

BMJ Open

BMJ Open is committed to open peer review. As part of this commitment we make the peer review history of every article we publish publicly available.

When an article is published we post the peer reviewers' comments and the authors' responses online. We also post the versions of the paper that were used during peer review. These are the versions that the peer review comments apply to.

The versions of the paper that follow are the versions that were submitted during the peer review process. They are not the versions of record or the final published versions. They should not be cited or distributed as the published version of this manuscript.

BMJ Open is an open access journal and the full, final, typeset and author-corrected version of record of the manuscript is available on our site with no access controls, subscription charges or pay-per-view fees (<http://bmjopen.bmj.com>).

If you have any questions on BMJ Open's open peer review process please email info.bmjopen@bmj.com

BMJ Open

A common methodology for validation of the Qcovid algorithm across the four UK nations

Journal:	<i>BMJ Open</i>
Manuscript ID	bmjopen-2021-050994
Article Type:	Protocol
Date Submitted by the Author:	06-Mar-2021
Complete List of Authors:	Kerr, Steven; The University of Edinburgh Usher Institute of Population Health Sciences and Informatics, Robertson, Chris; University of Strathclyde, Department of Mathematics and Statistics Nafilyan, Vahe; Office for National Statistics Lyons, Ronan; University of Wales Swansea, Swansea Clinical School Kee, Frank; Queen's University Belfast, UKCRC Centre of Excellence for Public Health (NI) Cardwell, Christopher; Queen's University Belfast, School of Medicine, Dentistry and Biomedical Sciences Coupland, Carol; University of Nottingham, Division of Primary Care Lyons, Jane; Swansea University Medical School Humberstone, Ben; Office for National Statistics Hippisley-Cox, Julia; University of Oxford, Nuffield Department of Primary Care Sciences Sheikh, Aziz; University of Edinburgh, Division of Community Health Sciences
Keywords:	COVID-19, Epidemiology < INFECTIOUS DISEASES, Public health < INFECTIOUS DISEASES

SCHOLARONE™
Manuscripts

A common methodology for validation of the QCOVID algorithm across the four UK nations

Steven Kerr, stevenkerr2@gmail.com (corresponding author) [1]

Chris Robertson [2,3]

Vahe Nafilyan [4]

Ronan A Lyons [5]

Frank Kee [6]

Chris Cardwell [6]

Carol Coupland [7]

Jane Lyons [5]

Ben Humberstone [4]

Julie Hippisley-Cox [8]

Aziz Sheikh [1]

[1] Usher Institute, The University of Edinburgh, NINE Edinburgh BioQuarter, Edinburgh, UK, EH16 4UX.

[2] University of Strathclyde, Glasgow, UK.

[3] Public Health Scotland, UK

[4] Office for National Statistics, UK.

[5] Population Data Science, Swansea University Medical School, Swansea, UK.

[6] Queen's University Belfast, Belfast, UK.

[7] University of Nottingham, Nottingham, UK.

[8] University of Oxford, Oxford, UK.

Keywords: Covid-19, Coronavirus, QCOVID, Epidemiology, Public Health.

Word count: 3,094

ABSTRACT

Introduction:

The QCOVID algorithm is a risk prediction tool for infection and subsequent hospitalisation/death due to SARS-Cov-2. At the time of writing, it is being used in important policymaking decisions by the UK and devolved governments for combatting the Covid-19 pandemic, including deliberations on shielding and vaccine prioritisation. There are four statistical validation exercises currently planned for the QCOVID algorithm, using data pertaining to England, Northern Ireland, Scotland and Wales respectively. This paper presents a common procedure for conducting and reporting on validation exercises for the QCOVID algorithm.

Methods and Analysis:

We will use open, retrospective cohort studies to assess the performance of the QCOVID risk prediction tool in each of the four UK nations. Linked datasets comprising of primary and secondary care records, virological testing data and death registrations will be assembled in trusted research environments in England, Scotland, Northern Ireland and Wales. We will seek to have population level coverage as far as possible within each nation. The following performance metrics will be calculated by strata: Harrell's C, Brier score, R^2 and Royston's D.

Ethics and dissemination:

Approvals have been obtained from relevant ethics bodies in each UK nation. Findings will be made available to national policymakers, presented at conferences and published in peer reviewed journal

Strengths and limitations of this study:

- We will use national level data within each UK nation
- There are potential issues with missing data and differences in the way data is recorded in each country.
- We will evaluate the performance of the algorithm according to several relevant metrics.

INTRODUCTION:

The QCOVID algorithm [1] has been developed to help identify adults at high risk of being hospitalised or dying following infection with SARS-Cov-2. The algorithm takes as input a total of 40 variables including age, sex, ethnicity, Townsend deprivation score [2] and housing category, as well as clinical information including body mass index (BMI) and 33 variables related to medical conditions and treatments. It outputs the predicted probability that an individual will be infected with SARS-Cov-2 and then hospitalised, and the predicted probability that an individual will be infected with SARS-Cov-2 and then die, over a 90 day period. The algorithm was trained using information from the QResearch database [3], which as of April 2020 contained routinely collected data from 1205 General Practices across England, covering 10.5 million patients. The initial training dataset comprised of a cohort of 6.08 million individuals tracked from the 24 January 2020 to 30 April 2020, and was validated on a subset of 2.17 million individuals tracked from 1 May 2020 to 30 June 2020. The research protocol for the development of the QCOVID algorithm can be found in [4].

The QCOVID algorithm was commissioned by the Chief Medical Officer for England on behalf of the UK government. The algorithm is currently being used to inform UK and devolved government policy on combatting the SARS-Cov-2 pandemic, including guidance on social-distancing and shielding measures, as well vaccine prioritisation. [5] It is therefore of great importance to validate the predictions of the algorithm in sub-populations of the UK that were not in the initial training set, but will potentially be subject to those policies.

Four separate validation exercises for the QCOVID algorithm are planned – one for each of England, Northern Ireland, Scotland and Wales. In order to facilitate useful comparison of the results of the separate validation exercises, it is necessary to establish a consistent set of procedures. The purpose of this paper is to explicate a common methodology for the validation of the QCOVID algorithm across the four nations of the UK.

METHODS AND ANALYSIS:

Study design:

Open, retrospective cohort study designs will be employed, making use of routinely collected data from General Practices as well as linked datasets on hospital admissions, reverse-transcription polymerase chain reaction (RT-PCR) testing for Covid-19, and registered deaths. We will aim to have national coverage as far as is possible within each of the four nations of the UK.

Data Sources:

Box 1 contains a brief summary of the main datasets that will be used in the validation exercise for each nation

Box 1: Main datasets to be used

England: Office for National Statistics (ONS) Public Health Linked Data Asset. This dataset is based on the 2011 Census in England covering 40.1 million people, linked at individual level using the NHS number to mortality records, Hospital Episode Statistics (HES) and the General Practice Extraction Service (GPES) data for pandemic planning and research. The data covers 80% of the population of England aged 19 and over.

Northern Ireland: National Health Application and Infrastructure Services (NHAIS) will be used for demographic information. The Patient Administration System (PAS) will be used for data on hospital admissions. Death data will be drawn from the Registrar General, and identified as Covid-19 related through the official Northern Ireland Statistics and Research Agency (NISRA) dashboard. The General Practice Information Platform (GPIP) will bring together GP records from practices across Northern Ireland into a single dataset for use in the validation. As this is not held in the Honest Broker Service, a separate request to its governance board is being made. The Electronic Prescribing Database (EPD) will be used to access information on prescriptions.

Scotland: EAVE II dataset [6]. Contains primary health care records for 5.4 million people covering 99% of the population of Scotland, linked with secondary care data from Scottish Morbidity Record (SMR), Covid-19 test results from Electronic Communication of Surveillance Scotland (ECOSS), and mortality data from National Records Scotland.

Wales: Secure Anonymised Information Linkage (SAIL system) [7]. This will utilise the Controlling Covid (ConCOV) platform linking records on 3.2 million people from the NHS population spine with hospital (Patient Episode Database for Wales), Welsh Longitudinal GP record (WLGP), Covid-19 test results from the Laboratory Information Management System (LIMS), and mortality and 2011 Census data from the Office for National Statistics (ONS) [8]

Selection criteria:

Any individual in the relevant linked dataset between the ages of 19 and 100 will be included.

Individuals who had an event (hospitalisation or death) in the first period (24 January 2020 – 30

April 2020) will be excluded from any analysis in the second period (1 May 2020 – 30 June 2020).

Exposure and Outcomes:

Table 1 and 2 list all exposure and outcomes variables respectively for the QCOVID algorithm, along with a description, variable type (e.g. integer, real, categorical) and possible values.

