QResearch

Research Protocol for an Ethical High-Quality Database for Medical Research

Author: Julia Hippisley-Cox Version: 15.00 Date: 28/10/2025

Full Title: *QResearch Medical Research Database*

Short Title: *QResearch*

Sponsor: Queen Mary University of London (QMUL)

Contact person:

Dr Mays Jawad
Research & Development Governance Operations Manager
Joint Research Management Office
Research Services, Dept. W, 69-89 Mile End Rd, London, E1 4UJ
Email: research.governance@qmul.ac.uk

IRAS Number 329187

REC Reference 23/EM/0166

Chief Investigator (CI)
Professor Julia Hippisley-Cox
Professor of Clinical Epidemiology & Predictive Medicine
Abernethy Building
Whitechapel
London
E1 2AD
Email Julia.hippisley-cox@phc.ox.ac.uk (until 31.01.2025)
Email Julia.hippisley-cox@qmul.ac.uk (from 01.02.2025)

1 Contents

2 Amendment history

Version	Date Issued	Author	Details of Changes Made
1.0	23.01.2003	Julia Hippisley-Cox	Original version submitted to REC and approved 03.04.2003 **MREC/03/4/021**
2.0	10.07.2007	Julia Hippisley-Cox	Updated for substantial amendment 2.
3.0	10.08.2015	Julia Hippisley-Cox	Updated for substantial amendment 3.
4.0	19.11.2018	Julia Hippisley-Cox	Updated references to Nottingham to refer to Oxford instead. Minor updates to formatting and insertion of version 3 of the patient information notice approved in Dec 2017
4.1	28.02.2019	Julia Hippisley-Cox	Minor changes following Oxford internal review: (a) Correct typos, grammar, hyperlinks (b) Add Oxford Logo (c) Add new REC number 18/EM/0400 (d) Amended section 3.4 to match REC application form for data controller and data custodian (e) more detail added to section 3.11, 5.3 and section 8 so that the document reflects existing technical and web-based information. The changes highlighted in yellow
5	26.03.2020	Julia Hippisley-Cox	In response to urgent COVID-19 research (a) added ICNARC data linkage (b) Added TPP data feed from GP practices using this system to create a parallel database called Pan-GP
5.1	01.08.2020	Julia Hippisley-Cox	$2 – updated recruitment figures $5.3.1 Clarification of information regarding the terms of reference for the scientific committee. reviewed by CRTG, changes not considered substantive requiring REC review.
6	11.10.2020	Julia Hippisley-Cox	$3.2 removal of reference to the THIN and ResearchOne database which no longer exists. Addition of the RCGP and OpenSafely databases which now exist $3.2 addition of linkage to blood transfusion and transplant database for COVID research $3.4 added reference to insurance provided indemnity arrangements (PID 14147) $4 clarification that the process includes eligibility check at box two $5.1 clarification to constitution of QResearch Management Board $5.3.1 clarifications regarding roles and decision making. $7 updated to the patient information privacy notice with details on where to find further information
7.0	19.01.2021	Julia Hippisley-Cox	$2 update to purpose of the management board $3.2 and $3.5 addition of data linkages to COVID-19 vaccination; ONS occupation data; cancer screening data, $7 addition of the QCovid algorithm to the patient information sheet Review by Oxford CTRG
7.1	20.01.2021	Julia Hippisley-Cox	Clarification that the data linkage includes COVID-19 test results
7.2	03.02.2021	Julia Hippisley-Cox	Typos corrected
7.3	08.02.2021	Julia Hippisley-Cox	Replies to REC review with typos corrected
8.0	08.02.2021	Julia Hippisley-Cox	$5.1 Update to terms of reference for the QResearch Management Board. $5.3.1 Updated to clarify decision making process for Science Committee $6 minor updates to practice information sheet to match website
8.1	24.02.2021	Julia Hippisley-Cox	$6 typo corrected
9.0	13.12.2021	Julia Hippisley-Cox	$section 3.2 and 3.5 Additional linked data for specialised commissioning data on antiviral and monoclonal antibodies for COVID-19.

			Non-substantial amendment agreed with CRTG 13.12.2021 to include on the REC annual report.
10.0	26.07.2022	Julia Hippisley-Cox	Section 3.2 and 3.5 and section 7. Addition of new data linkages to undertake research into safety of medicines in pregnancy • Maternal confidential enquiry data (lead Marian Knight, MBRRACE-UK) • UK teratology dataset (lead Ken Hodson, UK Teratology Information Service) • NHS number for babies (NHS Digital) • Specialised prescribing for other hospital-only prescriptions (NHS England) • Civil registration births (ONS) • Maternity dataset (NHS digital)
11	08.10.2022	Julia Hippisley-Cox	Update to Patient Information Sheet regarding data linkage and cost recovery model (amendment submitted as part of 2022 Annual Report as agreed with RGEA).
12	28.06.2023	Julia Hippisley-Cox	Removed patient and practice information sheets following RGEA review and advice on 08 June 2023
13	17.11.2024	Julia Hippisley-Cox	Amended University of Oxford to Queen Mary University London (QMUL) as the Data Controller (Julia Hippisley-Cox) is moving from Oxford to QMUL Addition of Signature Page and other sections to align to QMUL JRMO template SOP13a 2b for Research Databases as advised and reviewed by QMUL • Section 3 Glossary and Definitions • Section 4 Abbreviations • Section 5 Signature Page • Section 9 Purpose (now includes the existing data summary in the main text rather than an attachment) • Section 6 Summary table • Section 16 Finance and funding
14	03.10.2025	Julia Hippisley-Cox	Section 11 Amended for CAG s251 application for IM1 data extraction Section 12 added reference to latest DSPT Changed Optum to Optum as Optum has been acquired by Optum/United Health.
15	28.10.2025	Julia Hippisley-Cox	Changes made after CAG review - Correction to date for version 5 in the version history table to 26.03.2020 - Section 9 Amended to confirm which personal data items are removed or de-identified from GP systems under s251 support and which are stored. - Section 15: Amended to confirm that NHS England, rather than PHE and ONS, as the source of secondary care data.

3 Glossary and Definitions

Research database: "A research database is a structured collection of individual-level personal information, which is stored for potential research purposes beyond the life of a specific research study with defined endpoints. Research purposes in this context refers to analysis of data to answer research questions in multiple studies" (Source: HRA).

