What happens between first symptoms and first acute exacerbation of COPD? Mapping and prediction study

Version control			
Version	Author	Date	Changes
1.0	Alex Bottle	Dec 3 2018	Version approved by NIHR Board
1.1	Alex Bottle	May 3 2019	Timelines updated. Logos and acknowledgment added. This version is awaiting ethical approval from CPRD
1.2	Alex Bottle	Jun 21 2019	Ethical approval received from CPRD, with some small suggestions for the analysis regarding lung function tests (incorporated into this version): ISAC number 19_116

2 Summary of Research

Background: Chronic obstructive pulmonary disease (COPD) is the fourth leading cause of death and affects nearly 400 million worldwide. In the UK over a million people live with it, but, despite the National Audit and local prevalence estimates, we have limited understanding of how people are diagnosed, how long this takes, and the different approaches to clinical management taken in primary care. In addition, management should be tailored to each patient to improve outcomes, for instance by stratifying patients according to the risk of acute exacerbations of COPD (AECOPD). A recent systematic review of risk prediction models for AECOPD stated that “exacerbations are an ideal target for risk-stratified treatment, but…none of the existing models fulfilled the requirements for risk-stratified treatment to personalise COPD care”. We plan to overcome the statistical shortcomings of these models to produce a risk prediction model for the first AECOPD, which is a better point at which to intervene than afterwards due to the cumulative lung damage caused by repeat exacerbations.

Aims and objectives: To describe and model the patient journey from symptom presentation to diagnosis and first acute exacerbation for COPD patients in England. This will include examining variations by Clinical Commissioning Group (CCG), GP practice and time period, followed by the construction and validation of a risk prediction or risk trajectory model for the first AECOPD.

Methods: Using the Clinical Practice Research Datalink (CPRD) and two cohorts ten years apart, we will describe the management of the patient following initial presentation with symptoms through to their diagnosis of COPD and their first AECOPD, which for some patients will be the same event. Given that COPD can present differently depending on comorbidity, the mapping will be described separately for people with asthma and heart failure in particular. The second part will model the risk of the first AECOPD using factors such as airways obstruction, age, smoking, BMI, gender, comorbidities and public data on temperature and pollution. This will use logistic regression, random forests and cause-specific hazards modelling. Predictors will be ranked in importance from GP, patient and system perspectives. Models will be externally validated using CPRD “Aurum” practices.

Timelines for delivery: Mar 19 – May 19 (months -3 to 0): study set up; Jun 19 – May 20 (months 1-12): data extraction and objectives 1 and 2; Jun 20 – May 21 (months 13-24): objectives 3 and 4; Jun 21 – Aug 21 (months 25-27): report writing

Anticipated impact and dissemination: This study will fill key gaps in our understanding of how patients obtain their COPD diagnosis (their "route to diagnosis"), how they are managed in primary care, and how they get their first AECOPD. Comparisons between the two time periods will highlight what has changed and inform NHS preparation for future needs regarding COPD. A risk prediction model for first acute exacerbation will aid shared decision-making between GPs and patients and facilitate early intervention; ranking the predictors will suggest priorities for action. Dissemination will be to clinicians, patients and policy-makers, through academic outputs and relevant charities such as our local Breathe Easy patient group and the British Lung Foundation. Our approach and results will inform the National Audit and also NICE guidelines on diagnosis and management.

3 Background and Rationale

Chronic obstructive pulmonary disease (COPD) affects over a million people in the UK and nearly 400 million worldwide (see next section for more epidemiology data). National audits and other studies have documented variations in the quality of care and in outcomes for many conditions; for COPD and primary care, however, the most recently published national audit for England and Wales to go beyond Quality and Outcomes Framework data only covered Wales and does not distinguish between new and existing cases [1]. Patients can either be diagnosed in primary care or in hospital, the latter often by emergency admission. We have limited understanding of how this happens, how long it takes, and the different approaches to clinical management taken by primary care professionals. This is particularly true with regard to people with comorbidities such as asthma and heart failure that can also cause breathlessness. An analysis of primary care records for 2000-2009 found some small improvements in prescribing (rises in oral corticosteroids and triple therapy in accordance with guidelines, though another analysis concluded that COPD is not being treated in accordance with GOLD or NICE guidelines in the UK [2]), a reduction in COPD severity and age of diagnosis, and a rise in the mean number of GP consultations per patient-year [3]. However, they concluded that improvements in diagnosis were only modest during the period, and they were unable to look at exacerbations or hospitalisations. Local estimates of COPD prevalence are now published online by Public Health England, but overall we have little information on to what extent the NHS is meeting the needs of current patients with COPD and how well it might meet those of future patients.

To improve outcomes, management should be better tailored to each patient, as recommended by the National Audit [1]. One approach for personalising COPD treatment is to stratify patients according to the risk of acute exacerbations of COPD ("AECOPD") in order to prescribe treatments such as inhaled corticosteroids or phosphodiesterase-4 inhibitors earlier. AECOPDs are important predictors of mortality and reduced quality of life, and more timely intervention to prevent these would therefore be of significant benefit for patients. As the second most common reason for emergency hospital admission overall, they are also of great public health and financial importance [4,5]. For many patients, it is also when they are diagnosed with COPD. Information on who is at higher risk of AECOPD would help shared decision-making between the practice clinicians (GPs and practice nurses) and the patient. A well-performing risk prediction model that uses information available to the clinician would inform clinical decision-making and timely management. This is particularly important before the patient's first AECOPD, as each exacerbation damages the lungs and treatment is less effective thereafter [5,6,7]. Previous attempts to model the risk of AECOPD have for several reasons not been successful (see later section). The concept of a risk score based on measurements taken at a single time point may in any case

be too simplistic, depending on how patient risk factors change over time, but this is not known.

Whilst it has been recently announced that the NHS is to receive extra money, it is well recognised that it must do things differently, including better cooperation between community and hospital services, and with greater efficiency [8]. In short, to help achieve this for COPD patients and to improve clinical management, shared decision-making and patient outcomes, we need to understand this first key part of the "patient journey" from symptoms through to diagnosis and first AECOPD.