Table 1: Exposure variables in QCOVID algorithm.

Variable:	Description/Question:	Value:
Demographic:		
age	Age in years	Integer: 19-100
sex	Biological sex at birth	Categorical: female, male
town	Townsend Deprivation Score	Real number
ethnicity	Ethnicity	Categorical: White, Indian, Pakistani, Bangladeshi, Other Asian, Caribbean, Black African, Chinese, other ethnic group
homecat	What is your housing category - care home or homeless or neither?	Categorical: neither, care home, homeless

Clinical:		
bmi	Body Mass Index (kg/m ²)	Positive real number
chemocat	Have you had chemotherapy in the last 12 months?	Categorical: none, group A, group B, group C

learncat	Do you have a learning disability or Down's Syndrome?	Categorical: learning disability, Down syndrome
renalcat	Chronic Kidney Disease (CKD) stage	Categorical: No serious kidney disease, CKD stage 3, CKD stage 4, CKD stage 5 without dialysis or transplant, CKD stage 5 with dialysis in last 12 months, CKD stage 5 with transplant
diabetescat	Do you have diabetes?	Categorical: none, type 1, type 2
b2_82	Have you been prescribed immunosuppressants four or more times in the previous 6 months?	Categorical: yes, no
b2_leukolaba	Have you been prescribed anti-leukotriene or long acting beta2-agonists (LABA) four or more times in the previous 6 months?	Categorical: yes, no
b2_prednisone	Have you been prescribed oral prednisolone containing preparations prescribed four or more times in the previous 6 months?	Categorical: yes, no
b_AF	Do you have atrial fibrillation?	Categorical: yes, no
b_CCF	Do you have heart failure?	Categorical: yes, no
b_asthma	Do you have asthma?	Categorical: yes, no
b_bloodcancer	Have you a cancer of the blood or bone marrow such as leukaemia, myelodysplastic syndromes,	Categorical: yes, no

	lymphoma or myeloma and are at any stage of treatment?	
b_cerebralpalsay	Do you have cerebral palsy?	Categorical: yes, no
b_chd	Do you have coronary heart disease?	Categorical: yes, no
b_cirrhosis	Do you have cirrhosis of the liver?	Categorical: yes, no
b_congenheart	Do you have congenital heart disease or have you had surgery for it in the past?	Categorical: yes, no
b_copd	Do you have chronic obstructive pulmonary disease (COPD)?	Categorical: yes, no
b_dementia	Do you have dementia?	Categorical: yes, no
b_epilepsy	Do you have epilepsy?	Categorical: yes, no
b_fracture4	Have you had a prior fracture of hip, wrist, spine or humerus?	Categorical: yes, no
b_neurorare	Do you have motor neurone disease, multiple sclerosis, myaesthesia, or Huntingtons's Chorea?	Categorical: yes, no
b_parkinsons	Do you have Parkinson's disease?	Categorical: yes, no
b_pulmhyper	Do you have pulmonary hypertension or pulmonary fibrosis?	Categorical: yes, no
b_pulmrare	Do you have cystic fibrosis or bronchiectasis or alveolitis?	Categorical: yes, no
b_pvd	Do you have peripheral vascular disease?	Categorical: yes, no
b_ra_sle	Do you have rheumatoid arthritis or SLE?	Categorical: yes, no
b_respcancer	Do you have lung or oral cancer?	Categorical: yes, no
b_semi	Do you have severe mental illness?	Categorical: yes, no

b_sicklecelldisease	Do you have sickle cell disease or severe combined immune deficiency syndromes?	Categorical: yes, no
b_stroke	Have you had a stroke or TIA?	Categorical: yes, no
b_vte	Have you had a thrombosis or pulmonary embolus?	Categorical: yes, no
p_marrow6	Have you had a bone marrow or stem cell transplant in the last 6 months?	Categorical: yes, no
p_radio6	Have you had radiotherapy in the last 6 months?	Categorical: yes, no
p_solidtransplant	Have you had a solid organ transplant (lung, liver, stomach, pancreas, spleen, heart or thymus)?	Categorical: yes, no

Table 2: Outcomes variables in QCOVID algorithm.

Variable:	Description/Question:	Value:
Time to Covid-19 hospitalisation	Time to hospitalisation with RT-PCR confirmed Covid-19 infection in the cohort period in days.	Real number: 0-91
Time to Covid-19 death	Time to death with Covid-19 confirmed or suspected on their death certificate, or confirmed by RT-PCR test, in the cohort period in days.	Real number: 0-91

Whenever available, all variables will be taken as the most recent recorded value in the relevant dataset at the date of entry into the cohort. The Townsend Deprivation Score (TDS) will be determined by matching available residential location information with output area and the

1
2 corresponding TDS from the 2011 UK census [9]. Categories for the variable chemocat will be
3 determined using the lookup table in the supplemental materials.
4
5
6

7 **Data cleaning:**

8 The following procedures will be used for data cleaning:
9

- 10 • **diabetes_cat:** If the most recent entry has both type 1 and types 2 recorded, diabetes_cat
11 will be set to type 2.
12
- 13 • **BMI:** The most recently recorded patient BMI within the last 5 years. If the most recently
14 recorded BMI is from more than 5 years ago at the search date, bmi will be set to missing
15 value.
16
- 17 • **learncat:** If a patient is recorded as having both learning disability and Down's syndrome,
18 learncat will be set to Down's syndrome.
19
20
21
22
23

24 **Missing data:**

25 For comorbidities and medication use and treatments, missing values will be taken to mean absence
26 of that factor. Missing values for ethnicity will be set to "White". For any other missing values, a
27 single imputation will be considered. The following methods may be considered for use in the
28 imputation: predictive mean matching, least squares, logistic and multinomial models, imputation
29 by chained equations.
30
31
32
33
34
35
36
37
38
39
40
41

42 **Statistical Analysis:**

43 Each validation exercise will report a table of cohort characteristics, following Table 2 in [1]. The
44 main performance metrics that will be calculated are R^2 [9], Harrell's C, Royston's D [10] and the
45 Brier score. Different stratifications for these statistics will be considered, including by age, sex and
46 time period. 95% confidence intervals will be reported for R^2 , Harrell's C and Royston's D. Graphs
47 of observed and predicted probability of hospital admission and death by vigintile for stratified
48 subgroups will be reported, following [1]. Other analyses/reporting measures will also be
49 considered.
50
51
52
53
54
55
56

57 **Sample Size:**

58 A preliminary sample size calculation can be done using figures from the original paper [1]. Using
59 the estimated standard deviation of Harrell's C for females in the first time period and assuming
60

1
2 Harrell's C is asymptotically normally distributed implies that a sample size of approximately 5,714
3 would be sufficient to detect a true value for Harrell's C that is greater than or equal to 0.8 with
4 80% power. Repeating this calculation for other population subgroups and time periods yields
5 results of a similar magnitude. The samples sizes in the planned studies will be on the order of
6 hundreds of thousands or millions.
7
8
9

10 11 12 **Ethics, reporting and dissemination**

13
14 The ethics approval for the development and validation of QCOVID in England was granted by the
15 East Midlands-Derby Research Ethics Committee [reference 18/EM/0400]. For Scotland, approvals
16 have been obtained by the National Research Ethics Service Committee (REC), South East Scotland
17 02 (REC number: 12/SS/0201) and the Public Benefit and Privacy Panel for Health and Social Care
18 (reference number: 1920-0279). The data to be used in this study for Wales are available in the
19 SAIL Databank at Swansea University, Swansea, UK. All proposals to use SAIL data are subject to
20 review by an independent Information Governance Review Panel (IGRP). Before any data can be
21 accessed, approval must be given by the IGRP. The IGRP gives careful consideration to each
22 project to ensure proper and appropriate use of SAIL data. When access has been approved, it is
23 gained through a privacy-protecting safe haven and remote access system referred to as the SAIL
24 Gateway. SAIL has established an application process to be followed by anyone who would like to
25 access data via SAIL.[7] Findings will be presented at conferences, published in peer-reviewed
26 journals and to the funders and government COVID-19 advisory bodies as appropriate.
27

28 Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) and Reporting
29 of studies Conducted using Observational Routinely-collected Data (RECORD) (via the COVID-19
30 extension) checklists will guide our study findings reporting. The Northern Ireland validation study
31 proposal is under review by the NITRE for HSC data accessed via Northern Ireland Honest Broker
32 Service; an Ethics application has been submitted through IRAS.
33
34
35

36 37 **Author's Contributions:**

38 AS conceived this protocol. CR, VH, FK, TC, JHC, BH, CC, RAL and JL provided country
39 specific information about available data and analysis plans. SK wrote drafts of this protocol. All
40 authors gave final approval of the version to be published.
41
42
43
44
45

