Cohort Profile: Extended Cohort for E-health, Environment and DNA (EXCEED)

Author List

Catherine John[1], Nicola F. Reeve[1*], Robert C. Free[2], Alexander T. Williams[1], Aliki-Eleni Farmaki[1,3], Jane Bethea[1], Nick Shrine[1], Chiara Batini[1], Richard Packer[1], Sarah Terry[2], Beverley Hargadon[2], Qingning Wang[1], Carl Melbourne[1], Kyla Harrington[1], Nigel J. Brunskill[4], Christopher E. Brightling[2], Julian Barwell[5], Susan E. Wallace[1], Ron Hsu[1], David J. Shepherd[1], Edward J. Hollox[5], Louise V. Wain[1], Martin D. Tobin[1]

[1] Department of Health Sciences, University of Leicester, Leicester, LE1 7RH, UK

[2] Leicester NIHR BRC, Institute for Lung Health, Department of Infection, Immunity & Inflammation, University of Leicester, Leicester, LE1 7RH, UK

[3] Department of Population Science and Experimental Medicine, Institute of Cardiovascular Science, University College London, London, UK

[4] Infection, Immunity & Inflammation, University of Leicester, Leicester, LE1 7RH, UK

[5] Department of Genetics and Genome Biology, University of Leicester, Leicester, United Kingdom

Why was the cohort set up?

EXCEED aims to develop understanding of the genetic, environmental and lifestyle-related causes of health and disease. Cohorts of this kind, with broad consent to study multiple phenotypes related to onset and progression of disease and drug response have a role to play in medicines development, by providing genetic evidence that can identify, support or refute putative drug efficacy or identify possible adverse effects [1]. Furthermore, such cohorts are well suited to the study of multimorbidity – another key aim of EXCEED.

Multimorbidity describes the presence of multiple diseases or conditions in one patient, though definitions in the literature vary widely [2-4]. It demands a holistic approach to optimise care and avoid iatrogenic complications, such as drug interactions. In the context of increasing specialisation of many healthcare systems and high healthcare utilisation amongst people with multimorbidity, providing such care poses a complex challenge [5-7]. In high-income countries multimorbidity is particularly common amongst more deprived socioeconomic groups and may even be considered as the norm amongst older people [8, 9], whilst an ageing global population and a growing burden of non-communicable diseases in low- and middle-income countries emphasise its global importance [10]. An expert working group convened by the UK Academy of Medical Sciences recently highlighted the lack of available evidence relating to the burden, determinants, prevention and treatment of multimorbidity, and recommended the prioritisation of research on multimorbidity spanning the translational pathway from understanding of its biological mechanisms to health services research [11].

Studies designed to investigate multimorbidity, rather than considering individual conditions in relative isolation, are therefore vital [6, 7]. Linkage to electronic health records (EHR) has enabled information on a broad range of diseases and risk factors to be studied in EXCEED and places multimorbidity at the study's heart. The EHR linkage also facilitates longitudinal follow-up over an extended period, enabling, for example, the investigation of lifestyle factors and other exposures on healthy ageing and outcomes in later life.

Combining wide-ranging data from EHR with genome-wide genotyping is also central to EXCEED's purpose. In recent years, our understanding of which genes are associated with both rare and common diseases has advanced rapidly as available sample sizes for genome-wide association studies (GWAS) have grown rapidly [12]. For example, there are now over 140 genetic variants associated with lung function and COPD [13-24]. However, in many cases, our understanding of the mechanisms through which these variants influence disease risk – and which could therefore be therapeutic targets – is relatively limited. An efficient design to inform this understanding is to stratify participants based on available study data on their health status (phenotype) or genetic risk factors (genotype) to recall them for further detailed investigations which would be impracticable across a whole cohort. EXCEED was purposely designed as a resource for recall-by-genotype sub-studies and all participants have consented to be recalled on this basis.

The study is led by the University of Leicester, in partnership with University Hospitals of Leicester NHS Trust and in collaboration with Leicestershire Partnership NHS Trust, local general practices and smoking cessation services.

Who is in the cohort?

Recruitment to the cohort to date has taken place primarily from the general population through local general practices in Leicester City, Leicestershire and Rutland, with 9,840 participants recruited to date. 441 participants were recruited through smoking cessation services in Leicester City, Leicestershire and Rutland, and a further 44 through targeted recruitment of those with a recorded diagnosis of COPD in their electronic primary care record. All tables and figures present participants whose data was collected and quality control undertaken at 03/01/2018 (8,993 participants).

In the UK, over 98% of the population is registered with a National Health Service (NHS) general practitioner[25]. For recruitment through primary care, all registered patients aged between 40 and 69 years in participating general practices were eligible for recruitment. Exclusion criteria were minimal: those receiving palliative care, those with learning disabilities or dementia and those whose records indicated they had declined consent for record sharing for research. All eligible patients identified through primary care were sent an initial letter with brief information about the study and a reply slip to indicate their interest.

For participants recruited via smoking cessation services, the lower age limit was reduced to 30 years because of the higher risk of respiratory disease amongst smokers. Initial eligibility screening and information provision was either undertaken through electronic client records followed by a letter to the client (as in primary care) or face-to-face by a smoking cessation advisor during a routine appointment. Additionally, patients with a recorded diagnosis of COPD were invited from four local general practices with higher prevalence of COPD, to boost the numbers available for study of respiratory disease. For this group, the lower age limit was 30 years, and all other exclusion criteria were identical to the main primary care recruitment.

All those who responded to indicate they were interested in taking part were sent full written information on the study. They then participated via one of two routes depending on their location and personal preference: a face-to-face appointment with a research professional, or by post. The flow of participants through the main primary care recruitment route is illustrated in Figure 1. Approximately 8% of those who received an initial invite via primary care completed recruitment.