3a Evidence explaining why this research is needed now

The problem being addressed is our limited knowledge of the patient journey from symptom presentation to diagnosis and first acute exacerbation for COPD patients. We also lack a validated risk-prediction model for this first exacerbation to help with shared decision-making between clinicians and patients. The case for need for this project covers a number of areas.

HEALTH NEED
COPD is the fourth leading cause of death worldwide and accounts for 30,000 deaths each year in the UK [9]. According to the Global Burden of Disease project, COPD caused 2.6% of global disability-adjusted life years (DALYs) in 2015 – the eighth most important disease on this measure [10]. The UK is among the top 20 countries for COPD mortality worldwide. In Europe, only Denmark and Hungary have higher death rates. In the UK, 115,000 people are diagnosed with COPD each year – one every 5 minutes on average [11]. Numbers of people living with the disease are rising due to people living longer in general. Furthermore, significant regional variation exists in COPD, with prevalence varying nearly twofold across the UK, and age-standardised hospital admission rates varying about eightfold among local authorities [11].

AECOPDs are responsible for the majority of the disease burden, contribute to the progressive decline in lung function and reduce patients' quality and quantity of life [4,5]. Earlier diagnosis and a reduction in AECOPDs through more timely intervention could have a significant impact on the burden borne by patients and the NHS. The first step is to better understand how the diagnosis is made in practice, how long it takes and what are the main predictors of the first AECOPD.

EXPRESSED NEED
The importance of assessing how COPD is managed is reflected in the setting up of the National COPD Audit Programme. Now called the National Asthma and COPD Audit Programme, NACAP, this reported on the quality of primary COPD care in 2016 and 2017 in England and Wales; for Wales, this used an analysis of electronic health records but for England this was derived largely only from Quality and Outcomes Framework (QOF) data (see Existing Evidence below). Amongst the Audit's key recommendations were early and accurate diagnosis and a personalised approach to treatment, both of which are covered in our proposal. In addition, our proposed assessment of the quality of primary care in England and the details of how to do it using primary care electronic health records will be able to feed directly into the next iteration of the Audit. This will be greatly aided by the role of applicant JKQ as Analysis Lead to the Audit.

SUSTAINED INTEREST AND INTENT
According to the Global Burden of Disease report [10], COPD affected an estimated 104·7 million males and 69·7 million females in 2015 worldwide, with many more potentially undiagnosed. From 1990 to 2015, the age-standardised prevalence decreased by 14·7% but the crude prevalence of COPD increased by 44·2%. The global number of deaths rose by 11.6% to 3.2 million people between 1990 and 2015 [10]. The falls in the age-standardised

prevalence and death rates were more than compensated for by population growth and ageing. Annual direct healthcare costs of COPD in England have been estimated to increase from £1.5 billion in 2011 to £2.3 billion in 2030 [12]; nearly £1 billion was spent on respiratory inhalers in 2011, a significant proportion of which will have been for COPD [13].

The 2010 NICE guideline on diagnosis and management was reviewed in April 2016 and will report in December 2018 [14]. The results of this proposal will feed into the next review, just as our work on the "routes to diagnosis" for heart failure [15] is being used by NICE as they currently update their guidelines for that condition.

GENERALISABLE FINDINGS

UK electronic healthcare records (EHR) are becoming an increasingly important resource for evidence from real life research as they do not suffer from the selection biases inherent in randomised controlled trial populations. One such large EHR database is the Clinical Practice Research Datalink (CPRD), much used in research [16]. This contains anonymised, coded patient records from about 600 GP practices in the UK, with five million currently registered patients who are representative of the UK population in age, sex and ethnicity. This proposal will analyse CPRD and hence the results should be generalizable to the whole of England.

CAPACITY TO GENERATE NEW KNOWLEDGE

There is a lack of knowledge on patients' "route to diagnosis" for COPD in this country. There is also limited knowledge on the predictors of the first AECOPD, as most AE prediction models are for all AEs combined and find that the main predictor is the number of previous AEs. The relative importance of those predictors is particularly unknown. We also know little about how much this has changed over time. This proposal will address all of these gaps.

This project will also give us further experience with using primary care EHRs such as CPRD to evaluate the patient journey and quality of care. While CPRD is in many ways an excellent and much-used resource, all databases have limitations, and we pay particular attention in this document to potential data quality issues.

EXISTING EVIDENCE

Our recent analysis of patients with heart failure found that there are many routes to diagnosis of that condition, with different combinations of tests, specialist referral, medications and hospitalisation [15]. Comparable work has not been done for COPD. For instance, the current National COPD Audit includes an audit of primary care for the first time, but the diagnosis section in the November 2016 report covering England [17] focused largely on comparisons with figures from the Quality and Outcomes Framework; the patient-level data analysed in the December 2017 report came only from Wales [1]. We plan to go much further and fill these important knowledge gaps in the first part of this project.

Our HS&DR-funded study 14/19/50 "What are the determinants of variations in emergency readmission rates and one-year mortality in patients hospitalized with heart failure or chronic obstructive pulmonary disease?" (report currently in press) brought together a number of data sources that included aggregated practice-level information. We found that predictors for mortality in COPD patients included the number of GPs per 1000 patients (odds ratio 0.89 per extra GP, p=0.004) but none of the QOF indicators or COPD prevalence was significant for either mortality or all-cause readmission. The number of outpatient appointments attended and missed were strong predictors of both outcomes and will be considered as predictors for AECOPD. That study was limited to using practice-level information on primary care management, but the current proposal will benefit from patient-level data.

A recent systematic review of risk prediction models for AECOPD stated that, "Exacerbations are an ideal target for risk-stratified treatment since it is one of the most important outcomes for COPD patients, and avoiding them is likely to lead to a higher health-related quality of life, longer life and less healthcare cost." [18] The review found that "only two out of 25 studies validated the developed model, only one out of 27 models provided estimates of individual exacerbation risk, and only three…used high-quality statistical approaches for model development and evaluation." It concluded that "Overall, none of the existing models fulfilled the requirements for risk-stratified treatment to personalise COPD care." Nine models included previous exacerbations as a predictor and are therefore of limited relevance to the prediction of the first one. The main statistical limitations were problematic variable-selection procedures and the lack of external validation. Our proposed design overcomes these shortcomings.