46 47 **Acknowledgments**

48 This work will use data provided by patients and collected by a number of organisations. We would
49 like to acknowledge all patients who shared their information as well as all data providers who
50 make anonymised data available for research. In particular, Public Health Scotland, Public Health
51
52
53
54
55
56
57
58
59
60

1
2 Wales, Public Health England, the NHS, the SAIL databank, and the Office for National Statistics.
3
4

5 **Funding:**

6
7 The validation in England will be funded by a grant from the National Institute for Health Research
8 following a commission by the Chief Medical Officer for England. In Scotland, EAVE II is funded
9 by the Medical Research Council [MR/R008345/1] and supported by the Scottish Government. In
10 Wales, ConCOV is supported by the Medical Research Council [MR/V028367/1].
11
12
13
14

15 **Competing interests:**

16
17 AS reports grants from NIHR, grants from MRC, and grants from HRR UK, during the conduct of
18 the study. JL and RAL report grants from UKRI Medical Research Council, during the conduct of
19 the study. JHC reports grants from John Fell Oxford University Press Research Fund, grants from
20 Cancer Research UK (CR-UK) grant number C5255/A18085, through the Cancer Research UK
21 Oxford Centre, grants from the Oxford Wellcome Institutional Strategic Support Fund
22 (204826/Z/16/Z), grants from NIHR, during the conduct of the study; personal fees and other from
23 ClinRisk Ltd, outside the submitted work; and JHC is an unpaid director of QResearch, a not-for-
24 profit organisation which is a partnership between the University of Oxford and EMIS Health who
25 supply the QResearch database used for this work. Carol Coupland reports personal fees from
26 ClinRisk Ltd, outside the submitted work. JHC, AS, and Carol Coupland were members of the
27 research team involved in the development of the QCOVID risk prediction algorithm. All other
28 authors report no conflict of interest.
29
30
31
32
33
34
35
36
37
38
39

40 **Patient and Public Involvement:**

41 There are no plans for Patient and Public Involvement in this research.
42
43
44

45 **References:**

- 46 1. Clift AK, Coupland CAC, Keogh RH, Diaz-Ordaz K, Williamson E, Harrison EM, Hayward A,
47 Hemingway H, Horby P, Mehta N, Bengler J, Khunti K, Spiegelhalter D, Sheikh A, Valabhji J,
48 Lyons RA, Robson J, Semple MG, Kee F, Johnson P, Jebb S, Williams T, Hippisley-Cox J. Living
49 risk prediction algorithm (QCOVID) for risk of hospital admission and mortality from coronavirus
50 19 in adults: national derivation and validation cohort study. *BMJ*. 2020 Oct 20;371:m3731. doi:
51 10.1136/bmj.m3731. PMID: 33082154; PMCID: PMC7574532.
- 52 2. Townsend, P., Phillimore, P. and Beattie, A. (1988) *Health and Deprivation: Inequality and the*
53 *North*. Routledge, London. doi: 10.7748/ns.2.17.34.s66. PMID: 27415096.
- 54 3. Qresearch, <https://www.qresearch.org/>
55
56
57
58
59
60

- 1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
4. Hippisley-Cox J, Clift AK, Coupland CAC, et al. Protocol for the development and evaluation of a tool for predicting risk of short-term adverse outcomes due to COVID-19 in the general UK population. *MedRxiv* 2020:2020.06.28.20141986-2020.06.28
 5. <https://www.gov.uk/government/news/new-technology-to-help-identify-those-at-high-risk-from-covid-19>
 6. EAVE II, The Usher Institute, <https://www.ed.ac.uk/usher/eave-ii>
 7. The Secure Anonymised Information Linkage (SAIL) databank. Available at: <https://saildatabank.com>
 8. Lyons J, Akbari A, Torabi F, Davies G, North L, Griffiths R, Bailey R, Hollinghurst J, Fry R, Turner S, Thompson D, Rafferty J, Mizen A, Orton C, Ellwood-Thompson S, Au-Yeung L, Cross L, Gravenor M, Brophy S, Lucini B, John A, Szakmany T, Davies J, Davies C, Williams C, Emmerson C, Cottrell S, Connor T, Taylor C, Pugh R, Diggle PJ, John G, Scourfield S, Hunt J, Cunningham AM, Helliwell K, Lyons RA. (2020) Understanding and responding to COVID19 in Wales: protocol for a privacy protecting data platform for enhanced epidemiology and evaluation of interventions. *BMJ Open* 2020;10:e043010. doi:10.1136/bmjopen-2020-043010
 9. 2011 UK Census, Townsend Deprivation Scores, <https://www.statistics.digitalresources.jisc.ac.uk/dataset/2011-uk-townsend-deprivation-scores>
 10. Royston P, Sauerbrei W. A new measure of prognostic separation in survival data. *Stat Med*. 2004 Mar 15;23(5):723-48. doi: 10.1002/sim.1621. PMID: 14981672.

Group	Drug	Final classification	map
1	Group A	1.8.2	
2	Group B	3.8.4	
3	Group C	5	
4			
5			
6			
7			
8			
9			
10			
11			
12			
13			
14			
15			
16			
17			
18			
19			
20			
21			
22			
23			
24			
25			
26			
27			
28			
29			
30			
31			
32			
33			
34			
35			
36			
37			
38			
39			
40			
41			
42			
43			
44			
45			
46			
47			
48			
49			
50			
51			
52			
53			
54			
55			
56			
57			
58			
59			
60			

BMJ Open

A common methodology for validation of the Qcovid algorithm across the four UK nations

Journal:	<i>BMJ Open</i>
Manuscript ID	bmjopen-2021-050994.R1
Article Type:	Protocol
Date Submitted by the Author:	07-Dec-2021
Complete List of Authors:	Kerr, Steven; The University of Edinburgh Usher Institute of Population Health Sciences and Informatics, Robertson, Chris; University of Strathclyde, Department of Mathematics and Statistics Nafilyan, Vahe; Office for National Statistics Lyons, Ronan; University of Wales Swansea, Swansea Clinical School Kee, Frank; Queen's University Belfast, UKCRC Centre of Excellence for Public Health (NI) Cardwell, Christopher; Queen's University Belfast, School of Medicine, Dentistry and Biomedical Sciences Coupland, Carol; University of Nottingham, Division of Primary Care Lyons, Jane; Swansea University Medical School Humberstone, Ben; Office for National Statistics Hippisley-Cox, Julia; University of Oxford, Nuffield Department of Primary Care Sciences Sheikh, Aziz; University of Edinburgh, Division of Community Health Sciences
Primary Subject Heading:	Epidemiology
Secondary Subject Heading:	Infectious diseases, Respiratory medicine
Keywords:	COVID-19, Epidemiology < INFECTIOUS DISEASES, Public health < INFECTIOUS DISEASES

SCHOLARONE™
Manuscripts

A common methodology for validation of the QCOVID algorithm across the four UK nations

Steven Kerr, stevenkerr2@gmail.com (corresponding author) [1]

Chris Robertson [2,3]

Vahe Nafilyan [4]

Ronan A Lyons [5]

Frank Kee [6]

Chris Cardwell [6]

Carol Coupland [7]

Jane Lyons [5]

Ben Humberstone [4]

Julie Hippisley-Cox [8]

Aziz Sheikh [1]

[1] Usher Institute, The University of Edinburgh, NINE Edinburgh BioQuarter, Edinburgh, UK, EH16 4UX.

[2] University of Strathclyde, Glasgow, UK.

[3] Public Health Scotland, UK

[4] Office for National Statistics, UK.

[5] Population Data Science, Swansea University Medical School, Swansea, UK.

[6] Queen's University Belfast, Belfast, UK.

[7] University of Nottingham, Nottingham, UK.

[8] University of Oxford, Oxford, UK.

Keywords: Covid-19, Coronavirus, QCOVID, Epidemiology, Public Health.

Word count: 3,148

ABSTRACT

Introduction:

The QCOVID algorithm is a risk prediction tool for infection and subsequent hospitalisation/death due to SARS-Cov-2. At the time of writing, it is being used in important policymaking decisions by the UK and devolved governments for combatting the Covid-19 pandemic, including deliberations on shielding and vaccine prioritisation. There are four statistical validation exercises currently planned for the QCOVID algorithm, using data pertaining to England, Northern Ireland, Scotland and Wales respectively. This paper presents a common procedure for conducting and reporting on validation exercises for the QCOVID algorithm.

Methods and Analysis:

We will use open, retrospective cohort studies to assess the performance of the QCOVID risk prediction tool in each of the four UK nations. Linked datasets comprising of primary and secondary care records, virological testing data and death registrations will be assembled in trusted research environments in England, Scotland, Northern Ireland and Wales. We will seek to have population level coverage as far as possible within each nation. The following performance metrics will be calculated by strata: Harrell's C, Brier score, R^2 and Royston's D.

