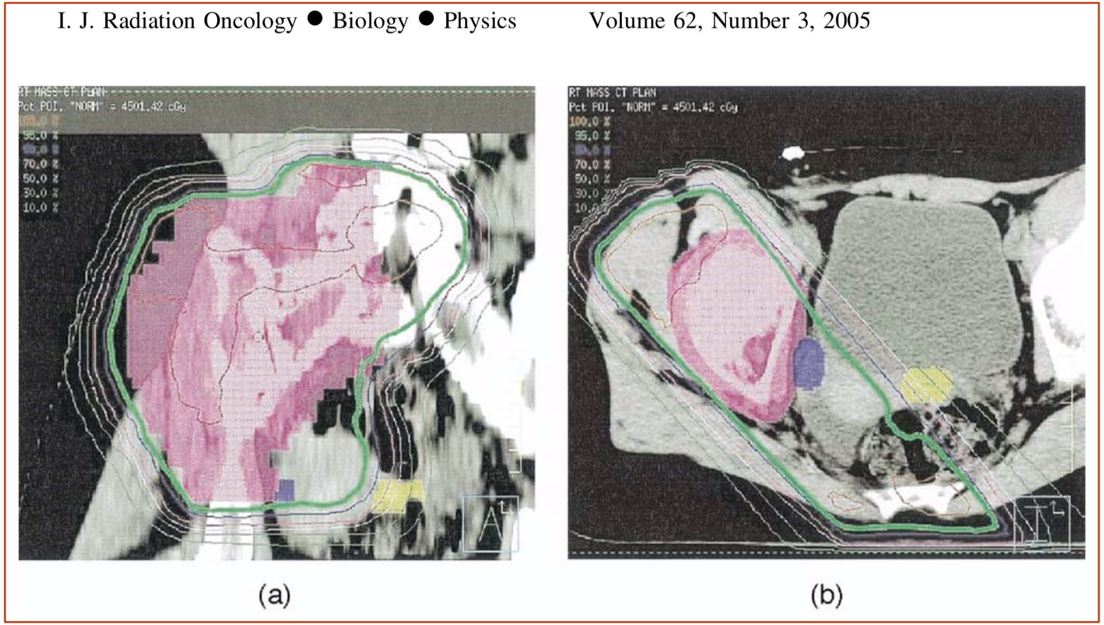
Random walk approach

- Same data
- Good agreement with Wallace-Kelsey
  - In particular, accurate prediction of ages at menopause
- More modern technique (arguably)
  - Machine learning/AI method to remember or forget elements to optimise results
  - Stochastic gradient descent
- Can be used to test two important assumtions made by Wallace-Kelsey
  - High/low population at birth means late/early menopause
  - Radio- and/or chemotherapy moves a patient to an older age in terms of fertility, and decline is at the rate for the older healthy woman

J. Johnson, JW. Emerson, SD. Lawley; https://doi.org/10.1101/2022.05.31.494088

University of St Andrews

Fertility after Radiotherapy



I. J. Radiation Oncology ● Biology ● Physics Volume 62, Number 3, 2005

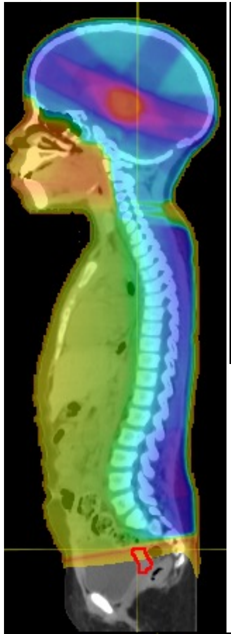
- Estimate LD_{50} for the human oocyte
- Use to plan conformal RXT to optimise dose to the least-affected ovary
- Calculate window of opportunity for fertility
- Calculate the age-related effective sterilising dose
- Use to inform fertility preservation decision making

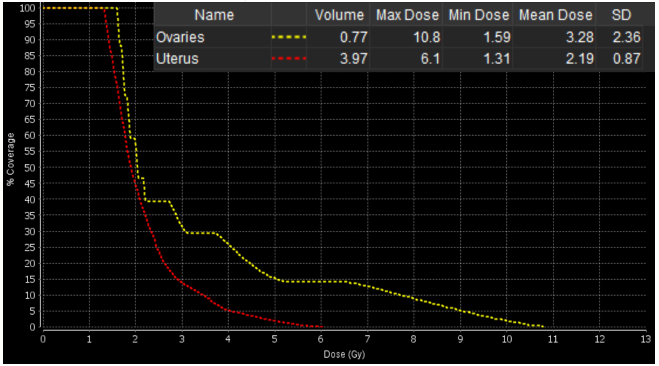
- Minimise the long-term effects of radiotherapy on healthy tissue
- Whilst maintaining cure rates

