# ISAC APPLICATION FORM

# PROTOCOLS FOR RESEARCH USING THE CLINICAL PRACTICE RESEARCH DATALINK (CPRD)

|  |  |  |
| --- | --- | --- |
| ISAC use only:  Protocol Number  Date submitted | .............................  ............................. | **IMPORTANT**  **If you have any queries, please contact ISAC Secretariat:** ISAC[@cprd.com](mailto:Annalisa.Rubino@gprd.com) |

|  |  |  |
| --- | --- | --- |
| **Section A: The study** | | |
| 1. **Study Title**   Disease clusters and multimorbidity patterns: a UK population cohort study | | |
| 1. **Has any part of this research proposal or a related proposal been previously submitted to ISAC?**   Yes No  *If Yes, please provide previous protocol numbers*: | | |
| 1. **Has this protocol been peer reviewed by another Committee? (e.g. grant award or ethics committee)**   Yes No  *If Yes, please state the name of the reviewing Committee(s) and provide an outline of the review process and outcome:* | | |
| 1. **Type of Study** (please tick all the relevant boxes which apply)   Adverse Drug Reaction/Drug Safety Drug Utilisation  Disease Epidemiology  Drug Effectiveness  Pharmacoeconomics  Methodological  Health/Public Health Services Research  Post-authorisation Safety  **Other\***  \*Please specify the type of study in the lay summary | | |
| 1. **This study is intended for** (please tick all the relevant boxes which apply)**:**   Publication in peer reviewed journals  Presentation at scientific conference  Presentation at company/institutional meetings  Regulatory purposes  Other | | |
| **Section B: The Investigators** | | |
| 1. **Chief Investigator** (full name, job title, organisation name & e-mail address for correspondence- see guidance notes for eligibility)   Prof Aroon Hingorani  Chair of Genetic Epidemiology (Honorary Consultant)  Institute of Cardiovascular Science  Faculty of Pop Health Sciences  University College London  [a.hingorani@ucl.ac.uk](mailto:a.hingorani@ucl.ac.uk)  CV has been previously submitted to ISAC   **CV number:**        A new CV is being submitted with this protocol  An updated CV is being submitted with this protocol | | |
| 1. **Affiliation** (full address)   Farr Institute of Health Informatics Research,  University College London  222 Euston Road London, NW1 2DA, United Kingdom | | |
| 1. **Corresponding Applicant**   Dr Valerie Kuan  v.kuan@ucl.ac.uk  Same as chief investigator  CV has been previously submitted to ISAC  **CV number:**        A new CV is being submitted with this protocol  An updated CV is being submitted with this protocol | | |
| 1. **List of all investigators/collaborators** (*please list the full names, affiliations and e-mail addresses\* of all collaborators*, *other than the Chief Investigator*)   Other investigator:  Prof JP Casas Romero  Farr Institute of Health Informatics Research  University College London  jp.casas@ucl.ac.uk  CV has been previously submitted to ISAC  **CV number:**        A new CV is being submitted with this protocol  An updated CV is being submitted with this protocol  Other investigator:  Dr David Prieto-Merino  Farr Institute of Health Informatics Research,  University College London  d.prieto-merino@ucl.ac.uk  CV has been previously submitted to ISAC  **CV number:**        A new CV is being submitted with this protocol  An updated CV is being submitted with this protocol  Other investigator:  Dr Spiros Denaxas  Farr Institute of Health Informatics Research,  University College London  s.denaxas@ucl.ac.uk  CV has been previously submitted to ISAC  **CV number:** 244\_15E       A new CV is being submitted with this protocol  An updated CV is being submitted with this protocol  Other investigator:  Professor Harry Hemingway  Farr Institute of Health Informatics Research,  University College London  h.hemingway@ucl.ac.uk  CV has been previously submitted to ISAC  **CV number:** 016\_15EL       A new CV is being submitted with this protocol  An updated CV is being submitted with this protocol  Other investigator:  Dr Alireza Moayyeri  Farr Institute of Health Informatics Research,  University College London  a.moayyeri@ucl.ac.uk  CV has been previously submitted to ISAC  **CV number:** 308\_15S  A new CV is being submitted with this protocol  An updated CV is being submitted with this protocol  [Please add more investigators as necessary]*\*Please note that your ISAC application form and protocol* ***must*** *be copied to all e-mail addresses listed above at the time of submission of your application to the ISAC mailbox. Failure to do so will result in delays in the processing of your application.* | | |
| 1. **Conflict of interest statement**\* (please provide a draft of the conflict (or competing) of interest (COI) statement that you intend to include in any publication which might result from this work)   All authors declare no competing interests.  *\*Please refer to the International Committee of Medical Journal Editors (ICMJE) for guidance on what constitutes a COI* | | |
| 1. **Experience/expertise available** (please complete the following questions to indicate the experience/expertise available within the team of investigators/collaborators actively involved in the proposed research, including the analysis of data and interpretation of results   **Previous GPRD/CPRD Studies** **Publications using GPRD/CPRD data**  None  1-3  > 3 | | |
|  | **Yes** | **No** |
| **Is statistical expertise available within the research team?**  *If yes, please indicate the name(s) of the relevant investigator(s)*  Dr David Prieto-Merino |  |  |
| **Is experience of handling large data sets (>1 million records) available within the research team?**  *If yes, please indicate the name(s) of the relevant investigator(s)*  Dr David Prieto-Merino  Dr Spiros Denaxas |  |  |
| **Is experience of practising in UK primary care available within the research team?**  *If yes, please indicate the name(s) of the relevant investigator(s)*  Dr Valerie Kuan |  |  |
| 1. **References relating to your study**   Please list up to 3 references (most relevant) relating to your proposed study:   1. Denaxas SC, George J, Herrett E, et al. Data Resource Profile: Cardiovascular disease research using linked bespoke studies and electronic health records (CALIBER). International Journal of Epidemiology. 2012;41(6):1625-1638. 2. Barnett K, Mercer S, Norbury M, Watt G, Wyke S, Guthrie B. The epidemiology of multimorbidity in a large cross-sectional dataset: implications for health care, research and medical education. Lancet 2012;380:37-43. 3. Hidalgo CA, Blumm N, Barabási A-L, Christakis NA (2009) A Dynamic Network Approach for the Study of Human Phenotypes. PLoS Comput Biol 5(4): e1000353. | | |
| **Section C: Access to the data** | | |
| 1. **Financial Sponsor of study**   Pharmaceutical Industry  *Please specify:*      Academia  *Please specify:*UCL  Government / NHS  *Please specify:*      Charity *Please specify:* Wellcome Trust  Other  *Please specify:*      None | | |
| 1. **Type of Institution carrying out the analyses**   Pharmaceutical Industry *Please specify:*      Academia *Please specify:* UCL  Government Department *Please specify:*      Research Service Provider *Please specify:*       NHS *Please specify:*      Other  *Please specify:* | | |

