International Journal of Population Data Science

Journal Website: www.ijpds.org

Validating the QCOVID risk prediction algorithm for risk of mortality from COVID-19 in the adult population in Wales, UK

Jane Lyons[1,*], Vahé Nafilyan[2], Ashley Akbari[1], Gareth Davies[1], Rowena Griffiths[1], Ewen M Harrison[3], Julia Hippisley-Cox[4], Joe Hollinghurst[1], Kamlesh Khunti[5], Laura North[1], Aziz Sheikh[6], Fatemeh Torabi[1], and Ronan A Lyons[1]

Submission History

Submitted:	26/08/2021
Accepted:	20/12/2021
Published:	16/02/2022

[1]Population Data Science, Health Data Research UK, Swansea University Medical School, Swansea, SA2 8PP
[2]Health Analysis and Life Events Division, Office for National Statistics, NP10 8XG
[3]Centre for Medical Informatics, Usher Institute, University of Edinburgh, EH16 4SA
[4]Nuffield Dept of Primary Care Health Sciences, University of Oxford, OX2 6GG
[5]Diabetes Research Centre, University of Leicester, Leicester LE5 4PW
[6]Usher Institute and Health Data Research UK BREATHE Hub, University of Edinburgh, Edinburgh EH8 9AG

Abstract

Introduction

COVID-19 risk prediction algorithms can be used to identify at-risk individuals from short-term serious adverse COVID-19 outcomes such as hospitalisation and death. It is important to validate these algorithms in different and diverse populations to help guide risk management decisions and target vaccination and treatment programs to the most vulnerable individuals in society.

Objectives

To validate externally the QCOVID risk prediction algorithm that predicts mortality outcomes from COVID-19 in the adult population of Wales, UK.

Methods

We conducted a retrospective cohort study using routinely collected individual-level data held in the Secure Anonymised Information Linkage (SAIL) Databank. The cohort included individuals aged between 19 and 100 years, living in Wales on 24th January 2020, registered with a SAIL-providing general practice, and followed-up to death or study end (28th July 2020). Demographic, primary and secondary healthcare, and dispensing data were used to derive all the predictor variables used to develop the published QCOVID algorithm. Mortality data were used to define time to confirmed or suspected COVID-19 death. Performance metrics, including R^2 values (explained variation), Brier scores, and measures of discrimination and calibration were calculated for two periods (24th January–30th April 2020 and 1st May–28th July 2020) to assess algorithm performance.

Results

1,956,760 individuals were included. 1,192 (0.06%) and 610 (0.03%) COVID-19 deaths occurred in the first and second time periods, respectively. The algorithms fitted the Welsh data and population well, explaining 68.8% (95% CI: 66.9-70.4) of the variation in time to death, Harrell's C statistic: 0.929 (95% CI: 0.921-0.937) and D statistic: 3.036 (95% CI: 2.913-3.159) for males in the first period. Similar results were found for females and in the second time period for both sexes.

Conclusions

The QCOVID algorithm developed in England can be used for public health risk management for the adult Welsh population.

Keywords

COVID-19 outcomes; QCOVID algorithm; risk prediction models; SAIL Databank; population data-linkage

[*] Corresponding Author:
Email Address: j.lyons@swansea.ac.uk (Jane Lyons)

https://doi.org/10.23889/ijpds.v5i4.1697

Introduction

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection was first identified in Wuhan, China [1]. On the 24th January 2020, the UK recorded its first case of SARS-CoV-2 and as of 22nd August 2021, there have been 6,492,906 confirmed cases with 131,640 COVID-19-related deaths in the UK [2, 3]. Research has shown that increased age, being male, certain minority ethnic groups, and having pre-existing conditions such as diabetes, cardiovascular disease, and obesity are associated with serious adverse COVID-19 outcomes, including hospitalisation and death [4–9].

To protect the most vulnerable, and to minimise the burden on the National Health Service (NHS) and its staff, it is important to identify those at greatest risk of serious adverse COVID-19 outcomes [10, 11]. COVID-19 risk prediction algorithms can be used to identify and prioritise at-risk individuals for targeting vaccination and treatments as well as to inform risk management decisions and policy as the pandemic evolves [12].

The New and Emerging Respiratory Virus Threats Advisory Group (NERVTAG)'s effort to develop a population risk assessment framework led to the development and validation of the QCOVID tool, a population-based prediction algorithm to predict the risk of being admitted to hospital or dying from COVID-19 across an adult population [3, 13, 14]. The algorithm was initially developed and validated on a cohort of six million primary care patients from 1,205 English practices contributing to the QResearch database, which allows linkage at the individual-level to general practitioner (GP) primary care data, death records, hospital admissions data and COVID-19 test results. Predictive demographic, clinical, and pharmaceutical variables (Box 1) were based on the clinical vulnerability group criteria used to identify those advised to shield at the start of the pandemic, and risk factors associated with adverse outcomes for respiratory diseases [15, 16].

Replication of results in diverse populations is an important component of scientific research and is especially important for validation of prediction algorithms generated using routine data where the results may be used to plan clinical management of individual patients. It was decided to replicate and compare the performance of the algorithm in each of the four nations in the UK to ensure validity and contribute to the application of the algorithm in managing responses to the outbreak. A recent published study validated the QCOVID predictive algorithm in estimating the risk of mortality from COVID-19 in 35 million adult residents of England by the Office for National Statistics using linked Census 2011 data [17]. The aim of our study was to externally validate the QCOVID risk prediction algorithm to estimate mortality outcomes from COVID-19 in adults in Wales, UK. This paper replicates the English validation study and follows the RECORD and TRIPOD reporting guidelines [18, 19].

Methods

Study design and data sources

This study used routinely collected anonymised health and demographic data held in the Secure Anonymised Information Linkage (SAIL) Databank to create a retrospective population-based individual-level linked e-cohort. The SAIL Databank is a Trusted Research Environment (TRE), which hosts linkable anonymised individual and household-level health, demographic, administrative and environmental data for the population of Wales [20, 21].

Following the emergence of the SARS-CoV-2 infection and the subsequent COVID-19 pandemic, two population-level cohorts (known as C16 and C20) were created to support rapid analysis, provide evidence in understanding the evolving pandemic, and evaluate national interventions attempting to reduce the spread of infection [22]. The C20 contains all individuals alive and living in Wales from 1st January 2020 and followed up until death, emigration/break in Welsh residency, or cohort end date (currently 30th June 2021). This cohort is updated on a monthly basis to extend the available follow-up time. The C16 acts as a contextual comparative cohort and contains all individuals alive and living in Wales on 1st January 2016 and followed up until death, emigration/break in Welsh residency, or 31st December 2019.

For this study, we used the C20 to create a cohort of all individuals aged 19–100 years, living in Wales and registered with a SAIL providing general practice on 24th January 2020. The 24th January 2020 was chosen as the cohort entry as this is the date of the first confirmed COVID-19 case in the UK. Individuals were followed up until death or study end date (28th July 2020), with the study divided into two time periods, 24th January 2020–30th April 2020 and 1st May 2020–28th July 2020, to match the English validation study [17]. Individuals who had died prior to 1st May 2020 were excluded from the second time period analysis.

Predictor variables

To validate the QCOVID algorithm, the C20 cohort was linked to the Welsh Longitudinal General Practice (WLGP), Patient Episode Database for Wales (PEDW), Wales Dispensing DataSet (WDDS), and Office for National Statistics (ONS) Census 2011 (CENW) data [23] to derive the pre-existing conditions and demographic characteristics that were used to develop the QCOVID algorithm (Box 1).

The C20 cohort was used to define age, sex, and Townsend score. Townsend score is a measure of deprivation, based on the area of residence, and a higher score implies a higher level of deprivation. The CENW is linked to derive ethnicity (i.e. Bangladeshi, Black African, Black Caribbean, Chinese, Indian, Pakistani, Mixed, Other, and White) [24]. The ethnicity variable had a category corresponding to 'not recorded/unknown'. This category was used whenever the corresponding value was missing.

The majority of pre-existing conditions were identified in the WLGP primary care data source using Read codes version 2 (CTV2). Where no timeframe was stated, a lookback period from 1st January 1998 to 24th January 2020 was used. For body mass index (BMI), the latest BMI measurement within 5 years to 24th January 2020 was used. BMI records outside this time period as well as BMIs <15 and >47 were set to missing. If an individual had multiple BMI records on the latest date, the highest BMI was included. Predicted values using all QCOVID predictor variables with age interactions from linear regression models, were used to impute any missing BMI values. Recorded BMI is dependent on the condition of interest and healthcare utilisation activity of the individual, therefore, it is possible to have individuals with no BMI recorded when using routinely collected healthcare data. For diabetes, if the latest health record had defined an individual with both type 1 and type 2 diabetes, type 2 took precedence [3]. For the housing covariate, if the latest record defined an individual being homeless and living in a care home, then living in a care home took precedence. For the learning disabilities covariate, if the latest record identified an individual as having learning disabilities and Down's syndrome, then Down's syndrome was prioritised.

Office of Population Censuses and Surveys (OPCS) Classification of Interventions and Procedures version 4 (OPCS-4) coded conditions in the inpatient (PEDW) data were used to identify chemotherapy status, Chronic Kidney Disease (CKD) stages, congenital heart disease surgery, bone marrow or stem cell transplant, radiotherapy, and solid organ transplant.

DMD (Dictionary of Medical Devices) coded prescriptions in the WDDS were used to identify individuals who had been dispensed immunosuppressants, anti-leukotriene or long acting beta2-agonists (LABA), or oral prednisolone at least four or more times within 6-months prior to 24th January 2020.

Outcome of interest – death involving COVID-19

We utilised a combination of data held in ONS Annual District Death Extract (ADDE) and Annual District Death Daily (ADDD), Welsh Demographic Service Dataset (WDSD) and Consolidated Death Data Source (CDDS) to identify all deaths, inclusive of in-hospital and out of hospital deaths, of Welsh residents. Deaths involving COVID-19 (confirmed or suspected) were identified using the tenth revision of the International Classification of Diseases (ICD-10) codes U07.1 or U07.2, or from text fields containing the causes of death within the data sources. Time to death from COVID-19 was calculated separately in the first period (24*th* January 2020–30*th* April 2020) and the second period (1*st* May 2020–28*th* July 2020).