Table 1 gives an overview of the demographic characteristics of the primary care population sampled, compared with the characteristics of those recruited to the study via primary care. Table 1 shows that participants in the study were older and more likely to be female than the wider primary care population from which they were drawn. This reflects well-known patterns of participation in similar cohorts [26, 27]. The local primary care population includes a large proportion of minority ethnic groups, especially Asian and Asian British. Notably, these groups are under-represented among study participants, although the proportion of study participants of Asian and Asian British ethnicity (5%) is higher than many UK cohorts, including UK Biobank [27]. This under-representation of minority ethnic groups is a well-known phenomenon for which explanations include language barriers, inequitable access to healthcare services, cultural sensitivities and a lack of awareness of medical research and its purpose [28, 29].

Figure 1 Flowchart of recruitment steps via primary care

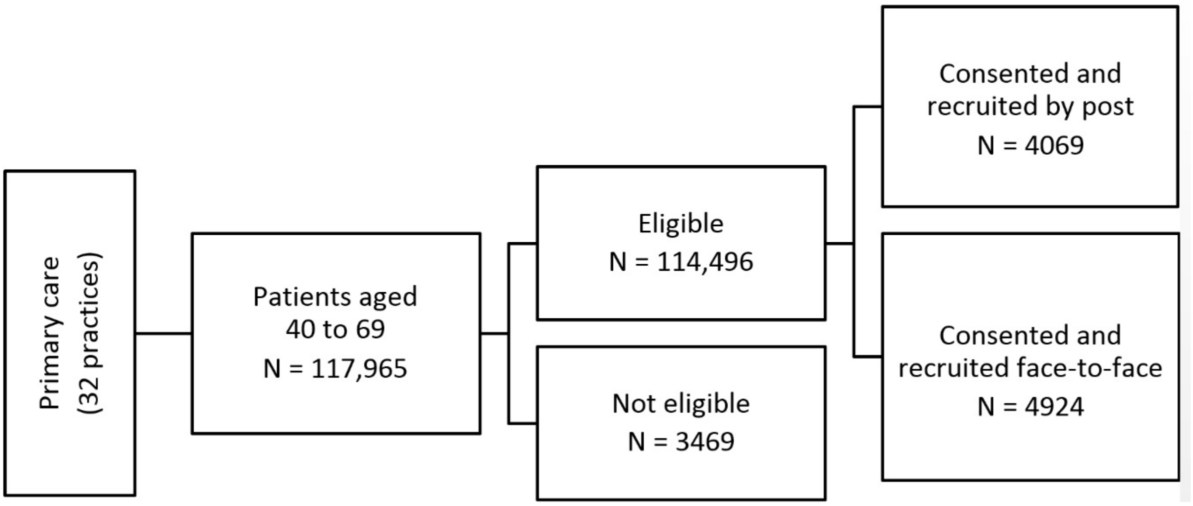


Table 1 Demographic characteristics of the primary care population sampled for the study and those who participated (via the primary care recruitment route only)

	Primary care population*		**Recruited**	
Age	**n (N = 117, 965)**	**%**	**n (N = 8576)**	**%**
(<45)	21057	17.9	628	7.3
(45 – 54)	44559	37.8	2163	25.2
(55 – 64)	36133	30.6	3200	37.3
(≥ 65)	16216	13.7	2585	30.1
Sex	**n (N = 117,965)**	**%**	**n (N = 8576)**	**%**
Male	59003	50.0	3840	44.8
Female	58962	50.0	4736	55.2
Ethnicity	**n (N = 81,947)**	**%**	**n (N = 8541)**	**%**
White	59576	72.7	7903	92.5
Asian/Asian British	17670	21.6	423	5.0
Black/African/Caribbean/Black British	2763	3.4	12	0.1
Mixed	686	0.8	93	1.1
Chinese	301	0.4	53	0.6
Other	951	1.2	61	0.7

*Primary care population is all patients within the eligible age range in the practices sampled, and includes those who were excluded at the next step (codes for palliative care, dementia, learning disability, or lack of consent to share data for research).

How often have they been followed up?

Participants have consented to follow-up through linkage to EHR and other health care records for up to 25 years. Linkage to electronic primary care records (i.e. records from the participant's general practice) has been completed for 8,442 participants.

As participants are prospectively followed up we expect losses due to deaths (to date less than 1% of participants), withdrawals (to date less than 0.1% of participants), relatively few losses due to house moves within the UK or changing general practitioner as NHS patients retain the same NHS number and their electronic records move with them, and some losses due to emigration. Analyses of historical healthcare records to track disease development and progression may be subject to selection bias, in particular survivor bias.

What has been measured?

There are several phases of data collection, summarised in Table 2. Linked primary care data provides historic cohort data. Since the mid-1990s, prospectively recorded consultations enable the retrieval of information not only on symptoms for which participants have visited their general practitioner and diagnoses which have been made, but also on examination findings (including blood pressure readings and spirometry results, for example), laboratory test results, drug prescriptions and secondary care referrals. Major diagnoses recorded on paper records prior to the mid-1990s were retrospectively coded at the time of computerisation and so can also be retrieved.

Baseline data collection for all participants included a self-completion questionnaire which collected detailed information on current and past smoking habits, smoking cessation attempts, e-cigarette and shisha usage, environmental tobacco smoke (second-hand smoke) exposure and alcohol use. For those recruited via a face-to-face appointment, this was undertaken during the appointment. Those participating by post completed the questionnaire online using their own computer, with a paper version available if necessary. Height, weight and waist circumference were either measured by a research professional or self-reported by postal participants. Those recruited face-to-face also had their hip circumference measured and underwent spirometric measurement of lung function.

Finally, a DNA saliva sample was collected from all participants either at their appointment or returned by post. The samples are stored at the NIHR Biocentre (Milton Keynes, UK), providing industrial-scale laboratory information management and automated robotic systems which have been shown to facilitate efficient error-free sample storage and extraction from freezers in the UK Biobank study [30]. To date, genome-wide genotyping has been undertaken for 6,178 samples using the Affymetrix UK Biobank Axiom Array, enabling analysis of over 40 million variants after imputation [31]. Planned quarterly updates to linked primary care records enables longitudinal tracking of health. There is also ongoing linkage to other sources of health data including Admissions, Accident and Emergency attendances and Outpatient appointments via Hospital Episode Statistics; Pathology Data (East Midlands Pathology Service), and the Myocardial Ischaemia national Audit (MINAP).