An advantage of CPRD is that it contains information entered by clinicians, giving a reliable indication of what they considered the important clinical problems at the consultation. However, this coding is complex and user-dependent, and so care needs to be taken and sensitivity analyses run regarding the choice of codes. Our group has validated this database in a series of studies on COPD, including the recording of the diagnosis, spirometry results and AECOPD events [19,20,21]. CPRD has been used to estimate the costs associated with treating COPD and its exacerbations [22]. Excluding non exacerbation-related medications, that study estimated an average of around £2,000 per patient overall (over 50% more than this for people with two or more AECOPDs), with exacerbations accounting for around 20% of the costs. A similar but smaller database in Scotland was one of those used to derive predictors for hospitalisation for AECOPD [23]. The study team concluded that "there is greater potential for primary care to prevent or delay the initial admission through appropriate disease management". The authors did not handle the competing risk of death and did not report model performance statistics, but their study demonstrates the potential for using CPRD to model AECOPD.

In summary, our evaluation of how patients present, are diagnosed and are managed in primary care will feed into future rounds of the National Audit and iterations of NICE guidelines. AECOPDs are common, serious, costly, and ripe for risk prediction to aid prevention, and CPRD represents an appropriate database to address this.

4 Aims and objectives

Our aim is to describe and model the patient journey from symptom presentation to diagnosis and first acute exacerbation for COPD patients in England.

Specific objectives:
1. Map out the clinical management and NHS contacts from symptom presentation to COPD diagnosis and first AECOPD (for some patients the latter two will be the same event)
2. Investigate whether and how this varies by Clinical Commissioning Group (CCG), GP practice and time over the last ten years
3. Rank predictors of the first AECOPD in importance from the GP, patient and health system perspectives and assess whether and how this has changed during the last ten years
4. Construct and externally validate risk prediction or risk trajectory model for first AECOPD.

5 Research Plan / Methods

Study design and outline

The proposal has two parts: (1) map out the routes to diagnosis and first AECOPD and (2) predict the first AECOPD. Both have an observational study design (cross-sectional and

cohort approaches). We first describe the project data set and patient cohorts before giving the plans for data management and analysis.

The database: CPRD

The Clinical Practice Research Datalink (CPRD) contains electronic patient records from participating general practices across the UK using Vision software ("CPRD Gold"); it has recently also incorporated practices using EMIS ("CPRD Aurum"). CPRD practices based in England are eligible to be linked with Hospital Episodes Statistics (HES), the national hospital administrative database covering all A&E attendances, outpatient appointments, day cases and inpatient stays. Approximately 75% of English practices have consented to participate in this linkage. Records are also linked to the national deaths registry and, via the practice and/or the patient ID, to area-level deprivation indices such as IMD.

We have previously constructed a data set of mean monthly temperature and pollution levels. For each practice, CPRD staff have linked the data from the nearest monitor to the main data set. This information covers the period 2004-2014 but is currently being updated (see section below for "PART 2" of the analysis).

Selection of patient cohort

We will include patients in CPRD registered at practices in England for at least three years and consenting to linkage to HES, who are aged over 35, who have any of the NICE-listed symptoms and with a code for COPD, either in the primary care records via Read codes or in hospital admissions data via ICD10 J40-J44. The symptoms of interest are exertional breathlessness, chronic cough, regular sputum production, frequent winter 'bronchitis' and wheeze. We will apply these criteria to each of two time periods: the most recent two years (likely 2016-7) and two years a decade earlier (i.e. 2006-7). A minimum of three years' registration will allow at least one year to track back from the date of first symptom presentation (or date of diagnosis if it is made during a hospital admission) for prior mentions of COPD, comorbidities and medication history and at least two years to track forward from symptom presentation or diagnosis and pick up management, mortality and AE information. We have previously found that the AE rate in one year strongly predicts the long-term AE rate [24]. Patients with a diagnosis of COPD recorded before their spirometry are likely to be existing cases who have transferred in from another practice and therefore will be excluded. Patients with a NICE symptom but no COPD diagnosis code will also be excluded; whilst those who do not receive a diagnosis within two years of symptom presentation and then transfer out of the practice after two years might go on to receive a COPD diagnosis later, this censoring means that we do not know their subsequent diagnosis.

Definition of AECOPD

When AECOPD is detected during a primary care visit, we will apply our published algorithm for Read codes [20]. When the AECOPD results in hospitalisation, we will use ICD10 codes. The specific ones are J44.0 and J44.1, but we have found before that the best approach is to use one of these specific AECOPD codes or a lower respiratory tract infection (LRTI) code (J22) in any diagnosis field or a COPD code in the primary diagnosis field in any consultant episode during a hospitalization. This had a sensitivity of 88% when compared against the discharge summaries [25]. Patients who present with an AECOPD to the emergency department or a walk-in centre but who are not hospitalised at that visit are a small minority; as well as clinical reasons, the decision to admit is partly influenced by the pressure of the four-hour A&E target. Those who are not admitted will not be captured in HES because HES A&E diagnosis information is neither complete nor specific enough for our purpose. Many of these will, however, be captured by Read codes in CPRD because of hospital letters that are scanned and coded by GP practices, although in practice some may be missed.

Sample size considerations

A previous CPRD analysis study [26] found 44,000 patients with COPD at Jan 2011 and at least two years' CPRD records at HES-linked practices, 37,000 (84%) of whom had a MRC dyspnoea grade recorded. This will be more than sufficient for descriptive analysis (overall and by CCG) and regression modelling; robust analysis by practice may require the inclusion of further years of data.

Data management

CPRD data will come from the Big Data Analysis Unit at Imperial. They are a fully certified ISO 27001:2013 research environment within Imperial College and are 100% compliant with NHS IG Toolkit Level 3 (EE133887). They hold a CPRD GOLD licence, and use of this will incur a cost (see Justification of Resources). CPRD data extracts from the BDAU can either remain on the BDAU server or more flexibly be accessed via secure CPRD key fobs. Both departments involved in this proposal hold key fobs, from which data extracts will be prepared and taken. These extracts will then be held on the secure server at the Department of Primary Care and Public Health, where the research associate who will analyse the data will be based. Thanks to our GOLD licence, we will obtain patient-level deprivation and ONS mortality data from CPRD staff. For the external validation of the AECOPD model, we will apply for CPRD Aurum records, which will also be held by the Department of Primary Care and Public Health.