Algorithm validation

The QCOVID risk equations (version 1) reported in the original study were fitted for males and females separately [3, 14]. The original paper utilised the Fine-Gray sub-distribution hazard model which is commonly used to estimate incidence of outcomes where competing risks exist. It relates covariates to the cumulative incidence function (CIF) of the outcome of interest [25, 26]. The following modifications for the Welsh adult population were required due to data issues. At the time of analysis, Systemic Anti-Cancer Therapy (SACT) data were not available, therefore, anyone receiving chemotherapy within 12-months of 24th January 2020 was assigned the chemotherapy group B (middle severity group) coefficients from the original study [27]. Due to low cohort numbers and subsequent outcome numbers for some ethnic groups, we collapsed ethnic groups to ensure ethnic minority populations or groups were not excluded from our study. Black Caribbean individuals were assigned Black African coefficients, Chinese individuals were assigned the coefficients for the Other ethnic group, and, all White ethnic groups were assigned the White British coefficients.

Performance metrics, including measures of discrimination and calibration, were calculated to validate the predicted risk of death from COVID-19 using the QCOVID algorithm at 97 days for the first period and 88 days for the second period [28–30]. We calculated R^2 values, D statistic, Harrell's C statistic and Brier scores with corresponding 95% confidence intervals for the total cohort by sex and over the two time periods. The performance measurements were also calculated by age bands, ethnicity and Townsend deprivation quintiles. The R^2 values refer to the proportion of variation in survival time explained by the model while the Brier score measures predictive accuracy. The D statistic and Harrell's C statistic are discrimination measures that quantify the separation in survival between patients with different levels of predicted risks, and the extent to which people with higher risk scores have earlier events, respectively. To measure calibration, we compared the mean observed and predicted risks within each twentieths of predicted risk (20 groups) for the two time periods. Observed risks were derived in each of the 20 groups using non-parametric estimates of the cumulative incidences.

Results

Overall, there were 1,956,760 individuals aged 19-100 years included in the final analysis for Wales. Of these, 967,975 (49.5%) were male with a mean age of 50.8 (SD 18.7) and the majority of individuals were from White ethnic backgrounds (1,741,527, 89.0%) (Table 1). In comparison with the English validation cohort and original cohort (Supplementary Table 1), these distributions of demographic characteristics were similar except for ethnicity with a lower proportion of individuals from ethnic minority backgrounds in Wales, but also a higher proportion (6.5%) of individuals missing this information (Table 1). The Welsh cohort had similar prevalence of pre-existing conditions when compared to the English validation cohort and original cohort. However, the proportion of people with higher BMI, CKD, respiratory cancer, venous thromboembolism (VTE), coronary heart disease (CHD) and osteoporotic fractures was slightly higher in the Welsh data and slightly lower for immunosuppressant use, dementia, or a serious mental illness compared to the English validation cohort. The proportions of people with missing BMI values, pulmonary hypertension and VTE were slightly higher in the Welsh data compared to original cohort.

In total, there were 1,192 (0.06%) COVID-19 deaths during the first period and 610 (0.03%) in the second period, which was similar to the English validation (0.08% and 0.04%, respectively) [16]. In general, individuals who died from COVID-19 during the first period were more likely to be male (674, 56.5%), aged 70 years and older (976, 81.9%), with diabetes, CKD, obesity, and cardio-pulmonary diseases being the pre-existing conditions with the highest proportions of death (Table 1). Individuals who died from COVID-19 during the second period had similar characteristics to the first period, Box 1: List of predictor variables for the QCOVID risk equations

Demographic

- Age in years on 24th January 2020
- Biological sex at birth
- Townsend Deprivation Score
- Ethnicity
- What is your housing category - care home, homeless or neither?

Lifestyle

- Body Mass Index

Conditions on current shielding patient list

- Have you had chemotherapy in the last 12 months?
- Have you had radiotherapy in the last 6 months?
- Have you had a bone marrow or stem cell transplant in the last 6 months?
- Have you had a solid organ transplant (lung, liver, stomach, pancreas, spleen, heart or thymus)?
- Do you have sickle cell disease or severe combined immune deficiency syndromes?
- Do you have cystic fibrosis, bronchiectasis or alveolitis?
- 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?
- Do you have lung or oral cancer?
- Do you have congenital heart disease or have you had surgery for it in the past?

Conditions moderately associated with increased risk of complications as per current NHS guidance

- Do you have a learning disability or Down's Syndrome?
- Chronic Kidney Disease (CKD) stage
- Do you have asthma?
- Do you have diabetes?
- Do you have Parkinson's disease?
- Do you have cerebral palsy?
- Do you have epilepsy?
- Do you have rheumatoid arthritis or Systemic lupus erythematosus?
- Do you have dementia?
- Do you have chronic obstructive pulmonary disease (COPD)?
- Do you have motor neurone disease, multiple sclerosis, myasthenia, or Huntington's chorea?
- Do you have coronary heart disease?
- Do you have heart failure?

Other medical conditions that investigators hypothesized to confer elevated risk

- Do you have peripheral vascular disease?
- Do you have severe mental illness?
- Have you had a prior fracture of hip, wrist, spine or humerus?
- Do you have atrial fibrillation?
- Do you have cirrhosis of the liver?
- Do you have pulmonary hypertension or pulmonary fibrosis?
- Have you had a thrombosis or pulmonary embolus?
- Have you had a stroke or transient ischaemic attack?

Concurrent medications

- Have you been prescribed immunosuppressants four or more times in the previous 6 months?
- Have you been prescribed anti-leukotriene or long acting beta2-agonists (LABA) four or more times in the previous 6 months?
- Have you been prescribed oral prednisolone containing preparations prescribed four or more times in the previous 6 months?

Table 1: Demographic and clinical characteristics for the total cohort and those who died with COVID-19 in the two time periods

	Overall cohort		**COVID-19 deaths in first period (24th Jan–30th Apr 2020)**		**COVID-19 deaths in second period (1st May–28th Jul 2020)**	
	N	**%**	**N**	**%**	**N**	**%**
Overall	1,956,760		1192		610	
Sex						
Male	967,975	49.47	674	56.54	299	49.02
Female	988,785	50.53	518	43.46	311	50.98
Age, years	50.8	18.7	79.4	11.8	81.0	11.1
Age group, years						
19-29	318,681	16.29	*		*	
30-39	313,802	16.04	*		*	
40-49	304,363	15.55	16	1.34	*	
50-59	353,539	18.07	61	5.12	28	4.59
60-69	291,042	14.87	132	11.07	49	8.03
70-79	240,840	12.31	305	25.59	136	22.30
80-89	111,631	5.70	429	35.99	250	40.98
≥90	22,862	1.17	242	20.30	138	22.62
Ethnicity						
Bangladeshi	7,011	0.36	*		*	
Black^	8,312	0.42	*		*	
Indian	8,885	0.45	*		*	
Mixed	27,582	1.41	*		*	
Other^	27,786	1.42	*		*	
Pakistani	7,688	0.39	*		0	0.00
White	1,741,527	89.00	1113	93.37	579	94.92
Not recorded	127,969	6.54	52	4.36	19	3.11
Townsend deprivation quintile						
1 (most affluent)	335,459	17.14	156	13.09	98	16.07
2	413,486	21.13	221	18.54	129	21.15
3	559,024	28.57	369	30.96	179	29.34
4	453,474	23.17	304	25.50	141	23.11
5 (most deprived)	195,317	9.98	142	11.91	63	10.33
Accommodation						
Neither homeless nor care home	1,940,224	99.15	987	82.80	476	78.03
Care home or nursing home	16,536	0.85	205	17.20	134	21.97
Body-mass index, kg/m2						
<18.5	21,944	1.12	53	4.45	33	5.41
18.5 to <25	316,569	16.18	277	23.34	161	26.39
25 to <30	375,501	19.19	300	25.17	154	25.25
≥30	403,871	20.64	294	24.66	114	18.69
Not recorded	838,875	42.87	268	22.48	148	24.26
Chronic kidney disease						
No Chronic Kidney disease	1,874,451	95.79	869	72.90	412	67.54
Stage 3	72,669	3.71	252	21.14	165	27.05
Stage 4	3,928	0.20	30	2.52	20	3.28
Stage 5	5,712	0.29	41	3.44	13	2.13
Learning disability						
No learning disability	1,928,040	98.53	1163	97.57	587	96.23
Learning disability	28,486	1.46	29	2.43	23	3.77
Down Syndrome	234	0.01	0	0.00	0	0.00

(Continued.)

Table 1: Continued

	Overall cohort		COVID-19 deaths in first period (24th Jan–30th Apr 2020)		COVID-19 deaths in second period (1st May–28th Jul 2020)	
	N	%	N	%	N	%
Chemotherapy						0.00
No chemotherapy in past 12-months	1,949,761	99.64	1167	97.90	597	97.87
Chemotherapy in past 12-months	6,999	0.36	25	2.10	13	2.13
Cancer and immunosuppression						
Blood cancer	10,547	0.54	38	3.19	14	2.30
Respiratory cancer	5,691	0.29	20	1.68	10	1.64
Radiotherapy in past 6-months	1,827	0.09	*		*	
Bone marrow transplant in past 6-months	56	0.00	0	0	0	0.00
Solid organ transplant	806	0.04	*		*	
Prescribed immunosuppressant medication by GP	2,884	0.15	*		*	
Prescribed leukotriene or LABA	38,658	1.98	59	4.95	42	6.89
Prescribed regular prednisolone	15,819	0.81	61	5.12	28	4.59
Other comorbidities						
Diabetes	161,227	8.24	359	30.12	178	29.18
COPD	66,937	3.42	209	17.53	100	16.39
Asthma	290,490	14.85	186	15.60	109	17.87
Rare pulmonary diseases	9,471	0.48	26	2.18	12	1.97
Pulmonary hypertension or pulmonary fibrosis	3,741	0.19	17	1.43	14	2.30
Coronary heart disease	89,686	4.58	239	20.05	137	22.46
Stroke	55,336	2.83	233	19.55	121	19.84
Atrial fibrillation	62,712	3.20	253	21.22	140	22.95
Congestive cardiac failure	30,937	1.58	151	12.67	99	16.23
Venous thromboembolism	43,708	2.23	111	9.31	54	8.85
Peripheral vascular disease	18,639	0.95	77	6.46	36	5.90
Congenital heart disease	17,071	0.87	30	2.52	12	1.97
Dementia	18,840	0.96	304	25.50	160	26.23
Parkinson's disease	5,717	0.29	40	3.36	32	5.25
Epilepsy	26,112	1.33	31	2.60	19	3.11
Rare neurological conditions	5,789	0.30	*		*	
Cerebral palsy	1,318	0.07	0	0.00	0	0.00
Severe mental illness	282,709	14.45	209	17.53	109	17.87
Osteoporotic fracture	73,679	3.77	154	12.92	96	15.74
Rheumatoid arthritis or SLE	22,485	1.15	35	2.94	16	2.62
Cirrhosis of the liver	7,210	0.37	17	1.43	*	
Sickle cell disease	1,094	0.06	0	0	0	0.00

Data are n (%) or mean (SD). * represents values which have been suppressed due to small numbers <10. ^represents collapsing of categories to suppress small numbers.

however, with a slight change to the sex ratio (56.5% of deaths in first period were in males compared to 51.0% deaths in the second period were in females).