Wallace *et al.*, 2003; Anderson *et al.*, Lancet Diabetes Endocrinol. 2015

University of St Andrews

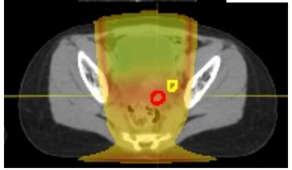
Fertility after Radiotherapy



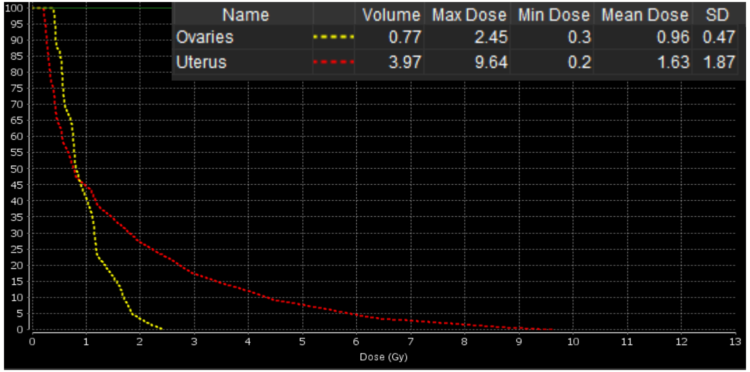


Name		Volume	Max Dose	Min Dose	Mean Dose	SD
Ovaries	- - - - -	0.77	10.8	1.59	3.28	2.36
Uterus	- - - - -	3.97	6.1	1.31	2.19	0.87

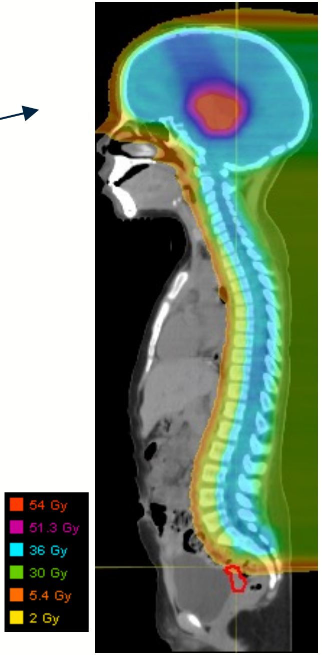
Photon Plan

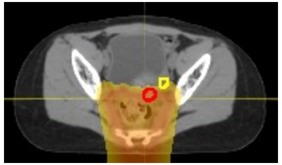


Proton Plan



Name		Volume	Max Dose	Min Dose	Mean Dose	SD
Ovaries	- - - - -	0.77	2.45	0.3	0.96	0.47
Uterus	- - - - -	3.97	9.64	0.2	1.63	1.87