Table 2 Summary of data collected at each phase

Phase	Measurements
Historic cohort data	Historically coded data (transferred from paper records at the time of practice computerisation, approximately mid-1990s) and since mid-1990s prospectively recorded consultations, with coded: • symptoms; • diagnoses; • measurements, such as blood pressure and spirometry • laboratory test results • drug prescriptions; • secondary care referrals
Baseline	All participants: Questionnaire, including smoking and alcohol use. DNA saliva sample Examination by research professional only: height, weight, waist circumference, hip circumference and spirometry Postal participants only: self-measured anthropometry and omitted spirometry
Ongoing	Planned quarterly updates to primary care record linkage (detailed above), with consent to follow-up for 25 years, to track health longitudinally Ongoing linkage to: • Admissions, Accident and Emergency attendances and Outpatient appointments via Hospital Episode Statistics • Pathology Data (East Midlands Pathology Service) • Myocardial Ischaemia National Audit (MINAP)

What has it found? Key findings and publications

Table 3 shows that, in general, our cohort is slightly healthier than average for common health risk factors and behaviours. This is similar to findings by other cohort studies [27]. For example, the total proportion of participants who were overweight or obese (64.4%) was slightly lower than similar age groups in Health Survey for England 2016, where it was above 70% for all ages from 45 upwards [32]. Alcohol intake for our cohort is comparable to that of similar age groups in Health Survey for England 2016 [33]. Only 25.9% of participants are in the two most deprived national quintiles and 29.2% are in the least deprived quintile. For Leicester City, 75.9% of the population are in the two most deprived quintiles and only 1.4% are in the least deprived quintile [34]. Though this reflects the whole Leicester population, not just those aged 40-69 and registered with the GP practices that agreed to take part in EXCEED, it indicates that the most deprived communities are under-represented in the cohort.

Similarly, the proportion of EXCEED participants who currently smoke is 10.1%, considerably lower than the national average (15.8%) and comparable only to the oldest age group (65 and over) in the national Annual Population Survey, amongst whom smoking prevalence was 8.3%. Smoking prevalence amongst all younger age groups nationally is 15% or above. On the other hand, the proportion of people who report never smoking is also lower than in national population surveys. This may be influenced by question wording and interpretation: while the relevant national survey asked if people had ever "regularly" smoked, the EXCEED questionnaire included occasional use in the definition of ever smokers [35]. Table 4 presents more detailed information on smoking habits amongst current and ex-smokers. The vast majority of both reported smoking cigarettes, but cigar/cigarillo and pipe smoking was less common amongst current than ex-smokers.

The Quality and Outcomes Framework (QOF), introduced in 2004, aims to improve the quality of care patients are given by rewarding practices for the quality of care they provide. Prevalence of 16 chronic conditions prioritised for management in primary care by the QOF is presented in Table 5, and the number of conditions per individual is summarised in Table 6. We found that, overall, 25.6% of our participants had a recorded diagnostic code for more than one QOF condition. This is in line with findings from a large study of almost 100,000 individuals in the Clinical Practice Research Database by Salisbury and colleagues, who used a similar approach to define multimorbidity [5]. They found that 16% of their population had a code for more than one QOF condition, but this rose sharply with age, reaching around 20% amongst 55- to 64-year-olds and over 30% in 65- to 74-year-olds. Two further large UK-based studies have used more comprehensive lists of conditions to define multimorbidity, but limited to active morbidity only, and found prevalence of multimorbidity between 23.2% and 27.2% across all ages, rising substantially with age to 50% or more amongst 65- to 74-year-olds [8, 36].

We specifically examined primary care diagnoses of one condition, COPD, for which we had independent diagnostic information from baseline spirometry. Diagnosis of COPD defined by presence of COPD code in primary care data compared with COPD defined by baseline spirometry results indicates underdiagnosis of COPD in primary care records in EXCEED was higher than in previous reports [37-39]: 86.3% of GOLD stage 1-4 COPD and 74.1% of GOLD stage 2-4 was undiagnosed (Table 7).

Table 8 provides an example of some of the most common measures available in the primary care data, and the numbers of participants with more than one, two and three recordings of these measures. Table 9 shows the average values of these measures. This demonstrates the utility of EXCEED for enabling cross-sectional or longitudinal studies of quantitative traits. For example, 91.1% of participants have more than three recordings of blood pressure and 64.8% have at least one quantitative blood pressure value recorded. Mean systolic blood pressure was 129.5 (sd 13.6) and mean diastolic blood pressure was 78.1 (sd 8.6) (Table 9).

Table 3 Prevalence of risk factors and health behaviours

	n	**%**
Deprivation* (N = 8782)		
1 (most deprived)	1182	13.5
2	1091	12.4
3	1613	18.4
4	2332	26.6
5 (least deprived)	2564	29.2
BMI (N = 8897)		
Underweight (<18.5)	93	1.0
Normal (18.5 – 24.9)	3080	34.6
Overweight (25 – 29.9)	3421	38.5
Obese (30 – 39.9)	2013	22.6
Morbidly obese (≥40)	290	3.3
Waist circumference (N = 8726)		
Low risk (males<94cm, females<80cm)	2852	32.7
Increased risk (males 94-102cm; females 80-88cm)	2330	26.7
High risk (males>102cm; females>88cm)	3544	40.6
Smoking status*^{2} (N = 8990)		
Current smoker	904	10.1
Ex-smoker (regular or occasional)	3516	39.1
Never smoker	4570	50.8
Alcohol intake (units/week) (N = 8955)		
None	1721	19.2
Lower risk (females<14u; males<21u)	4912	54.9
Increasing risk (females 14-35u; males 21-50u)	1735	19.4
Higher risk (females>35u; males>50u)	587	6.6

*Index of multiple deprivation national quintiles by postcode

*^{2}Includes cigarettes, cigars, cigarillos, pipes or shisha

Table 4 Smoking history (self-reported by current or ex-smokers)