The performance metrics calculated to validate the predicted risk of death from COVID-19 using the QCOVID algorithm are presented in Table 2 [3, 14]. The metrics have been provided for both sexes and time periods. In the first time-period for males, the algorithm explained 68.8% (95% CI: 66.9–70.4) of the variation in time to death, the Harrell's C statistic was 0.929 (95% CI: 0.921–0.937), the D statistic was 3.036 (95% CI: 2.913–3.159) and Brier score was 0.0007. Similar results were found for females and in the second time period. Similar results were also found in the English validation, the D statistics was 3.761 (3.732–3.789), Harrell's C statistic was 0.935 (95% CI: 0.933–0.937) and Brier score was 0.0013 in males in the first period, with similar results found in females and in the second time period [17]. Performance metrics by age band, ethnicity and Townsend deprivation quintile can be found in the Appendices (Supplementary Tables 2–5).

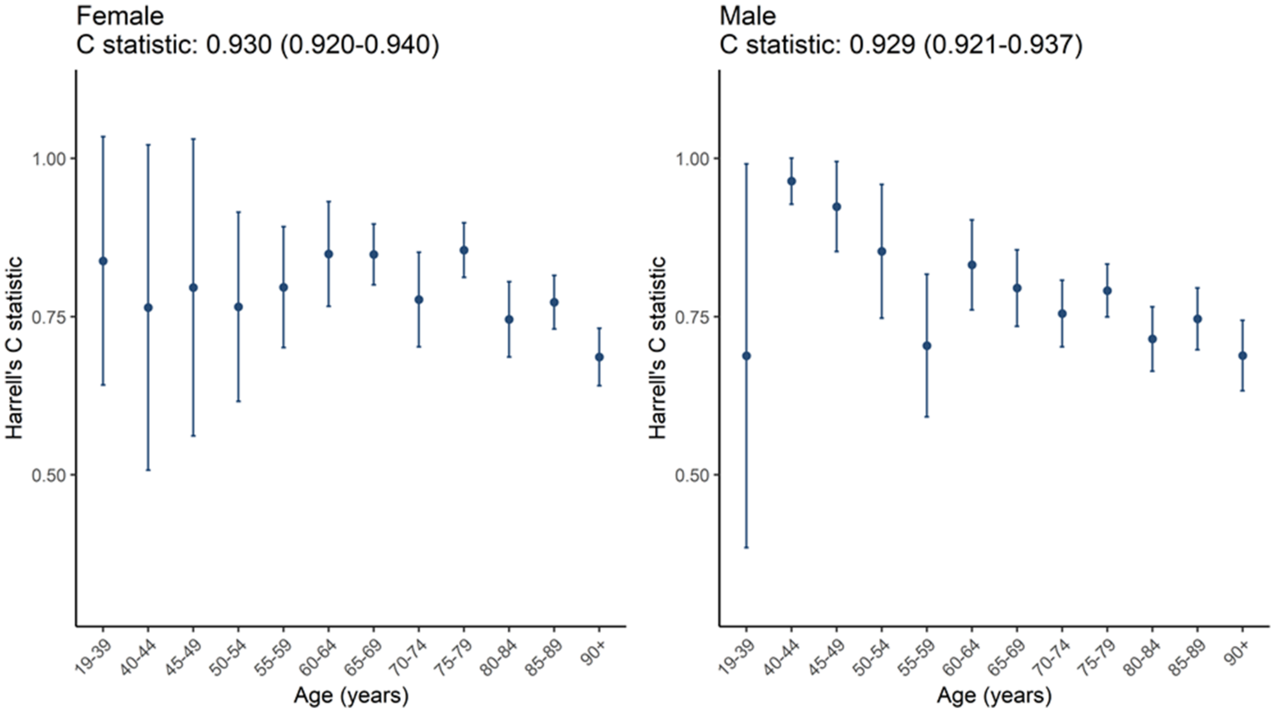
The Harrell's C statistic varied across the age bands and time periods (Figures 1, 2), with acceptable discrimination (>0.7) in both time periods for males and females, and across age groups. The oldest group (90+ years old) yielded poorer discrimination for both males and females as well as the youngest male group in the first time period. In the second

Table 2: Performance of the risk models to predict risk of COVID-19 death by sex and time period for total cohort

	First period (24th January 2020–30th April 2020)		Second period (1st May 2020–28th July 2020)	
	COVID-19 death in females	**COVID-19 deaths in males**	**COVID-19 death in females**	**COVID-19 deaths in males**
R-squared statistic	0.691 (0.671–0.710)	0.688 (0.669–0.704)	0.721 (0.698–0.742)	0.711 (0.686–0.733)
D statistic	3.062 (2.922–3.202)	3.036 (2.913–3.159)	3.293 (3.113–3.472)	3.207 (3.024–3.390)
Harrell's C statistic	0.930 (0.920–0.940)	0.929 (0.921–0.937)	0.950 (0.942–0.959)	0.933 (0.921–0.945)
Brier score	0.0005	0.0007	0.0003	0.0003

Data are estimated (95% CI).

Figure 1: The concordance index by sex and age group in the first time period (24$^{\text{th}}$ January–30$^{\text{th}}$ April 2020)



Bars represent 95% CI.

time period, it was not possible to plot the Harrell's C statistic for the youngest age groups for females (19-39 and 40-44 years) or for 19-39 years in males due to low numbers. Whilst the Harrell's C statistic was slightly lower in Wales compared to England across sex and age groups, the pattern of reduced discrimination for certain age groups was similar.

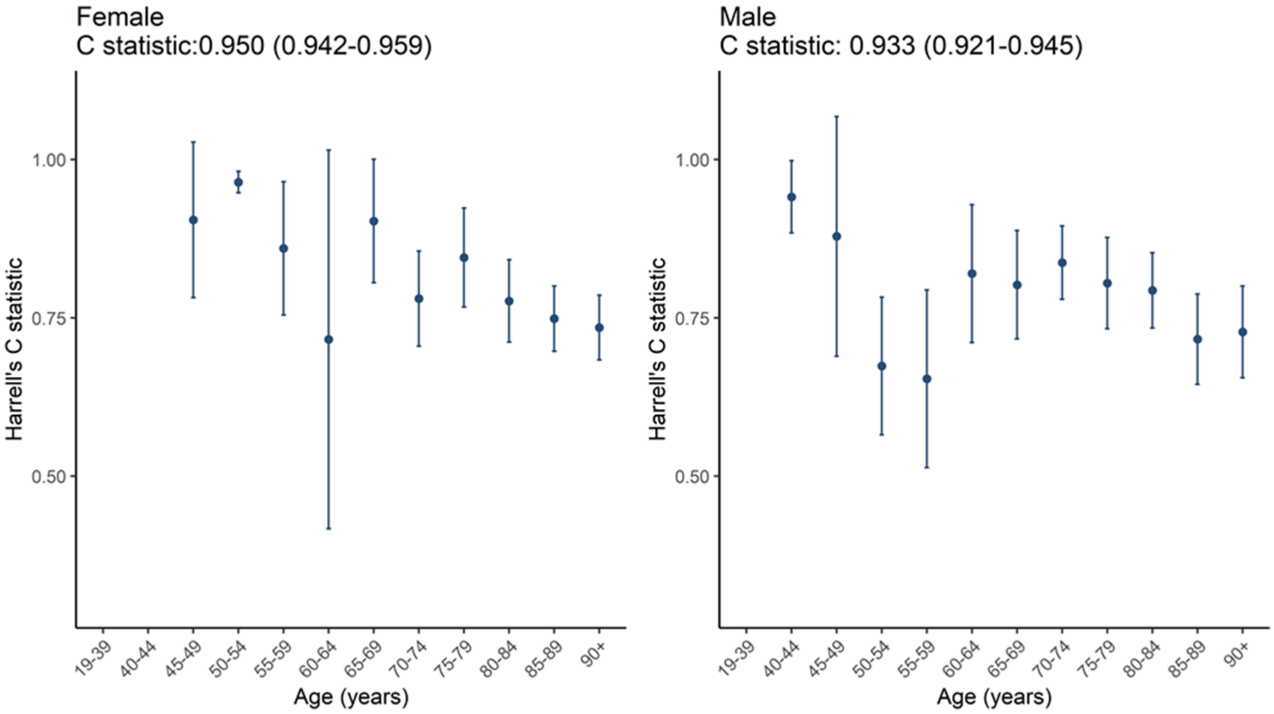
The calibration plots in Figure 3 showed that the predicted and observed risks of COVID-19 related death were similar for both males and females in the first time period, demonstrating the QCOVID equations were well calibrated. However, there was slight under-prediction in the highest risk category for COVID-19 death which was also demonstrated in the English validation and original cohorts [3, 17]. Predicted and observed risks of COVID-19 related death in the second time period can be found in Supplementary Figure 1.

Figure 4 demonstrates that the sensitivity at different absolute risk thresholds for COVID-19-related deaths was higher for females in the top 13 centiles compared to males in the first period and was higher in females than males across the second period. 60.2% and 65.4% of deaths occurred in those in the top 5% for predicted absolute risk of death from COVID-19 in the first time period for males and females respectively; 64.9% and 72.0% of deaths occurred in those in the top 5% for predicted absolute risk of death from COVID-19 in the second time period for males and females, respectively (Supplementary Table 6).