Statistical analysis

PART 1: routes to diagnosis and first AECOPD

NICE Clinical Guideline 101 states that a diagnosis of COPD should be considered in patients over the age of 35 who have a risk factor (generally smoking) and who present with one or more of the five symptoms listed earlier. These patients should also be asked about other factors such as weight loss and fatigue. The diagnosis is suggested by these factors and supported by spirometry (NICE Quality Standard 10), plus other investigations for some people. The combinations of symptoms, investigations, referrals and prescriptions in patients who go on to receive a COPD diagnosis will be mapped out descriptively as we did for heart failure (HF) [15], noting what proportion followed the NICE guideline and how long pathway elements took. Patients can be referred to a range of people, including specialist physicians or nurses, physiotherapists, dieticians and occupational therapists. We will particularly note whether patients are investigated for asthma or HF (e.g. fractional exhaled nitric oxide (FeNO); echocardiography / BNP testing; and specialist i.e. cardiology referral). It will be useful to understand chest infections and antibiotic prescribing before diagnosis as they will be managed differently (see next paragraph). Given that COPD can present differently depending on what comorbidities are already present, the mapping will be described separately for people with asthma and HF in particular; where possible we will also assess the severity and management of those comorbidities, e.g. asthma severity as per Bloom et al [27]. It will be important to look at mental health conditions and consultations, particularly for anxiety. For patients receiving their diagnosis during an acute exacerbation, we will describe what primary care contacts and tests such as for lung function they had in the previous year to assess the potential for earlier diagnosis.

For patients with five years of data before diagnosis, another way to estimate the potential for earlier diagnosis is to identify chest infections and episodes of bronchitis during this time to see whether the patient had more of these than the "average" patient did. We will define "more" as at least twice the average rate. This background rate for an "average" patient of

the same age and gender will be estimated from a random sample of CPRD patients with the same number of years of registration as the COPD patient.

The next step is to describe what happens between diagnosis and the first AECOPD and how long this period lasts; for some people, the two events will be the same. This will be done descriptively, stratifying by key patient characteristics such as airflow obstruction grade (we have algorithms to describe this), age, smoking history, comorbidity (especially asthma and heart failure) and initial management (NICE or otherwise). We will also look for differences by CCG in crude and risk-adjusted rates via funnel plots and via multilevel models with patient factors as level 1 predictors, practices as level 2 and CCGs as level 3. The latter will allow us to see how much variation exists by practice and by CCG after accounting for differences in patient mix. Timings between events will be summarised using descriptive statistics and cumulative incidence function plots with death as a censoring event.

All analyses in Part 1 will be run firstly using the most recent two years of records (allowing for sufficient follow-up time to capture AEs) and then using data from ten years earlier in order to see what has changed.

We also plan to capture patients' experiences of their route to diagnosis more directly using a multi-pronged approach to take account of the different types of people who are likely to respond to each option. First, we will take advantage of our existing links with our local Breathe Easy group to make contact with their counterpart groups in other parts of the country. A member of the research team will attend two of their meetings, explain our research and ask people about their diagnosis "journey". Second, we propose two focus groups, one in a large city other than London and one in a rural area (and in different towns from the Breathe Easy group meetings): this will involve a facilitator and a note-taker, both from the research team, with help from Imperial's experienced Patient Experience Research Centre as necessary. Third, with help from our PPI co-applicants, we will design a short survey to be spread on the British Lung Foundation's patient forum (we will ask the BLF to post this on our behalf), Twitter (one of our PPI co-Is has a list of relevant active users who tweet about COPD) and INVOLVE's People in Research. We will ask our PPI co-Is for help with summarising this information. We will also promote these activities via local media, Twitter and with a video on our project website made by a patient.

PART 2: modelling the risk of the first AECOPD

The literature suggests that candidate predictors would likely include airways obstruction, age, smoking, BMI and gender. Others that are available in CPRD include comorbidities, prescriptions, area-level socio-economic deprivation (IMD via linkage), long-term oxygen, lung function tests, hospital admission for other reasons and hospital outpatient appointments (attended and missed, via HES linkage [28]). Comorbidities of interest will include heart failure, coronary heart disease, heart valvular disease, asthma, diabetes, cancer (particularly in the lung), venous thromboembolism, anxiety, depression, dementia, osteoporosis, obesity, underweight (BMI<20), metabolic syndrome and anaemia. Some studies have found associations between cold weather and AECOPD [e.g. 29]; one of our patient representatives finds cold weather difficult and is keen to look at this. We will include temperature and pollution data from the nearest monitor in the models (see below). Candidate predictors will be checked for frequency and data quality. Some imputation may be necessary for continuous variables; missing values for categorical variables will be retained as an extra category – see section on "data quality considerations and sensitivity analyses" below.

Modelling will begin with time-to-event analysis using cause-specific hazards to deal with the competing risk of death and/or transfer out of the practice. Patients who are diagnosed via

their first AECOPD will be modelled as a separate group, with time=0 set to their first presentation for COPD symptoms rather than COPD diagnosis: identifying this t=0 event will likely need some iteration as we did for HF [15]. We will assess which predictors change over time, e.g. the development of new comorbidities, some of which can be caused by COPD [30]. Coronary heart disease events such as admission for acute myocardial infarction that occur after the COPD diagnosis can indicate previously undiagnosed cardiovascular disease, which is known to be a predictor of adverse outcomes. We will also assess which change effects over time. Using the model's predicted values, we can observe how AECOPD risk changes over time since diagnosis. We will likely need to adjust for clustering by practice, for instance by using robust standard errors. It may be possible to simplify the modelling by fitting a logistic model. We will investigate whether it is worthwhile implementing random forests with cross-validation, which can perform better than other machine learning approaches and logistic regression [31], its risk of overfitting mitigated by our large sample size and external validation data set (see below). Random forests can be extended to the time-to-event framework and have the advantages of not having to impose constraints on the underlying distributions and of providing a way to automatically deal with high-level interactions and higher-order terms in variables [32].