4 Abbreviations

APR	Annual Progress Report
Barts Health	Barts Health NHS Trust
CI	Chief Investigator
DCC	Data Collection Centres
DI	Designated Individual
GCP	Good Clinical Practice
GDPR	General Data Protection Regulation
HES	Hospital Episode Statistics
HRA	Health Research Authority
HTA	Human Tissue Authority
IRAS	Integrated Research Applications Systems
JRMO	Joint Research Management Office
MBRRACE-UK	Mothers and Babies: Reducing Risk through Audits and Confidential Enquiries across the UK
MHRA	Medicines and Healthcare Products Regulatory Agency
ONS	Office of National Statistics
PI	Principal Investigator
PPI	Patient and Public Involvement
Queen Mary	Queen Mary University of London
RD	Research Database
REC	Research Ethics Committee
RTB	Research Tissue Bank
TCC	Tissue Collection Centres

5 Signature page

CI Agreement

The QResearch Database as detailed within this Protocol, will be conducted in accordance with the principles of Good Clinical Practice (GCP), the UK Policy Framework for Health and Social Care Research, and the Declaration of Helsinki, Human Tissue Act (HTA) (RTBs only), GAMP 5 (RDB only) and any other applicable regulations. I agree to take responsibility for the oversight of QResearch (the Research Database)

CI Name: Julia Hippisley-Cox

Signature: 

Date: 03.10.2025

6 Summary

RTB/ RD Title	QResearch
Data/ tissue collection methods	Computerised uploads of electronic health records. Data will be stored at a QMUl data centre in London.
Type of Data/Tissue	Anonymised linked electronic health records
Aims	QResearch is a high-quality clinical database derived from UK general practices which started in 2003. The aim of QResearch is to develop and maintain a high-quality database of general practice derived data linked to secondary care data for use in ethical medical research. The database is used for medical research into the causes of disease, history, treatment and outcomes. In particular, it has been used for research into health inequalities; safety of medicines; the development of risk prediction tools to identify those at risk of a future disease, such as cardiovascular disease, (www.qrisk.org) or those at risk of a current diagnosis of a disease such as cancer (www.qcancer.org).
Number of samples / data points	40 million patient records
Inclusion and exclusion criteria	**Inclusion**: all patients of all ages registered with approximately 1500 GP practices who have consented to share anonymised medical data for research. Exclusion: Patients who have opted out via the National Data Opt Out (NDOP)
Planned approval duration	**QResearch** was established in 2003 as a REC approved database. The REC approval is requested for 5 years when a renewal will be submitted. The intention is for the approval to be ongoing and long term.

The special features of QResearch are:

- A nationally representative sample of approximately 1500 general practices which use Optum software contribute their data to QResearch.
- A sample of general practices which use TPP SystmOne may also contribute data QResearch.
- The data extracted contains no strong patient identifiers since the data are de-identified at source.
- Practices display a notice in their surgery waiting rooms informing patients that the practice contributes anonymised data for research.
- Patients are able to opt out if they would prefer their pseudonymised data not to be included in the upload.
- The database is open to bona fide researchers employed by UK universities who must be able to publish whatever the findings.
- The researchers will be provided with samples of anonymised GP data that are required to answer their research question (not the whole dataset).
- The data from practices are assessed including their quality (in terms of completeness and accuracy).
- Following review by the QResearch Advisory Board, Derby Research Ethics Committee and the Ethics and Confidentiality Committee of the National Information Governance Board, the entire database has been linked to cause of death data, cancer and hospital data at individual patient level with linkages extending back as far as 1993. The linked data are only available at QMUL.
- Within the aggregated full database, a unique number is assigned to each patient and to each practice to allow longitudinal tracking. Researchers will only have access to fully anonymised data (without the anonymised identifying code).
- Each use of the database is reviewed by the QResearch Scientific Committee.
- The researchers are only allowed to use the data to answer their research question and must return to gain permission for use of the data for other purposes.
- The researchers may not pass on patient data to anybody else.
- The costs of use of the data will be carefully controlled to allow the scheme to be self-funding but to allow good access to bone fide researchers.
- Tabular analyses are undertaken to demonstrate the accuracy and completeness of the data and made available for morbidity analyses.
- Data quality checks will include comparison of birth rates, death rates and prescribing patterns with other aggregated sources.

The key organisational features are:

- QResearch is non-profit-making.
- Management decisions (strategy, vision, pricing, use of funds) are taken by a board representing the interests of patients, practices, researchers and QMUL .
- A QResearch Advisory Group sets policy and oversees the operating of the database.
- A QResearch Scientific Committee approves access to data.

7 Introduction

7.1 Aim

The aim of QResearch is to develop and maintain a high-quality database of general practice-derived data linked to secondary care data for use in ethical medical research. QResearch has created a new, high-quality, primary care-derived database that contains descriptive coded data on the health needs, risks, care and outcomes for a large population. It is one of the highest-quality and largest such datasets in Europe. Research projects suitable for undertaking on the database are those which generate or test hypotheses and are intended for publication in peer-reviewed academic journals. This also includes pilot work to determine sample size calculations and feasibility of specific analyses.

7.2 Background:

While the introduction of IT into the clinical environment in hospitals has been slow and unsatisfactory, general practices have developed clinical systems that are widespread (99%) and well used (over 30% of practices only use their computer for recording their clinical records), and they are often highly accurate and complete. Researchers have begun to exploit the potential of the clinical databases in general practice. Professor Julia Hippisley-Cox leads a team that has published widely using general practice data as their main source. Originally such research was done using high-recording practices in Trent, sometimes recruited through the Trent Focus' Collaborative Research Network, with data being extracted using a program called MIQUEST[1-7]. For research using MIQUEST report style queries this involves writing queries (a skilled and time-consuming task), visiting the practices to run the queries and then integrating the data into a common file (which is again a skilled and time-consuming task). If a search has not run correctly, the practice needed to be revisited.

There are alternatives, CPRD, RCGP and OpenSafely extract GP data. Although useful for surveillance and/or research, these datasets have limitations including:

- Geographical coverage is limited
- There is limited recent information on data quality
- The access costs can be high
- Some are mainly used by the pharmaceutical industry for post-marketing surveillance
- Some do not have Research Ethics Committee approval as research databases
- Some lack a clear application process
- Some have time-limited legal limitations on what research can be done (e.g. may be limited to COVID-19 research under emergency COPI notices or only allow surveillance under regulation 3).