Discussion

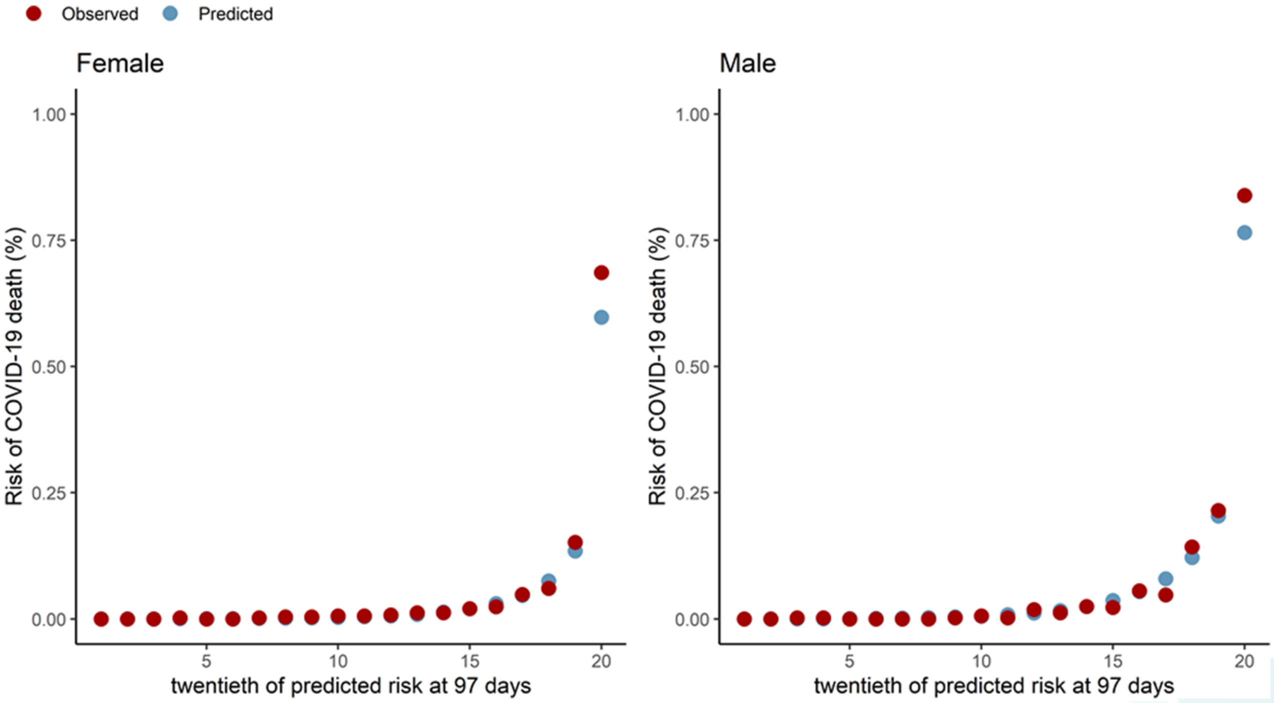
The results from this validation of the QCOVID risk prediction algorithm show that the models fit the Welsh population data well and yielded similar results, but with less precision (predictably, given the smaller population size) compared to the English validation and original study. This study used individual-level linked data on the adult population of Wales, registered with a SAIL providing general practice, which is independent of the original and validation study populations [22]. Use of SAIL Databank allowed linkage across primary and secondary health care data with mortality outcome data to

Figure 2: The concordance index by sex and age group in the second time period (1st May–28th July 2020)



Bars represent 95% CI.

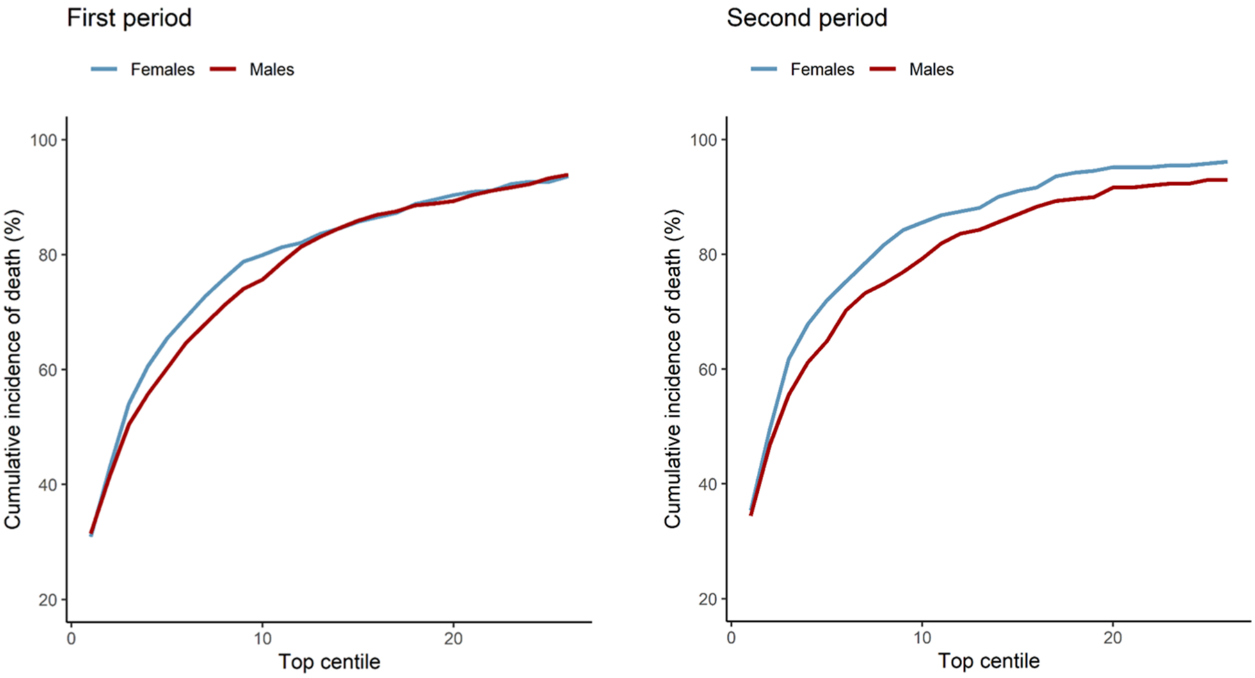
Figure 3: Predicted and observed risk of COVID-19-related death in the first time period (24th January–30th April 2020)



allow replication of the original and English validation studies and inclusive of all predictor variables [3, 17].

The risk models from the original QCOVID and English validation paper were based on GP data largely from England [3, 17]. Age standardised death rates in Wales pre-pandemic were about 6% higher than in England [31]. Some differences in prediction accuracy are expected and this is consistent with the higher observed to predicted mortality numbers at the higher end of risk in Figure 3 [32]. The predicted and observed risks of COVID-19-related death were similar across most of the predicted risk distribution, demonstrating the models were well calibrated, (60.2-72.0% of deaths occurred in the top 5% for predicted absolute risk of death), apart from the highest 20*th* of risk where the risk of death was higher in Wales, as shown in Figure 3. This is similar to the English validation study, which demonstrated 65.9-77.2% of death occurred in

Figure 4: Sensitivity for COVID-19-related death in the first (24th January–30th April 2020) and second (1st May–28th July 2020) time periods



Centiles were based on predicted absolute risks in males and females in each period.

individuals in the top 5% for predicted absolute risk of death [17].

The overall Harrell's C statistic was >0.9 for males and females for both time periods, demonstrating good overall discrimination of the models. Lower and more varied Harrell's C statistics across the age bands are likely due to a smaller population and more deaths occurring in the first period during the first peak of the UK pandemic [33].

Despite the predictive model performance metrics indicating that the algorithm performed well on the Welsh data, there are a number of limitations. The Welsh cohort was restricted to individuals registered to a SAIL providing general practice, therefore, results are based on 80% population coverage (330/412, of all general practices in Wales). This restriction was necessary due to the amount of predictor variables that required primary care GP data. Whilst we were able to calculate all predictor variables required, 42.8% of our cohort did not have a BMI recorded in the previous five years, therefore, missing observations were imputed. Also, this study was designed to replicate the English validation study and therefore focussed on COVID-19-related deaths, COVID-19-related hospital admissions will be presented in a subsequent paper. Additionally, as highlighted in the English validation study, testing for COVID-19 was limited in the early stages of the pandemic and therefore some of the early deaths might not be recorded as being COVID-19-related. As this study period covers the start of the pandemic, outcomes relate to the COVID-19 Wild type triggered wave and does not include subsequent Alpha and Delta variant waves. Finally, it was not possible to calculate performance metrics for some age groups and ethnic groups. Due to low numbers of some ethnic groups and consequent death we collapsed some ethnic groups to ensure privacy protection whilst including them in our study. We combined Black African and Black Caribbean groups, and Chinese and Other groups. This analysis was carried out on a smaller and less ethnically diverse population compared to the original studies [3, 17].

Conclusion

This validation of the QCOVID algorithm indicates that the risk prediction models are applicable on a population independent of the original study, which has not been reported before. Our validation is based on Welsh primary care registered patients, for whom the QCOVID algorithm was not modelled on, whereas the original study was based on English primary care registered patients. The Welsh validation offers evidence that the QCOVID algorithm can be used for public health risk management and also could be applied to other populations. This study covered the first wave of the pandemic in Wales/the UK; however, with the emergence of new variants of concern, subsequent new waves of infection and changes in presentation in symptoms of SARS-CoV-2 it is important to adapt these algorithms over longer periods and assess their predictive ability in the context of the evolving pandemic. Further work will include applying an updated algorithm to assess the predictive risk of COVID-19 death and hospitalisation over a longer period of time. We will also assess the impact of the national vaccination program to see how changes in immunity level have impacted adverse COVID-19 outcomes.

Acknowledgments

This study makes use of anonymised data held in the SAIL Databank. This work uses data provided by patients and collected by the NHS as part of their care and support and the Understanding Patient Data initiative. We would also like to acknowledge all data providers who make anonymised data available for research. We wish to acknowledge the collaborative partnership that enabled acquisition and access to the de-identified data, and sharing of necessary methodological documentation and scripts which led to this output. This is a collaboration between colleagues at University of Oxford, University of Edinburgh, University of Nottingham, Office for National Statistics, London School of Hygiene and Tropical Medicine, University College London, Office of the Chief Medical Officer, Department of Health and Social Care, NHS Digital, University of Leicester, University of Cambridge, NHS England, Queen Mary University of London, University of Liverpool, Queen's University Belfast, Association of Local Authority Medical Advisors, Imperial College London, and Swansea University Health Data Research UK. Swansea University Health Data Research UK team is under the direction of the Welsh Government Technical Advisory Cell (TAC) and includes the following groups and organizations: the SAIL Databank, Administrative Data Research (ADR) Wales, Digital Health and Care Wales (DHCW), Public Health Wales, NHS Shared Services Partnership (NWSSP) and the Welsh Ambulance Service Trust (WAST). All research conducted has been completed under the permission and approval of the SAIL independent Information Governance Review Panel (IGRP) project number 0911. KK is supported by the National Institute for Health Research (NIHR) Applied Research Collaboration East Midlands (ARC EM) and the NIHR Leicester Biomedical Research Centre (BRC).