These models will be first run without incorporating the temperature or pollution information. This is because the information comes largely from urban areas and may not be generalizable to rural ones. We have obtained daily temperature data from the British Atmospheric Data Centre and daily pollution data from the Department for Environment, Food and Rural Affairs (DEFRA). The latter source covers major pollutants believed to contribute to the risk of AECOPD: nitrogen dioxide, sulphur dioxide, ozone, carbon monoxide, particulate matter of size below 10 micrometres (PM10), and particulate matter of size below 2.5 micrometres (PM2.5). These data will be integrated with CPRD patient data to enable the analysis, with mapping to each person's nearest general practice by via the eastings and northings of the monitoring station and the postcodes of the practice. CPRD have agreed to perform this linkage. If the pollution means derived from the monitoring stations are unique enough so that they could identify an individual practice, they will be converted into deciles or quintiles by CPRD staff. To incorporate this information, we propose to employ a case-crossover design rather than time series models [33]. This design has been widely used to study the association between short-term air pollution exposure and the risk of an acute adverse health event. It uses cases only, i.e. those with AECOPDs: for each individual case, exposure just before the event is compared with exposure at other control (or "referent") times. In the resulting conditional logistic regression model, cases act as their own controls and thereby control for time-invariant confounders. Covariates that do vary with time can also be controlled for by design by matching referents to the index time. The selection of referent times is therefore important in order to avoid bias – recommendations are given by Janes et al [33]. The day of the AECOPD will be taken as the case day and the same day of the week in the same month and year as control days. We will stratify by season by splitting the data into colder months (September to February) and warmer ones (March to August); some studies in Europe shift these forward one month, which we will include as a sensitivity analysis. Pollutants will be tried as linear terms initially, one at a time, with non-linear effects tested via generalised additive models with smoothing splines. There is no standard way to choose lag periods to look for lagged effects, so we will try lags up to five days and cumulative means up to five days before the AE and be guided by the Akaike Information Criterion (AIC) as is commonly done [34]. We will look for effect modification by age group and gender to see if some patient groups are more susceptible than others, as has been found for acute myocardial infarction, for instance [34].

We note that it can take months or years to obtain a diagnosis, for example due to comorbidities or chest infections. The GP may prescribe antibiotics, inhalers or even sometimes steroids before investigating for and diagnosis COPD; we will therefore

incorporate such pre-diagnosis prescribing as predictors in the model. This will mean that the route to diagnosis will be a potential predictor.

HF is a common comorbidity in people with COPD and, due to the overlap of symptoms noted earlier, can make COPD harder to diagnose. We will therefore stratify the analysis by the presence of HF; it is also possible that some AECOPD predictors will differ in people with HF, who are on average older and frailer. Treatment for HF can also have an impact on COPD: for example, beta-blockers improve survival for HF but are often not prescribed to patients with co-existing asthma or COPD because of concerns over adverse lung effects (though there is now some evidence that these drugs might reduce the AECOPD risk). In this group, the electronic frailty index will be tried as a predictor. For similar reasons, we will also stratify the analysis by the presence of asthma. Predictors will be entered at once into the model and the non-significant ones dropped only if this does not affect the coefficients of the remaining variables. Model performance will be assessed using standard measures such as discrimination, calibration and residuals.

We will rank predictors by importance by considering the perspectives of the primary care clinician, the patient and the health system separately. Ranking by p values is common but unhelpful as it is largely dependent on the size of the sample rather than the size of the effect. For the clinician with a patient in front of them, relative risk type measures such as the odds ratio or hazard ratio will be used and only information that they will have available to them will be considered. For the patient, odds ratios are also important even though they will likely not know the term. Risk factors such as age and gender are not modifiable, but knowing that e.g. adherence to medication would reduce one's risk more than attending pulmonary rehabilitation (PR) would be actionable and therefore useful (adherence, estimated by the medication possession ratio (see section on data quality considerations), and PR use can both be examined using CPRD). Finally, a system perspective needs to consider not just odds ratios but also prevalences, e.g. via population attributable risks (PARs). One would expect quitting smoking to have the greatest impact, but these calculations might show that more AECOPDs could be avoided nationally by e.g. people reducing their weight more than by the GP helping people control their diabetes. This perspective has the most relevance for policy. Measures such as PARs allow a direct estimation of the economic impact of reducing AEs. For example, York Health Economics Consortium give some background figures for a CCG of a population of 250,000 people when estimating the potential cost savings of a web-based self-management programme: 4,000 AEs per year, of which 500 resulted in hospitalisation at a mean cost of £1,590 each for a total of about £800,000 [35]. If 10% of hospitalised AEs were associated with exposure to a given risk factor and half of those 10% were avoided in practice, there would be a gross saving of 5% times £800,000 = £40,000 for the CCG on hospital admissions alone. We will obtain such estimates from the PARs obtained from the logistic regression model rather than the time-to-event ones: the difference between PARs with odds ratios obtained from logistic regression and those from hazard ratios obtained from time-to-event analysis varies with the distribution of the events, the presence of competing risks and the change in the prevalence of the risk factor with time. PARs also assume causal relations with the outcome, but they are nonetheless useful with this caveat borne in mind. We will obtain estimates of AEs and their associated financial costs for an average CCG and for England as a whole. National reference costs for the relevant financial year will be applied to hospital admissions and costs for other AEs taken from the literature.

These models will be run on the most recent two years of data and then on data from ten years earlier to see what has changed. This will give an indication of how future-proof such models are likely to be, for instance regarding whether any predicators have changed in importance and to what extent recalibration is necessary. Having developed the models using data from Vision practices, we will apply them to data for the same time period from EMIS practices (other than those that simply switch from Vision) for external validation. In

this, we will pay particular attention to the predictive performance, calibration and model fit using measures including the c statistic, Hosmer-Lemeshow statistics but particularly the associated plots to see where any recalibration is needed, and residuals. The measures of model performance and their exact definitions will depend on the exact form of the final model.

Data quality considerations and sensitivity analyses

Any database study requires an assessment of the quality of the important data items. The key items in this project include initial presentation with symptoms, spirometry testing, medications, COPD symptom severity, and AECOPD events and their predictors such as comorbidity and smoking. Much of the relevant groundwork has already been done, and this proposal will provide further valuable experience with using primary care electronic health records for assessing the quality of care in a low-cost way.