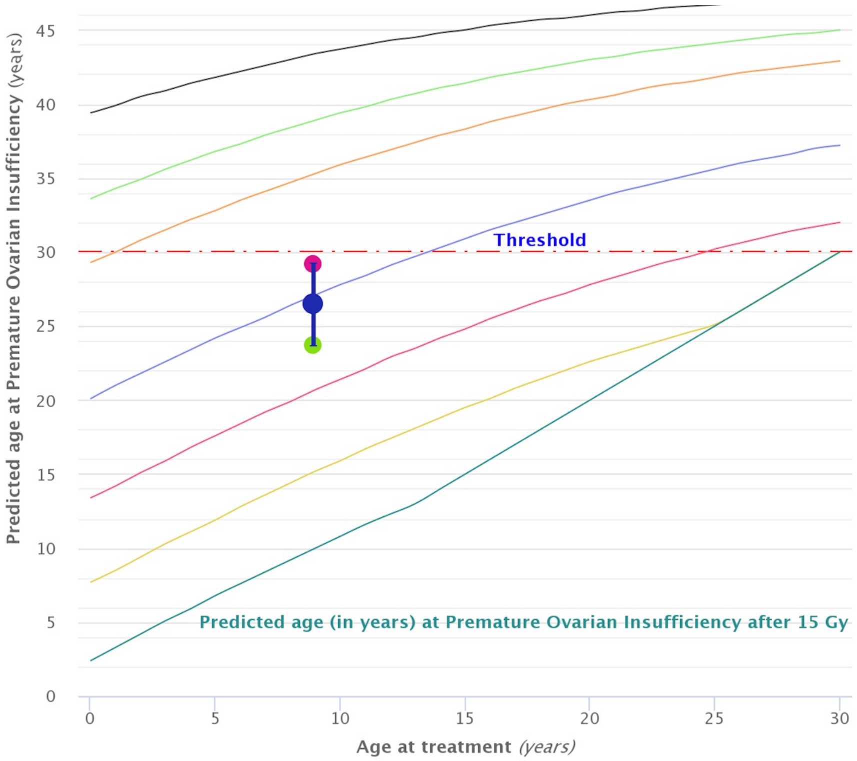


Source: Danny Indelicato, MD - University of Florida Proton Therapy Institute

University of St Andrews

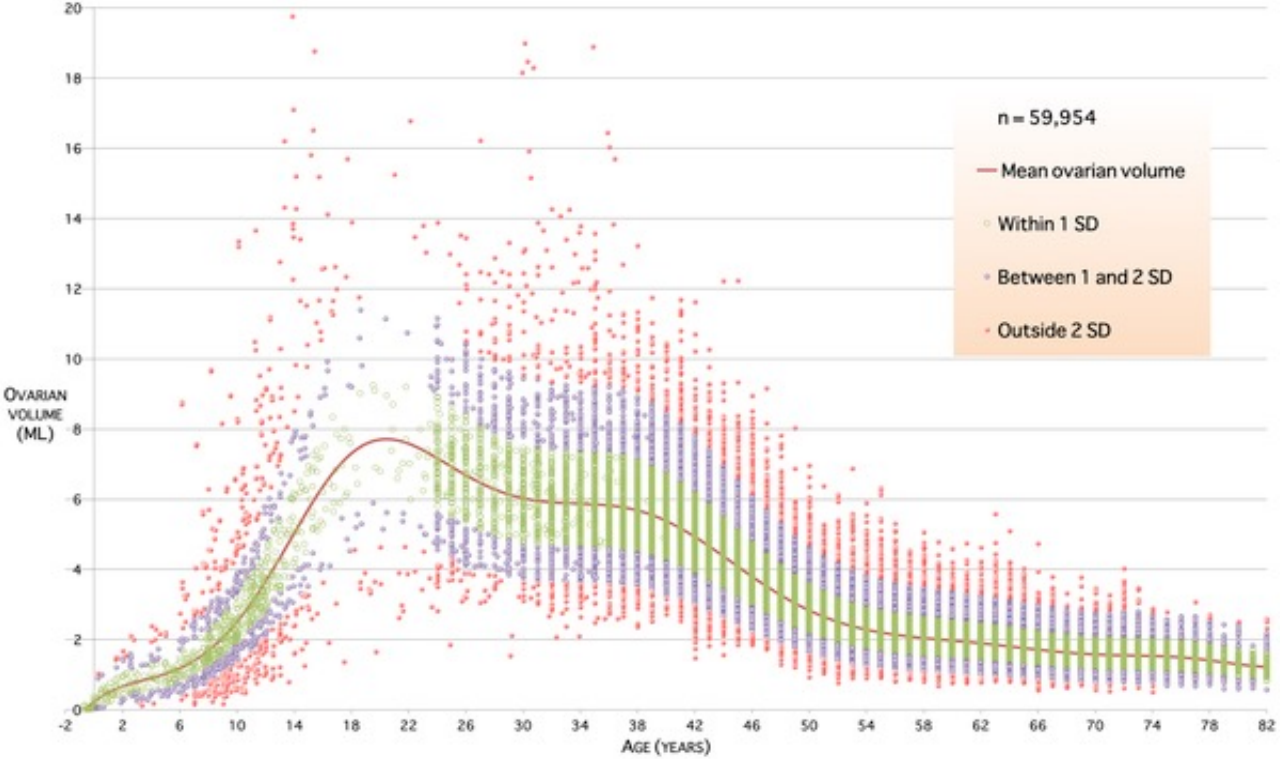
Fertility after Radiotherapy

- 8.9 year old patient (say)
- CSI plan for Ewing sarcoma treatment with calculated min, max and mean dose to the least affected ovary
- Revised radiosensitivity modelling using externally validated model of ovarian reserve and best current estimate of LD_{50}
- Estimate range of ages at premature ovarian insufficiency
- Useful for treatment planning
- Also an illustrative tool informing fertility preservation discussions



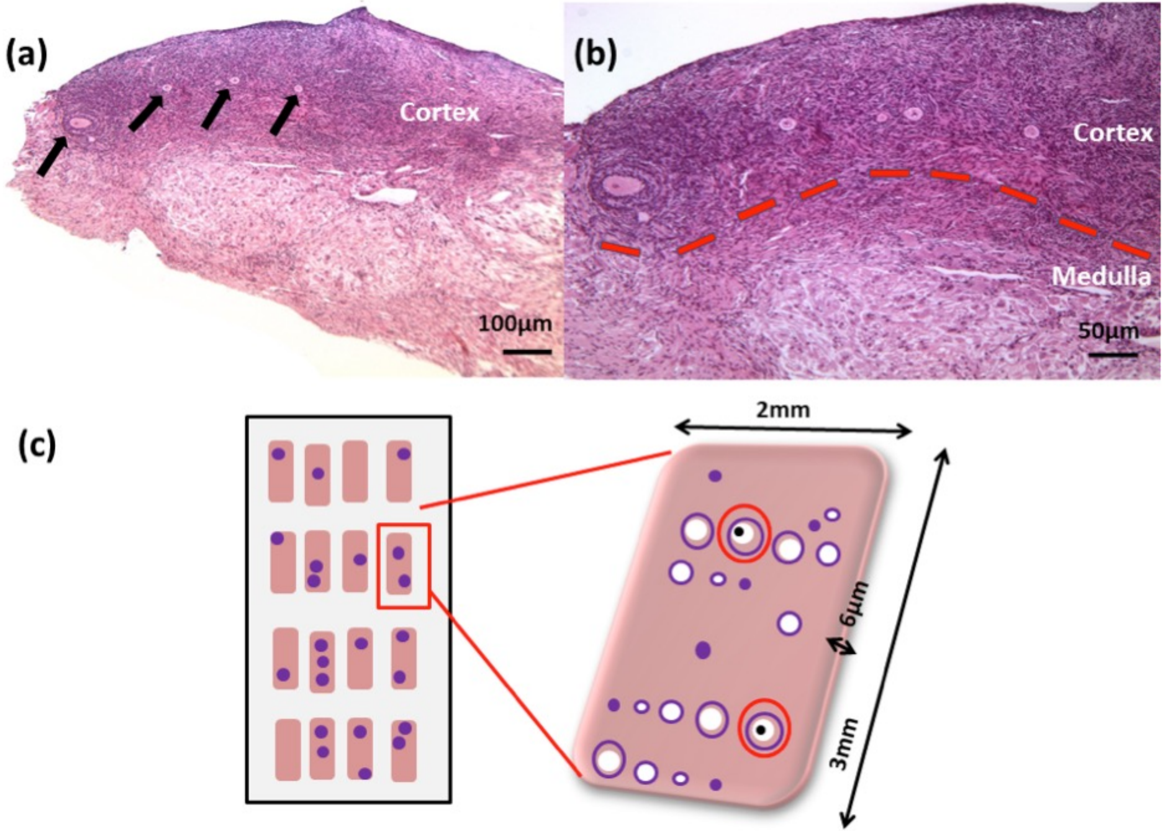
Kelsey *et al.*, 2022 PLOS ONE (under review)