This work was supported by the Con-COV team funded by the Medical Research Council (grant number: MR/V028367/1). This work was supported by Health Data Research UK, which receives its funding from HDR UK Ltd (HDR-9006) and the Medical Research Council (MR/ S027750/1). HDR UK Ltd is funded by the UK Medical Research Council, Engineering and Physical Sciences Research Council, Economic and Social Research Council, Department of Health and Social Care (England), Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Health and Social Care Research and Development Division (Welsh Government), Public Health Agency (Northern Ireland), British Heart Foundation (BHF) and the Wellcome Trust. This work was supported by the ADR Wales programme of work. The ADR Wales programme of work is aligned to the priority themes as identified in the Welsh Government's national strategy: Prosperity for All. ADR Wales brings together data science experts at Swansea University Medical School, staff from the Wales Institute of Social and Economic Research, Data and Methods (WISERD) at Cardiff University and specialist teams within the Welsh Government to develop new evidence which supports Prosperity for All by using the SAIL Databank at Swansea University, to link and analyse anonymized data. ADR Wales is part of the Economic and Social Research Council (part of UK Research and Innovation) funded ADR UK (grant ES/S007393/1). This work was supported by the Wales COVID-19 Evidence Centre, funded by Health and Care Research Wales.

Conflicts of interest

AS is a member of the Scottish Government's COVID-19 Chief Medical Officer's Advisory Group and its Standing Committee on Pandemics; he is also a member of NERVTAG's Risk Stratification Subgroup. KK is member of NERVTAG subgroup and member of the Scientific Advisory Group for Emergencies (SAGE). JHC reports grants from National Institute for Health Research (NIHR) Biomedical Research Centre, Oxford, grants from John Fell Oxford University Press Research Fund, grants from Cancer Research UK (CR-UK) grant number C5255/A18085, through the Cancer Research UK Oxford Centre, grants from the Oxford Wellcome Institutional Strategic Support Fund (204826/Z/16/Z) and other research councils, during the conduct of the study. JHC is an unpaid director of QResearch, a not-for-profit organisation which is a partnership between the University of Oxford and EMIS Health who supply the QResearch database used for this work. JHC is a founder and shareholder of ClinRisk ltd and was its medical director until 31st May 2019. ClinRisk Ltd produces open and closed source software to implement clinical risk algorithms (outside this work) into clinical computer systems. JHC is chair of the NERVTAG risk stratification subgroup and a member of SAGE COVID-19 groups and the NHS group advising on prioritisation of use of monoclonal antibodies in COVID-19 infection. RAL is a member of the Welsh Government COVID-19 Technical Advisory Group.

Ethics statement

The data used in this study are available in the SAIL Databank at Swansea University, Swansea, UK, but as restrictions apply they are not publicly available. All proposals to use SAIL data are subject to review by an independent Information Governance Review Panel (IGRP). Before any data can be accessed, approval must be given by the IGRP. The IGRP contains a multidisciplinary professional group, including members of the public, and it gives careful consideration to each project to ensure proper and appropriate use of SAIL data. When access has been granted, it is gained through a privacy protecting safe haven and remote access system referred to as the SAIL Gateway. SAIL has established an application process to be followed by anyone who would like to access data via SAIL at https://www.saildatabank.com/application-process. Participant consent was not required for this study as all data is anonymised and further encrypted.

References

1. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. Clinical features of patients infected With 2019 novel coronavirus in Wuhan, China. The Lancet. 2020;395(10223):497–506. https://doi.org/10.1016/S0140-6736(20)30183-5

2. Coronavirus cases: [Internet]. Worldometer. [cited 2021Aug23]. Available from: https://www.worldometers.info/coronavirus/

3. Clift AK, Coupland CA, Keogh RH, Diaz-Ordaz K, Williamson E, Harrison EM, et al. Living risk prediction algorithm (QCOVID) for risk of hospital admission and mortality from coronavirus 19 in adults: National derivation and Validation cohort study. BMJ. 2020;371:m3731. https://doi.org/10.1136/bmj.m3731

4. Zhou F, Yu T, Du R, Fan G, Liu Y, Liu Z, et al. Clinical course and risk factors for mortality of adult inpatients with Covid 19 in Wuhan, China A retrospective cohort study. The Lancet. 2020;395(10229):1054–62. https://doi.org/10.1016/S0140-6736(20)30566-3

5. Harrison SL, Fazio-Eynullayeva E, Lane DA, Underhill P, Lip GY. Comorbidities associated with mortality in 31,461 adults with COVID-19 in the United states: A federated electronic medical record analysis. PLOS Medicine. 2020;17(9). https://doi.org/10.1371/journal.pmed.1003321

6. Richardson S, Hirsch JS, Narasimhan M, Crawford JM, McGinn T, Davidson KW, et al. Presenting Characteristics, Comorbidities, and Outcomes Among 5700 Patients Hospitalized With COVID-19 in the New York City area. JAMA. 2020;323(20):2052. https://doi.org/10.1001/jama.2020.6775

7. Singh AK, Gillies CL, Singh R, Singh A, Chudasama Y, Coles B, et al. Prevalence of co-morbidities and their association with mortality in patients with COVID-19: A systematic review and meta-analysis. Diabetes, Obesity and Metabolism. 2020;22(10):1915–24. https://doi.org/10.1111/dom.14124

8. Sattar N, McInnes IB, McMurray JJV. Obesity is a risk factor for severe covid-19 infection: Multiple potential mecahanisms. Circulation. 2020;142(1):4–6. https://doi.org/10.1161/circulationaha.120.047659

9. Docherty AB, Harrison EM, Green CA, Hardwick HE, Pius R, Norman L, et al. Features of 20 133 UK patients in hospital With Covid-19 using the ISARIC WHO clinical Characterisation Protocol: Prospective observational cohort study. BMJ. 2020;369:m1985. https://doi.org/10.1136/bmj.m1985

10. Smith GD, Spiegelhalter D. Shielding from covid-19 should be stratified by risk. BMJ. 2020;369:m2063. https://doi.org/10.1136/bmj.m2063

11. Hollinghurst J, Lyons J, Fry R, Akbari A, Gravenor M, Watkins A, et al. The impact of COVID-19 on adjusted mortality risk in care homes for older adults in Wales, UK: a retrospective population-based cohort study for mortality in 2016–2020. Age and Ageing. 2020;50(1):25–31. https://doi.org/10.1093/ageing/afaa207

12. Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, et al. Prediction models for diagnosis and prognosis of covid-19: Systematic review and critical appraisal. BMJ. 2020;369:m1328. https://doi.org/10.1136/bmj.m1328

13. New and Emerging Respiratory Virus Threats Advisory Group [Internet]. GOV.UK. GOV.UK; 2021 [cited 2021Nov10]. Available from: https://www.gov.uk/government/groups/new-and-emerging-respiratory-virus-threats-advisory-group

14. Welcome to The Qcovid®risk calculator [Internet]. University of Oxford. [cited 2021Aug18]. Available from: https://qcovid.org/

15. Shielded Patient List [Internet]. Nhs choices. NHS; [cited 2021Aug3]. Available from: https://digital.nhs.uk/coronavirus/shielded-patient-list

16. Who is at high risk from coronavirus (clinically extremely vulnerable) [Internet]. Nhs choices. NHS; [cited 2021Aug3]. Available from: https://www.nhs.uk/conditions/coronavirus-covid-19/people-at-higher-risk/who-is-at-high-risk-from-coronavirus-clinically-extremely-vulnerable/

17. Nafilyan V, Humberstone B, Mehta N, Diamond I, Coupland C, Lorenzi L, et al. An external validation of the QCovid risk prediction algorithm for risk of mortality from COVID-19 in adults: a national validation cohort study in England. The Lancet Digital Health. 2021;3(7). https://doi.org/10.1016/S2589-7500(21)00080-7

18. Benchimol EI, Smeeth L, Guttmann A, Harron K, Moher D, Petersen I, et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement. PLOS Medicine. 2015;12(10). https://doi.org/10.1371/journal.pmed.1001885

19. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. Ann Intern Med 2015;162:55-63. https://doi.org/10.7326/M14-0697

20. Lyons RA, Jones KH, John G, Brooks CJ, Verplancke J-P, Ford DV, et al. The SAIL databank: Linking multiple health and social care datasets. BMC Medical Informatics and Decision Making. 2009;9(1). https://doi.org/10.1186/1472-6947-9-3

21. Ford DV, Jones KH, Verplancke J-P, Lyons RA, John G, Brown G, et al. The SAIL databank: Building a national architecture for e-health research and evaluation. BMC Health Services Research. 2009;9(1). https://doi.org/10.1186/1472-6963-9-157

22. Lyons J, Akbari A, Torabi F, Davies GI, North L, Griffiths R, et al. Understanding and responding To COVID-19 in Wales: Protocol for a privacy-protecting data platform for enhanced epidemiology and evaluation of interventions. BMJ Open. 2020;10(10). https://doi.org/10.1136/bmjopen-2020-043010

23. Gateway HI [Internet]. 2021 [cited 2021Oct26]. Available from: https://web.www.healthdatagateway.org/dataset/

24. UK data Service: Census data [Internet]. 2011 UK Townsend Deprivation Scores |UK Data Service |Census Data. 2017 [cited 2021Aug18]. Available from:

https://statistics.ukdataservice.ac.uk/dataset/2011-uk-townsend-deprivation-scores

25. Austin PC, Steyerberg EW, Putter H. Fine-Gray subdistribution hazard models to simultaneously estimate the absolute risk of different event types: Cumulative total failure probability may exceed 1. Statistics in Medicine. 2021;40(19):4200–12. https://doi.org/10.1002/sim.9023

26. Fine JP, Gray RJ. A Proportional Hazards Model for the Subdistribution of a Competing Risk. J Am Stat Assoc 1999;94:496-509. https://doi.org/10.1080/01621459.1999.10474144

27. Coronavirus (COVID-19) risk assessment; [cited 2021Aug18]. Available from: https://digital.nhs.uk/coronavirus/risk-assessment

28. Royston P. Explained Variation for Survival Models. The Stata Journal: Promoting communications on statistics and Stata. 2006;6(1):83–96. https://doi.org/10.1177/2F1536867X0600600105

29. Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Statistics in Medicine. 1996;15(4):361–87. https://doi.org/10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4

30. Royston P, Sauerbrei W. A new measure of prognostic separation in survival data. Statistics in Medicine. 2004;23(5):723–48. https://doi.org/10.1002/sim.1621

31. Cornish D. Monthly Mortality Analysis, England and Wales: July 2021 [Internet]. Monthly mortality analysis, England and Wales - Office for National Statistics. Office for National Statistics; 2021 [cited 2021Nov16]. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/bulletins/monthlymortalityanalysisenglandandwales/july2021

32. Avoidable mortality in the uk: 2019 [Internet]. Avoidable mortality in the UK - Office for National Statistics. Office for National Statistics; 2021 [cited 2021Aug18]. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/causesofdeath/bulletins/avoidablemortalityinenglandandwales/2019