The main symptoms of COPD are breathlessness, cough and sputum. In order to make the patient summary neat and easier to search for diagnoses, GPs may choose to write the symptoms as free text rather than use Read codes. This free text is no longer available in CPRD. We previously showed that at least one of the main symptoms of heart failure was in fact coded for 80% of patients before diagnosis [15]. For patients without COPD symptoms coded, we will take the date of referral for spirometry (if recorded) or the date of their spirometry itself as time zero to mark the start of follow-up. This is typically within 2-3 weeks or so of the GP consultation, as the GP will ask the patient to come back and see the practice nurse for the spirometry.

The diagnosis of COPD depends on clinical judgment and confirmation of the presence of airflow obstruction using spirometry, which is now routinely available. Our group has validated the database for the key elements of COPD diagnosis, spirometry and AECOPD [19,20,21]. For a sample of CPRD patients, two chest physicians re-read all spirometric readings for both quality of the procedure and interpretation; of those conducted in primary care, 98.6% (n=218) of spirometry traces were of adequate quality, which allows us to use spriometric values when identifying COPD diagnosis in electronic records [20]. For AECOPD, we compared 15 algorithms and found that a combined strategy of antibiotic and oral corticosteroids prescriptions for 5–14 days, or lower respiratory tract infection or AECOPD code resulted in a high PPV of 85.5% [19].

Prescription data in CPRD provide information about whether a GP prescription has been issued and is captured automatically and accurately at the point of issue. It thereby tells us about GP compliance with guidelines. It does not tell us whether the prescription has been redeemed, although according to NHS Digital over 98% of GP prescriptions are actually redeemed. It does not tell us whether the patient has taken it as recommended. Common ways of trying to capture whether the patient has been taking the medication include the medication possession ratio (MPR) and proportion of days covered (PDC). The MPR is the sum of the days' supply for all fills of a given drug in a particular time period, divided by the number of days in the time period. Both can over- or under-estimate adherence. Despite this, many studies use MPR for several reasons. First, it is easy to calculate. Second, studies have shown similar non-adherence rates when comparing MPRs from GP-issued prescriptions and from pharmacy dispensing records [36]. Finally, it is the method of estimating adherence recommended by the International Society for Pharmacoeconomics and Outcomes Research group. We will also employ MPR but our discussion will acknowledge that medication adherence for COPD, which is generally treated with inhalers, is notoriously low. A previous CPRD study showed that less than half of the patients were adherent (MPR>=80%) [37], and a study using an objective measure showed that only 6% were actually adherent [38]. While we can estimate adherence in the sense of the patient having enough of the drug at home during the period in which they have been prescribed it,

it will be more difficult to assess the effectiveness of patients' inhaler technique. However, there are some potential approaches to this, which we will investigate, such as the prescription of a spacer and inhaler technique Read codes following COPD and asthma reviews.

AECOPDs are well recorded in primary care data, which are used not just by GPs but by organisations such as Public Health England to compare data on AECOPD incidence and management across localities and by clinical commissioning groups to inform delivery of care and design of services [25]. If the patient attends a walk-in centre or emergency department, and is hospitalised at that visit, then we can use the ICD10 of the resulting inpatient record in HES to identify COPD as a diagnosis and AEs. The primary diagnosis in hospital administrative data has been shown in general to be 96% accurate [39]. If they are not hospitalised, which happens in a minority of cases, then CPRD may still capture the event from the hospital level, though for some people this will not be scanned and coded as mentioned earlier. The diagnostic coding in HES A&E is not of sufficient completeness or granularity to be useful.

AECOPD predictors will include comorbidity and smoking. The main relevant comorbidities have been associated with high positive predictive values (PPV) in CPRD, e.g. 98.6% for diabetes [40] and 86.4% for asthma [41]. Booth et al compared age- and sex-standardised rates in CPRD with equivalent rates from national health survey data (Health Survey for England, HSE) for 2007-2011 [42]. The sets of current smoking rates matched very closely, but the rates of former smokers were a little lower than expected from the survey: estimated prevalence in CPRD for men was 26.7% (HSE 31.3%) and in CPRD for women was 22.9% (HSE 25.0%). A similar study argued that those who had quit smoking a long time ago and/or before the age of 30 were more likely to be miscategorised as non-smokers than those who quit recently and/or at an older age [43]. This misclassification represents a potential source of bias but is likely to be of modest importance, as the authors acknowledged that the HSE definition of ex-smoker was highly sensitive and clearly defined, including those who only smoked a small daily amount or for a short period. They also found that former or non-smokers were more likely to have missing data than current smokers. One of their suggestions was to dichotomise smoking status into current and non-current smokers, with missing data assumed to correspond to non-current smokers. However, this could be problematic where former smokers have quit only recently and so are at greater risk of poor outcomes such as AECOPD, so a missing indicator may be more appropriate. We will therefore use four categories: current, former, never and unknown; if the model coefficients for the unknown and one or more other categories are similar, it will shed light on who is represented by the unknown group. If, in the year before first presentation with COPD symptoms the patient's recorded status changes from "current" to "former" smoker or becomes missing, we will use their earliest recorded status.

Several aspects of other key covariates such as disease management for COPD, HF, and asthma have been included in the NICE Quality and Outcomes Framework. This means that the recording of key indicators and common variables such as BP, HbA1c and BMI is high and have increased substantially in CPRD. A recent study showed that percent predicted FEV1 was available for 80.9% of patients and symptoms for 75.6% of patients in CPRD [44]. In addition, they found that patients with and without available data for spirometry were similar across all demographic and most clinical characteristics. We will assess the severity of COPD by the airflow obstruction grade, using an algorithm previously described by our group [21], rather than the commonly used combined measures such as the GOLD classification or BODE index. Although the BODE index has superior predictive ability for adverse outcomes, its recording is impractical in primary care because of the need for the 6-minute walking test [45]. On the other hand, the GOLD classification has shown poor predictive ability, and the most recent update now assesses airflow obstruction grade separately [46].

BMI is recorded for most patients but may be missing in a non-random pattern. Analytical methods commonly used to overcome missing data include complete-case analyses, missing indicator method (i.e. the use of a category for missing values), single value imputation, and sensitivity analysis incorporating worst- and best-case scenarios. These methods are often used for their simplicity, but they can provide biased estimates if variables are not missing at random. Instead, the use of multiple imputation is often recommended as it can yield estimates closer to those calculated from full data [47-50]. We will investigate the use of multiple imputation when variables are missing at random. For BMI we will also use WHO categories plus a missing category.