Optum (formerly EMIS) is a general practice computer supplier with their system installed in 5,400 practices (over half of the practices in the UK). Optum has previously explored the creation of a large database with other partners, but discussions have never come to fruition. The QResearch team has the technical and research skills to create and run a large general practice database; it also has good relationships with general practices keen to participate in research.

The combination of Optum and the QResearch team offered sufficient national standing to ensure ownership by the profession. In 2002, we co-opted leading national figures to a pilot advisory group. We undertook a successful pilot project involving 22 practices which are participating in other studies approved by Trent MREC (MREC/01/04/012 and MREC 02/4/052) and Nottingham LREC (P2100201). These studies have compared data extracted by the QResearch methodology with that obtained from MIQUEST and found no important discrepancies. In 2003, following the successful conclusion of the pilot project, the national project was begun and has continued since then with favourable annual reviews by the ethics committee and national QResearch advisory board.

In 2018 Professor Julia Hippisley-Cox was appointed as Professor of Clinical Epidemiology and General Practice at the University of Oxford so QResearch moved to Oxford. In 2025, QResearch moved to QMUL following Professor Hippisley-Cox's appointment as Professor there. It has been agreed that QResearch will relocate to QMUL. Therefore, this protocol was updated to reflect the new location from February 2025.

In 2020, in response to the COVID-19 pandemic, we are expanding the network of practices contributing to QResearch to:

a. include data from GP practices using the TPP SystmOne software in addition to practices using Optum. TPP is used by approximately 40% of GP practices in England and so provides a wider coverage to rapidly enable research into COVID-19 for public benefit; and

b. link to the national Intensive care database (ICNARC) which covers all 185 ITUs in the UK as this enables research into severe outcomes of COVID-19.

c. link to the national Blood Transfusion and Transplant database to allow better understanding of risk factors for COVID and development of risk stratification processes

In response to the COVID-19 vaccination programme in 2021, additional linkages are being undertaken to quantify risk of severe outcomes for COVID-19 patients and research the effectiveness, safety and uptake of therapeutics including COVID-19 vaccinations, antivirals and monoclonal antibodies, including by occupational group (a question of high public interest and potential benefit in the pandemic). This includes

a. COVID-19 test results (antibodies and virus).
b. National Immunisation Database of COVID-19 vaccinations.
c. ONS census occupation data.
d. Specialised commissioning data from NHS England on COVID-19 therapeutics including antivirals and monoclonal antibodies.

e. Maternal confidential enquiry data (lead Marian Knight, MBRRACE-UK)
f. UK teratology dataset (lead Ken Hodson, UK Teratology Information Service)
g. NHS number for babies (NHS Digital)
h. Specialised prescribing for other community and hospital-only prescriptions (NHS England)
i. Civil registration births (ONS)
j. Maternity dataset (NHS digital)

QResearch will also be linked to NHS England's national lung cancer screening data to undertake research regarding the risks and benefits of the new programme.

8 Aims

The aim of QResearch is to develop and maintain a high-quality database of general practice-derived data linked to secondary care data for use in ethical medical research. QResearch has created a new, high-quality, primary care-derived database that contains descriptive coded data on the health needs, risks, care and outcomes for a large population. It is one of the highest-quality and largest such datasets in Europe. Research projects suitable for undertaking on the database are those which generate or test hypotheses and are intended for publication in peer-reviewed academic journals. This also includes pilot work to determine sample size calculations and feasibility of specific analyses.

9 Purpose/Description of QResearch

- QResearch consists of anonymised health records from patients registered general practices.
- QResearch also includes data from hospitals, cancer registry, and mortality
- Data is collected on an ongoing basis
- Data is collected electronically from NHS England and from GP practices (via the GP system supplier).
- Data is stored on servers owned by QMUL.
- Down steam processing will be undertaken to limit datasets for individual research projects to ensure that the dataset provided to the researchers includes the relevant subsets of patients and has the necessary fields.
- Subsets of the original databases are created for individual research projects
- The Research database is used for medical research.
- Data is collected by the research team at QMUL on an ongoing basis.

The data which will be stored in the QResearch includes the following

GP Data

Confidential data items from GP systems (Optum and TPP) are de-identified at the point of extraction. This step is supported by section 251 of the 2002 Control of Patient Information Regulations to allow the disclosure of confirmation information from the GP system to Queen Mary University of London.

Confidential data items which are removed include name, address, postcode, consultation free text. Personal data items which are anonymised include the NHS number, date of birth (which is converted to year of birth) date of death (which is converted to month and year of death). For further details see section 15.

The resulting de-identified data tables are then stored in the QResearch database.

- Demographics
- Prescriptions
- Diagnoses and problems
- Laboratory tests
- Clinical values
- Symptoms
- Consultations
- Referrals

Hospital Data

QResearch is linked to hospital episode statistics provided by NHS England including Admitted Patients Care

- Emergency care
- Critical Care
- Outpatients
- Maternity
- COVID-19 infection, vaccination and therapeutics

Cancer Registry Data

Anonymised information is provided by NHS England on diagnoses of cancer. This is a subset of the information on the National Cancer Registry and includes

- Cancer diagnosis
- Cancer type
- Location of cancer
- Behaviour
- Date of diagnosis
- Histology
- Grade at diagnosis

- Stage at diagnosis
- Treatment (chemotherapy, hormone, radiotherapy)
- Route to diagnosis (e.g. screening, two-week wait)

Mortality data

Anonymised information of deaths which have been recorded on the National Death Register. This includes the

- Date of death,
- Underlying cause of death
- Contributory causes of death
- Country of birth
- Place of death

Intensive Care National Audit & Research Centre (ICNARC)

Anonymised information is provided by ICNARC from their Case Mix Programme (CMP). The CMP is a recognised national audit of patient outcomes from adult, general critical care units (intensive care and combined intensive care/high dependency units).

9.1 Inclusion criteria

QResearch includes data from patients of all ages registered with GP practices who have consented to contribute anonymised data to QResearch

9.2 Exclusion criteria

QResearch excludes any patient who has registered an objection via the National Data Out are excluded.

10 Screening and Recruitment

10.1 Practice recruitment

Optum and TPP practices are invited to participate in the national scheme. See page **Error! Bookmark not defined.** for a copy of the practice information sheet which is also available on the practice website. We have recruited approximately 1500 practices geographically

dispersed throughout the UK. If more practices volunteer than are required, we will randomly select practices, stratifying for partnership size and deprivation.