33. Person. Excess deaths in your neighbourhood during the coronavirus (covid-19) pandemic [Internet]. Excess deaths in your neighbourhood during the coronavirus (COVID-19) pandemic - Office for National Statistics. Office for National Statistics; 2021 [cited 2021Aug23]. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/articles/excessdeathsinyourneighbourhoodduringthecoronaviruscovid19pandemic/2021-08-03

Abbreviations

SAIL:	Secure Anonymised Information Linkage
NERVTAG:	New and Emerging Respiratory Virus Threats Advisory Group
RECORD:	REporting of studies Conducted using Observational Routinely-collected health Data
WLGP:	Welsh Longitudinal General Practice
PEDW:	Patient Episode Database for Wales
WDDS:	Wales Dispensing DataSet
CENW:	Census 2011 data
BMI:	Body Mass Index
ICD-10:	International Classification of Diseases, Tenth Revision
OPCS-4:	OPCS Classification of Interventions and Procedures version 4
CKD:	Chronic Kidney Disease
ONS:	Office for National Statistics
ADDE:	Annual District Death Extract
ADDD:	Annual District Death Daily
WDSD:	Welsh Demographic Service Dataset
CDDS:	Consolidated Death Data Source
DMD:	Dictionary of Medicines and Devices
SACT:	Systemic Anti-Cancer Therapy

Appendix

Supplementary table 1: Demographic and clinical characteristics for the Welsh and English validation cohorts and for those who died with COVID-19 in the two time periods

	Welsh validation study						**English validation study**					
	Overall cohort		**COVID-19 death in first period**		**COVID-19 death in second period**		**Overall cohort**		**COVID-19 death in first period**		**COVID-19 death in second period**	
	N	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**
Overall	1,956,760		1192	0.06	610	0.03	34897648		26,985	0.08	13,177	0.04
Sex												
Male	967,975	49.47	674	56.54	299	49.02	16,599,875	47.57	15,334	56.82	6,617	50.22
Female	988,785	50.53	518	43.46	311	50.98	18,297,773	52.43	11,651	43.18	6,560	49.78
Age, years (mean)	50.8		79.4		81.0		51.09		79.98		82.13	
Age group, years												
19–29	318,681	16.29	*		*		5,601,475	16.05	44	0.16	13	0.10
30–39	313,802	16.04	*		*		5,268,030	15.10	116	0.43	30	0.23
40–49	304,363	15.55	16	1.34	*		5,625,225	16.12	364	1.35	125	0.95
50–59	353,539	18.07	61	5.12	28	4.59	6,435,204	18.44	1,196	4.43	400	3.04
60–69	291,042	14.87	132	11.07	49	8.03	5,185,917	14.86	2,727	10.11	962	7.30
70–79	240,840	12.31	305	25.59	136	22.30	4,225,729	12.11	6,280	23.27	2,695	20.45
80–89	111,631	5.70	429	35.99	250	40.98	2,093,545	6.00	10,841	40.17	5,580	42.35
≥90	22,862	1.17	242	20.30	138	22.62	462,523	1.33	5,417	20.07	3,372	25.59
Ethnicity												
Bangladeshi	7,011	0.36	*		*		258,053	0.74	179	0.66	29	0.22
Black^	8,312	0.42	*		*		895,529	2.56	1,130	4.18	186	1.41
Indian	8,885	0.45	*		*		931,247	2.67	800	2.96	216	1.64
Mixed	27,582	1.41	*		*		551,567	1.58	184	0.68	67	0.51
Other^	27,786	1.42	*		*		1,021,472	2.92	697	2.59	157	1.19
Pakistani	7,688	0.39	*		0	0.00	679,062	1.95	426	1.58	123	0.93
White	1,741,527	89.00	1113	93.37	579	94.92	30,560,718	87.58	23,569	87.34	12,399	94.09
Not recorded	127,969	6.54	52	4.36	19	3.11						
Townsend deprivation quintile												
1 (most affluent)	335,459	17.14	156	13.09	98	16.07	7,491,652	21.47	4,993	18.50	2,842	21.57
2	413,486	21.13	221	18.54	129	21.15	7,738,292	22.17	5,326	19.74	2,967	22.52
3	559,024	28.57	369	30.96	179	29.34	6,834,804	19.58	5,111	18.94	2,647	20.09
4	453,474	23.17	304	25.50	141	23.11	6,467,204	18.53	5,365	19.88	2,472	18.76
5 (most deprived)	195,317	9.98	142	11.91	63	10.33	6,366,096	18.24	6,190	22.94	2,249	17.07
Accommodation												
Neither homeless nor care home	1,940,224	99.15	987	82.80	476	78.03	34,667,007	99.34	19,995	74.10	9,039	68.60
Care home or nursing home	16,536	0.85	205	17.20	134	21.97	230,641	0.66	6,990	25.90	4,138	31.40
Body-mass index, kg/m2												
<18.5	21,944	1.12	53	4.45	33	5.41	393,928	1.13	983	3.64	614	4.66
18.5 to <25	316,569	16.18	277	23.34	161	26.39	6,658,276	19.08	5,776	21.40	2,965	22.50
25 to <30	375,501	19.19	300	25.17	154	25.25	6,661,721	19.09	5,552	20.57	2,385	18.10
≥30	403,871	20.64	294	24.66	114	18.69	5,661,007	16.22	5,540	20.53	2,066	15.68
Not recorded	838,875	42.87	268	22.48	148	24.26	15,522,716	44.48	9,134	33.85	5,147	39.06

(Continued)

Supplementary table 1: Continued

	Welsh validation study						English validation study					
	Overall cohort		COVID-19 death in first period		COVID-19 death in second period		Overall cohort		COVID-19 death in first period		COVID-19 death in second period	
	N	%	N	%	N	%	N	%	N	%	N	%
Chronic kidney disease												
No Chronic Kidney disease	1,874,451	95.79	869	72.90	412	67.54	34,392,544	9855	24,425	90.51	11,939	90.60
Stage 3	72,669	3.71	252	21.14	165	27.05	436,595	1.25	1,820	6.74	914	6.94
Stage 4	3,928	0.20	30	2.52	20	3.28	45,638	0.13	452	1.68	205	1.56
Stage 5	5,712	0.29	41	3.44	13	2.13	22,871	0.07	288	1.07	119	0.90
Learning disability												
No learning disability	1,928,040	98.53	1163	97.57	587	96.23	34,393,288	98.55	25,300	93.76	12,386	94.00
Learning disability	28,486	1.46	29	2.43	23	3.77	490,357	1.41	1,616	5.99	*	
Down Syndrome	234	0.01	0	0.00	0	0.00	14,003	0.04	69	0.26	*	
Chemotherapy												
No chemotherapy in past 12-months	1,949,761	99.64	1167	97.90	597	97.87	34,776,317	99.65	26,472	98.10	12,908	97.96
Chemotherapy in past 12-months	6,999	0.36	25	2.10	13	2.13	121,331	0.35	513	1.9	269	2.04
Cancer and immunosuppression												
Blood cancer	10,547	0.54	38	3.19	14	2.30	336,990	0.97	897	3.32	465	3.53
Respiratory cancer	5,691	0.29	20	1.68	10	1.64	9,720	0.03	142	0.53	66	0.50
Radiotherapy in past 6-months	1,827	0.09	*		*		56,252	0.16	174	0.64	100	0.76
Bone marrow transplant in past 6-months	56	0.00	0	0	0	0.00						
Solid organ transplant	806	0.04	*		*		3,488	0.01	26	0.10	*	
Prescribed immunosuppressant medication by GP	2,884	0.15	*		*		7,237	0.02	20	0.07	*	
Prescribed leukotriene or LABA	38,658	1.98	59	4.95	42	6.89	2,362,855	6.77	4,956	18.37	2,319	17.60
Prescribed regular prednisolone	15,819	0.81	61	5.12	28	4.59	404,467	1.16	2,124	7.87	1,028	7.80
Other comorbidities												
Diabetes	161,227	8.24	359	30.12	178	29.18	3,087,792	8.85	8,700	32.24	3,650	27.70
COPD	66,937	3.42	209	17.53	100	16.39	1,053,783	3.02	3,814	14.13	1,809	13.73
Asthma	290,490	14.85	186	15.60	109	17.87	4,382,954	12.56	3,344	12.39	1,504	11.41
Rare pulmonary diseases	9,471	0.48	26	2.18	12	1.97	373,807	1.07	1,707	6.33	734	5.57
Pulmonary hypertension or pulmonary fibrosis	3,741	0.19	17	1.43	14	2.30	127,760	0.37	1,158	4.29	502	3.81
Coronary heart disease	89,686	4.58	239	20.05	137	22.46	1,549,243	4.44	5,946	22.03	2,861	21.71

(Continued)

Supplementary table 1: Continued

	Welsh validation study						**English validation study**					
	Overall cohort		**COVID-19 death in first period**		**COVID-19 death in second period**		**Overall cohort**		**COVID-19 death in first period**		**COVID-19 death in second period**	
	N	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**
Stroke	55,336	2.83	233	19.55	121	19.84	902,277	2.59	5,086	18.85	2,685	20.38
Atrial fibrillation	62,712	3.20	253	21.22	140	22.95	1,096,209	3.14	5,237	19.41	2,894	21.96
Congestive cardiac failure	30,937	1.58	151	12.67	99	16.23	545,617	1.56	3,739	13.86	1,830	13.89
Venous thromboembolism	43,708	2.23	111	9.31	54	8.85	8,878	0.03	35	013	*	
Peripheral vascular disease	18,639	0.95	77	6.46	36	5.90	303,118	0.87	1,588	5.88	771	5.85
Congenital heart disease	17,071	0.87	30	2.52	12	1.97	359	<0.01	*		0	
Dementia	18,840	0.96	304	25.50	160	26.23	414,540	1.19	8,293	30.73	4,699	35.66
Parkinson's disease	5,717	0.29	40	3.36	32	5.25	113,647	0.33	1,021	3.78	573	4.35
Epilepsy	26,112	1.33	31	2.60	19	3.11	405,047	1.16	797	2.95	387	2.94
Rare neurological conditions	5,789	0.30	*		*		27,583	0.08	149	0.55	48	0.36
Cerebral palsy	1,318	0.07	0	0.00	0	0.00	4,350	0.01	31	0.11	*	
Severe mental illness	282,709	14.45	209	17.53	109	17.87	6,574,526	18.84	5,341	19.79	2,541	19.28
Osteoporotic fracture	73,679	3.77	154	12.92	96	15.74	29,153	0.08	194	0.72	92	0.70
Rheumatoid arthritis or SLE	22,485	1.15	35	2.94	16	2.62	315,431	0.90	696	2.58	369	2.80
Cirrhosis of the liver	7,210	0.37	17	1.43	*		8,753	0.23	241	0.89	114	0.87
Sickle cell disease	1,094	0.06	0	0.00	0	0.00						

*represents values which have been suppressed due to small numbers <10. ^ represents collapsing of categories to suppress small numbers.