|  |
| --- |
| 1. **Data source**   The sponsor has direct access to CPRD GOLD and will extract the relevant data\*    A data set will be supplied by CPRD\*\*  CPRD has been commissioned to extract the relevant data and to perform the analyses  Other *Please specify:*         \*If data sources other than CPRD GOLD are required, these will be supplied by CPRD  \*\* Please note that datasets provided by CPRD are limited in size. Applicants should contact CPRD ([KC@CPRD.com](mailto:KC@CPRD.com)) if a dataset of >300,000 patients is required. |
| 1. **Primary care data** (please specify which primary care data set(s) are required)   Vision only (Default for CPRD studies)  EMIS® only\*  Both Vision and EMIS®\*  *Note: Vision and EMIS are different clinical systems, Vision data has traditionally been used for CPRD, EMIS is currently undergoing beta-testing.*  *\*Investigators requiring the use of EMIS data must discuss the study with a member of CPRD staff before submitting an ISAC application*  Please list below the name of the person/s at the CPRD with whom you have discussed your request for EMIS data: |
| **Section D: Data linkage** |
| 1. **Does this protocol also seek access to data held under the CPRD Data Linkage Scheme?**   Yes\*  No  If No, please move to section E.  *\*Investigators requiring linked data must discuss the study with a member of CPRD staff. It is important to be aware that linked data are not available for all patients in CPRD GOLD, the coverage periods for each data source may differ and charges may be applied. Please contact the CPRD Research Team on +44 (20) 3080 6383 or email* [*kc@cprd.com*](mailto:kc@cprd.com) *to discuss your requirements before submitting your application.*  Please list below the name of the person/s at the CPRD with whom you have discussed your request:  Rachael Williams (emails on 3 Dec 2015 and 23 Nov 2015)       *Please note that as part of the ISAC review of linkages, the protocol may be shared - in confidence - with a representative of the requested linked data set(s) and summary details may be shared - in confidence - with the Confidentiality Advisory Group of the Health Research Authority.* |