	Current smokers		**Ex-smokers**	
	n	**%**	**n**	**%**
Type of tobacco used*[1]	**(N = 895)**		**(N = 3507)**	
Cigarettes*[2]	848	93.8	3439	97.8
Shisha	1	0.1	6	0.2
Cigars/cigarillos	44	4.9	243	6.9
Pipe	15	1.7	145	4.1
Other	12	1.3	2	0.1
Use of electronic cigarettes	**(N = 901)**		**(N = 3513)**	
Ever	234	26.0	202	5.8
Never	667	74.0	3311	94.2
Smoking cessation aids used (ever)*[3]	**(N = 374)**		**(N = 3476)**	
NRT	117	31.3	456	13.0
Bupropion	5	1.3	39	1.1
Varenicline	55	14.7	225	6.4
Other	47	12.6	296	8.4
None	189	50.5	2519	71.6
	Mean	**SD**	**Mean**	**SD**
Pack-years of smoking*[4]	**(N = 519)**		**(N = 3026)**	
	27.4	19.2	18.4	20.4
Cigarettes per day	**(N = 561)**		**(N = 3048)**	
	13.6	8.7	14.7	11.8
Age at smoking initiation (years)	**(N = 770)**		**(N = 3484)**	
	18.4	5.7	17.1	3.8

*[1] People may use more than one type of tobacco, so percentages will not add up to 100.

*[2] Filtered, unfiltered and handrolled.

*[3] only for quit attempts lasting at least 6 months. Denominator for percentages is current smokers who have made a quit attempt lasting at least 6 months, or total number of ex-smokers. People may have used more than one aid, so percentages will not add up to 100.

*[4] only for cigarette smokers

Table 5 Prevalence of chronic conditions

Condition	n*	%
Atrial Fibrillation	204	2.4
Asthma	1035	12.3
Cancer	513	6.1
Coronary Heart Disease	351	4.2
Chronic Kidney Disease (3-5)	235	2.8
Chronic Obstructive Pulmonary Disease	262	3.1
Depression	1841	21.8
Diabetes	730	8.6
Epilepsy	92	1.1
Heart Failure	76	0.9
Hypertension	2280	27.0
Mental Health (psychosis, schizophrenia and bipolar affective disorder)	62	0.7
Osteoarthritis	234	2.8
Peripheral Arterial Disease	44	0.5
Rheumatoid Arthritis	105	1.2
Stroke	95	1.1

* Number of participants with one occurrence at any time of a diagnostic code listed in the Quality and Outcomes Framework for that condition. % is out of all participants for whom primary care data was available (8442).

Table 6 Proportion of participants with multiple chronic conditions[1]*

Number of chronic conditions	n	%*[2]
1	2781	32.9
2	1402	16.6
3	519	6.1
4	184	2.2
5	42	0.5
6 or more	11	0.1

*[1] 16 chronic conditions prioritised for management in primary care by the Quality and Outcomes Framework (see Table 5)
*[2] of participants with primary care data

Table 7 Comparison of diagnosis of COPD, using COPD codes in primary care data vs COPD defined by baseline spirometry

		COPD defined by baseline spirometry using GOLD criteria							
		GOLD 1 - 4				GOLD 2 - 4			
		Yes		No		Yes		No	
		n	%	n	%	n	%	n	%
COPD code in primary care	Yes	73	13.7	14	0.6	66	25.9	21	0.8
	No	461	86.3	2452	99.4	189	74.1	2724	99.2

*for participants with linked primary care data and baseline spirometry measures (n=3,000)
*all percentages are column percentages

Table 8 Numbers of participants with more than one, two or three recordings of selected measures

	>1 record	%	>2 records	%	>3 records	%
O/E - blood pressure reading	8267	97.9	8007	94.8	7693	91.1
Serum creatinine	6720	79.6	5668	67.1	4797	56.8
Serum sodium	6698	79.3	5641	66.8	4778	56.6
Serum potassium	6676	79.1	5609	66.4	4752	56.3
Serum urea level	6671	79.0	5607	66.4	4739	56.1
eGFR*	6395	75.8	5255	62.2	4372	51.8
Serum triglyceride levels	6200	73.4	4862	57.6	3953	46.8
Serum cholesterol level	6170	73.1	4862	57.6	4001	47.4
Platelet count	6156	72.9	5039	59.7	4085	48.4
Serum HDL cholesterol level	5978	70.8	4585	54.3	3690	43.7
Serum LDL cholesterol level	5609	66.4	4224	50.0	3382	40.1
Serum bilirubin level	5163	61.2	4070	48.2	3211	38.0
Haemoglobin A1c level	4158	49.3	2843	33.7	2015	23.9
Total white blood count	6031	71.4	4855	57.5	3931	46.6
Eosinophil count	6092	72.2	4930	58.4	3965	47.0

* Glomerular filtration rate calculated by abbreviated Modification of Diet in Renal Disease Study Group calculation [40]

Table 9 Summary of values for selected measures

Term	**n*¹**	**%**	**Mean**	**sd**
O/E - blood pressure reading Systolic (mmHg)	5468	64.8	129.5	13.6
O/E - blood pressure reading Diastolic (mmHg)	5468	64.8	78.1	8.6
Serum creatinine level (umol/L)	7876	93.3	73.6	19.4
Serum sodium level (mmol/L)	7871	93.2	140.3	2.2
Serum potassium level (mmol/L)	7820	92.6	4.4	0.4
Serum urea level (mmol/L)	7861	93.1	5.6	1.5
eGFR*² (mL/min/1.73m^2)	7620	90.3	82.7	10.5
Serum triglyceride levels (mmol/L)	7856	93.1	1.5	0.8
Serum cholesterol level (mmol/L)	7597	90.0	5.2	1.0
Platelet count - observation (x10^9/L)	7437	88.1	253	63.9
Serum HDL cholesterol level (mmol/L)	7776	92.1	1.6	0.5
Serum LDL cholesterol level (mmol/L)	7480	88.6	2.96	0.9
Serum bilirubin level (umol/L)	6488	76.9	10.5	5.5
Haemoglobin A1c level (%)	5806	68.8	5.7	0.7
	n*¹	**%**	**Median**	**IQR**
Total white blood count (x10^9/L)	7345	87.0	6.2	5.2-7.4
Eosinophil count - observation (x10^9/L)	7417	87.9	0.16	0.10 - 0.24

* Where participants have more than one recording of a measure, the most recent value for each participant was used.