Supplementary table 2: Performance of the risk models to predict risk of COVID-19 death in Wales by sex, age group, ethnicity, deprivation and time period using Harrell's C statistic

	First period (24th Jan 2020–30th Apr 2020)		**Second period (1st May 2020–28th Jul 2020)**	
	COVID-19 death in females	**COVID-19 deaths in males**	**COVID-19 death in females**	**COVID-19 deaths in males**
Age band				
19–39	0.838 (0.642 to 1.034)	0.688 (0.385 to 0.992)	*	*
40–44	0.764 (0.507 to 1.022)	0.964 (0.928 to 1.001)	*	0.941 (0.884 to 0.998)
45–49	0.796 (0.561 to 1.031)	0.924 (0.853 to 0.995)	0.905 (0.782 to 1.028)	0.879 (0.689 to 1.068)
50–54	0.765 (0.616 to 0.915)	0.853 (0.748 to 0.959)	0.964 (0.948 to 0.981)	0.674 (0.565 to 0.783)
55–59	0.797 (0.701 to 0.892)	0.704 (0.591 to 0.817)	0.860 (0.755 to 0.965)	0.654 (0.513 to 0.794)
60–64	0.849 (0.766 to 0.932)	0.832 (0.760 to 0.903)	0.716 (0.417 to 1.015)	0.820 (0.711 to 0.929)
65–69	0.848 (0.800 to 0.896)	0.795 (0.735 to 0.856)	0.903 (0.805 to 1.000)	0.802 (0.716 to 0.888)
70–74	0.777 (0.702 to 0.852)	0.755 (0.702 to 0.808)	0.781 (0.705 to 0.856)	0.837 (0.779 to 0.895)
75–79	0.855 (0.812 to 0.898)	0.791 (0.749 to 0.833)	0.845 (0.767 to 0.924)	0.805 (0.733 to 0.877)
80–84	0.746 (0.686 to 0.805)	0.715 (0.664 to 0.766)	0.777 (0.712 to 0.842)	0.793 (0.734 to 0.853)
85–89	0.773 (0.731 to 0.815)	0.746 (0.697 to 0.795)	0.749 (0.697 to 0.800)	0.716 (0.645 to 0.787)
90+	0.686 (0.641 to 0.732)	0.688 (0.633 to 0.744)	0.735 (0.684 to 0.786)	0.728 (0.655 to 0.800)
Ethnicity				
Bangladeshi	*	0.996 (0.994 to 0.998)	*	0.785 (0.772 to 0.798)
Black African	0.686 (0.671 to 0.701)	0.978 (0.950 to 1.007)	*	0.999 (0.997 to 1.001)
Indian	*	0.920 (0.809 to 1.031)	*	0.989 (0.986 to 0.992)
Mixed	0.977 (0.961 to 0.993)	0.974 (0.953 to 0.995)	0.989 (0.979 to 1.000)	0.992 (0.991 to 0.994)
Not recorded	0.951 (0.923 to 0.979)	0.984 (0.977 to 0.991)	0.953 (0.900 to 1.006)	0.974 (0.958 to 0.990)
Other	0.902 (0.749 to 1.055)	0.721 (0.469 to 0.973)	0.890 (0.884 to 0.895)	0.970 (0.932 to 1.009)
Pakistani	0.913 (0.842 to 0.983)	0.846 (0.754 to 0.939)	*	*
White	0.929 (0.919 to 0.940)	0.925 (0.916 to 0.933)	0.949 (0.941 to 0.958)	0.929 (0.916 to 0.942)
Townsend quintile				
1	0.933 (0.901 to 0.965)	0.930 (0.911 to 0.949)	0.967 (0.950 to 0.984)	0.940 (0.914 to 0.966)
2	0.951 (0.931 to 0.972)	0.931 (0.913 to 0.949)	0.957 (0.943 to 0.970)	0.947 (0.922 to 0.971)
3	0.928 (0.913 to 0.942)	0.926 (0.911 to 0.941)	0.940 (0.921 to 0.958)	0.942 (0.922 to 0.962)
4	0.921 (0.898 to 0.943)	0.928 (0.912 to 0.945)	0.948 (0.931 to 0.965)	0.910 (0.883 to 0.938)
5	0.922 (0.893 to 0.951)	0.930 (0.903 to 0.958)	0.951 (0.921 to 0.982)	0.937 (0.899 to 0.974)

*unable to calculate metrics.

Supplementary table 3: Performance of the risk models to predict risk of COVID-19 death in Wales by sex, age group, ethnicity, deprivation and time period using D statistic

	First period (24th Jan 2020–30th Apr 2020)		**Second period (1st May 2020–28th Jul 2020)**	
	COVID-19 death in females	**COVID-19 deaths in males**	**COVID-19 death in females**	**COVID-19 deaths in males**
Age band				
19–39	2.015 (0.208 to 3.823)	1.855 (0.290 to 3.421)	4.662 (1.318 to 8.006)	*
40–44	1.493 (−0.314 to 3.300)	3.193 (1.601 to 4.786)	2.701 (−0.453 to 5.856)	2.710 (0.480 to 4.940)
45–49	1.773 (−0.035 to 3.581)	2.668 (1.375 to 3.962)	2.705 (0.475 to 4.934)	3.418 (1.567 to 5.269)
50–54	1.716 (0.724 to 2.708)	2.281 (1.409 to 3.154)	3.067 (1.647 to 4.486)	0.761 (−0.638 to 2.160)
55–59	1.797 (1.013 to 2.582)	1.382 (0.681 to 2.083)	2.190 (0.786 to 3.593)	0.932 (0.064 to 1.799)
60–64	2.317 (1.562 to 3.071)	2.181 (1.619 to 2.743)	1.331 (−0.235 to 2.897)	2.094 (1.187 to 3.002)
65–69	2.156 (1.581 to 2.731)	1.977 (1.539 to 2.415)	2.959 (2.004 to 3.914)	1.779 (1.110 to 2.447)
70–74	1.821 (1.375 to 2.267)	1.548 (1.183 to 1.914)	1.603 (1.058 to 2.149)	2.104 (1.549 to 2.659)
75–79	2.318 (1.922 to 2.714)	1.796 (1.499 to 2.094)	2.269 (1.695 to 2.842)	2.098 (1.607 to 2.589)
80–84	1.510 (1.130 to 1.890)	1.255 (0.976 to 1.534)	1.718 (1.274 to 2.162)	1.840 (1.447 to 2.233)
85–89	1.524 (1.223 to 1.825)	1.566 (1.276 to 1.856)	1.470 (1.111 to 1.830)	1.555 (1.147 to 1.964)
90+	1.081 (0.816 to 1.347)	1.122 (0.805 to 1.438)	1.358 (1.032 to 1.684)	1.435 (0.963 to 1.907)
Ethnicity				
Bangladeshi	*	4.819 (0.734 to 8.904)	*	1.262 (−1.878 to 4.402)
Black African	0.779 (−2.369 to 3.927)	4.103 (1.515 to 6.692)	*	7.105 (2.571 to 11.639)
Indian	*	4.018 (1.460 to 6.577)	*	3.840 (0.307 to 7.373)
Mixed	3.414 (1.768 to 5.059)	4.042 (2.033 to 6.050)	4.704 (2.840 to 6.569)	4.025 (0.556 to 7.493)
Not recorded	3.326 (2.688 to 3.964)	4.104 (3.423 to 4.785)	3.400 (2.470 to 4.330)	3.449 (2.234 to 4.664)
Other	3.908 (1.938 to 5.877)	1.120 (−1.103 to 3.344)	1.961 (−1.187 to 5.108)	3.620 (1.255 to 5.985)
Pakistani	2.614 (0.993 to 4.235)	1.661 (0.125 to 3.197)	*	*
White	3.049 (2.903 to 3.194)	2.985 (2.858 to 3.112)	3.263 (3.078 to 3.447)	3.160 (2.973 to 3.347)
Townsend quintile				
1	3.116 (2.696 to 3.536)	2.929 (2.605 to 3.252)	3.563 (3.086 to 4.040)	3.221 (2.782 to 3.660)
2	3.485 (3.141 to 3.828)	3.222 (2.933 to 3.512)	3.446 (3.065 to 3.826)	3.370 (2.955 to 3.785)
3	2.936 (2.695 to 3.177)	2.914 (2.681 to 3.146)	3.062 (2.735 to 3.390)	3.342 (2.995 to 3.689)
4	3.083 (2.807 to 3.358)	3.064 (2.821 to 3.308)	3.122 (2.734 to 3.511)	2.978 (2.611 to 3.345)
5	2.942 (2.523 to 3.361)	3.173 (2.820 to 3.526)	3.694 (3.156 to 4.231)	3.333 (2.716 to 3.951)

*unable to calculate metrics.

Supplementary table 4: Performance of the risk models to predict risk of COVID-19 death in Wales by sex, age group, ethnicity, deprivation and time period using r^2 (explained variation)