|  |
| --- |
| 1. **Please select the source(s) of linked data being requested:**   ONS Mortality Data  NCDR Cancer Registry Data\*  Inpatient Hospital Episode Statistics  MINAP  Outpatient Hospital Episode Statistics  Mother Baby Link    Index of Multiple Deprivation  Townsend Score  Other\*\* *Please specify:*       *\*Please note that applicants seeking access to cancer registry data must provide consent for publication of their study title and study institution on the UK Cancer Registry website. They must also complete a* ***Cancer Dataset Agreement Form*** *(available from CPRD) and provide a* ***System level Security Policy*** *for each organisation involved in the study.*  *\*\* If “Other” is specified, please name an individual in CPRD that this linage has been discussed with.* |
| 1. **Total number of linked datasets requested including CPRD GOLD**: 4 |
| 1. **Is linkage to a local dataset with <1 million patients being requested?**   Yes\*  No  *\* If yes, please provide further details:* |
| 1. **If you have requested linked data sets, please indicate whether the Chief Investigator or any of the collaborators listed in response to question 5 above, have access to any of the linked datasets in a patient identifiable form, or associated with a patient index.**   Yes\*  No  *\* If yes, please provide further details:* |
| 1. **Does this study involve linking to patient *identifiable* data from other sources?**   Yes  No |
| **Section E: Validation/verification** |
| 1. **Does this protocol describe a purely observational study using CPRD data (this may include the review of anonymised free text)?**   Yes\*  No\*\*  *\* Yes: If you will be using data obtained from the CPRD Group, this study does not require separate ethics approval from an NHS Research Ethics Committee.*  *\*\* No: You may need to seek separate ethics approval from an NHS Research Ethics Committee for this study. The ISAC will provide advice on whether this may be needed.* |
| 1. **Does this study require anonymised free text?**   Yes\*  No  *\*Please note that work involving free text can only be performed on the July 2013 CPRD GOLD database build or earlier versions. CPRD can provide further advice on the use of anonymised free text.* |
| 1. **Does this protocol involve requesting any additional information from GPs?**   Yes\*  No  \* *Please indicate what will be required:*  Completion of questionnaires by the GP** Yes  No  Provision of anonymised records (e.g. hospital discharge summaries) Yes  No  Other (please describe)        * Any questionnaire for completion by GPs or other health care professional must be approved by ISAC before circulation for completion.* |
| 1. **Does this study require contact with patients in order for them to complete a questionnaire?**   Yes\*  No  *\*Please note that any questionnaire for completion by patients must be approved by ISAC before circulation for completion.* |
| 1. **Does this study require contact with patients in order to collect a sample?**   Yes\*  No  *\* Please state what will be collected:* |
| **Section F: Signatures** |
| 1. **Signature from the Chief Investigator**   I confirm that the above information is to the best of my knowledge accurate, and I have read and understood the guidance to applicants.  Name: Prof Aroon Hingorani  Date: 19.1.16  E. signature (type name): Prof Aroon Hingorani |

**Protocol Section**

1. **Lay Summary (Max. 200 words)**

In an ageing population, and higher survival rates for many conditions more people will develop and live with multiple diseases, leading to decreased quality of life, increased numbers of prescribed medications and higher health care costs. Currently, most clinical guidelines and research focus on single diseases in isolation. This makes it difficult for clinicians to treat patients with co-existing diseases as there is insufficient evidence of which medications for one condition may be beneficial or harmful for others. We propose to study disease patterns and clusters in the UK population to identify key combinations of illnesses which lead to severe adverse outcomes such as death. Patients with or at high risk of these illness clusters can provide the basis for future clinical trials to determine the efficacy and safety profiles of medications in the multiple disease setting. These disease patterns and clusters might also help predict which patients have a higher risk of mortality. This work will help scope the possibility of preventing progression to severe disease outcomes by prescribing the right medications at the right time for the right patients, helping to address the feasibility of the emerging stratified / precision medicine agenda.

1. **Technical Summary (Max. 200 words)**

We aim to describe disease patterns and within-person disease clusters in a representative sample of the UK population using electronic health records from 1997 to 2010 in the CALIBER (CArdiovascular research using LInked Bespoke studies and Electronic health Records) programme. Diseases will be aggregated into broad diagnosis groups modified from the Clinical Classifications Software (CCS) categorisation scheme developed by the Agency for Healthcare Research and Quality (AHRQ). Single and co-occurring prevalence rates for major diseases will be described by age, gender, ethnicity and deprivation index. Comorbidity scores that quantify the strength of disease co-occurrence will be calculated for disease pairs. Using the comorbidity score and p values based on the Benjamini-Hochberg false discovery rate, we will rank and select disease pairs which capture the highest correlations between different disorders. Patients will be stratified based on the similarity of their disease classifications using novel unsupervised machine learning methods involving clustering algorithms. This will allow us to identify the key disease characteristics in each cluster. We will then investigate which combination of key diagnoses lead to severe outcomes in the disease clusters using network analytic measures such as connectivity and lethality.

1. **Objectives, Specific Aims and Rationale**

The broad objective is to characterise age-specific multimorbidity disease clusters in patients from a UK study population using electronic health records in order to identify diseases that occur together more commonly than expected by chance. Significant disease combinations can inform the selection of patients for inclusion in future randomised controlled trials (RCTs) so that these trials properly reflect the multimorbidity in patient populations, and so their results can be readily translated into clinical practice. Examining these clusters by age will allow us to follow the progression of disease clusters and identify patients with precursor conditions that put them at risk of severe outcomes. From a population health perspective, optimal preventative and therapeutic intervention strategies targeting these precursor conditions could limit progression to mortality and severe morbidity.