Observational data model – ovarian volume



Kelsey TW et al. (2013) PLoS ONE 8(9): e71465

Follicle density



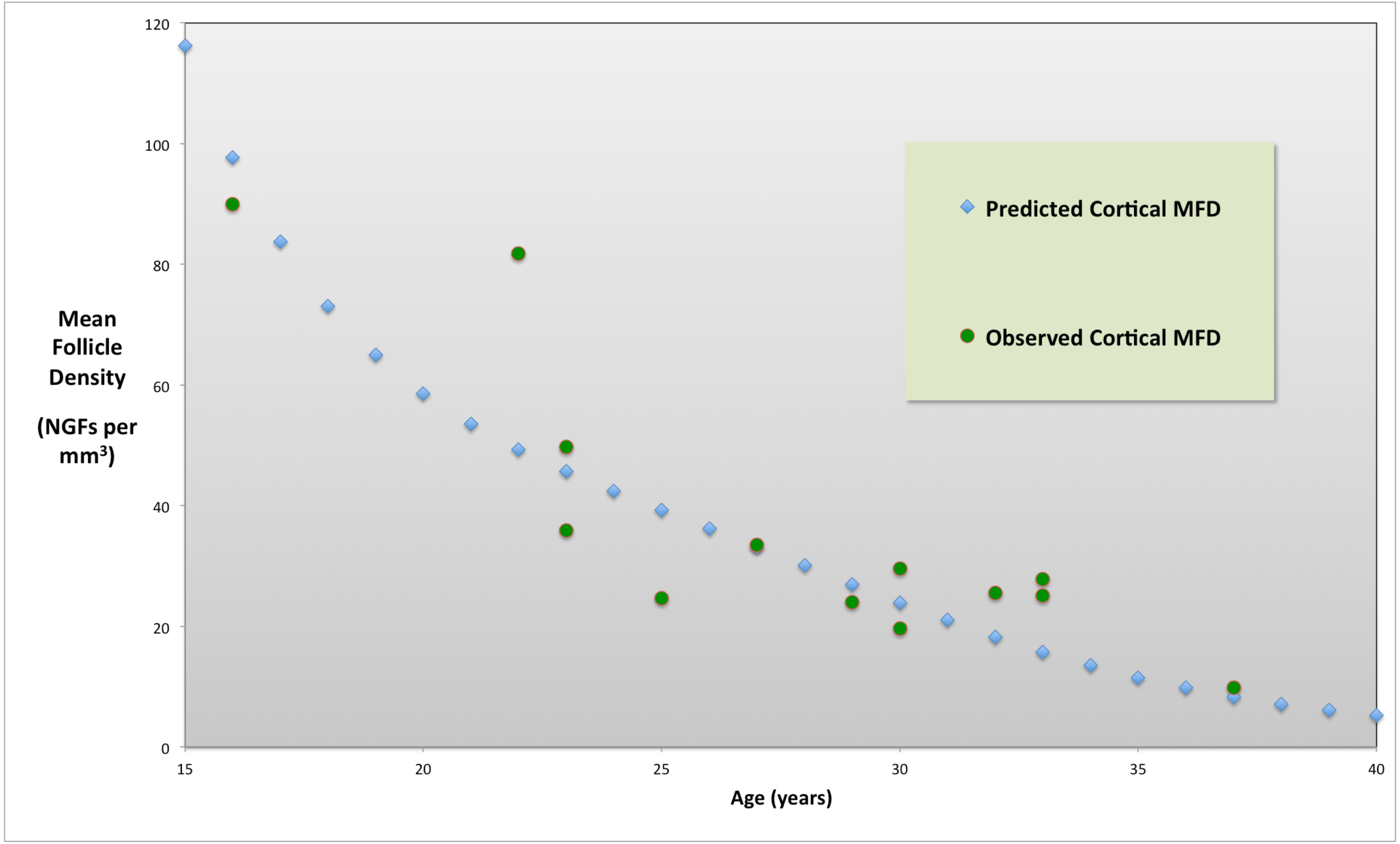
University of St Andrews

A model to predict Mean Follicle Density for healthy females aged 15 – 37 years

- Assumption: a large ovary contains more eggs than a small one
- NGFs: use the Wallace-Kelsey model to estimate NGF population, giving NGF(age)
- Volume: use our model to estimate the ovarian volume, giving Volume(age)
- Predicted MFD(age) is then NGF(age) divided by Volume(age)
- (We have to adjust for the proportion of a typical ovary that consists of cortical tissue)
- Simple arithmetic – no AI, no advanced statistics
- Sophisticated and modern techniques are not always required