The following apply to all uses of the data:

- The QResearch databases operate as not-for-profit at QMUL.

- The accounts will be transparent to the parent organisations and the QResearch Advisory Group; access fees will be agreed in order to ensure fair and reasonable access to researchers and other users, while ensuring the efficient operation of the databases.

- The extraction of data from GP practices and the general methodology used will be covered by research ethics committee approval.

- The QResearch Advisory Board will oversee the operation of the databases, including setting the criteria for access by prospective users.

- The QResearch Scientific Committee will be required to give prospective consent for access to the databases for all users

- All users will be provided with data that are appropriate to their requirements. For researchers that will normally mean a patient-level analysis (a file that contains records of anonymised individual patient data), only containing variables that relate to their hypothesis and a sample size sufficient to answer the research question. For others it will mean a tabular output containing no patient-level data. For further details, see below.

- Requests for the provision of data will be risk-assessed by the QResearch Scientific Committee to ensure the highest protection for patient confidentiality. Users will be required to sign an undertaking that they will not try to attempt to identify any patient(s) or practices. All users will give signed assurance on the following questions:

  - To your knowledge is this work original and capable of publication as original research in a peer-reviewed journal?
  - Are you free to undertake this study and publish its findings without needing to clear it with the funding source or any other organisation?
  - Do you agree to acknowledge the source of QResearch data in any publication, paper, report or software/tool?
  - Do you agree NOT to attempt to identify patient(s) or practice(s)?
  - Do you undertake to provide a copy of the final report of the project and copies of any publications within one year of the project completion?
  - Do you agree NOT to release the data to any third party including the funder, sponsor or other such body?

- Do you agree not to use the data for any other project except that which is expressly described in your protocol?
- Do you undertake to check the data you are given within a month of receipt and report back any problems within that time?
- Do you have a statistician on the project team who has contributed to the design of the study and will advise on the analysis?
- Do you agree to have a project summary on the QResearch website once the project starts?

11 Ethics and Research Governance

11.1 Ethics Approval and annual reporting

- The Chief Investigator and controller of the data for the ethics committee as lead applicant is Professor Julia Hippisley-Cox - Professor of General Practice and Epidemiology within the Wolfson Institute of Population Health, QMUL and an NHS General Practitioner. QResearch protocol has been reviewed by QMUL and has insurance-provided indemnity arrangements.

- QMUL is responsible for ensuring the research governance required.

- The Derby Research Ethics Committee is responsible for ethical approval. Along with the QResearch Advisory Board, the ethics committee is in receipt of a report from QResearch on an annual basis.

11.2 Confidentiality Advisory Group Approval

- 2003: In order to determine whether Section 60 support was necessary to cover the process of anonymisation/pseudonymisation in 2003, we contacted Sean Kirwan from the Department of Health with a copy of the protocol and details of the processes to be used. He advised us that Section 60 support was necessary only when patient-identifiable information is required and it is not practicable to either obtain patient consent or use anonymised/pseudonymised data. With the process of pseudonymisation employed in QResearch, no patient-identifiable information is shared with, or processed by, a third party (i.e. an individual or organisation not employed by the GP practice) and hence Section 60 support is not required for the QResearch database.

- 2011: Advice regarding the need for section 251 support (which replaced section 60 approval in 2008) was sought in 2011 from the Ethics and Confidentiality Committee (ECC) of the National Information Governance Board. This was in order to link general practice data to the HES, Cancer Registry and mortality data at individual patient level. The linked cancer, mortality and hospital data enables researchers to analyse additional information on patient characteristics, treatment and outcomes which will improve the epidemiological analyses of studies since the data will be more complete (the QResearch database without linkage does not capture all this information

reliably). The linked data also contain additional detail to allow research into the causes and outcomes for major diseases including the development and validation[8] of tools designed to assess risk of cardiovascular disease[9-11], osteoporotic fracture[12], risk of current cancer[13-20] and future cancer[21]. Without the data linkage, for example, research may underestimate the incidence of cancer[21], or the risk associated with interventions such as prescribed medicines[22-27].

- 2013: Following a detailed review, the Confidentiality Advisory Committee concluded that, given that (a) no strong identifiers were extracted from the system, (b) there was an irreversible pseudonymisation of the NHS number prior to disclosure from the source system, and (c) the strong IG controls in place at the University, the extraction did not constitute identifiable data and hence section 251 support was not required. For further details of the method for data linkage see page 27.

- 2025: Due to a technical change in platform used to extract GP data from GP system suppliers (Optum and TPP), section 251 support is likely to be required for access to identifiable GP data at the point of data extraction from the IM1 platform. Previously the deidentification step was undertaken by the GP System Supplier immediately prior to data extraction but this is no longer technically possible.

- De-identification software will be run in the secure environment to de-identify key data items from the GP data required for data linkage at the point of data extraction (NHS Number, date of birth, date of death, address and postcode). This process is likely to be undertaken once or twice a year.

- The data items will be de-identified immediately the data are extracted using strong cryptographic techniques which apply a one-way project specific hashing algorithm to these data items. No identifiers will be stored into the same data as the clinical data.

- The encryption key would be held by NHS England in Azure key safe and only available to QMUL whilst the software run.

- Only the minimum necessary deidentified/anonymised data items will be stored in the QResearch database. S251 support is then not required for linkage or analysis as this is done using the anonymised data. The exit strategy is for NHS England to enable the incorporation of the deidentification/anonymisation software process into the IM1 platform itself.

12 Computer systems

This section describes the computer systems and arrangements for the handling of the data. Technical details are available in the document entitled “Systems Levels Security Policy QResearch Data Linkage” and Data Security and Protection Tool Kit (8HX86) which are reviewed annually by QMUL, and NHS England. The next section provides a lay summary.

There are several main computers (servers) involved in the QResearch project.

- The data collection servers at the GP system suppliers. These servers will be linked to practices via the NHSnet in order to undertake the triggered upload ONLY after the practice has authorised the upload by activating the QResearch module within its surgery system.
- The research servers, which house the resulting aggregated databases, one for each database, and which will be located at QMUL. QMUL will be the single point of access to the data collected by QResearch.

Each of the servers at the GP system suppliers and at QMUL, is used solely for the purposes of QResearch.

GP data from practices using TPP systems will be de-identified and transferred as anonymised data to QMUL using the same principles and strong ethical, information governance and security controls as are in place for the data feed from Optum.