The specific aims and rationale of this study are:

1. To calculate the age-specific prevalence of multimorbidity, single and co-occurring diseases, stratified by sex in order to outline the epidemiology of the individual and comorbid diseases.

2. To quantify the strength of relatedness between diseases using comorbidity measures such as relative risk and φ-correlation. To rank disease pairs according to their association p-values. A cut-off point will be imposed based on the comorbidity metric and Benjamini-Hochberg false discovery rate. The result will be a list of candidate diagnostic pairs that occur together more often than expected by chance.

3. To calculate similarity measures such as the cosine distance metric between the disease vectors for all possible pairs of patients and use this as the basis for a hierarchical clustering of patients. This will allow us to identify the key disease characteristics in each cluster.

4. To use network analytic measures such as connectivity and lethality to investigate which combination of key diagnoses lead to severe outcomes and are associated with higher risk of mortality in the disease clusters. Interventions that target these key diagnoses could lead to improved population health.

Validation will be performed using data from the UK Biobank. Directions for future research could include studying whether there is a causal relationship between a key diagnosis and other diseases in a cluster. Randomised controlled trials to investigate the efficacy and safety of interventions to prevent or treat the causal condition at an early stage could be designed based on groups of patients with the diagnosis.

1. **Background**

Multimorbidity, defined as the presence of two or more chronic conditions in an individual, is increasingly becoming the norm rather than the exception in ageing populations[[1]](#endnote-1), with associated costs in human and economic terms. Management of patients with several chronic diseases is therefore an important focus for health care systems in developed countries. This presents a challenge to the single-disease focus that pervades medicine[[2]](#endnote-2), with clinical guidelines based on evidence from randomised controlled trials (RCTs) that routinely exclude older patients and those with multiple chronic diseases[[3]](#endnote-3),[[4]](#endnote-4).

Patients with multimorbidities would benefit from a more individualised approach to medicine, with a paradigm shift from the mindset of 'one drug fits all' to ‘the right drug for the right patient at the right dose and time’[[5]](#endnote-5). For this to be economically and logistically feasible, tools need to be developed that accurately group patients into clusters that predict disease progression and treatment outcome. RCTs based on groups of patients with key diagnoses which lead to severe outcomes could inform health policy decisions regarding the optimal stage at which to instigate preventive or therapeutic interventions. The benefits and harms of medical interventions in patients within these disease clusters could be assessed, leading to more rational prescribing in patients with multiple morbidities.

A Scottish cross-sectional study has looked at the prevalence and determinants of multimorbidity with suggestions for changes in the priorities and organisation of health care delivery, clinical guidelines, research and medical education[[6]](#endnote-6). However, this study did not analyse disease correlations or clusters. A large Danish cohort study has analysed temporal disease trajectories using hospital contact data from the Danish National Patient Registry[[7]](#endnote-7). In the UK, diagnoses of many chronic conditions are made or curated in primary care. Hence, analysis of hospital data alone would not be representative of disease patterns in the UK. We propose an investigation of disease patterns using a rich resource consisting of a combination of primary care, secondary care and disease registries in the UK.

1. **Study Type**

The study will be exploratory with a methodological component.

1. **Study Design**

This will be a retrospective cohort study.

1. **Sample Size**

The number of eligible patients in the CALIBER (CArdiovascular research using LInked Bespoke studies and Electronic health Records) platform[[8]](#endnote-8) is estimated at 3,994,541 (see section I). The population age structure of England for 2005 from the Office of National Statistics for males and females in nine age categories is shown in Table 1 in Annex A. Table 2 transposes this age and sex structure to the CALIBER data set with its population of 3,994,541, so that each cell represents the estimated number of individuals within those age and sex categories. Herrett, et al. have shown that the CPRD data set is representative of the UK population in terms of age and sex[[9]](#endnote-9).

We use the χ2 test to assess deviations from the null hypothesis that diseases A and B are independent[[10]](#endnote-10). Statistical power computations are based on the noncentral χ2 distribution. Its noncentrality parameter

λ = *w*2 x *N*

is the product of the sample size *N* and the squared effect size index

*w*=

where

m = number of cells in the contingency table

*p*0i = expected proportion in cell i under H0

*p*1i = observed proportion in cell i under H1

Let the following 2x2 contingency table represent binary comorbidity prevalence rates for diseases A and B for j=0,1, where j=0 represents the expected prevalence rates and j=1 represents the observed prevalence rates:

| Comorbidity rates | Disease A | No Disease A |
| --- | --- | --- |
| **Disease B** | *pj*1 | *pj*2 |
| **No Disease B** | *pj*3 | *pj*4 |

From the definition of w above, the required sample size depends on the prevalences of the individual diseases A and B, as well as the comorbid prevalence of A and B. Given the notation in the contingency table above, the relative risk (RR) as defined in Section L can be rewritten as