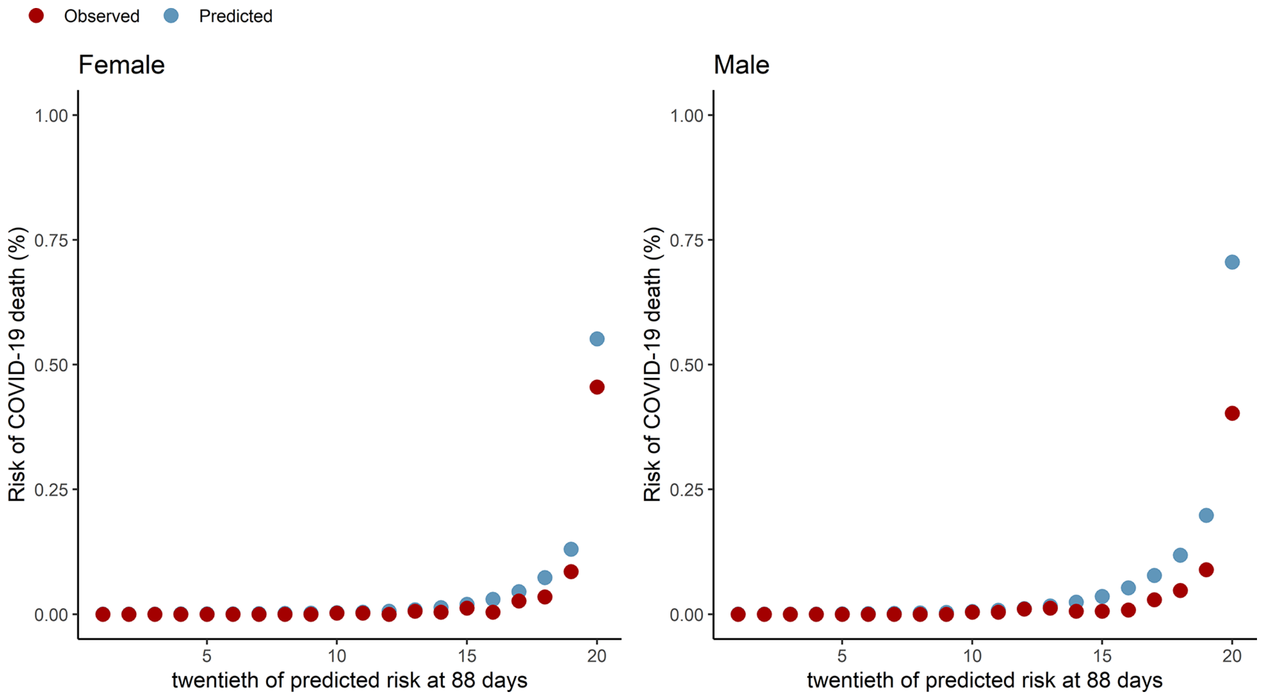
	First period		**Second period**	
	(24th Jan 2020–30th Apr 2020)		**(1st May 2020–28th Jul 2020)**	
	R2	**R2**	**R2**	**R2**
	COVID-19 death in females	**COVID-19 deaths in males**	**COVID-19 death in females**	**COVID-19 deaths in males**
Age band				
19–39	0.492 (0.010 to 0.777)	0.451 (0.020 to 0.736)	0.838 (0.293 to 0.939)	*
40–44	0.347 (0.023 to 0.722)	0.709 (0.380 to 0.845)	0.635 (0.047 to 0.891)	0.637 (0.052 to 0.854)
45–49	0.429 (0.000 to 0.754)	0.630 (0.311 to 0.789)	0.636 (0.051 to 0.853)	0.736 (0.370 to 0.869)
50–54	0.413 (0.111 to 0.636)	0.554 (0.321 to 0.704)	0.692 (0.393 to 0.828)	0.121 (0.089 to 0.527)
55–59	0.435 (0.197 to 0.614)	0.313 (0.100 to 0.509)	0.534 (0.129 to 0.755)	0.172 (0.001 to 0.436)
60–64	0.562 (0.368 to 0.692)	0.532 (0.385 to 0.642)	0.297 (0.013 to 0.667)	0.512 (0.252 to 0.683)
65–69	0.526 (0.374 to 0.640)	0.483 (0.361 to 0.582)	0.676 (0.489 to 0.785)	0.430 (0.227 to 0.588)
70–74	0.442 (0.311 to 0.551)	0.364 (0.250 to 0.466)	0.380 (0.211 to 0.524)	0.514 (0.364 to 0.628)
75–79	0.562 (0.469 to 0.637)	0.435 (0.349 to 0.511)	0.551 (0.407 to 0.659)	0.512 (0.381 to 0.615)
80–84	0.352 (0.234 to 0.460)	0.273 (0.185 to 0.360)	0.413 (0.279 to 0.527)	0.447 (0.333 to 0.543)
85–89	0.357 (0.263 to 0.443)	0.369 (0.280 to 0.451)	0.340 (0.228 to 0.444)	0.366 (0.239 to 0.479)
90+	0.218 (0.137 to 0.302)	0.231 (0.134 to 0.331)	0.306 (0.203 to 0.404)	0.330 (0.181 to 0.465)
Ethnicity				
Bangladeshi	*	0.847 (0.114 to 0.950)	*	0.276 (0.457 to 0.822)
Black African	0.127 (0.573 to 0.786)	0.801 (0.354 to 0.914)	*	0.923 (0.612 to 0.970)
Indian	*	0.794 (0.337 to 0.912)	*	0.779 (0.022 to 0.928)
Mixed	0.736 (0.427 to 0.859)	0.796 (0.497 to 0.897)	0.841 (0.658 to 0.912)	0.795 (0.069 to 0.931)
Not recorded	0.725 (0.633 to 0.790)	0.801 (0.737 to 0.845)	0.734 (0.593 to 0.817)	0.740 (0.544 to 0.839)
Other	0.785 (0.473 to 0.892)	0.231 (0.225 to 0.727)	0.479 (0.252 to 0.862)	0.758 (0.273 to 0.895)
Pakistani	0.620 (0.191 to 0.811)	0.397 (0.004 to 0.709)	*	*
White	0.689 (0.668 to 0.709)	0.680 (0.661 to 0.698)	0.718 (0.693 to 0.739)	0.705 (0.678 to 0.728)
Townsend quintile				
1	0.699 (0.634 to 0.749)	0.672 (0.618 to 0.716)	0.752 (0.694 to 0.796)	0.712 (0.649 to 0.762)
2	0.744 (0.702 to 0.778)	0.713 (0.672 to 0.746)	0.739 (0.692 to 0.778)	0.731 (0.676 to 0.774)
3	0.673 (0.634 to 0.707)	0.670 (0.632 to 0.703)	0.691 (0.641 to 0.733)	0.727 (0.682 to 0.765)
4	0.694 (0.653 to 0.729)	0.692 (0.655 to 0.723)	0.699 (0.641 to 0.746)	0.679 (0.619 to 0.728)
5	0.674 (0.603 to 0.729)	0.706 (0.655 to 0.748)	0.765 (0.704 to 0.810)	0.726 (0.638 to 0.788)

*unable to calculate metrics.

Supplementary table 5: Performance of the risk models to predict risk of COVID-19 death in Wales by sex, age group, ethnicity, deprivation and time period using Brier score

	First period (24th **Jan 2020–**30th **Apr 2020)**		**Second period** (1st **May 2020–**28th **Jul 2020)**	
	COVID-19 death in females	**COVID-19 deaths in males**	**COVID-19 death in females**	**COVID-19 deaths in males**
Age band				
19–39	0.00001	0.00001	0.00000	0.00000
40–44	0.00004	0.00005	0.00001	0.00003
45–49	0.00004	0.00007	0.00003	0.00004
50–54	0.00011	0.00016	0.00006	0.00006
55–59	0.00018	0.00024	0.00006	0.00015
60–64	0.00024	0.00043	0.00005	0.00016
65–69	0.00042	0.00076	0.00015	0.00032
70–74	0.00071	0.00111	0.00046	0.00048
75–79	0.00118	0.00241	0.00056	0.00088
80–84	0.00179	0.00414	0.00131	0.00210
85–89	0.00439	0.00710	0.00317	0.00373
90+	0.00912	0.01333	0.00647	0.00658
Ethnicity				
Bangladeshi	0.00000	0.00029	0.00000	0.00030
Black African	0.00027	0.00045	0.00001	0.00035
Indian	0.00000	0.00031	0.00000	0.00021
Mixed	0.00028	0.00023	0.00026	0.00009
Not recorded	0.00057	0.00031	0.00026	0.00009
Other	0.00021	0.00015	0.00007	0.00015
Pakistani	0.00109	0.00125	0.00000	0.00002
White	0.00053	0.00075	0.00033	0.00034
Townsend quintile				
1	0.00035	0.00059	0.00026	0.00032
2	0.00043	0.00063	0.00033	0.00029
3	0.00062	0.00070	0.00033	0.00031
4	0.00058	0.00075	0.00030	0.00033
5	0.00061	0.00083	0.00037	0.00027

Supplementary Figure 1: Predicted and observed risk of COVID-19-related death in the second time period (1st May–28th July 2020)



Supplementary table 6: Sensitivity for COVID-19-related death in Wales by sex at different absolute risk thresholds for the first time period

Top centile	Absolute risk centile cut-off (%)	Total deaths in each absolute risk centile	Total number of patients in each centile	Cumulative deaths	Cumulative % deaths based on absolute risk	sex
1	0.7952	212	9679	212	31.45	Males
2	0.5133	68	9679	280	41.54	Males
3	0.3952	60	9679	340	50.45	Males
4	0.3241	36	9679	376	55.79	Males
5	0.2759	30	9679	406	60.24	Males
6	0.2399	30	9679	436	64.69	Males
7	0.2115	22	9679	458	67.95	Males
8	0.1883	22	9679	480	71.22	Males
9	0.1687	19	9679	499	74.04	Males
10	0.1520	11	9679	510	75.67	Males
11	0.1381	20	9679	530	78.64	Males
12	0.1256	18	9679	548	81.31	Males
13	0.1146	12	9679	560	83.09	Males
14	0.1051	10	9679	570	84.57	Males
15	0.0967	*	9679	*	*	Males
16	0.0891	*	9679	*	*	Males
17	0.0821	*	9679	*	*	Males
18	0.0759	*	9679	*	*	Males
19	0.0703	*	9679	*	*	Males
20	0.0651	*	9679	*	*	Males
21	0.0603	*	9679	*	*	Males
22	0.0558	*	9679	*	*	Males
23	0.0517	*	9679	*	*	Males
24	0.0478	*	9679	*	*	Males
25	0.0443	*	9679	*	*	Males
26	0.0409	*	9680	633	93.92	Males
1	0.6338	160	9887	160	30.89	Females
2	0.3704	63	9887	223	43.05	Females
3	0.2757	57	9887	280	54.05	Females
4	0.2222	34	9887	314	60.62	Females
5	0.1866	25	9887	339	65.44	Females
6	0.1604	19	9887	358	69.11	Females
7	0.1396	19	9887	377	72.78	Females
8	0.1228	16	9887	393	75.87	Females
9	0.1089	15	9887	408	78.76	Females
10	0.0971	*	9887	*	*	Females
11	0.0870	*	9887	*	*	Females
12	0.0782	*	9887	*	*	Females
13	0.0706	*	9887	*	*	Females
14	0.0639	*	9887	*	*	Females
15	0.0580	*	9887	*	*	Females
16	0.0528	*	9888	*	*	Females
17	0.0482	*	9888	*	*	Females
18	0.0441	*	9888	*	*	Females
19	0.0404	*	9888	*	*	Females
20	0.0371	*	9888	*	*	Females
21	0.0342	*	9888	*	*	Females
22	0.0315	*	9888	*	*	Females
23	0.0290	*	9888	*	*	Females
24	0.0268	*	9888	*	*	Females
25	0.0248	*	9888	*	*	Females
26	0.0228	*	9888	485	93.63	Females

*represents values which have been suppressed to mask small numbers.