The GP system supplier only transfers data under this agreement to one organisation, namely QMUL. The data transfer is secure as the data are de-identified prior to extraction and the data are also encrypted.

The GP system supplier and QMUL are contractually bound not to use the data collected by QResearch for any other purpose than that stated within this protocol.

a. The QResearch database consists of a triggered upload of all coded data from patients registered with the participating practices. By coded data, we mean all computer entries which have been coded using the Read or other similar code classifications. No clinical free text will be extracted.
b. The GP system supplier patches a look-up table to the practice which maps patients' postcodes to the census variables (such as deprivation scores and rurality) associated with the relevant electoral ward or enumeration district. These data are uploaded into the patients' records within the practice system and hence they can be extracted without the need for any postcode information to ever leave the practice.
c. The uploads are undertaken by the GP system supplier after the practices have given informed consent and activated the upload of QResearch data from within their own practice system.
d. De-identification of strong patient identifiers is undertaken at the point of data extraction. No strong patient identifiers are then stored from the general practices contributing to QResearch. Each patient will be assigned a unique code (anonymised) in order to maintain the chronological integrity of the database and to allow follow up of individuals and cohorts. Full details of anonymisation are given below.
e. Since 2011, QResearch has been linked to three additional data sources - Hospital Episodes Statistics database, cancer registries and ONS mortality registrations. The data linkage is undertaken using an identifier on the QResearch database without any strong patient identifiers being supplied from any of the source systems (see further details of the data linkage methodology in section 8).

f. In 2020, QResearch was linked to the national ITU (ICNARC) database to support COVID-19 research by including data from severely ill patients admitted to Intensive Care.
g. In 2021, QResearch was linked to data from ONS, transplant data from NHS BT, COVID-19 vaccination data from the National Immunisation Database, specialised therapeutics antivirals and monoclonal antibodies for COVID-19 and to the lung cancer screening database.
h. In 2023, QResearch was linked to additional datasets to enable evaluation of safety of medication in pregnancy including the Maternity Services Dataset.
i. Future linkages include Maternal confidential enquiry data (lead Marian Knight, MBRRACE-UK); UK teratology dataset (lead Ken Hodson, UK Teratology Information Service); NHS number for babies (NHS Digital); Specialised prescribing for other hospital-only prescriptions (NHS England); Civil registration births (ONS);
j. Incremental uploads are undertaken by the GP system supplier and new versions of the databases are transferred to QMUL at intervals as required to meet the needs of the research.
k. Patients will be able to request that their data are not included in the anonymised upload from the practice. This is implemented by the National Data Opt Out system implemented by the GP system supplier which allows the data from individual patients to be filtered out of the data collection.

There are three main issues regarding the security of the data on the servers. These are:

(1) process of de-identification – i.e. the measures which are taken to ensure complete confidentiality of patients and also of participating general practices
(2) the physical security of the server - measures to restrict physical access and prevent theft
(3) electronic security of the server - measures to prevent unauthorised access and monitor authorised access

12.1 De-identification and anonymisation

As noted in section 11, de-identification software will be run in the secure environment to de-identify key data items required for data linkage at the point of data extraction (NHS Number, date of birth, date of death, address and postcode). This process is likely to be undertaken once or twice a year. The data items will be de-identified immediately the data are extracted using strong cryptographic techniques which apply a one-way project specific hashing algorithm to these data items. The encryption key would be held by NHS England and not available to QMUL except whilst the software run. Only the minimum necessary deidentified/anonymised data items will be stored in the QResearch database. S251 support is then not required for linkage or analysis as this is done using the anonymised data. Each patient is allocated a unique number (known as a GUID). This GUID is used by the practice system to allocate later data to the same patient file. The collection server cannot re-identify which patient the GUID refers to. This additional protection prevents the potential for the GUID from the research database being taken back to the practice, the database being illegally accessed and the GUID cross-referenced back to the patient.

Researchers, having gone through the process of approval, will be given, if appropriate, access to the GP records which contain records for individual patients. However, these records will not contain a GUID.

When the database is interrogated for information for morbidity analyses, the results will not contain any records for individual patients. The outputs will be in tables or graphs and we refer to these as tabular analyses.

Further detail of the anonymization process is included below.

12.2 Security arrangements

12.2.1 Physical security

The QMUL data collection server has two critical security-related roles.

1. It de-identifies data from GP practices at the point of extraction from the IM1 system.
2. It assembles the Qresearch database to be used by all subsequent users of the data for secure onward transmission

Therefore, it is a security requirement that the servers are hosted in a secure data centre with full NHS security clearance, personnel access restrictions and physical access obstacles. This includes steel doors, ID cards, closed-circuit television etc. Physical access to each server is restricted to specifically named engineering staff who use strong authentication to gain access for the purpose of hardware repair.

12.2.2 Ensuring authorised access

Only a limited number of named support personnel have access to the QMUL dedicated servers. This includes the relevant software support manager and up to two support staff. All databases are password protected, and in addition hardware access authentication is used. Logs of all access to the servers are maintained.

12.3 Security arrangements

The main issues for data security in are guaranteeing physical security and preventing unauthorised access.

12.3.1 Physical security

The physical servers will be in a locked room at QMUL with restricted access (named key holders). There is CCTV in the building. The computer itself is in a metal cage which is locked and secured to the floor.

12.3.2 Ensuring authorised access

Named individuals who have access to the computer are bound by confidentiality clauses in their contracts. Only Julia Hippisley-Cox and the IT support John Croasdale of Dancing Houses Ltd. have access. They will control all accesses to the database on a daily basis as described in the QResearch Systems Level Security Policy (SLSP). The Advisory Board are notified if a new member of staff is required to access the database directly. The data on the research computer are encrypted and passwords are required to access the data. As with the Optum and TPP servers, all accesses to the data will be logged (time, user) using electronic tracking software.

12.3.3 Practice or patient identification

One named member of staff in QMUL and one in each GP system supplier will have a list of the practices which have given informed consent to participate in QResearch. This list is be kept on a separate computer from the GP system supplier file server or the QResearch server in QMUL, and will be encrypted. The list of participating practices will not be released to other individuals or organisations. There are no patient identifiers on the databases because of the anonymisation process outlined above. In this way, patient confidentiality is completely secure.