As well as missing data, another important issue is censoring due to patients leaving the practice before obtaining their COPD diagnosis or before their first AECOPD. We will include only patients registered for at least two years following their presentation with symptoms, and will therefore exclude those who change GP practice and are diagnosed elsewhere. To assess the potential bias, we will compare patient characteristics of those who receive a diagnosis and then transfer out during the first two years with those who do not transfer out. Patients who do not receive a diagnosis within two years of symptom presentation and then transfer out after two years clearly represent opportunities for improvement regarding early diagnosis, but the transfer out results in censoring and so we do not know if they were subsequently diagnosed with COPD elsewhere.

Similarly, patients who have not had an AECOPD by the time the study period ends or they transfer out of their practice will be compared with those who do not transfer out. If the patient characteristics of the two groups are similar, this will support the generalisability of the results to the whole population. The model coefficient for a given predictor and therefore the estimated risk for a given patient will be biased if the AE rate for those transferring out differs from the AE rate for those remaining.

Patients who die before their first AECOPD will be handled using the competing risks framework of the time-to-event analyses.

6 Outputs, Dissemination and Anticipated Impact

We intend to produce the following as specific outputs:

- Description of the routes to COPD diagnosis both now and ten years ago
- Assessment of variations in diagnosis and management by CCG and GP practice both now and ten years ago
- Ranking of predictors of the first AECOPD from GP, NHS and patient perspectives and estimates of NHS costs of these AEs associated with each predictor
- Validated risk prediction model for the first AECOPD

Through these, we will highlight potential opportunities for improvement and an assessment of whether things have improved or got worse. This will be useful for GPs, commissioners and NICE. Both the methods of measurement using electronic health records and the results could feed into future iterations of the National Audit, aided by co-applicant JKQ's role as Analysis Lead for the Audit, and future updates of NICE guidelines on diagnosis and management. Further, more formal modelling could extend this into the future to see if we will be able to meet the demand with current practice and to establish the economic case for change. Our results could also feed into existing and planned economic modelling work. Some examples of this are: the Socio-Technical Allocation of Resource (STAR), designed to enable stakeholders to explore how to improve the value of health care given constrained resources; published prevalence, healthcare costs and number of deaths in England and Scotland 2011–2030 [51]; and our research group's planned health impact assessment

analysis on the effect of tobacco pricing on smoking and then on COPD incidence, prevalence and outcomes.

A risk prediction model for the first AECOPD would be useful for GPs and patients as part of shared and informed decision-making and early intervention to improve patients' outcomes. Firstly, the predicted risks need to be incorporated into primary care systems. This should be feasible, like the nationally mandated electronic frailty index (eFI) and QRISK algorithm to identify patients at high risk of admission. If we find that the temperature and/or pollution effects are significant, then these would need incorporating either into the predicted risk calculated by the software, which would require an as yet unknown quantity of historical and/or forecasted information, or into the advice given by the clinician during the consultation with the patient. The most practical option will be the latter: the risk model without temperature or pollution data would be used to provide an estimated AECOPD risk and therefore an indication of whether the patient was at low, medium or high risk, and the patient could be advised that their risk would increase on days of e.g. high pollution or cold weather. Further research would be needed to trial the use of the prediction tool in practice. This would need funding from NIHR or industry and buy-in from IT vendors and GPs / practice nurses. A study of the views of primary care practitioners in Wales regarding how predictive models might be used to identify patients for case management interventions to prevent readmission is encouraging [52]. Staff could see possibilities to use the models to offer care more proactively but only if interventions existed to reduce the admission risk of the identified patients and the surrounding support services were available. It has been argued that the necessary linkage between predictive models and actionable opportunities for improving care will most likely be identified through close collaboration between analysts, healthcare practitioners and patients [53]. The distinction between high risk and "impactible" is useful here: impactibility models identify the subset of at-risk patients for whom interventions to prevent disease or poor outcomes such as AEs is expected to be successful, for instance because patients are amenable to behaviour change [54].

The results of the study will be disseminated to all stakeholders including primary care and respiratory clinicians, NHS managers, patient groups (such as the British Lung Foundation and local Breathe Easy groups – our local group has already requested that we present our findings to them), and policy makers. This will be done via conferences (e.g. Health Services Research UK and the International Society for Quality in Healthcare), publications, presentations (e.g. at knowledge exchange events such as the Collaboration for Leadership in Applied Health Research and Care (CLAHRC) for Northwest London Monthly Research Meeting and Collaborative Learning and Delivery events) and targeted messages relevant to each stakeholder in print, electronic and social media.

7 Project / research timetable

Set-up phase: recruit RA; submit ISAC application; apply for mapped temperature and pollution data to be linked to each practice
Year 1 (months 1-12): obtain and prepare CPRD extract; objectives 1 and 2; submit academic outputs
Year 2 (months 13-24): objectives 3 and 4; submit academic outputs
Write-up phase (months 25-27): write report for NIHR; further dissemination

8 Project management

The PI will oversee RA recruitment, data preparation, analysis and write-up. He will be responsible for project management and line management of the RA. The project team, which consists of the academic investigators and the two named PPI representatives, will meet as a whole group at the project start and 2-3 times each year to review progress and

outputs. The PI and co-applicants JKQ and BH will meet with the RA more frequently as required, particularly during the data extract preparation and early analysis phase. The RA will produce the analysis for discussion at the team meetings. The project will have an Advisory Group / Steering Committee, comprising representatives from primary care, commissioning and patients plus an external statistician. This group will meet three times during the project and will provide advice in general but particularly on public and professional engagement and dissemination. The following people have already agreed to join the group:
Noel Baxter, a GP involved in commissioning at NHS Southwark CCG. He will chair the group.
Rishi Kanapathipillai, GP and GP trainer, Lewisham CCG.
David Dullaghan, Operational Manager - Specialist Input, Wandsworth Community Healthcare, Central London Community Healthcare NHS Trust. David is a physiotherapist by background but has taken on a managerial role. He will also provide a CCG perspective.
Mark Joy, external statistician (senior research fellow) at the University of Surrey.
We have recruited one and are in the process of recruiting a further patient representative who are not our co-Is.
Meetings will be held at or near the Royal Brompton Hospital to minimise costs and travel distance for the patients (we are now requesting money for their time and travel).
We will also approach a representative from one of the IT vendors such as EMIS to join the group for the second year of the project to discuss potential ways to implement any resulting model(s) in practice.