Ethics and dissemination:

Approvals have been obtained from relevant ethics bodies in each UK nation. Findings will be made available to national policymakers, presented at conferences and published in peer reviewed journal

Strengths and limitations of this study:

- We will use national level data within each UK nation
- There are potential issues with missing data and differences in the way data is recorded in each country.
- We will evaluate the performance of the algorithm according to several relevant metrics.

INTRODUCTION:

The QCOVID algorithm [1] has been developed to help identify adults at high risk of being hospitalised or dying following infection with SARS-Cov-2 (Severe acute respiratory syndrome coronavirus 2). The algorithm takes as input a total of 40 variables including age, sex, ethnicity, Townsend deprivation score [2] and housing category, as well as clinical information including body mass index (BMI) and 33 variables related to medical conditions and treatments. It outputs the predicted probability that an individual will be infected with SARS-Cov-2 and then hospitalised, and the predicted probability that an individual will be infected with SARS-Cov-2 and then die, over a 90 day period. The algorithm was trained using information from the QResearch database [3], which as of April 2020 contained routinely collected data from 1205 General Practices across England, covering 10.5 million patients. The initial training dataset comprised of a cohort of 6.08 million individuals tracked from the 24 January 2020 to 30 April 2020, and was validated on a subset of 2.17 million individuals tracked from 1 May 2020 to 30 June 2020. The research protocol for the development of the QCOVID algorithm can be found in [4].

The QCOVID algorithm was commissioned by the Chief Medical Officer for England on behalf of the UK government. The algorithm has been used to inform UK and devolved government policy on combatting the SARS-Cov-2 pandemic, including guidance on social-distancing and shielding measures, as well vaccine prioritisation. [5] It is therefore of great importance to validate the predictions of the algorithm in sub-populations of the UK that were not in the initial training set, but will potentially be subject to those policies.

At the time of writing, there are validation exercises planned in Scotland, Northern Ireland and Wales, and a validation exercise underway in England. Validation work has been expedited in order to support national decision making. In order to facilitate useful comparison of the results of the separate validation exercises, it is necessary to establish a consistent set of procedures. The purpose of this paper is to explicate a common methodology for the validation of the QCOVID algorithm across the four nations of the UK.

METHODS AND ANALYSIS:

Study design:

1
2 Open, retrospective cohort study designs will be employed, making use of routinely collected data
3 from General Practices for clinical and demographic information, as well as linked datasets on
4 hospital admissions, reverse-transcription polymerase chain reaction (RT-PCR) testing for Covid-
5 19, and registered deaths. We will aim to have national coverage as far as is possible within each of
6 the four nations of the UK.
7
8
9

10 11 12 **Data Sources:**

13
14 Box 1 contains a brief summary of the main datasets that will be used in the validation exercise for
15 each nation
16

17 **Box 1: Main datasets to be used**

18
19
20
21 **England:** Office for National Statistics (ONS) Public Health Linked Data Asset. This dataset is
22 based on the 2011 Census in England covering 40.1 million people, linked at individual level
23 using the NHS number to mortality records, Hospital Episode Statistics (HES) and the General
24 Practice Extraction Service (GPES) data for pandemic planning and research. The data covers
25 80% of the population of England aged 19 and over.
26
27

28
29
30 **Northern Ireland:** National Health Application and Infrastructure Services (NHAIS) will be used
31 for demographic information. The Patient Administration System (PAS) will be used for data on
32 hospital admissions. Death data will be drawn from the Registrar General, and identified as
33 Covid-19 related through the official Northern Ireland Statistics and Research Agency (NISRA)
34 dashboard. The General Practice Information Platform (GPIP) will bring together GP records
35 from practices across Northern Ireland into a single dataset for use in the validation. As this is not
36 held in the Honest Broker Service, a separate request to its governance board is being made. The
37 Electronic Prescribing Database (EPD) will be used to access information on prescriptions.
38
39

40
41
42 **Scotland:** EAVE II dataset [6]. Contains primary health care records for 5.4 million people covering
43 99% of the population of Scotland, linked with secondary care data from Scottish Morbidity
44 Record (SMR), Covid-19 test results from Electronic Communication of Surveillance Scotland
45 (ECOSS), and mortality data from National Records Scotland.
46
47

48
49
50
51 **Wales:** Secure Anonymised Information Linkage (SAIL system) [7]. This will utilise the Controlling
52 Covid (ConCOV) platform linking records on 3.2 million people from the NHS population spine
53 with hospital (Patient Episode Database for Wales), Welsh Longitudinal GP record (WLGP),
54 Covid-19 test results from the Laboratory Information Management System (LIMS), and
55 mortality and 2011 Census data from the Office for National Statistics (ONS) [8]
56
57
58
59
60

Selection criteria:

Any individual in the relevant linked dataset between the ages of 19 and 100 will be included.

Individuals who had an event (hospitalisation or death) in the first period (24 January 2020 – 30

April 2020) will be excluded from any analysis in the second period (1 May 2020 – 30 June 2020).

Exposure and Outcomes:

Table 1 and 2 list all exposure and outcomes variables respectively for the QCOVID algorithm, along with a description, variable type (e.g. integer, real, categorical) and possible values.

Table 1: Exposure variables in QCOVID algorithm.

Variable:	Description/Question:	Value:
Demographic:		
age	Age in years	Integer: 19-100
sex	Biological sex at birth	Categorical: female, male
town	Townsend Deprivation Score	Real number
ethnicity	Ethnicity	Categorical: White, Indian, Pakistani, Bangladeshi, Other Asian, Caribbean, Black African, Chinese, other ethnic group
homecat	What is your housing category - care home or homeless or neither?	Categorical: neither, care home, homeless

Clinical:		
bmi	Body Mass Index (kg/m ²)	Positive real number
chemocat	Have you had chemotherapy in the last 12 months?	Categorical: none, group A, group B, group C

learncat	Do you have a learning disability or Down's Syndrome?	Categorical: learning disability, Down syndrome
renalcat	Chronic Kidney Disease (CKD) stage	Categorical: No serious kidney disease, CKD stage 3, CKD stage 4, CKD stage 5 without dialysis or transplant, CKD stage 5 with dialysis in last 12 months, CKD stage 5 with transplant
diabetescat	Do you have diabetes?	Categorical: none, type 1, type 2
b2_82	Have you been prescribed immunosuppressants four or more times in the previous 6 months?	Categorical: yes, no
b2_leukolaba	Have you been prescribed anti-leukotriene or long acting beta2-agonists (LABA) four or more times in the previous 6 months?	Categorical: yes, no
b2_prednisone	Have you been prescribed oral prednisolone containing preparations prescribed four or more times in the previous 6 months?	Categorical: yes, no
b_AF	Do you have atrial fibrillation?	Categorical: yes, no
b_CCF	Do you have heart failure?	Categorical: yes, no
b_asthma	Do you have asthma?	Categorical: yes, no
b_bloodcancer	Have you a cancer of the blood or bone marrow such as leukaemia, myelodysplastic syndromes,	Categorical: yes, no

	lymphoma or myeloma and are at any stage of treatment?	
b_cerebralpalsay	Do you have cerebral palsy?	Categorical: yes, no
b_chd	Do you have coronary heart disease?	Categorical: yes, no
b_cirrhosis	Do you have cirrhosis of the liver?	Categorical: yes, no
b_congenheart	Do you have congenital heart disease or have you had surgery for it in the past?	Categorical: yes, no
b_copd	Do you have chronic obstructive pulmonary disease (COPD)?	Categorical: yes, no
b_dementia	Do you have dementia?	Categorical: yes, no
b_epilepsy	Do you have epilepsy?	Categorical: yes, no
b_fracture4	Have you had a prior fracture of hip, wrist, spine or humerus?	Categorical: yes, no
b_neurorare	Do you have motor neurone disease, multiple sclerosis, myasthenia, or Huntingtons's Chorea?	Categorical: yes, no
b_parkinsons	Do you have Parkinson's disease?	Categorical: yes, no
b_pulmhyper	Do you have pulmonary hypertension or pulmonary fibrosis?	Categorical: yes, no
b_pulmrare	Do you have cystic fibrosis or bronchiectasis or alveolitis?	Categorical: yes, no
b_pvd	Do you have peripheral vascular disease?	Categorical: yes, no
b_ra_sle	Do you have rheumatoid arthritis or SLE?	Categorical: yes, no
b_respcancer	Do you have lung or oral cancer?	Categorical: yes, no
b_semi	Do you have severe mental illness?	Categorical: yes, no

b_sicklecelldisease	Do you have sickle cell disease or severe combined immune deficiency syndromes?	Categorical: yes, no
b_stroke	Have you had a stroke or TIA?	Categorical: yes, no
b_vte	Have you had a thrombosis or pulmonary embolus?	Categorical: yes, no
p_marrow6	Have you had a bone marrow or stem cell transplant in the last 6 months?	Categorical: yes, no
p_radio6	Have you had radiotherapy in the last 6 months?	Categorical: yes, no
p_solidtransplant	Have you had a solid organ transplant (lung, liver, stomach, pancreas, spleen, heart or thymus)?	Categorical: yes, no

Table 2: Outcomes variables in QCOVID algorithm.