RR= =

We are interested in detecting significant disease comorbidities with high public health impact. Therefore we aim to identify pairwise comorbid diseases which occur together more often than expected by chance with prevalence rates > 0.5% and relative risk (RR) >2 (the comorbidity score cut-off point chosen by Roque et al[[11]](#endnote-11); Hidalgo et al used a cut-off value of 20). Table 3 in Annex A demonstrates how the value of w, and hence the sample size, changes with relative risk and observed comorbidity prevalence rates. The function defined in R for this calculation is in Annex B. It is modified from that in the pwr package developed by Stephane Champely. Inputs to the pwr.chisq.test.mod function are:

*w* as in Table 3

df = 1 (for a 2x2 contingency table)

sig.level = 0.05 x 2 /259/258/18 = 8.3e-08, Bonferroni corrected for 259 disease groups (see Section K for further details) and 18 subgroups (nine age categories, two sex categories)

power = 0.8

We can see that the sample size required to identify pairwise comorbid diseases which occur together more often than expected by chance with comorbid prevalence rates > 0.5% and relative risk (RR) > 2 is 13,885 (row highlighted in red and bold). This corresponds to an effect size, w, of approximately 0.05263. Given that each of the subcategories in Table 2 have individuals well above this number, the CALIBER data set is well powered to detect significant disease correlations with high public health impact.

Annex A shows the prevalence rates from 2006 for single and pairwise long term conditions identified as important for health service planning by NHS Scotland[[12]](#endnote-12) (Tables 4 & 5). The chronic disease prevalences reported in the Quality and Outcomes framework for England in 2013-14 are in Table 6. These are very similar to those reported for NHS Scotland. Hence it would be reasonable to assume that the comorbidity probabilities are similar for the England population.

Table 7 in Annex A contains the required sample sizes to detect whether two diseases occur together more often than expected by chance, calculated from Tables 4 and 5. Sample sizes corresponding to w > 0.05263 , and hence to comorbid prevalences > 0.5% and RR > 2 are highlighted in red and bold. As demonstrated before, these are all well below the number of individuals in the age and sex subgroups calculated in Table 2.

These calculations are conservative for several reasons. Firstly, although the Bonferroni adjustment has been used in our sample size calculations, we will apply the Benjamini Hochberg method in our analysis, which is less stringent than the Bonferroni adjustment but will control for the False Discovery Rate.

Secondly, Table 2 gives a snapshot of the number of individuals in a given age and sex category at a single point in time. As the mean follow-up time in CALIBER is estimated to be just above 5 years8, an individual will on average be contributing to 5 such time points over the the study period. Therefore, the estimated number of individuals in each subgroup should be multiplied by 5 to take this into account.

Thirdly, the prevalences in Tables 4 to 6 are based on the whole population. We expect the age and sex stratified prevalences to be significantly higher for chronic diseases and diseases of high public health impact with increasing age.

Finally, we have based the sample size calculations on 259 diseases from Section K. However, the final disease list will be shorter, as the preliminary code list contains administrative codes, codes specific to symptoms and signs instead of diseases, and codes relating to heterogeneous and low prevalence disease groups which will be excluded from the final list.

1. **Data Linkage Required (if applicable)**

As this is an exploratory study involving the investigation of multimorbidity and related outcomes, including mortality, we require data covering the widest possible range of diseases. The entire study will be based on the CALIBER (CArdiovascular research using LInked Bespoke studies and Electronic health Records) platform8, which is linked to CPRD, Office of National Statistics (ONS) mortality and Index of Multiple Deprivation (IMD) and Hospital Episode Statistics (HES) inpatient datasets. This application is for access to CPRD linked data within the CALIBER dataset. We will require CPRD data for our analysis given that diagnosis and management of most chronic diseases occur in primary care. Previous studies have found that the primary care practices in CPRD and the subset of linked practices are representative of the UK primary care setting and have been validated for epidemiological research[[13]](#endnote-13),[[14]](#endnote-14). HES inpatient data will provide information on serious acute illness episodes which may precipitate or be a consequence of chronic disease. ONS mortality data will give us information regarding causes of death, and hence is important in determining disease progression to severe outcomes. IMD data from the CALIBER dataset will allow us to stratify our analyses by deprivation, an important determinant of comorbidity6.

1. **Study Population**

The study period will be from 1 January 1998 to 25 March 2010. All individuals in the CALIBER programme will be eligible for inclusion in the study if they have been registered for at least a year before the start date, their practice is deemed to be contributing ‘up-to-standard' (UTS) data and their individual record is of ‘acceptable' research quality as verified by the CPRD. These patients will be followed up until the earliest of date of death, the date a patient's care was transferred out of a CPRD practice or the last date of the study.