*¹Number of participants for whom values were available.

*² Glomerular filtration rate calculated by abbreviated Modification of Diet in Renal Disease Study Group calculation [40]

Recall-by-phenotype study

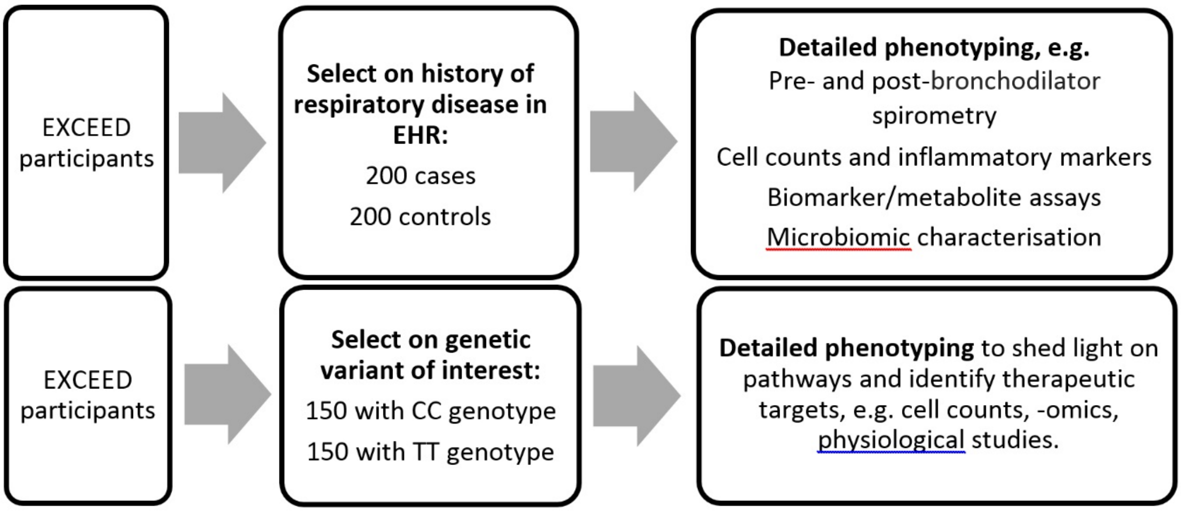
Recalling by phenotype facilitates in-depth study of disease mechanisms, with a reduced risk of bias as with nested case-control studies [41]. One such study has recalled EXCEED participants to take part in a study examining the microbiome in COPD cases and in smoking and non-smoking controls.

Potential for recall-by-genotype studies

Future recall-by-genotype studies are expected to contribute to a deeper understanding of genetic variants which may be potential therapeutic targets, by bringing back participants for detailed assessments on the basis of the known or suspected mechanism of the relevant gene. Such recall-by-genotype sub-studies may investigate disease susceptibility, disease progression or drug response and whilst they could be interventional in design, most will be observational studies [42]. Observational studies of this kind can provide evidence which is not susceptible to reverse causation and to confounding by lifestyle factors given Mendelian randomisation [43].

Nested designs are also feasible which do not rely on recall of participants but which could be undertaken quickly and inexpensively using stored biological samples and linked electronic data, and such sub-studies could select samples based on either phenotype or genotype. Small-scale intervention-by-genotype studies could, for example, evaluate response to a treatment with a known safety profile in participants with a specific genetic variant.

Figure 2 Examples of potential recall-by-phenotype (top) and recall-by-genotype studies (bottom)



EXCEED participants

Select on history of respiratory disease in EHR:

200 cases

200 controls

Detailed phenotyping, e.g.

Pre- and post-bronchodilator spirometry

Cell counts and inflammatory markers

Biomarker/metabolite assays

Microbiomic characterisation

EXCEED participants

Select on genetic variant of interest:

150 with CC genotype

150 with TT genotype

Detailed phenotyping to shed light on pathways and identify therapeutic targets, e.g. cell counts, -omics, physiological studies.

What are the main strengths and weaknesses?

Linkage to EHR and other health care records is a key strength of EXCEED, enabling the study of a wide range of risk factors and diseases, even where data has not been specifically collected at baseline or precedes enrolment as a study participant. UK general practice has had over 20 years of near-universal computerised records [44]. These records have been further enhanced with the introduction of the QOF in 2004, which incentivised GPs to keep comprehensive records of several chronic diseases [45]. Some of these indicators incentivise the recording of quantitative traits relevant to the chronic disease diagnosed, such as blood pressure, lung function, estimated glomerular filtration rate, glycated haemoglobin (HbA1c) and cholesterol measures. That these are expected to be recorded approximately annually means that registered patients often have many repeat measures within linked EHRs, providing an excellent opportunity to study trends in control of conditions such as hypertension or progression of diseases such as COPD. Previous studies have validated some of these primary care measures – for example, routinely recorded spirometry has shown good validity when compared to study specific measures [46]. Other more complex longitudinal outcomes for example, related to healthy ageing can also be measured using EHRs.

The use of EHR can have limitations. Misclassification and miscoding of diagnoses may occur, and it is particularly likely that the true prevalence of many diseases will be underestimated (the "clinical iceberg"), as demonstrated by a comparison of COPD diagnoses in primary care data in EXCEED with COPD from spirometry (Table 7). However, the availability of repeat recordings and multiple types of data (including examination findings, pathology results and onwards referrals) over a long period of time can be used to improve and validate the classification of diagnoses and other important exposures and outcomes. Many disease definitions have been validated already – for example, definitions of COPD and asthma in the GOLD-CPRD database – and EXCEED will contribute further to this important area of study [47-53]. In addition to disease status validation, combining records of drug prescriptions, diagnostic and symptom codes can be used define complex phenotypes that have not been possible to study previously.

The utility of combining EHR and genetic data for efficient and flexible genetic studies has been highlighted by the eMERGE network of biobanks and Geisinger MyCode [54, 55]. The comprehensive nature and near-universal coverage of NHS health records adds further strength to this study design. In particular the ability to capture virtually all primary and secondary care contacts over decades of the lifespan enables longitudinal studies with a depth of data available in relatively few studies.