Members of the Advisory Board will undertake, as a minimum, annual site visits to check the adequacy of the security measures. The Advisory Board will decide the frequency of the site visits. Logs of all access to the computer will be made available to the Advisory Board on request.

13 Process for accessing data:

13.1 Users needing access to patient or practice level data

Users requiring access to patient-level GP data will need to fulfil the following criteria:

- To your knowledge is this work original and capable of publication as original research in a peer-reviewed journal?
- Are you free to undertake this study and publish its findings without needing to clear it with the funding source or any other organisation?
- Do you agree to acknowledge the source of QResearch data in any publication, paper, report or software/tool?

- Do you agree NOT to attempt to identify patient(s) or practice(s)?
- Do you undertake to provide a copy of the final report of the project and copies of any publications within one year of the project completion?
- Do you agree NOT to release the data to any third party including the funder, sponsor or other such body?
- Do you agree not to use the data for any other project except that which is expressly described in your protocol?
- Do you undertake to check the data you are given within a month of receipt and report back any problems within that time?
- Do you have a statistician on the project team who has contributed to the design of the study and will advise on the analysis?
- Do you agree to have a project summary on the QResearch website once the project starts?

Users of output from the QResearch database must provide QResearch with copies of publications or reports. These will be made available to the QResearch Advisory Group and, unless confidential, to the QResearch practices on request.

The data linked to mortality, cancer registrations and hospital episode statistics are only available on the secure QResearch server hosted by QMUL. The linked data are not made available outside of QMUL servers.

13.2 Users needing access to tabular output

Most analyses resulting in tabular output will be either using it for establishing evidence to use in research applications including pilot studies and sample size calculations (referred to here as “pilot studies”) or for describing care need, care or outcomes in general practice (referred to here as “morbidity analyses”)

- Those wanting to undertake research studies must fulfil the following criteria:
    - There will be a named principal investigator and named co-investigators
    - There will be a written protocol with a clear statement of the intended research question, the pilot data required and an intention to develop a full research protocol
    - They will agree to acknowledge QResearch as the source of the pilot data in any application, publication or report
    - The QResearch Scientific Committee will need to give approval within the guidelines set out by the QResearch Advisory Board before analysis can occur and data can be supplied to the user
- The QResearch team undertake morbidity analyses in order to establish the accuracy, completeness and functionality of the QResearch database.

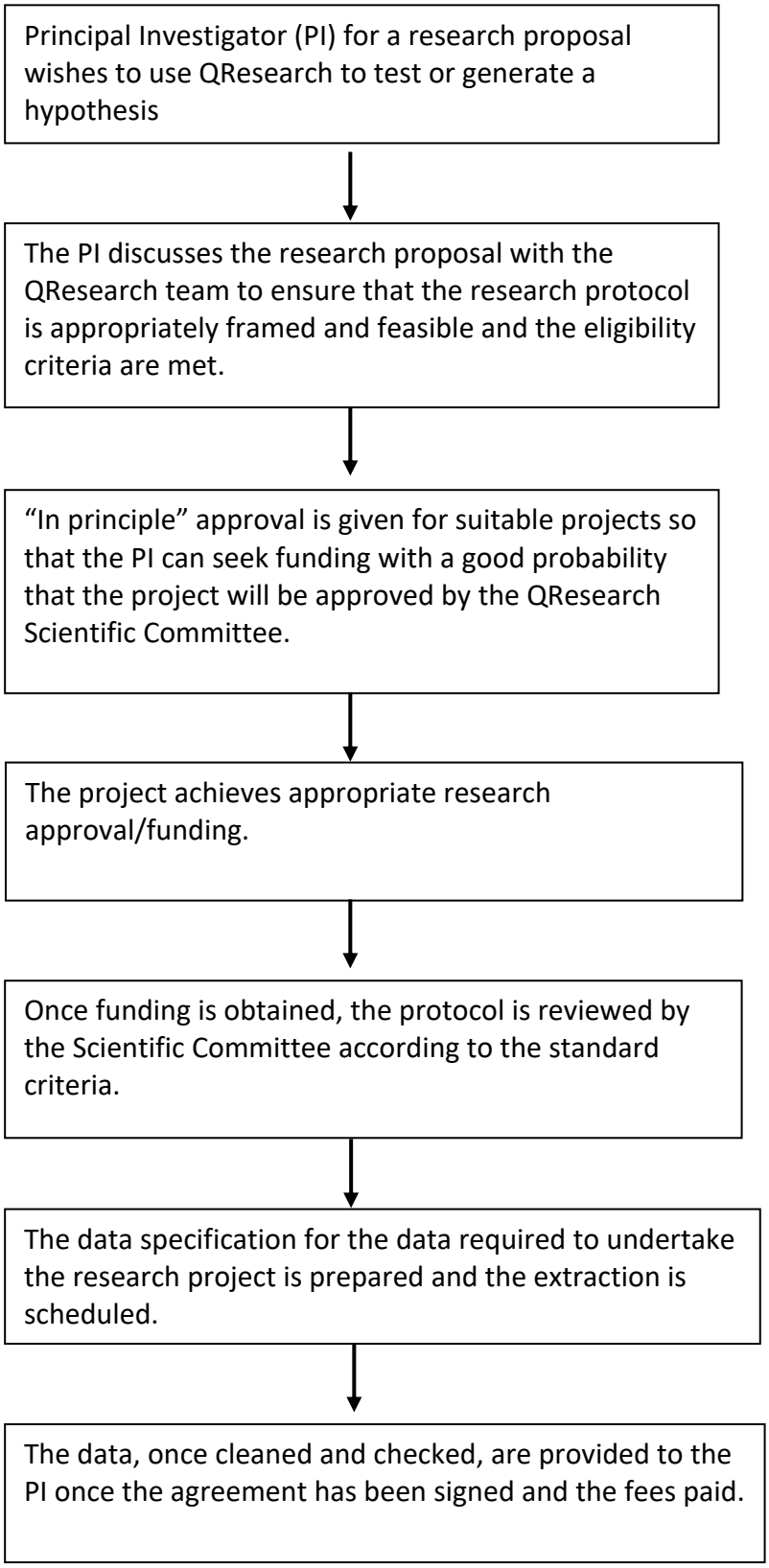
- It is recognised that such analyses may be of general value to a range of individuals and organisations. It is intended therefore to make such analyses, where they have been undertaken, publicly available through the related publications on the QResearch website.