9 Ethics / Regulatory Approvals

We have approval from the Secretary of State and the Health Research Authority under Regulation 5 of the Health Service (Control of Patient Information) Regulations 2002 to hold confidential data and analyse them for research purposes (CAG ref 15/CAG/0005). We have approval to use them for research and measuring quality of delivery of healthcare, from the London - South East Ethics Committee (REC ref 15/LO/0824): this will cover this proposal in terms of the analysis. During the project set-up phase, we submitted ISAC forms to seek approval from CPRD to obtain the linked data, link CPRD to the pollution monitoring sites and publish the project outputs: this was obtained in June 2019.

10 Patient and Public Involvement

For Stage 1 of the proposal, we asked for two members of our team's established local PPI group to give their views on the usefulness of the project and to review the lay summary. They both strongly supported the project and approved the lay summary. One was particularly interested in our plan to look at the effect of cold weather, as it is a problem for them. The other, whilst being happy with her GP, knows there are variations in care around the country and is keen for us to investigate this. They would also like the team to present the findings to the whole Breathe Easy group, which we have included as part of the dissemination plan. For the approved project, two other members have agreed to join the project team. They will take part in project meetings, particularly in the first year, to help map out the patient journeys and bring their experience of navigating the NHS and considering the impact of having multiple tests, medications and priorities on patients' health and lives. They will ensure that we capture comorbidities, competing priorities e.g. regarding medications, and other factors that are important to them, where possible with the data. As described in an earlier section, they will help summarise the information on patients' direct experiences of diagnosis that we will obtain from focus groups and other methods. They will also help with lay summaries in the study newsletter and other publications. Two additional patients will sit on the Advisory Group. Training and mentoring will be provided by our experienced local PPI group lead and the project team academics.

11 Project / research expertise

The PI for the project, AB is a senior statistician with expertise in quantitative health services research on measuring and monitoring the quality and safety of healthcare using large NHS databases, particularly HES and CPRD. He was PI on previous NIHR-funded projects on risk adjustment and on predicting outcomes in patients with heart failure and COPD.
JKQ will be the clinical lead for the project. She is a respiratory physician and Clinical Senior Lecturer in Respiratory Epidemiology. She is PI of a research group who have extensive experience in using electronic health records including CPRD to study COPD. She is Analysis lead for NACAP.
BH will provide the main primary care input. He is a practising GP and lecturer in primary care. His practical experience of general practice gives him detailed insight into the coding and interpretation of clinical data in primary care. This has informed recent primary care studies using CPRD and HES to investigate impact of an incentivization scheme for reduction in antibiotic prescribing and track the clinical pathways of patients with heart failure.
AM will provide additional primary care input. He is a practising GP and professor of primary care, heading the host department. He has contributed to a wealth of CPRD-based studies on the delivery of care for chronic diseases and evaluation of relevant policies.
PA will provide policy and epidemiology input and secures the infrastructure to hold and analyse large data extracts. He is a professor of epidemiology and public health at Imperial College London and director of the Dr Foster Unit. His unit investigates variations in performance in healthcare delivery, making extensive use of routinely collected health data such as HES and CPRD. He also leads a research theme within the NIHR funded Patient Safety Translational Research Centre and a work stream within the NIHR funded Health Protection Research Unit for Healthcare Associated Infection and Antimicrobial Resistance.
Lay members Roger Williams and Patricia Craik have had COPD for many years and will contribute their experience of being diagnosed with and living with the condition and navigating the NHS.

12 Success criteria and barriers to proposed work

We will consider the project a success if we meet the objectives, engage the academic community via accepted conference presentations and publications, engage the wider NHS and patient community through dissemination of our results via their platforms (charities, social media, other media etc) and produce actionable recommendations for GPs, practice nurses and policymakers. We particularly expect to influence the next iteration of the National Audit and hope to have an impact on NICE and other clinical guidelines. Other relevant organisations that we will contact are the Primary Care Respiratory Society, who provide quarterly respiratory learning updates and professional development tools and have an annual conference. The RCGP run clinical updates on various topics including respiratory, which we can target, and have e-learning modules on COPD. Some CCGs, though not all, have leads for respiratory medicine, so we will contact them to engage with local practices.

The main potential barrier is CPRD data quality for certain elements: see section “Data quality considerations” above for a full discussion. In brief, the stage 1 panel asked for specific information on the data quality of AECOPD diagnosis when the patient presents to walk-in centres or EDs, smoking, first diagnosis, and medication adherence. Our group has validated the database for the key elements of COPD diagnosis, spirometry and AECOPD, and we are used to running sensitivity analyses in CPRD, for example on the impact of different choices of Read codes and strategies to handle missing data. Regarding the diagnosis coding of walk-in centre or ED presentation, most patients are hospitalised at that visit and so we can use the ICD10 of the resulting inpatient record. If they are not hospitalised, then some cases will be captured in CPRD via the hospital letter, though some

will be missed this way. Smoking is well recorded in CPRD, particularly for current and never smokers, aided by QOF incentives (as are factors such as BMI, HbA1c and BP), but the use of an “unknown” group will likely be needed. CPRD includes information on all medications that are prescribed by the GP; there are some potential options in the data for assessing to what extent the patient actually takes them, which we will investigate. In terms of assessing the prescribing practices of GPs and compliance with guidelines, patient adherence does not matter. For the risk modelling, methods exist for estimating adherence as described earlier, but we acknowledge it as a possible limitation.

We hold ethics permission to hold and analyse CPRD and its linked elements. We will apply for ISAC approval during the project set-up in order to publish results; the team is very experienced with this process, and we do not expect this to be a problem. We have used the air temperature and pollution data before in another project and are familiar with their structure.

Funding acknowledgement and disclaimer

This study/project is funded by the National Institute for Health Research (NIHR) Health Services and Delivery Research (project reference 17/99/72). The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care.