Strengths of the study also include consent from all participants to be contacted to participate in recall-by-genotype studies, a type of consent which is not yet widely sought in cohort studies. Recall-by-genotype studies are expected to be highly valuable to identify and validate drug targets and to inform targeting of therapeutics in a precision medicine approach [42].

Some studies incorporating genetic analyses (such as Genomics England) actively seek clinically actionable variants, whilst most cohort studies may not seek to identify these but may discover them as incidental findings. Anticipating this potential, at the time of consent, we asked whether participants would wish to be notified about clinically actionable variants; 99.5% of participants stated that they would wish to be informed in this situation. Clinically actionable variants will be discussed with the regional clinical genetics department of University Hospitals Leicester NHS Trust and then reported back to participants on request for NHS validation. Understanding the reasons for participants' preferences, how these change over time and how these can best be supported by future policies and procedures will be of key importance for EXCEED and other longitudinal cohort studies.

Minority ethnic groups, notably Leicester's South Asian population, are currently underrepresented in EXCEED. This reflects the recruitment methods utilised to date. We have extended recruitment to the EXCEED study to increase ethnic minority participant numbers and have adapted our recruitment methods to achieve this, for example, by undertaking recruitment at community events. Minority ethnic groups are also substantially underserved in the availability of samples with genome-wide genotype data worldwide. Whilst the situation has improved in recent years for Asian populations, only 14% of individuals included in genome-wide association studies worldwide up to 2016 were from Asian backgrounds [56]. This situation is replicated in UK-based studies. In UK

Biobank, only 2% of participants are from Asian or Asian British ethnic groups, despite this group representing around 7% of the UK population. It is essential that representation of minority ethnic groups increases substantially in genomic studies if these communities are to realise the benefits of genomically–informed advances in precision medicine. EXCEED aims to contribute towards this important goal.

Can I get hold of the data? Where can I find out more?

Participants have consented to their pseudonymised data being made available to other approved researchers and we welcome requests for collaboration and data access. Access to the resource requires completion of a proposal form, including a lay summary of the proposed research. Applications to access the resource will be assessed for consistency with the data access policy and with the guidance of the Scientific Committee, which has participant representation. Access to the data will be subject to completion of an appropriate Data/Materials Transfer Agreement and to necessary funding being in place. Requests to collect new data or to utilise biological samples may be subject to additional requirements. Interested researchers are encouraged to contact the study management team via exceed@le.ac.uk.

Profile in a nutshell

- EXCEED is a longitudinal population-based cohort which facilitates investigation of genetic, environmental and lifestyle-related determinants of a broad range of diseases and of multiple morbidity through data collected at baseline and via electronic healthcare record linkage.
- Recruitment has taken place in Leicester, Leicestershire and Rutland since 2013 and is ongoing, with 9,840 participants aged 30-69 to date. The population of Leicester is diverse and additional recruitment from the local South Asian community is ongoing.
- Participants have consented to follow-up for up to 25 years through electronic health records and additional bespoke data collection is planned.
- Data available includes baseline demographics, anthropometry, spirometry, lifestyle factors (smoking and alcohol use) and longitudinal health information from primary care records, with additional linkage to other EHR datasets planned. Patients have consented to be contacted for recall-by-genotype and recall-by-phenotype sub-studies, providing an important resource for precision medicine research.
- We welcome requests for collaboration and data access by contacting the study management team via exceed@le.ac.uk.

Pocket Profile for Extended Cohort for E-health, Environment and DNA (EXCEED)

Title: Cohort Profile: Extended Cohort for E-health, Environment and DNA (EXCEED)
Authors*:* Catherine John[1], Nicola F. Reeve[1], Robert C. Free[2], Julian Barwell[5], Susan E. Wallace[1], Ron Hsu[1], David J. Shepherd[1], Edward J. Hollox[5], Louise V. Wain[1], Martin D. Tobin[1]

The complete author list is available in the full version of the profile
Keywords: EXCEED, cohort profile, multimorbidity, recall-by-genotype, recall-by-phenotype
Corresponding author*:* Nicola F. Reeve, nfr5@leicester.ac.uk

Cite this as: The full version of this profile is available at IJE online and should be used when citing this profile.

Cohort purpose: EXCEED facilitates investigation of genetic, environmental and lifestyle-related determinants of a broad range of diseases and of multiple morbidity via electronic healthcare record linkage, supplemented by baseline data collection.

Cohort Basics: Recruitment has taken place in Leicester, Leicestershire and Rutland since 2013 and is ongoing, with 9,840 participants aged 30-69 to date. The population of Leicester is diverse and additional recruitment from the local South Asian community is ongoing.

Follow-up and attrition: Participants have consented to follow-up for up to 25 years through electronic health records, and additional bespoke data collection is planned.

Design and Measures: EXCEED is a longitudinal population-based cohort study. Data available includes baseline demographics, anthropometry, spirometry, lifestyle factors (smoking and alcohol use) and longitudinal health information from primary care records, with additional linkage to other EHR datasets planned.

Unique features: Patients have consented to be contacted for recall-by-genotype and recall-by-phenotype studies, providing an important resource for precision medicine research. Consent for linkage to EHR and other health care records is also a key strength, enabling the study of the determinants of susceptibility, progression and treatment response pertinent to a wide range of diseases. It facilitates the study of longitudinal outcomes related to healthy ageing.

Reasons to be cautious: The use of EHR has limitations. Misclassification and miscoding of diagnoses may occur, the prevalence of diseases may be higher or lower than the general population and analyses based on historic data may be subject to survivor and other biases. However, the availability of repeat recordings and multiple types of data (including examination findings, pathology results and onwards referrals) over a long period of time can be used to improve and validate the classification of diagnoses and other important exposures and outcomes. Leicester's South Asian population has been underrepresented by the recruitment methods utilised to date, but planned recruitment should address this.

Collaboration and data access: We welcome requests for collaboration and data access by contacting the study management team via exceed@le.ac.uk.