There were 4,703,682 patients registered with 225 primary care practices in CALIBER during the study period. Of these, 135 will be excluded from the study for missing sex variable and 709,006 excluded for having less than one year registration prior to study entry. This will leave 3,994,541 patients included in the study.

1. **Selection of comparison group(s) or controls**

CALIBER phenotype algorithms will be used to define cases and controls. Patients who meet the criteria for a particular disease according to the phenotyping algorithms will be designated as cases for that disease, while those who do not will be defined as controls.

1. **Exposures, Outcomes and Covariates**

Pairwise disease correlations will be analysed for diagnoses from a final disease list, which we will construct from the preliminary code list shown in Annex C. This preliminary list aggregates ICD 10 disease codes into 259 broad diagnosis groups according to the Clinical Classifications Software (CCS) categorisation scheme developed by the Agency for Healthcare Research and Quality (AHRQ)[[15]](#endnote-15). Diseases A and B for which correlations and comorbidity scores are calculated can be considered exposures and outcomes interchangeably. The final list will reflect the granularity and public health importance of diseases based on prevalence and homogeneity as agreed by the clinicians in the research team. In addition, it will exclude administrative codes and codes specific to symptoms and signs instead of diseases. CALIBER phenotyping algorithms which exist for diseases in the final list combining Read, ICD-10, drug and procedure codes will be applied to extract patients for analyses from the study population. These are available on the CALIBER portal[[16]](#endnote-16). We will construct codelists for diseases which are not currently available in CALIBER using the CALIBERcodelists user guide[[17]](#endnote-17).

For morbidities which have lifelong implications such as coronary heart disease, we will define the presence of the morbidity on the basis of a relevant code ever being recorded. However, this is not appropriate for acute illnesses or conditions where full prolonged remission or cure is possible. For morbidities where this is judged to be the case, we will define the presence of the morbidity in terms of a code recorded in a defined period (for example, cancer recorded in the previous five years), or in terms of the presence of a code ever recorded and relevant prescribing in the previous year (which is the way in which the UK Quality and Outcomes Framework defines ‘active’ asthma and epilepsy for payment purposes).

CPRD provides primary care data on demographics, ethnicity, health behaviours, diagnoses, investigations, procedures and prescriptions. Diagnoses are coded using the Read Clinical Terminology. READ codes from CPRD data will be mapped to ICD 10 codes using cross map data files from the Technology Reference Data Update Distribution Service (TRUD)[[18]](#endnote-18).

HES provides information about ICD-10 diagnoses and medical procedures related to all elective and emergency hospital admissions across all National Health Service hospitals in England.

Date and ICD-10 coded cause of death will be obtained from the Office for National Statistics (ONS) death registry. ICD 9 disease codes in the ONS Mortality dataset prior to 2001 will be mapped to ICD 10 codes using General Equivalence Mappings from the Centers for Medicare and Medicaid Services (CMS) available online[[19]](#endnote-19). The index of multiple deprivation according to the patient's area of residence will also be obtained from ONS.

1. **Data/ Statistical analysis**

The analysis will be performed in R. In our main analyses, patients will be stratified into subgroups by age and sex. All the analyses as described below will be performed for each subgroup. In sensitivity analyses, we will additionally stratify by ethnicity and deprivation.

Descriptive analysis

We will calculate the age-specific prevalence of comorbidity (number of diseases in an individual), single and co-occurring diseases stratified by sex, ethnicity and deprivation status.

Disease comorbidity

Patients will be analysed according to predefined age-groups. We will use comorbidity scores to quantify the strength of relatedness between two different diseases A and B. An example of such a score is the relative risk as defined by Hidalgo et al[[20]](#endnote-20):

*RRAB* = , where

CAB = observed number of individuals with both diseases A and B

= expected number of individuals with both diseases A and B

= PAPB/N

N = total number of individuals in the sample,

PA = observed number of people with disease A,

PB = observed number of people with disease B

For each pair of diseases A and B, we will calculate the p values for their association using Fisher's exact test according to a 2x2 contingency table: A&B, A NOT B, B NOT A, NOT A NOT B.

The disease pairs will be sorted and filtered on a cut-off value of the comorbidity score. Several different comorbidity scores may be compared, to mitigate biases in the different methods. The Benjamini Hochberg false discovery rate will be used to correct for multiple testing.

Disease clusters from a patient-patient network

We will develop a clustering algorithm to discover disease patterns in the CALIBER dataset. Firstly, we will construct a patient-disease association matrix for patients with two or more diseases in each age and sex subgroup with binary variables indicating the presence or absence of a particular disease in an individual. In order to quantify the degree of association between patients based on their multimorbidities, we will derive proximity matrices based on (dis)similarity measures such as the Jaccard index, simple matching coefficient, Dice's coefficient or other measures appropriate for binary variables. We will then apply clustering methods based on models such as connectivity or density to these distance matrices. These include single linkage, complete linkage, group average linkage or other methods suitable for binary data[[21]](#endnote-21), [[22]](#endnote-22),[[23]](#endnote-23). This will result in different combinations of distances matrices and clustering methods, which will be applied to the patient-disease matrix.