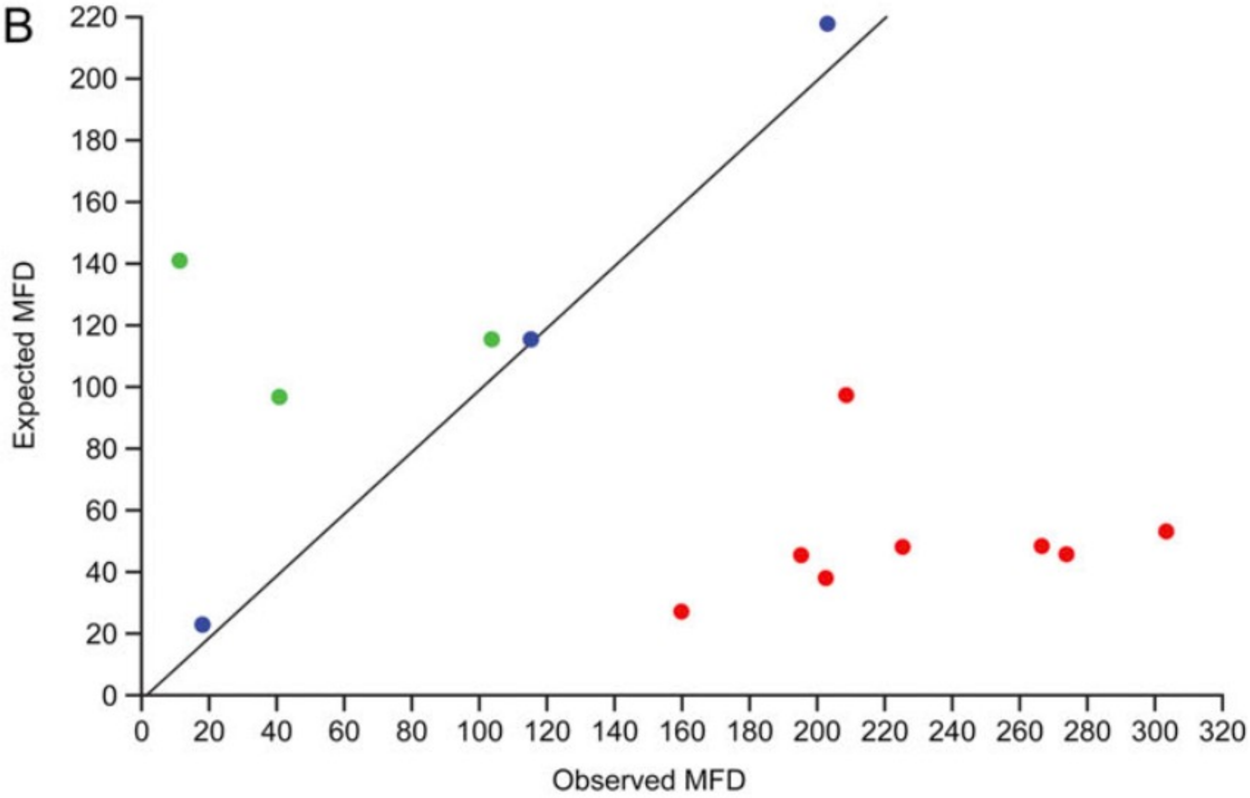
Fleming, Kelsey et al. Fertility & Sterility 2012

External Validation of NGF and OV Models



McGloughlin, Kelsey et al. Journal of Assisted Reproduction and Genetics **32; 2015**

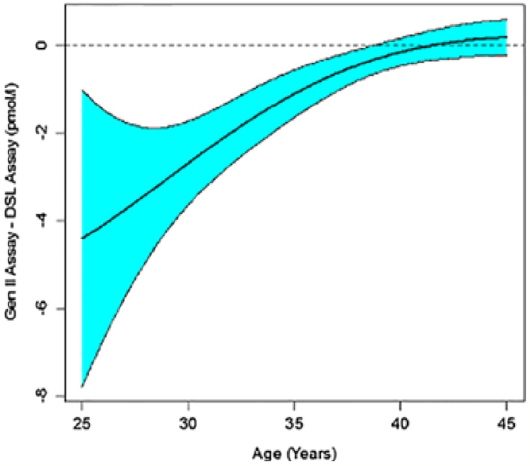
Using the models to assess NGF density after ABVD