Funding and competing interests: The study has been supported by the University of Leicester, the NIHR Leicester Respiratory Biomedical Research Centre, the NIHR Clinical Research Network East Midlands, Leicester City Council, the Medical Research Council (grant G0902313) and the Wellcome Trust (grant 202849).

Author affiliations:

[1] Department of Health Sciences, University of Leicester, Leicester, LE1 7RH, UK

[2] Leicester NIHR BRC, Institute for Lung Health, Department of Infection, Immunity & Inflammation, University of Leicester, Leicester, LE1 7RH, UK

[3] Department of Population Science and Experimental Medicine, Institute of Cardiovascular Science, University College London, London, UK

[4] Infection, Immunity & Inflammation, University of Leicester, Leicester, LE1 7RH, UK

[5] Department of Genetics and Genome Biology, University of Leicester, Leicester, United Kingdom

Table: Summary of data collected at each phase

Phase	**Measurements**
Historic cohort data	Historically coded data (transferred from paper records at the time of practice computerisation, approximately mid-1990s) and since mid-1990s prospectively recorded consultations, with coded: • symptoms; • diagnoses; • measurements, such as blood pressure and spirometry • laboratory test results • drug prescriptions; • secondary care referrals
Baseline	All participants: Questionnaire, including smoking and alcohol use. DNA saliva sample Examination by research professional only: height, weight, waist circumference, hip circumference and spirometry Postal participants only: self-measured anthropometry and omitted spirometry
Ongoing	Planned quarterly updates to primary care record linkage (detailed above), with consent to follow-up for 25 years, to track health longitudinally Ongoing linkage to: • Admissions, Accident and Emergency attendances and Outpatient appointments via Hospital Episode Statistics • Pathology Data (East Midlands Pathology Service) • Myocardial Ischaemia National Audit (MINAP)

Acknowledgements

This paper presents independent research funded partially by the National Institute for Health Research (NIHR) and the Leicester Biomedical Research centre (BRC). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health. We particularly acknowledge the contribution of the late Dr Corinne Camilleri-Ferrante, Director of Recruitment at the time of the study's launch. The EXCEED study gratefully acknowledges the support of all participants and staff who have contributed to the study.

References

1. Nelson, M.R., et al., *The support of human genetic evidence for approved drug indications.* Nat Genet, 2015. **47**(8): p. 856-60.
2. Huntley, A.L., et al., *Measures of multimorbidity and morbidity burden for use in primary care and community settings: a systematic review and guide.* Ann Fam Med, 2012. **10**(2): p. 134-41.
3. Almirall, J. and M. Fortin, *The coexistence of terms to describe the presence of multiple concurrent diseases.* J Comorb, 2013. **3**: p. 4-9.
4. Le Reste, J.Y., et al., *The European General Practice Research Network presents a comprehensive definition of multimorbidity in family medicine and long term care, following a systematic review of relevant literature.* J Am Med Dir Assoc, 2013. **14**(5): p. 319-25.
5. Salisbury, C., et al., *Epidemiology and impact of multimorbidity in primary care: a retrospective cohort study.* Br J Gen Pract, 2011. **61**(582): p. e12-21.
6. Mercer, S.W., et al., *Managing patients with mental and physical multimorbidity.* BMJ : British Medical Journal, 2012. **345**.
7. Smith, S.M., et al., *Managing patients with multimorbidity: systematic review of interventions in primary care and community settings.* BMJ : British Medical Journal, 2012. **345**.
8. Barnett, K., et al., *Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study.* Lancet, 2012. **380**(9836): p. 37-43.
9. National Guideline Centre, *Multimorbidity: clinical assessment and management. NICE guideline NG56.* 2016.
10. Garin, N., et al., *Global Multimorbidity Patterns: A Cross-Sectional, Population-Based, Multi-Country Study.* J Gerontol A Biol Sci Med Sci, 2016. **71**(2): p. 205-14.
11. The Academy of Medical Sciences, *Multimorbidity: a priority for global health research*. 2018.
12. Visscher, P.M., et al., *10 Years of GWAS Discovery: Biology, Function, and Translation.* Am J Hum Genet, 2017. **101**(1): p. 5-22.
13. Repapi, E., et al., *Genome-wide association study identifies five loci associated with lung function.* Nat Genet, 2010. **42**(1): p. 36-44.
14. Hancock, D.B., et al., *Meta-analyses of genome-wide association studies identify multiple loci associated with pulmonary function.* Nat Genet, 2010. **42**(1): p. 45-52.
15. Loth, D.W., et al., *Genome-wide association analysis identifies six new loci associated with forced vital capacity.* Nat Genet, 2014. **46**(7): p. 669-77.
16. Soler Artigas, M., et al., *Genome-wide association and large-scale follow up identifies 16 new loci influencing lung function.* Nature Genetics, 2011. **43**(11): p. 1082-1090.
17. Wilk, J.B., et al., *Genome-wide association studies identify CHRNA5/3 and HTR4 in the development of airflow obstruction.* Am J Respir Crit Care Med, 2012. **186**(7): p. 622-32.
18. Castaldi, P.J., et al., *The association of genome-wide significant spirometric loci with chronic obstructive pulmonary disease susceptibility.* Am J Respir Cell Mol Biol, 2011. **45**(6): p. 1147-53.
19. Cho, M.H., et al., *Risk loci for chronic obstructive pulmonary disease: a genome-wide association study and meta-analysis.* Lancet Respir Med, 2014. **2**(3): p. 214-25.
20. Hobbs, B.D., et al., *Genetic loci associated with chronic obstructive pulmonary disease overlap with loci for lung function and pulmonary fibrosis.* Nature Genetics, 2017. **49**: p. 426.
21. Hobbs, B.D., et al., *Exome Array Analysis Identifies a Common Variant in IL27 Associated with Chronic Obstructive Pulmonary Disease.* Am J Respir Crit Care Med, 2016. **194**(1): p. 48-57.
22. Soler Artigas, M., et al., *Effect of five genetic variants associated with lung function on the risk of chronic obstructive lung disease, and their joint effects on lung function.* Am J Respir Crit Care Med, 2011. **184**(7): p. 786-95.