Variable:	Description/Question:	Value:
Time to Covid-19 hospitalisation	Time to hospitalisation with RT-PCR confirmed Covid-19 infection in the cohort period in days.	Real number: 0-91
Time to Covid-19 death	Time to death with Covid-19 confirmed or suspected on their death certificate, or confirmed by RT-PCR test, in the cohort period in days.	Real number: 0-91

Whenever available, all variables will be taken as the most recent recorded value in the relevant dataset at the date of entry into the cohort. The Townsend Deprivation Score (TDS) will be determined by matching available residential location information with output area and the

1
2 corresponding TDS from the 2011 UK census [9]. Categories for the variable chemocat will be
3 determined using the lookup table in the supplemental materials.
4
5

6 7 **Data cleaning:**

8 The following procedures will be used for data cleaning:
9

- 10 • **diabetes_cat:** If the most recent entry has both type 1 and types 2 recorded, diabetes_cat
11 will be set to type 2.
12
- 13 • **BMI:** The most recently recorded patient BMI within the last 5 years. If the most recently
14 recorded BMI is from more than 5 years ago at the search date, BMI will be set to missing
15 value. Implausible values for BMI (<12 or >70) will be set to missing value.
16
17
- 18 • **learncat:** If a patient is recorded has having both learning disability and Down's syndrome,
19 learncat will be set to Down's syndrome.
20
21
22
23

24 25 **Missing data:**

26 For comorbidities and medication use and treatments, missing values will be taken to mean absence
27 of that factor. Modal substitution will be considered for missing values for ethnicity. For any other
28 missing values of predictor variables, a single imputation will be considered. Outcome variables
29 will not be imputed. The following methods may be considered for use in the imputation: predictive
30 mean matching, least squares, logistic and multinomial models, imputation by chained equations.
31
32
33
34
35
36
37
38
39
40
41

42 43 **Statistical Analysis:**

44 Each validation exercise will report a table of cohort characteristics, following Table 2 in [1]. The
45 main performance metrics that will be calculated are R^2 [9], Harrell's C, Royston's D [10] and the
46 Brier score. Different stratifications for these statistics will be considered, including by age, sex and
47 time period. 95% confidence intervals will be reported for R^2 , Harrell's C and Royston's D. Graphs
48 of observed and predicted probability of hospital admission and death by vigintile for stratified
49 subgroups will be reported, following [1].
50
51
52
53
54

55 56 **Sample Size:**

57 A preliminary sample size calculation can be done using figures from the original paper [1]. Using
58 the estimated standard deviation of Harrell's C for females in the first time period and assuming
59 Harrell's C is asymptotically normally distributed implies that a sample size of approximately 5,714
60

1
2 would be sufficient to detect a true value for Harrell's C that is greater than or equal to 0.8 with
3 80% power. Repeating this calculation for other population subgroups and time periods yields
4 results of a similar magnitude. The samples sizes in the planned studies will be on the order of
5 hundreds of thousands or millions.
6
7
8
9

10 **Ethics, reporting and dissemination**

11 The ethics approval for the development and validation of QCOVID in England was granted by the
12 East Midlands-Derby Research Ethics Committee [reference 18/EM/0400]. For Scotland, approvals
13 have been obtained by the National Research Ethics Service Committee (REC), South East Scotland
14 02 (REC number: 12/SS/0201) and the Public Benefit and Privacy Panel for Health and Social Care
15 (reference number: 1920-0279). The data to be used in this study for Wales are available in the
16 SAIL Databank at Swansea University, Swansea, UK. All proposals to use SAIL data are subject to
17 review by an independent Information Governance Review Panel (IGRP). Before any data can be
18 accessed, approval must be given by the IGRP. The IGRP gives careful consideration to each
19 project to ensure proper and appropriate use of SAIL data. When access has been approved, it is
20 gained through a privacy-protecting safe haven and remote access system referred to as the SAIL
21 Gateway. SAIL has established an application process to be followed by anyone who would like to
22 access data via SAIL.[7] Findings will be presented at conferences, published in peer-reviewed
23 journals and to the funders and government COVID-19 advisory bodies as appropriate.
24
25

26 Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) and Reporting
27 of studies Conducted using Observational Routinely-collected Data (RECORD) (via the COVID-19
28 extension) checklists will guide our study findings reporting. The Northern Ireland validation study
29 proposal is under review by the NITRE for HSC data accessed via Northern Ireland Honest Broker
30 Service; an Ethics application has been submitted through IRAS.
31
32
33
34

35 **Author's Contributions:**

36 AS conceived this protocol. CR, VH, FK, TC, JHC, BH, CC, RAL and JL provided country
37 specific information about available data and analysis plans. SK wrote drafts of this protocol. All
38 authors gave final approval of the version to be published.
39
40
41
42
43
44

45 **Acknowledgments**

46 This work will use data provided by patients and collected by a number of organisations. We would
47 like to acknowledge all patients who shared their information as well as all data providers who
48 make anonymised data available for research. In particular, Public Health Scotland, Public Health
49
50
51
52
53
54
55
56
57
58
59
60

1
2 Wales, Public Health England, the NHS, the SAIL databank, and the Office for National Statistics.
3
4

5
6 **Funding:**

7 The validation in England will be funded by a grant from the National Institute for Health Research
8 following a commission by the Chief Medical Officer for England. In Scotland, EAVE II is funded
9 by the Medical Research Council [MR/R008345/1] and supported by the Scottish Government. In
10 Wales, ConCOV is supported by the Medical Research Council [MR/V028367/1].
11
12
13
14

15
16 **Competing interests:**

17 AS reports grants from NIHR, grants from MRC, and grants from HRR UK, during the conduct of
18 the study. JL and RAL report grants from UKRI Medical Research Council, during the conduct of
19 the study. JHC reports grants from John Fell Oxford University Press Research Fund, grants from
20 Cancer Research UK (CR-UK) grant number C5255/A18085, through the Cancer Research UK
21 Oxford Centre, grants from the Oxford Wellcome Institutional Strategic Support Fund
22 (204826/Z/16/Z), grants from NIHR, during the conduct of the study; personal fees and other from
23 ClinRisk Ltd, outside the submitted work; and JHC is an unpaid director of QResearch, a not-for-
24 profit organisation which is a partnership between the University of Oxford and EMIS Health who
25 supply the QResearch database used for this work. Carol Coupland reports personal fees from
26 ClinRisk Ltd, outside the submitted work. JHC, AS, and Carol Coupland were members of the
27 research team involved in the development of the QCOVID risk prediction algorithm. All other
28 authors report no conflict of interest.
29
30
31
32
33
34
35
36
37
38
39

40 **Patient and Public Involvement:**

41 There are no plans for Patient and Public Involvement in this research.
42
43
44

45 **Data sharing:**

46 All code used in these analyses will be made publicly available online e.g. through GitHub.
47
48
49

50 **References:**

51 **1.** Clift AK, Coupland CAC, Keogh RH, Diaz-Ordaz K, Williamson E, Harrison EM, Hayward A,
52 Hemingway H, Horby P, Mehta N, Bengler J, Khunti K, Spiegelhalter D, Sheikh A, Valabhji J,
53 Lyons RA, Robson J, Semple MG, Kee F, Johnson P, Jebb S, Williams T, Hippisley-Cox J. Living
54 risk prediction algorithm (QCOVID) for risk of hospital admission and mortality from coronavirus
55 19 in adults: national derivation and validation cohort study. *BMJ*. 2020 Oct 20;371:m3731. doi:
56 10.1136/bmj.m3731. PMID: 33082154; PMCID: PMC7574532.
57
58
59
60