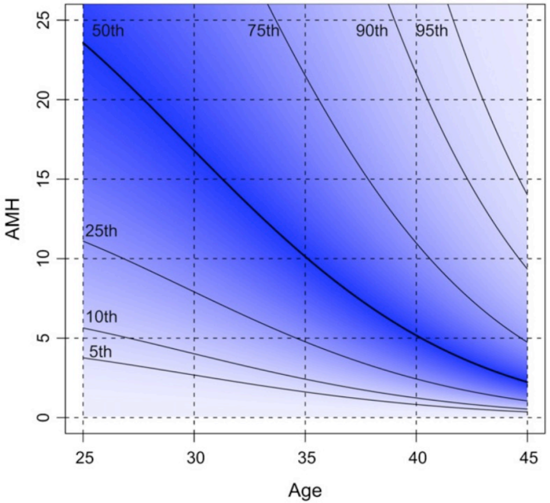
McLaughlin, Kelsey et al. Human Reproduction 32(1) 2016

University of
St Andrews

Observational data approaches have been critical to our understanding of AMH and its utility



Derived N = 5,492
Validated N = 5,492
Compared N = 9,601



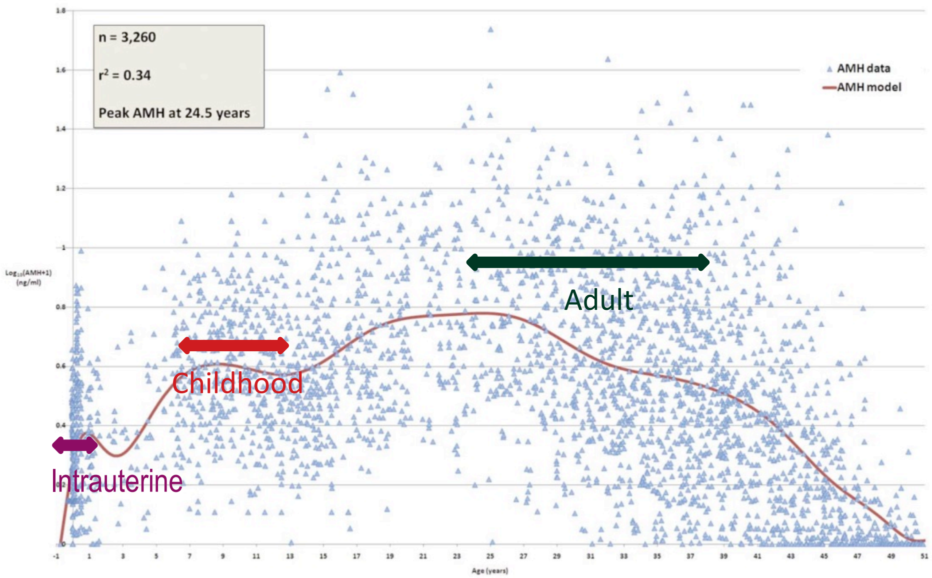
Derived = 9,601
Validated in N =15,834

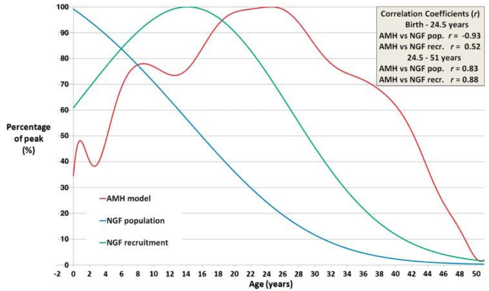
- AMH is a product preantral and small antral follicles in women
- As such, AMH is only present in the ovary until menopause
- **Can it be used as a biomarker for remaining ovarian reserve?**
- First studies are promising, but are based on infertile subjects

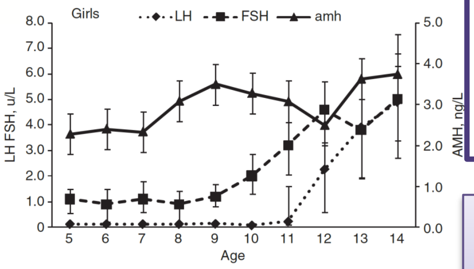
Nelson *et al* Fertil Steril 2011, Nelson *et al* RBM Online 2011

University of St Andrews

Observational data approaches have been critical to our understanding of AMH and its utility