23. Wain, L.V., et al., *Novel insights into the genetics of smoking behaviour, lung function, and chronic obstructive pulmonary disease (UK BiLEVE): a genetic association study in UK Biobank.* The Lancet Respiratory Medicine, 2015. **3**(10): p. 769-781.
24. Wain, L.V., et al., *Genome-wide association analyses for lung function and chronic obstructive pulmonary disease identify new loci and potential druggable targets.* Nat Genet, 2017. **49**(3): p. 416-425.
25. Herrett, E., et al., *Data Resource Profile: Clinical Practice Research Datalink (CPRD).* International Journal of Epidemiology, 2015. **44**(3): p. 827-836.
26. Smith, B.H., et al., *Cohort Profile: Generation Scotland: Scottish Family Health Study (GS:SFHS). The study, its participants and their potential for genetic research on health and illness.* International Journal of Epidemiology, 2013. **42**(3): p. 689-700.
27. Fry, A., et al., *Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population.* American Journal of Epidemiology, 2017. **186**(9): p. 1026-1034.
28. Gill, P.S., et al., *Under-representation of minority ethnic groups in cardiovascular research: a semi-structured interview study.* Family Practice, 2013. **30**(2): p. 233-241.
29. Hussain-Gambles, M., K. Atkin, and B. Leese, *Why ethnic minority groups are under-represented in clinical trials: a review of the literature.* Health Soc Care Community, 2004. **12**(5): p. 382-8.
30. Sudlow, C., et al., *UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age.* PLoS Med, 2015. **12**(3): p. e1001779.
31. McCarthy, S., et al., *A reference panel of 64,976 haplotypes for genotype imputation.* Nature Genetics, 2016. **48**(10): p. 1279-1283.
32. Baker, C. *House of Commons Library Briefing Paper Number 3336. Obesity Statistics.* 2018; Available from: http://researchbriefings.files.parliament.uk/documents/SN03336/SN03336.pdf.
33. NHS Digitial. *Health Survey for England – 2016 Adult Health Trends - tables, Table 10*. 2017 [cited 2018 12 June]; Available from: https://digital.nhs.uk/data-and-information/publications/statistical/health-survey-for-england/health-survey-for-england-2016.
34. Leicester City Council. *How Deprived is Leicester?* 2016 [cited 2018 6 August]; Available from: https://www.leicester.gov.uk/media/181928/indices-of-deprivation-in-leicester-september-2016.pdf.
35. Office for National Statistics, *Adult smoking habits in the UK: 2016.* 2017.
36. Cassell, A., et al., *The epidemiology of multimorbidity in primary care: a retrospective cohort study.* British Journal of General Practice, 2018.
37. Bernd, L., et al., *Determinants of underdiagnosis of COPD in national and international surveys.* Chest, 2015. **148**(4): p. 971-985.
38. Tinkelman, D.G., et al., *Misdiagnosis of COPD and Asthma in Primary Care Patients 40 Years of Age and Over.* Journal of Asthma, 2006. **43**(1): p. 75-80.
39. Hill, K., et al., *Prevalence and underdiagnosis of chronic obstructive pulmonary disease among patients at risk in primary care.* Canadian Medical Association Journal, 2010. **182**(7): p. 673-678.
40. Levey, A.S., et al., *A more accurate method to estimate glomerular filtration rate from serum creatinine: A new prediction equation.* Annals of Internal Medicine, 1999. **130**(6): p. 461-470.
41. Sedgwick, P., *Nested case-control studies: advantages and disadvantages.* BMJ : British Medical Journal, 2014. **348**.
42. Corbin, L.J., et al., *Formalising recall by genotype as an efficient approach to detailed phenotyping and causal inference.* Nature Communications, 2018. **9**(1): p. 711.

43. Smith, G.D. and S. Ebrahim, *'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease?* International Journal of Epidemiology, 2003. **32**(1): p. 1-22.
44. Benson, T., *Why general practitioners use computers and hospital doctors do not—Part 1: incentives.* BMJ, 2002. **325**(7372): p. 1086.
45. Roland, M., *Linking Physicians' Pay to the Quality of Care — A Major Experiment in the United Kingdom.* New England Journal of Medicine, 2004. **351**(14): p. 1448-1454.
46. Rothnie, K.J., et al., *P223 Validity and interpretation of spirometry for patients in primary care.* Thorax, 2015. **70**(Suppl 3): p. A188.
47. Nissen, F., et al., *Validation of asthma recording in the Clinical Practice Research Datalink (CPRD).* BMJ Open, 2017. **7**(8): p. e017474.
48. Nissen, F., et al., *Validation of asthma recording in electronic health records: a systematic review.* Clin Epidemiol, 2017. **9**: p. 643-656.
49. Quint, J.K., et al., *Validation of chronic obstructive pulmonary disease recording in the Clinical Practice Research Datalink (CPRD-GOLD).* BMJ Open, 2014. **4**(7): p. e005540.
50. Rothnie, K.J., et al., *Validation of the Recording of Acute Exacerbations of COPD in UK Primary Care Electronic Healthcare Records.* PLoS One, 2016. **11**(3): p. e0151357.
51. Moore, E., et al., *Effects of Pulmonary Rehabilitation on Exacerbation Number and Severity in People With COPD An Historical Cohort Study Using Electronic Health Records.* Chest, 2017. **152**(6): p. 1188-1202.
52. Morgan, A.D., et al., *COPD disease severity and the risk of venous thromboembolic events: a matched case-control study.* Int J Chron Obstruct Pulmon Dis, 2016. **11**: p. 899-908.
53. Windsor, C., et al., *No association between exacerbation frequency and stroke in patients with COPD.* International Journal of Chronic Obstructive Pulmonary Disease, 2016. **11**: p. 217-225.
54. Carey, D.J., et al., *The Geisinger MyCode community health initiative: an electronic health record–linked biobank for precision medicine research.* Genetics In Medicine, 2016. **18**: p. 906.
55. Gottesman, O., et al., *The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future.* Genetics In Medicine, 2013. **15**: p. 761.
56. Popejoy, A.B. and S.M. Fullerton, *Genomics is failing on diversity.* Nature, 2016. **538**(7624): p. 161-164.