13.3 Criteria for access to the QResearch databases

The QResearch Advisory Board draws up the criteria for access that will be applied by the QResearch Scientific Committee. It is anticipated that the QResearch Advisory Board may wish to refine the criteria set out in this document over time at the annual review. The up-to-date criteria are published in the QResearch website (www.qresearch.org).

External researchers applying for access will be academics employed by other UK universities where the university has signed a sublicense agreement and the individual researchers have signed data-sharing agreements and completed IG training.

We process requests for access to data as follows:



14 Management Committees

There are two boards/committees

(1) QResearch Advisory Board
(2) QResearch Scientific Committee

14.1 QResearch Advisory Board

In order to oversee the ongoing development of QResearch we have appointed an Advisory Board. Details of the advisory board are published here

https://www.qresearch.org/about/advisory-board/advisory-board-in-detail/

The minutes of the advisory board are published here

https://www.qresearch.org/about/advisory-board/

14.1.1 Terms of reference/remit

- To oversee the general working of QResearch including the handling of the data, the type of analyses undertaken and access to the database.
- To oversee communication with and benefits back to contributing patients and practices.
- To agree and update the criteria and principles for access to the QResearch database and oversee their application.
- To review any changes to the context of the data extracted for QResearch or changes to the terms of its use.
- To advise on policy for accessing data (i.e. how quickly should researchers get data; how much should it cost).
- To offer general advice on professional issues to the QResearch management team.

14.1.2 Membership

The board represents the key stakeholders in QResearch in order to gain and retain the respect of the public, the NHS, practices and the research community. Its membership includes representatives of:

- QMUL
- The Royal College of General Practitioners
- The British Medical Association itself and its General Practitioners Committee
- Society for Academic Primary Care
- EMIS National User Group
- Optum
- Patient representation including National Association for Patient Participation
- Independent patient representation

14.2 QResearch Scientific Committee

https://www.qresearch.org/about/scientific-committee/

The committee is a virtual committee consisting of the peer reviewers selected for each application. At least two peer reviewers are selected to review each application. Reviewers are asked to declare conflicts of interest and alternative reviewers are found where conflicts are present.

14.2.1 Terms of reference/remit

- To advise whether the research meets a minimum scientific standard and if not, what amendments are required
- To follow the criteria and principles set out by the QResearch Advisory Board in assessing and advising on applications to access the QResearch database.
- To assess the risks involved in each application for use of the QResearch database and to refer difficult areas for decision to the QResearch Advisory Board.
- To advise the QResearch team on technical issues.
- To advise on whether the research application has a clear research question or hypothesis which is likely to lead to generalisable findings capable of publication in a peer-reviewed medical journal.
- To advise whether the research team is likely to be able to conduct the study and its analysis.

Julia Hippisley-Cox is the Chief Investigator for QResearch and is responsible for ensuring that data access is provided in accordance with the protocol, ethics approval for the research database and following the advice of the Advisory and Scientific Committees.

The peer review forms used by the Scientific Committee include the following questions, with yes/no options and space for comments.

1. Is there a clear research question or hypothesis which is likely to lead to generalisable findings, capable of publication in a peer-reviewed medical journal?
2. Are the researchers likely to be able to conduct the study and its analysis?
3. Is QResearch the appropriate database to be used to conduct the research?
4. Is the methodology appropriate to answer the question (including the possibility of bias and confounding)?
5. Are there any potential risks to the ethical position of QResearch in undertaking this research (including the potential identification of patients or practices)?
6. Do you have any conflicts of interest in reviewing this application? If so, please give brief details
7. Overall, do you think QResearch should approve this research project? Please select one of the options

- Yes
- Yes, but needs modification (if so, please say what)
- No
- Refer to the QResearch Advisory Board

8. Please give details of any modifications or general comments

If modifications are suggested, the applicant will be given opportunity to make them and the revised protocol will be reviewed for a further time.

15 Data handling, storage and record keeping

QResearch adopts a 'anonymisation-at-source' approach in which the data are irreversibly anonymised before the data leave each of the four source systems (GP, ONS mortality, ONS cancer registration, Intensive Care and Hospital Episode Statistics).

NHS England (which is the source for the Hospital Episode Statistics, mortality, cancer registration, COVID, prescribing and maternity data) anonymises the NHS number on the full HES database before it leaves NHS Digital. The GP system suppliers, Optum Health and TPP are the source of the GP data. Exactly the same anonymisation technique within the source system at the point the the data are extracted from the system. This means that the same ID can be generated and can used to link the records from each data sources together.

The steps are as followed:

a. The software applies a project-specific 'salt code' (which is a random text word similar to a password) which is held securely by NHS England in a secure key safe in Amazon Web Services. This salt code is combined with the hashing algorithm to ensure that the resulting ID is unique to QResearch project. The 'salt code' and software is to be kept confidential by the source systems and not used for any other project.
b. The OpenPseudonymiser software is then applied to a comma separated file of source data using a secure one-way hashing algorithm to the NHS number which generates a unique ID.
c. The OpenPseudonymiser software then removes the NHS number from the data file so that resulting data file has ID but no NHS number. The batch processor will also replace the full date of birth with the year of birth and remove other strong identifiers such as postcode if present on the data file.
d. The de-identified data file is then encrypted by the source supplier before it is securely stored in QResearch at QMUL.
e. On receipt of the anonymised data at the University, the ID is used to link the data files together. The ID is further transformed to a second ID for use within the linked databases.

f. No other stronger identifiers are extracted from the any source system (i.e. no name, NHS number or postcode or full date of birth).
g. There is no disclosure by QResearch of any patient or practice information which could directly or indirectly lead to the identification of patients or practices contributing to the QResearch or Pan-GP databases.

The source identifiers are retained by the source systems and not sent to QMUL. Thus, on arrival at QMUL, data are anonymised. Each of the source systems retains the salt key. This salt key is not shared with QMUL.

This GUID is used by the source system to allocate later data to the same patient file.

The collection server cannot identify which patient the GUID refers to. This additional protection prevents the potential for the GUID from the research database being taken back to the practice, the database being illegally accessed and the GUID cross-referenced back to the patient. No-one is able to track back to the practice of origin from the resulting database. All these data are referred to as anonymised.

16 Finance and Funding

QResearch uses a cost recovery-based costing model to attribute the actual cost of providing QResearch data access between users according to the size and scale of their project.