- 1
- 2 **2.** Townsend, P., Phillimore, P. and Beattie, A. (1988) Health and Deprivation: Inequality and the
- 3 North. Routledge, London. doi: 10.7748/ns.2.17.34.s66. PMID: 27415096.
- 4
- 5 **3.** Qresearch, <https://www.qresearch.org/>
- 6
- 7 **4.** Hippisley-Cox J, Clift AK, Coupland CAC, et al. Protocol for the development and evaluation of
- 8 a tool for predicting risk of short-term adverse outcomes due to COVID-19 in the general UK
- 9 population. *MedRxiv* 2020:2020.06.28.20141986-2020.06.28
- 10
- 11 **5.** [https://www.gov.uk/government/news/new-technology-to-help-identify-those-at-high-risk-from-](https://www.gov.uk/government/news/new-technology-to-help-identify-those-at-high-risk-from-covid-19)
- 12 [covid-19](https://www.gov.uk/government/news/new-technology-to-help-identify-those-at-high-risk-from-covid-19)
- 13
- 14 **6.** EAVE II, The Usher Institute, <https://www.ed.ac.uk/usher/eave-ii>
- 15
- 16 **7.** The Secure Anonymised Information Linkage (SAIL) databank. Available at:
- 17 <https://saildatabank.com>
- 18
- 19 **8.** Lyons J, Akbari A, Torabi F, Davies G, North L, Griffiths R, Bailey R, Hollinghurst J, Fry R,
- 20 Turner S, Thompson D, Rafferty J, Mizen A, Orton C, Ellwood-Thompson S, Au-Yeung L, Cross
- 21 L, Gravenor M, Brophy S, Lucini B, John A, Szakmany T, Davies J, Davies C, Williams C,
- 22 Emmerson C, Cottrell S, Connor T, Taylor C, Pugh R, Diggle PJ, John G, Scourfield S, Hunt J,
- 23 Cunningham AM, Helliwell K, Lyons RA. (2020) Understanding and responding to COVID19 in
- 24 Wales: protocol for a privacy protecting data platform for enhanced epidemiology and evaluation of
- 25 interventions. *BMJ Open* 2020;10:e043010. doi:10.1136/bmjopen-2020-043010
- 26
- 27 **9.** 2011 UK Census, Townsend Deprivation Scores,
- 28 <https://www.statistics.digitalresources.jisc.ac.uk/dataset/2011-uk-townsend-deprivation-scores>
- 29
- 30 **10.** Royston P, Sauerbrei W. A new measure of prognostic separation in survival data. *Stat Med.*
- 31 2004 Mar 15;23(5):723-48. doi: 10.1002/sim.1621. PMID: 14981672.
- 32
- 33
- 34
- 35
- 36
- 37
- 38
- 39
- 40
- 41
- 42
- 43
- 44
- 45
- 46
- 47
- 48
- 49
- 50
- 51
- 52
- 53
- 54
- 55
- 56
- 57
- 58
- 59
- 60

BMJ Open

A common protocol for validation of the QCOVID algorithm across the four UK nations

Journal:	<i>BMJ Open</i>
Manuscript ID	bmjopen-2021-050994.R2
Article Type:	Protocol
Date Submitted by the Author:	14-May-2022
Complete List of Authors:	Kerr, Steven; The University of Edinburgh Usher Institute of Population Health Sciences and Informatics, Robertson, Chris; University of Strathclyde, Department of Mathematics and Statistics Nafilyan, Vahe; Office for National Statistics Lyons, Ronan; University of Wales Swansea, Swansea Clinical School Kee, Frank; Queen's University Belfast, UKCRC Centre of Excellence for Public Health (NI) Cardwell, Christopher; Queen's University Belfast, School of Medicine, Dentistry and Biomedical Sciences Coupland, Carol; University of Nottingham, Division of Primary Care Lyons, Jane; Swansea University Medical School Humberstone, Ben; Office for National Statistics Hippisley-Cox, Julia; University of Oxford, Nuffield Department of Primary Care Sciences Sheikh, Aziz; University of Edinburgh, Division of Community Health Sciences
Primary Subject Heading:	Epidemiology
Secondary Subject Heading:	Infectious diseases, Respiratory medicine
Keywords:	COVID-19, Epidemiology < INFECTIOUS DISEASES, Public health < INFECTIOUS DISEASES

SCHOLARONE™
Manuscripts

A common protocol for validation of the QCOVID algorithm across the four UK nations

Steven Kerr, stevenkerr2@gmail.com (corresponding author) [1]

Chris Robertson [2,3]

Vahe Nafilyan [4]

Ronan A Lyons [5]

Frank Kee [6]

Chris Cardwell [6]

Carol Coupland [7]

Jane Lyons [5]

Ben Humberstone [4]

Julie Hippisley-Cox [8]

Aziz Sheikh [1]

[1] Usher Institute, The University of Edinburgh, NINE Edinburgh BioQuarter, Edinburgh, UK, EH16 4UX.

[2] University of Strathclyde, Glasgow, UK.

[3] Public Health Scotland, UK

[4] Office for National Statistics, UK.

[5] Population Data Science, Swansea University Medical School, Swansea, UK.

[6] Queen's University Belfast, Belfast, UK.

[7] University of Nottingham, Nottingham, UK.

[8] University of Oxford, Oxford, UK.

Keywords: Covid-19, Coronavirus, QCOVID, Epidemiology, Public Health.

Word count: 1,653

ABSTRACT

Introduction:

The QCOVID algorithm is a risk prediction tool for infection and subsequent hospitalisation/death due to SARS-Cov-2. At the time of writing, it is being used in important policymaking decisions by the UK and devolved governments for combatting the Covid-19 pandemic, including deliberations on shielding and vaccine prioritisation. There are four statistical validation exercises currently planned for the QCOVID algorithm, using data pertaining to England, Northern Ireland, Scotland and Wales respectively. This paper presents a common procedure for conducting and reporting on validation exercises for the QCOVID algorithm.

Methods and Analysis:

We will use open, retrospective cohort studies to assess the performance of the QCOVID risk prediction tool in each of the four UK nations. Linked datasets comprising of primary and secondary care records, virological testing data and death registrations will be assembled in trusted research environments in England, Scotland, Northern Ireland and Wales. We will seek to have population level coverage as far as possible within each nation. The following performance metrics will be calculated by strata: Harrell's C, Brier score, R^2 and Royston's D.

Ethics and dissemination:

Approvals have been obtained from relevant ethics bodies in each UK nation. Findings will be made available to national policymakers, presented at conferences and published in peer reviewed journal

Strengths and limitations of this study:

- We will use national level data within each UK nation
- There are potential issues with missing data and differences in the way data is recorded in each country.
- We will evaluate the performance of the algorithm according to several relevant metrics.

INTRODUCTION:

The QCOVID algorithm [1] has been developed to help identify adults at high risk of being hospitalised or dying following infection with SARS-Cov-2 (Severe acute respiratory syndrome coronavirus 2). The algorithm takes as input a total of 40 variables including age, sex, ethnicity, Townsend deprivation score [2] and housing category, as well as clinical information including body mass index (BMI) and 33 variables related to medical conditions and treatments. It outputs the predicted probability that an individual will be infected with SARS-Cov-2 and then hospitalised, and the predicted probability that an individual will be infected with SARS-Cov-2 and then die, over a 90 day period. The algorithm was trained using information from the QResearch database [3], which as of April 2020 contained routinely collected data from 1205 General Practices across England, covering 10.5 million patients. The initial training dataset comprised of a cohort of 6.08 million individuals tracked from the 24 January 2020 to 30 April 2020, and was validated on a subset of 2.17 million individuals tracked from 1 May 2020 to 30 June 2020. The research protocol for the development of the QCOVID algorithm can be found in [4].

The QCOVID algorithm was commissioned by the Chief Medical Officer for England on behalf of the UK government. The algorithm has been used to inform UK and devolved government policy on combatting the SARS-Cov-2 pandemic, including guidance on social-distancing and shielding measures, as well vaccine prioritisation. [5] It is therefore of great importance to validate the predictions of the algorithm in sub-populations of the UK that were not in the initial training set, but will potentially be subject to those policies.

At the time of writing, there are validation exercises planned in Scotland, Northern Ireland and Wales, and a validation exercise underway in England. Validation work was considered urgent and has been expedited in order to support national decision making. In order to facilitate useful comparison of the results of the separate validation exercises, it is necessary to establish a consistent set of procedures. The purpose of this paper is to explicate a common methodology for the validation of the QCOVID algorithm across the four nations of the UK.