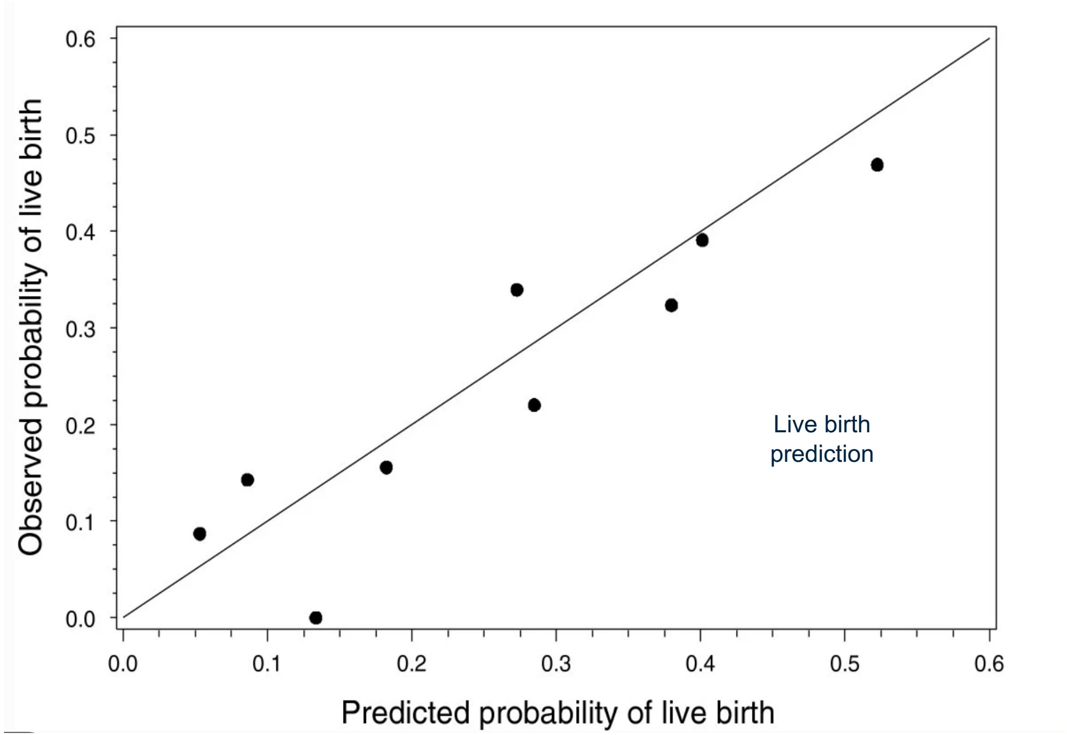
- AMH model from conception to menopause
- Validated for adult ages
- Validated for childhood/pubertal ages using 10-year longitudinal data

- AMH now accepted as biomarker

Kelsey et al PLOS One 2011, Kelsey et al Mol Hum Reprod 2012; Jeffery et al J Ped Endocrinol Metab 2015

University of St Andrews

Observational data approaches have been critical to our understanding of AMH and its utility



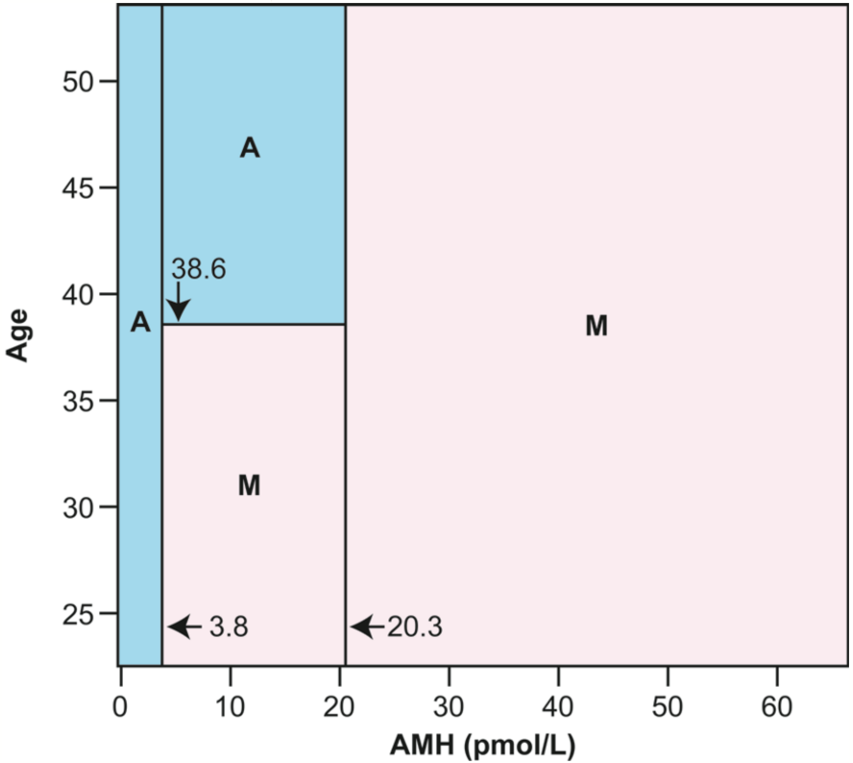
- "Our findings provide genetic support for the well-established use of AMH as a marker of ovarian reserve"
- AMH now routinely used as adjunct to the 2003 criteria for PCOS diagnosis
- Predicted live births based on AMH match observations

- AMH now used effectively as a biomarker
- Providing further validation of the underlying models

Perry et al,. Hum Mol Genet, 2016; Iliodromiti et al HRU 2014; Khader et al J Ovarian Res 2013

University of St Andrews

Using AMH to inform fertility preservation for survivors of cancer



Anderson et al 2013 Eur J Cancer; Anderson et al 2020 in prep.

- Pre-treatment AMH predicts for loss of ovarian function after chemotherapy for early breast cancer
- 6-month post-treatment AMH has high PPV for impaired fertility
- Pre- and post-chemo AMH combined with BMI, age, parity and endocrine factors has high diagnostic utility

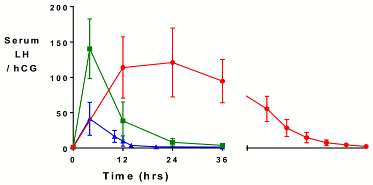
- We can optimise and personalise post-chemo endocrine therapy