The choice of (dis)similarity measures and clustering methods within the algorithms will take into account the type of data (binary), the presence of outliers and the natural structure of the clusters. Finch[[24]](#endnote-24) and Li[[25]](#endnote-25) have compared the performances of different clustering algorithms for binary data. In order to decide whether the results of the different algorithms truly characterise the data, the results from the different algorithms will be compared, interpreted and presented with the accompanying strengths and limitations of each algorithm. Different evaluation techniques will be considered in order to measure how well a given partitioning corresponds to the natural cluster structure of the data. Amongst these are internal validation criteria such as the Dunn index, the Davies Boudin index and the silhouette coefficient which can be used to compare different algorithms[[26]](#endnote-26),[[27]](#endnote-27). Binary external indices such as the Jaccard index will also be used to quantify analytically the similarity between clustering results (including those with different numbers of clusters). If conceptually different algorithms generate highly similar partitions, this is a good indicator that actual structure has been discovered.

Lethality

We will define the lethality of a disease as the proportion of people who die a pre-specified number of years (Hidalgo et al used eight years) after first diagnosis with the disease15. Centrality measures for each disease such as degree and connectivity will be obtained from the disease network constructed from the pairwise disease comorbidity analysis described above. In the disease network, nodes will be diseases, while links will connect diseases significantly associated with each other based on the comorbidity score and p-values as defined above.

The degree centrality of a disease is defined as the number of other diseases with which it is significantly associated. This measures the influence a disease has on the network, ie the more diseases it is significantly associated with, the more important it is. The connectivity of a disease is defined as the sum of the comorbidity score between the disease and all other significantly associated diseases. High connectivity values of a disease indicate that it has strong comorbidity to many other diseases in the network.

We will then explore whether the centrality of a disease correlates with higher lethality. The correlation between the various centrality measures and lethality will be calculated. Linear regression of lethality on the different centrality measures will also be performed to characterise diseases with high lethality.

1. **Plan for addressing confounding**

The main analyses will be performed for subgroups stratified by age and sex. Sensitivity analyses will additionally be stratified by ethnicity and deprivation status. However, we would like to emphasise that the analyses in this project are not designed to prove a causal relationship between diseases within a cluster, but aim to provide the basis for further research investigating causality using genetic approaches such as Mendelian Randomisation.

1. **Plan for addressing missing data**

The design of the study assumes that patients who do not meet the criteria according to the coding algorithm do not have the disease. This may introduce biases due to under-reporting or misreporting. Please see section O for further details.

In the main analyses, where patients will be stratified by age and sex, any individuals with missing data for either will be excluded. In addition, patients without data for ethnicity and deprivation will be excluded from sensitivity analyses.

1. **Limitations of the study design, data sources and analytical methods**

Analysis of disease progression might be limited by missed or late diagnosis of conditions. Conditions such as chronic obstructive pulmonary disease (COPD), dementia and mental health conditions could be under-reported, potentially causing underestimates of disease correlations and inaccurate age-related prevalence rates. Errors in disease coding and time of diagnosis may also present problems in data quality. Variation in coding across different practices and over time may introduce systematic biases.

Using different sources of data may result in conflicting data, which differ in diagnostic granularity and timing accuracy. It is difficult to ascertain if multiple records at similar time points reflect a single or recurrent event. There may be errors in data linkage, which can lead to significantly different conclusions about the associations of risk factors with outcomes. However, the linkage method used for linking CALIBER has been shown to yield reasonably high quality matches8.

1. **Patient or user group involvement (if applicable)**

This study has no direct patient involvement.

1. **Plans for disseminating and communicating study results, including the presence or absence of any restrictions on the extent and timing of publication**

Posters, presentations and publications in peer reviewed journals.

**Appendices**

Annex A: Tables and data for calculation of required sample size

Annex B: R code for pwr.chisq.test.mod function used to calculate required sample size

Annex C: Code list for diseases (exposures and outcomes) as aggregated from ICD-10 disease categories using the Clinical Classifications Software