METHODS AND ANALYSIS:

Study design:

1
2 Open, retrospective cohort study designs will be employed, making use of routinely collected data
3 from General Practices for clinical and demographic information, as well as linked datasets on
4 hospital admissions, reverse-transcription polymerase chain reaction (RT-PCR) testing for Covid-
5 19, and registered deaths. We will aim to have national coverage as far as is possible within each of
6 the four nations of the UK.
7
8
9

10 11 12 **Data Sources:**

13
14 Box 1 contains a brief summary of the main datasets that will be used in the validation exercise for
15 each nation
16

17 **Box 1: Main datasets to be used**

18
19
20
21 **England:** Office for National Statistics (ONS) Public Health Linked Data Asset. This dataset is
22 based on the 2011 Census in England covering 40.1 million people, linked at individual level
23 using the NHS number to mortality records, Hospital Episode Statistics (HES) and the General
24 Practice Extraction Service (GPES) data for pandemic planning and research. The data covers
25 80% of the population of England aged 19 and over.
26
27

28
29
30 **Northern Ireland:** National Health Application and Infrastructure Services (NHAIS) will be used
31 for demographic information. The Patient Administration System (PAS) will be used for data on
32 hospital admissions. Death data will be drawn from the Registrar General, and identified as
33 Covid-19 related through the official Northern Ireland Statistics and Research Agency (NISRA)
34 dashboard. The General Practice Information Platform (GPIP) will bring together GP records
35 from practices across Northern Ireland into a single dataset for use in the validation. As this is not
36 held in the Honest Broker Service, a separate request to its governance board is being made. The
37 Electronic Prescribing Database (EPD) will be used to access information on prescriptions.
38
39

40
41
42 **Scotland:** EAVE II dataset [6]. Contains primary health care records for 5.4 million people covering
43 99% of the population of Scotland, linked with secondary care data from Scottish Morbidity
44 Record (SMR), Covid-19 test results from Electronic Communication of Surveillance Scotland
45 (ECOSS), and mortality data from National Records Scotland.
46
47

48
49
50
51 **Wales:** Secure Anonymised Information Linkage (SAIL system) [7]. This will utilise the Controlling
52 Covid (ConCOV) platform linking records on 3.2 million people from the NHS population spine
53 with hospital (Patient Episode Database for Wales), Welsh Longitudinal GP record (WLGP),
54 Covid-19 test results from the Laboratory Information Management System (LIMS), and
55 mortality and 2011 Census data from the Office for National Statistics (ONS) [8]
56
57
58
59
60

Selection criteria:

Any individual in the relevant linked dataset between the ages of 19 and 100 will be included.

Individuals who had an event (hospitalisation or death) in the first period (24 January 2020 – 30

April 2020) will be excluded from any analysis in the second period (1 May 2020 – 30 June 2020).

These time periods were chosen to mirror the time periods in the original QCOVID paper. After the vaccination programme started in the UK on 8 December 2020, work had already begun on QCOVID 2&3, which will take into account vaccination status. Future validation work will focus on QCOVID 2&3 for more recent time periods.

Exposure and Outcomes:

Table 1 and 2 list all exposure and outcomes variables respectively for the QCOVID algorithm, along with a description, variable type (e.g. integer, real, categorical) and possible values.

Table 1: Exposure variables in QCOVID algorithm.

Variable:	Description/Question:	Value:
Demographic:		
age	Age in years	Integer: 19-100
sex	Biological sex at birth	Categorical: female, male
town	Townsend Deprivation Score	Real number
ethnicity	Ethnicity	Categorical: White, Indian, Pakistani, Bangladeshi, Other Asian, Caribbean, Black African, Chinese, other ethnic group
homecat	What is your housing category - care home or homeless or neither?	Categorical: neither, care home, homeless

Clinical:		
bmi	Body Mass Index (kg/m ²)	Positive real number

chemocat	Have you had chemotherapy in the last 12 months?	Categorical: none, group A, group B, group C
learncat	Do you have a learning disability or Down's Syndrome?	Categorical: learning disability, Down syndrome
renalcat	Chronic Kidney Disease (CKD) stage	Categorical: No serious kidney disease, CKD stage 3, CKD stage 4, CKD stage 5 without dialysis or transplant, CKD stage 5 with dialysis in last 12 months, CKD stage 5 with transplant
diabetescat	Do you have diabetes?	Categorical: none, type 1, type 2
b2_82	Have you been prescribed immunosuppressants four or more times in the previous 6 months?	Categorical: yes, no
b2_leukolaba	Have you been prescribed anti-leukotriene or long acting beta2-agonists (LABA) four or more times in the previous 6 months?	Categorical: yes, no
b2_prednisone	Have you been prescribed oral prednisolone containing preparations prescribed four or more times in the previous 6 months?	Categorical: yes, no
b_AF	Do you have atrial fibrillation?	Categorical: yes, no
b_CCF	Do you have heart failure?	Categorical: yes, no
b_asthma	Do you have asthma?	Categorical: yes, no

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

b_bloodcancer	Have you a cancer of the blood or bone marrow such as leukaemia, myelodysplastic syndromes, lymphoma or myeloma and are at any stage of treatment?	Categorical: yes, no
b_cerebralpalsay	Do you have cerebral palsy?	Categorical: yes, no
b_chd	Do you have coronary heart disease?	Categorical: yes, no
b_cirrhosis	Do you have cirrhosis of the liver?	Categorical: yes, no
b_congenheart	Do you have congenital heart disease or have you had surgery for it in the past?	Categorical: yes, no
b_copd	Do you have chronic obstructive pulmonary disease (COPD)?	Categorical: yes, no
b_dementia	Do you have dementia?	Categorical: yes, no
b_epilepsy	Do you have epilepsy?	Categorical: yes, no
b_fracture4	Have you had a prior fracture of hip, wrist, spine or humerus?	Categorical: yes, no
b_neurorare	Do you have motor neurone disease, multiple sclerosis, myasthenia, or Huntingtons's Chorea?	Categorical: yes, no
b_parkinsons	Do you have Parkinson's disease?	Categorical: yes, no
b_pulmhyper	Do you have pulmonary hypertension or pulmonary fibrosis?	Categorical: yes, no
b_pulmrare	Do you have cystic fibrosis or bronchiectasis or alveolitis?	Categorical: yes, no
b_pvd	Do you have peripheral vascular disease?	Categorical: yes, no
b_ra_sle	Do you have rheumatoid arthritis or SLE?	Categorical: yes, no
b_respcancer	Do you have lung or oral cancer?	Categorical: yes, no

b_semi	Do you have severe mental illness?	Categorical: yes, no
b_sicklecelldisease	Do you have sickle cell disease or severe combined immune deficiency syndromes?	Categorical: yes, no
b_stroke	Have you had a stroke or TIA?	Categorical: yes, no
b_vte	Have you had a thrombosis or pulmonary embolus?	Categorical: yes, no
p_marrow6	Have you had a bone marrow or stem cell transplant in the last 6 months?	Categorical: yes, no
p_radio6	Have you had radiotherapy in the last 6 months?	Categorical: yes, no
p_solidtransplant	Have you had a solid organ transplant (lung, liver, stomach, pancreas, spleen, heart or thymus)?	Categorical: yes, no

Table 2: Outcomes variables in QCOVID algorithm.

Variable:	Description/Question:	Value:
Time to Covid-19 hospitalisation	Time to hospitalisation with RT-PCR confirmed Covid-19 infection in the cohort period in days.	Real number: 0-91
Time to Covid-19 death	Time to death with Covid-19 confirmed or suspected on their death certificate, or confirmed by RT-PCR test, in the cohort period in days.	Real number: 0-91

Whenever available, all variables will be taken as the most recent recorded value in the relevant dataset at the date of entry into the cohort. The Townsend Deprivation Score (TDS) will be

1
2 determined by matching available residential location information with output area and the
3 corresponding TDS from the 2011 UK census [9]. Categories for the variable chemocat will be
4 determined using the lookup table in the supplemental materials.
5
6
7

8 **Data cleaning:**

9 The following procedures will be used for data cleaning:

- 10 • **diabetes_cat:** If the most recent entry has both type 1 and types 2 recorded, diabetes_cat
11 will be set to type 2.
- 12 • **BMI:** The most recently recorded patient BMI within the last 5 years. If the most recently
13 recorded BMI is from more than 5 years ago at the search date, BMI will be set to missing
14 value. Implausible values for BMI (<12 or >70) will be set to missing value.
- 15 • **learncat:** If a patient is recorded has having both learning disability and Down's syndrome,
16 learncat will be set to Down's syndrome.
17
18
19
20
21
22
23
24
25

26 **Missing data:**

27 For comorbidities and medication use and treatments, missing values will be taken to mean absence
28 of that factor. Modal substitution will be considered for missing values for ethnicity. For any other
29 missing values of predictor variables, a single imputation will be considered. Outcome variables
30 will not be imputed, and nor will they be included as predictors in the imputation. The following
31 methods may be considered for use in the imputation: predictive mean matching, least squares,
32 logistic and multinomial models, imputation by chained equations.
33
34
35
36
37
38
39
40
41
42
43
44