The cost of using QResearch includes:

- pre-application support from the QResearch team
- provision of support to make an application to the QResearch Scientific Committee
- support to develop the data specification
- review by the QResearch Scientific Committee
- extraction, manipulation and linkage of data to create the specified dataset
- support to meet the requirements to access the QResearch server
- access to the dataset and agreed software to work with the dataset on the QResearch server for the duration of the project
- operating costs including; fees from our data providers, server software and hardware costs, and server service costs.

Researchers recover the costs associated with database access via research grants. QResearch publishes a list of research funders here

https://www.qresearch.org/information/funding-sources/

17 Insurance and Indemnity

The insurance that Queen Mary has in place provides cover for the design and management of the study as well as "No Fault Compensation" for participants, which provides an indemnity to participants for negligent and non-negligent harm.

18 References

1. Hippisley-Cox, J. & Pringle, M. Are spouses of patients with hypertension at increased risk of hypertension? A population based case-control study. *British Journal of General Practice* **46**, 1580-1584 (1998).
2. Hippisley-Cox, J. & Pringle, M. Depression and risk of ischaemic heart disease in men: authors' reply. *BMJ* **317**, 1450 (1998).
3. Hippisley-Cox, J., Pringle, M., Crown, N., Meal, A. & Wynn, A. Sex inequalities in ischaemic heart disease in general practice: cross sectional survey. *BMJ* **322**, 832 (2001).
4. Hippisley-Cox, J*., et al.* Antidepressants as risk factor for ischaemic heart disease: case-control study in primary care. *BMJ* **323**, 666-669 (2001).
5. Hippisley-Cox, J., Coupland, C., Pringle, M., Crown, C. & Hammersley, V. Married couples risk of same disease: cross sectional study. *BMJ* **325**, 636-638 (2002).
6. Hippisley-Cox, J., Cater, R., Pringle, M. & Coupland, C. A cross-sectional survey of the effectiveness of lipid lowering drugs in lowering serum cholesterol in 17 general practices:how well do they work? *BMJ* **326**, 689-694 (2003).
7. Hippisley-Cox, J., Pringle, M., Crown, N. & Coupland, C. Does hormone replacement therapy protect against ischaemic heart disease? Evidence from a general practice based case control study. *BJGP* **53**, 191-196 (2003).
8. Hippisley-Cox, J., Coupland, C. & Brindle, P. The performance of seven QPrediction risk scores in an independent external sample of patients from general practice: a validation study. *BMJ Open* **4**, e005809 (2014).
9. Hippisley-Cox, J*., et al.* Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. *BMJ*, bmj.39609.449676.449625 (2008).
10. Hippisley-Cox, J*., et al.* Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study. *BMJ*, bmj.39261.471806.471855 (2007).
11. Hippisley-Cox, J., Coupland, C., Robson, J. & Brindle, P. Derivation, validation, and evaluation of a new QRISK model to estimate lifetime risk of cardiovascular disease: cohort study using QResearch database. *BMJ* **341**, c6624 (2010).
12. Hippisley-Cox, J. & Coupland, C. Derivation and validation of updated QFracture algorithm to predict risk of osteoporotic fracture in primary care in the United Kingdom: prospective open cohort study. *BMJ* **344**, e3427-e3427 (2012).
13. Hippisley-Cox, J. & Coupland, C. Identifying patients with suspected gastro-oesophageal cancer in primary care: derivation and validation of an algorithm. *British Journal of General Practice* **61**, e707-714 (2011).
14. Hippisley-Cox, J. & Coupland, C. Identifying patients with suspected lung cancer in primary care: derivation and validation of an algorithm. *British Journal of General Practice* **61**, e715-723 (2011).
15. Hippisley-Cox, J. & Coupland, C. Identifying women with suspected ovarian cancer in primary care: derivation and validation of algorithm. *BMJ* **344**(2012).

16. Hippisley-Cox, J. & Coupland, C. Identifying patients with suspected colorectal cancer in primary care: derivation and validation of an algorithm. *British Journal of General Practice* **62**, e29-e37 (2012).
17. Hippisley-Cox, J. & Coupland, C. Identifying patients with suspected pancreatic cancer in primary care: derivation and validation of an algorithm. *British Journal of General Practice* **62**, e38-e45 (2012).
18. Hippisley-Cox, J. & Coupland, C. Identifying patients with suspected renal tract cancer in primary care: derivation and validation of an algorithm. *Br J Gen Pract* **62**, e251-260 (2012).
19. Hippisley-Cox, J. & Coupland, C. Symptoms and risk factors to identify women with suspected cancer in primary care: derivation and validation of an algorithm. *Br J Gen Pract* **63**, 11-21 (2013).
20. Hippisley-Cox, J. & Coupland, C. Symptoms and risk factors to identify men with suspected cancer in primary care: derivation and validation of an algorithm. *Br J Gen Pract* **63**, 1-10 (2013).
21. Hippisley-Cox, J. & Coupland, C. Development and validation of risk prediction algorithms to estimate future risk of common cancers in men and women: prospective cohort study. *BMJ Open* **5**(2015).
22. Coupland C, Hill T, P, B., V, V. & Hippisley-Cox J. Thiazolidinedione drugs and the risk of cancer in patients with diabetes: nested case-control studies using a primary care database. *SAPC conference 2015* (2015).
23. Vinogradova, Y., Coupland, C. & Hippisley-Cox, J. Exposure to bisphosphonates and risk of common non-gastrointestinal cancers: series of nested case-control studies using two primary-care databases. *Br J Cancer* (2013).
24. Vinogradova, Y., Coupland, C. & Hippisley-Cox, J. Exposure to bisphosphonates and risk of gastrointestinal cancers: series of nested case-control studies with QResearch and CPRD data. *BMJ* **346**, f114 (2013).
25. Vinogradova, Y., Coupland, C. & Hippisley-Cox, J. Exposure to statins and risk of common cancers: a series of nested case-control studies. *BMC Cancer* **11**, 409 (2011).
26. Vinogradova, Y., Coupland, C. & Hippisley-Cox, J. Exposure to cyclooxygenase-2 inhibitors and risk of cancer: nested case-control studies. *Br J Cancer* **105**, 452-459 (2011).
27. Logan, R.F., Vinogradova, Y., Coupland, C.A. & Hippisley-Cox, J. Is Low-Dose Aspirin Use Associated With a Reduced Risk of Colorectal Cancer? A QResearch Primary Care Database Analysis. *Gastroenterology* **140**, S-402 (2011).