

## Enabling longitudinal follow up of cancer outcomes in the UK Biobank – a pilot linkage to the National Cancer Data Repository

Brannan R<sup>1</sup>, Chapman M<sup>1</sup>, Thomas J<sup>1,2</sup>, Cooper N<sup>1</sup>, Møller H<sup>3</sup>

**Rachael Brannan,**  
**Research Officer, National Cancer Intelligence Network**

The National Cancer Intelligence Network would like to thank UK Biobank,  
Dr. Cathie Sudlow, Dr. Michael Lay and Dr. Alan Young



## Background to the project: UK Biobank

- Prospective cohort study following 500,000 adults aged 40-69 years old to enhance our understanding of the natural history and risk factors of diseases of middle age and later life.
- Extensive core data set includes demographic and self reported outcomes: questionnaire, physical measures & specimens
- Currently some 20TB of data (including specimen metadata)
- General consent obtained to follow-up through all health-related records & for unspecified health research

## Outcome adjudication in UK Biobank

- Need to confirm and characterise the health outcomes at baseline and during follow up
  - Who is your case?
  - What is the definition of the case?
- Insufficient quality and detail would limit the scientific value
- Need to avoid statistical power being reduced by misclassification of cases and controls
- UK Biobank has established a series of pilot projects to identify mechanisms for accurately identifying and phenotyping disease outcomes (Cancer, Stroke, Diabetes, CVD)

## Importance of cancer outcomes adjudication

- More than 1 in 3 affected by cancer
- Cancer will be among the first diseases to accrue useful numbers for scientific inquiry
  - Nationwide, population based registration in place since 1971
- Amenable to case ascertainment with expert standardised classifications (including site & morphology)
- NCDR 1985-2009 contains **8,398,222** patient records

### Expected incident cancer cases in UK Biobank

	2012	2017	2022
Breast	2,500	6,000	10,000
Colorectal	1,500	3,500	7,000
Prostate	1,500	3,500	7,000
Lung	500	2,000	4,000

*Breast cancers will be the fourth largest group of diseases among incident cases*

## Establishing mechanisms for identifying and phenotyping cancer outcomes – Pilot Objectives



1. Demonstrate linkage of a subset of the UK Biobank cohort to the National Cancer Data Repository for England and assess the quality of this link.
2. Identify prevalent and incident cases since baseline visit.
3. Classify cases by the certainty of diagnosis.
4. Define an initial dataset for cancer cases in UK
5. Identify the hospital trusts treating cancer patients in the selected cohort and assess obtaining additional clinical data, diagnostic samples and imaging information.

## Test linkage for adjudication of cancer outcomes in the UK Biobank cohort using the National Cancer Data Repository (1985-2009)



Trusted 3<sup>rd</sup> party linkage to establish co-referent pairs in two databases using exact matching

- UK Biobank English residents (466,850 unique individuals)
- NCDR 1985-2009 (8,398,222 patient records)

Linkage was conducted using 3 levels;

- NHS number only
- NHS Number/DOB/Sex (NDS)
- NHS Number/DOB/Postcode/Sex (NDPS).

UK Biobank data quality was very good with data available for:

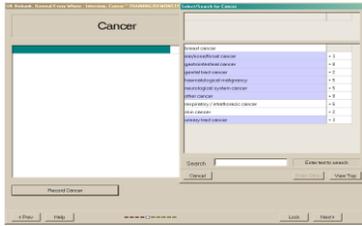
NHS Numbers	99.99%
Postcodes	99.79%

- Matching by (DOB/PCDR/Sex) where NHS Number is not available:
  - 31 Participants with zero matching

## Case ascertainment in UK Biobank – record of cancer diagnosis in the National Cancer Data Repository 1985-2009

### Prevalent cases in UK Biobank

- **41,712** participants self-reported diagnosis of a cancer



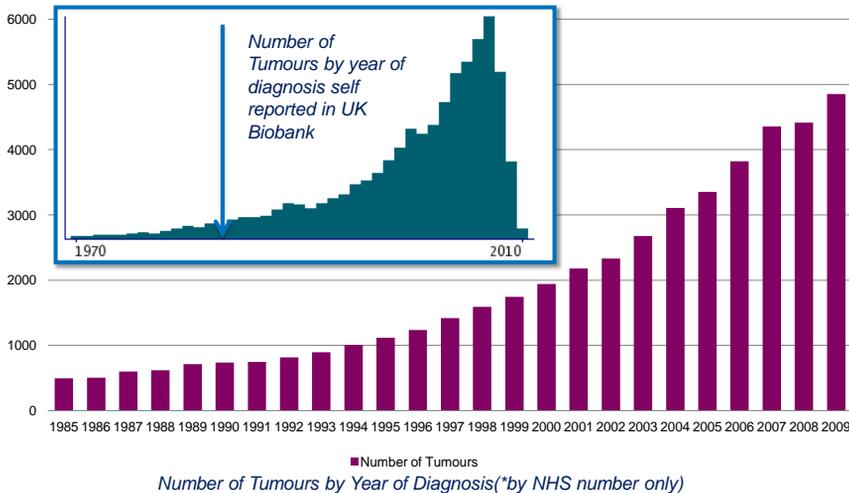
### Cancer cases recorded in the NCDR (1985-2009)

- **41,957 participants** identified as having an exact match in the NCDR (NHS Number only)
- **47,380 tumours** recorded
- NHS #/DOB/Sex identifiers
  - Exact matching decreased to 41,435
  - Individual tumours fell to 46,785

#### NCDR: Pre and Post 13-03-2006 Breakdown

Incident	16824
Prevalent	30556

## Cancer outcomes in UK Biobank – record of cancer diagnosis in the National Cancer Data Repository (1985-2009)



Site	Self reports in UK Biobank baseline questionnaire	Records in the NCDR (1985-2009)
Breast	11382	10939*
Colorectal	1590	1493**
Prostate	3467	3608
Lung	441	514

\*ICD 10 C50& D05  
 \*\* ICD 10 C18 & D12

•Includes ICD-10 codes C00-D48, O01 and Q85

•Self reports: this methodology susceptible to recall, ascertainment, and social-desirability biases.

•Participants may/may not have recalled in situ or benign tumours

•Healthy cohort effect?

## Conclusion

- Cancer case ascertainment is readily established in the UK Biobank using exact matching by NHS Number to the National Cancer Data Repository
- Linkage to the NCDR may also be used to phenotype cancers – this will be the next phase of our work
- What's next?

**Phase 1: NCIN has successfully demonstrated the usefulness of English cancer registration data in cancer case ascertainment using a pilot linkage of all English resident sample of UK Biobank participants**

**Phase 2: Identify and phenotype prevalent and incident cases and work with the research community to define the dataset for confirmed cancer cases from the National Cancer Data Repository plus Scottish and Welsh cancer registries**

**Phase 3: Obtain additional information plus tissue samples and images from treating trusts as required**

## What next?



### Cancer Outcomes in UK Biobank Workshop

Friday 6<sup>th</sup> July , 10am-4pm

Wellcome Trust, Euston Road

Registration Open - contact  
[rachael.brannan@ncin.org.uk](mailto:rachael.brannan@ncin.org.uk)



[www.ncin.org.uk](http://www.ncin.org.uk)