**References:**

1. WHO Global status report on noncommunicable diseases 2010: description of the global burden of NCDs, their risk factors and determinants World Health Organization, Geneva (2011) [↑](#endnote-ref-1)
2. Salisbury C. Multimorbidity: redesigning health care for people who use it. Lancet. 2012; 380(9836):7-9. [↑](#endnote-ref-2)
3. Guthrie B, Payne K, Alderson P, McMurdo MET, Mercer SW. Adapting clinical guidelines to take account of multimorbidity BMJ 2012; 345 :e6341 [↑](#endnote-ref-3)
4. Rothwell PM. External validity of randomised controlled trials: To whom do the results of this trial apply? Lancet 2005;365:82–93. [↑](#endnote-ref-4)
5. Sadée W., Dai Z. Pharmacogenetics/genomics and personalized medicine. Hum. Mol. Genet. 2005;14:R207–R214. [↑](#endnote-ref-5)
6. Barnett K, Mercer S, Norbury M, Watt G, Wyke S, Guthrie B. The epidemiology of multimorbidity in a large cross-sectional dataset: implications for health care, research and medical education. Lancet 2012;380:37-43. [↑](#endnote-ref-6)
7. Jensen AB, Moseley PL, Oprea TI, Ellesoe SG, Eriksson R, Schmock H, et al. Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. Nat Commun. 2014;5:4022. [↑](#endnote-ref-7)
8. Denaxas SC, George J, Herrett E, et al. Data Resource Profile: Cardiovascular disease research using linked bespoke studies and electronic health records (CALIBER). International Journal of Epidemiology. 2012;41(6):1625-1638. [↑](#endnote-ref-8)
9. Herrett E, Gallagher AM, Bhaskaran K, Forbes H, Mathur R, van Staa T, Smeeth L. Data resource profile: clinical practice research datalink (CPRD). Int J Epidemiol 2015;44:827–836. [↑](#endnote-ref-9)
10. Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale,NJ: Lawrence Erlbaum. [↑](#endnote-ref-10)
11. Roque FS, Jensen PB, Schmock H, et al. Using Electronic Patient Records to Discover Disease Correlations and Stratify Patient Cohorts. Ritchie MD, ed. PLoS Computational Biology. 2011;7(8):e1002141. doi:10.1371/journal.pcbi.1002141. [↑](#endnote-ref-11)
12. Information Services Division NHS Scotland. Measuring Long-Term Conditions in Scotland. Edinburgh: ISD Scotland http://www.isdscotland.org/isd/5658.html, 2008 [↑](#endnote-ref-12)
13. Gallagher AM, Puri S, van Staa TP. Linkage of the General Practice Research Database (GPRD) with other data sources. Pharmacoepidemiol Drug Saf. 2011;20:S230–S367. [↑](#endnote-ref-13)
14. Herrett E, Thomas SL, Schoonen WM, et al. Validation and validity of diagnoses in the General Practice Research Database: a systematic review. Br J Clin Pharmacol 2010;69:4–14. [↑](#endnote-ref-14)
15. Clinical Classifications Software for ICD-10 Data:. December 2012. Agency for Healthcare Research and Quality, Rockville, MD. <http://www.ahrq.gov/research/data/hcup/icd10usrgd.html> [↑](#endnote-ref-15)
16. <https://www.caliberresearch.org/portal> [↑](#endnote-ref-16)
17. Shah, A. CALIBERcodelists user guide. 2015. [https://r-forge.r-project.org/scm/viewvc.php/\*checkout\*/pkg/CALIBERcodelists/inst/doc/userguide.pdf?root=caliberanalysis](https://r-forge.r-project.org/scm/viewvc.php/*checkout*/pkg/CALIBERcodelists/inst/doc/userguide.pdf?root=caliberanalysis) [↑](#endnote-ref-17)
18. <https://isd.hscic.gov.uk/trud3/user/guest/group/0/home> [↑](#endnote-ref-18)
19. <https://www.cms.gov/medicare/coding/icd10/2015-icd-10-cm-and-gems.html> [↑](#endnote-ref-19)
20. Hidalgo CA, Blumm N, Barabási A-L, Christakis NA (2009) A Dynamic Network Approach for the Study of Human Phenotypes. PLoS Comput Biol 5(4): e1000353. [↑](#endnote-ref-20)
21. Anil K. Jain, Richard C. Dubes. Algorithms for Clustering Data. Prentice Hall (1988) [↑](#endnote-ref-21)
22. Everitt BS. Cluster Analysis (1993) Edward Arnold. [↑](#endnote-ref-22)
23. A.K. Jain, M.N. Murty, P.J. Flynn. Data clustering: a review. ACM Computing Surveys, 31 (2009), pp. 264–323 [↑](#endnote-ref-23)
24. Finch H. Comparison of distance measures in cluster analysis with dichotomous data. Journal of Data Science. 2005; 3: 85–100. [↑](#endnote-ref-24)
25. Tao Li, A general model for clustering binary data. KDD (2005), pp. 188–197 [↑](#endnote-ref-25)
26. Halkidi M, et al. On clustering validation techniques. J. Intell. Inform. Syst. (2001) 17:107–145. [↑](#endnote-ref-26)
27. Y. Liu, Z. Li, H. Xiong, X. Gao, J. Wu, Understanding of internal clustering validation measures, in: IEEE 10th International Conference on Data Mining (ICDM), 2010, pp. 911–916. [↑](#endnote-ref-27)