45 **Statistical Analysis:**

46 Each validation exercise will report a table of cohort characteristics, following Table 2 in [1]. The
47 main performance metrics that will be calculated are R^2 [9], Harrell's C, Royston's D [10] and the
48 Brier score. Different stratifications for these statistics will be considered, including by age, sex and
49 time period. 95% confidence intervals will be reported for R^2 , Harrell's C and Royston's D. Graphs
50 of observed and predicted probability of hospital admission and death by vigintile for stratified
51 subgroups will be reported, following [1].
52
53
54
55
56
57
58

59 **Sample Size:**

1
2 A preliminary sample size calculation can be done using figures from the original paper [1]. Using
3 the estimated standard deviation of Harrell's C for females in the first time period and assuming
4 Harrell's C is asymptotically normally distributed implies that a sample size of approximately 5,714
5 would be sufficient to correctly reject a null hypothesis of $C=0.5$ at significance level 0.05 with
6 probability 80% given a true value of $C=0.8$. Repeating this calculation for other population
7 subgroups and time periods yields results of a similar magnitude. The samples sizes in the planned
8 studies will be on the order of hundreds of thousands or millions.
9
10
11
12
13
14

15 **Ethics, reporting and dissemination**

16 The ethics approval for the development and validation of QCOVID in England was granted by the
17 East Midlands-Derby Research Ethics Committee [reference 18/EM/0400]. For Scotland, approvals
18 have been obtained by the National Research Ethics Service Committee (REC), South East Scotland
19 02 (REC number: 12/SS/0201) and the Public Benefit and Privacy Panel for Health and Social Care
20 02 (REC number: 1920-0279). The data to be used in this study for Wales are available in the
21 SAIL Databank at Swansea University, Swansea, UK. All proposals to use SAIL data are subject to
22 review by an independent Information Governance Review Panel (IGRP). Before any data can be
23 accessed, approval must be given by the IGRP. The IGRP gives careful consideration to each
24 project to ensure proper and appropriate use of SAIL data. When access has been approved, it is
25 gained through a privacy-protecting safe haven and remote access system referred to as the SAIL
26 Gateway. SAIL has established an application process to be followed by anyone who would like to
27 access data via SAIL.[7] Findings will be presented at conferences, published in peer-reviewed
28 journals and to the funders and government COVID-19 advisory bodies as appropriate.
29 Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) and Reporting
30 of studies Conducted using Observational Routinely-collected Data (RECORD) (via the COVID-19
31 extension) checklists will guide our study findings reporting. The Northern Ireland validation study
32 proposal is under review by the NITRE for HSC data accessed via Northern Ireland Honest Broker
33 Service; an Ethics application has been submitted through IRAS.
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49

50 **Author's Contributions:**

51 AS conceived this protocol. CR, VH, FK, TC, JHC, BH, CC, RAL and JL provided country
52 specific information about available data and analysis plans. SK wrote drafts of this protocol. All
53 authors gave final approval of the version to be published.
54
55
56
57
58

59 **Acknowledgments**

1
2 This work will use data provided by patients and collected by a number of organisations. We would
3 like to acknowledge all patients who shared their information as well as all data providers who
4 make anonymised data available for research. In particular, Public Health Scotland, Public Health
5 Wales, Public Health England, the NHS, the SAIL databank, and the Office for National Statistics.
6
7
8
9

10 **Funding:**

11
12 The validation in England will be funded by a grant from the National Institute for Health Research
13 following a commission by the Chief Medical Officer for England. In Scotland, EAVE II is funded
14 by the Medical Research Council [MR/R008345/1] and supported by the Scottish Government. In
15 Wales, ConCOV is supported by the Medical Research Council [MR/V028367/1].
16
17
18
19

20 **Competing interests:**

21
22 AS reports grants from NIHR, grants from MRC, and grants from HRR UK, during the conduct of
23 the study. JL and RAL report grants from UKRI Medical Research Council, during the conduct of
24 the study. JHC reports grants from John Fell Oxford University Press Research Fund, grants from
25 Cancer Research UK (CR-UK) grant number C5255/A18085, through the Cancer Research UK
26 Oxford Centre, grants from the Oxford Wellcome Institutional Strategic Support Fund
27 (204826/Z/16/Z), grants from NIHR, during the conduct of the study; personal fees and other from
28 ClinRisk Ltd, outside the submitted work; and JHC is an unpaid director of QResearch, a not-for-
29 profit organisation which is a partnership between the University of Oxford and EMIS Health who
30 supply the QResearch database used for this work. Carol Coupland reports personal fees from
31 ClinRisk Ltd, outside the submitted work. JHC, AS, and Carol Coupland were members of the
32 research team involved in the development of the QCOVID risk prediction algorithm. All other
33 authors report no conflict of interest.
34
35
36
37
38
39
40
41
42
43
44

45 **Patient and Public Involvement:**

46 There are no plans for Patient and Public Involvement in this research.
47
48
49

50 **Data sharing:**

51 All code used in these analyses will be made publicly available online e.g. through GitHub.
52
53
54

55 **References:**

56
57 1. Clift AK, Coupland CAC, Keogh RH, Diaz-Ordaz K, Williamson E, Harrison EM, Hayward A,
58 Hemingway H, Horby P, Mehta N, Benger J, Khunti K, Spiegelhalter D, Sheikh A, Valabhji J,
59 Lyons RA, Robson J, Semple MG, Kee F, Johnson P, Jebb S, Williams T, Hippisley-Cox J. Living
60

- 1
2 risk prediction algorithm (QCOVID) for risk of hospital admission and mortality from coronavirus
3
4 19 in adults: national derivation and validation cohort study. *BMJ*. 2020 Oct 20;371:m3731. doi:
5 10.1136/bmj.m3731. PMID: 33082154; PMCID: PMC7574532.
6
7 **2.** Townsend, P., Phillimore, P. and Beattie, A. (1988) *Health and Deprivation: Inequality and the*
8 *North*. Routledge, London. doi: 10.7748/ns.2.17.34.s66. PMID: 27415096.
9
10 **3.** Qresearch, <https://www.qresearch.org/>
11
12 **4.** Hippisley-Cox J, Clift AK, Coupland CAC, et al. Protocol for the development and evaluation of
13 a tool for predicting risk of short-term adverse outcomes due to COVID-19 in the general UK
14 population. *MedRxiv* 2020:2020.06.28.20141986-2020.06.28
15
16 **5.** [https://www.gov.uk/government/news/new-technology-to-help-identify-those-at-high-risk-from-](https://www.gov.uk/government/news/new-technology-to-help-identify-those-at-high-risk-from-covid-19)
17 [covid-19](https://www.gov.uk/government/news/new-technology-to-help-identify-those-at-high-risk-from-covid-19)
18
19 **6.** EAVE II, The Usher Institute, <https://www.ed.ac.uk/usher/eave-ii>
20
21 **7.** The Secure Anonymised Information Linkage (SAIL) databank. Available at:
22 <https://saildatabank.com>
23
24 **8.** Lyons J, Akbari A, Torabi F, Davies G, North L, Griffiths R, Bailey R, Hollinghurst J, Fry R,
25 Turner S, Thompson D, Rafferty J, Mizen A, Orton C, Ellwood-Thompson S, Au-Yeung L, Cross
26 L, Gravenor M, Brophy S, Lucini B, John A, Szakmany T, Davies J, Davies C, Williams C,
27 Emmerson C, Cottrell S, Connor T, Taylor C, Pugh R, Diggle PJ, John G, Scourfield S, Hunt J,
28 Cunningham AM, Helliwell K, Lyons RA. (2020) Understanding and responding to COVID19 in
29 Wales: protocol for a privacy protecting data platform for enhanced epidemiology and evaluation of
30 interventions. *BMJ Open* 2020;10:e043010. doi:10.1136/bmjopen-2020-043010
31
32 **9.** 2011 UK Census, Townsend Deprivation Scores,
33 <https://www.statistics.digitalresources.jisc.ac.uk/dataset/2011-uk-townsend-deprivation-scores>
34
35 **10.** Royston P, Sauerbrei W. A new measure of prognostic separation in survival data. *Stat Med*.
36 2004 Mar 15;23(5):723-48. doi: 10.1002/sim.1621. PMID: 14981672.
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

	Group	Final classification	map
1		Group A	1 & 2
2		Group B	3 & 4
3		Group C	5
4			
5			
6			
7			
8			
9			
10			
11			
12			
13			
14			
15			
16			
17			
18			
19			
20			
21			
22			
23			
24			
25			
26			
27			
28			
29			
30			
31			
32			
33			
34			
35			
36			
37			
38			
39			
40			
41			
42			
43			
44			
45			
46			
47			
48			
49			
50			
51			
52			
53			
54			
55			
56			
57			
58			
59			
60			