University of St Andrews

AI as a strategy to improve endocrine therapy after breast cancer

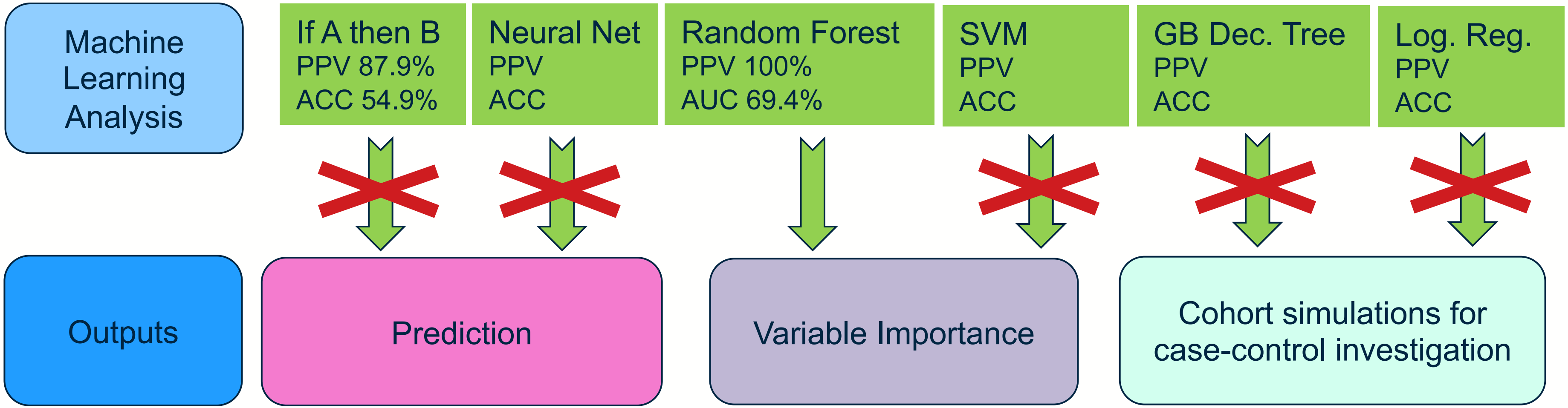
Inputs

Longitudinal endocrine data



Treatment data

Patient phenotype



Machine Learning Analysis

If A then B
PPV 87.9%
ACC 54.9%

Neural Net
PPV
ACC

Random Forest
PPV 100%
AUC 69.4%

SVM
PPV
ACC

GB Dec. Tree
PPV
ACC

Log. Reg.
PPV
ACC

Outputs

Prediction

Variable Importance

Cohort simulations for case-control investigation

Anderson et al Breast Cancer Research and Treatment **192, 2022**

University of
St Andrews

Where is the modern AI?

- Cohort studies
  - Regression models
  - Multivariable logistic regression
  - Dose-response models
- Oservational data studies
  - Life tables using Kaplan-Meier
  - PK style modelling using ODEs
  - Cox proportional hazards
  - Normative age-related models
- Meta analyses
  - Hierarchical summary ROC curves
  - Fixed & random effects meta-regression

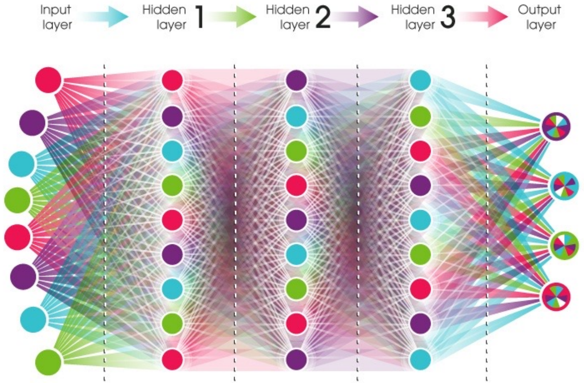
All of these are well-understood statistical and/or optimisation methods

AI methods are slowly and steadily being used to obtain publishable results and new insights

University of
St Andrews

Due process of AI studies still required

Deep neural network



- Publish in accordance with
- existing reporting standards

→

Clinical validation in real-world medicine

- Publish, RCTs showing benefit, Regulatory approval

→

Implementation in healthcare

- Cost of implementation

- how many workflows will be affected?
- Does the model increase the efficiency of existing workflows?
- Is the model being deployed within an existing digital workflow?

Slide taken from Scot Nelson, 2020

Nagendran et al BMJ 2020; Topol Nature Medicine 2019; Morse et al Nature Medicine 2020;

University of St Andrews

Conclusions

- Careful identification and analysis of biomedical data can lead to models
  - All of these models are wrong
  - Some of them are useful
- The key measure of utility is **external validation**
  - Predictions match observations for new and/or unseen data
- Many of the techniques used are old and well understood
- AI and machine learning techniques are becoming more prevalent, with notable improvements on existing knowledge
- But the specific method used is less important than validation

Colleagues

- Edinburgh
  - Hamish Wallace, Richard Anderson, Evelyn Telfer, …
- Copenhagen
  - Stine Gry Kristensen, Linn Mamsen, Claus Yding Andersen,…
- Imperial College
  - Ali Abbara, Waljit Dhillo,…
- Glasgow
  - Scott Nelson, Stamiatina Iliodromiti
- St Andrews
  - Gerry Humphris, Frank Sullivan,…

University of St Andrews | FOUNDED 1413

Thank you