BMJ Open

Development and validation of an algorithm to estimate the risk of severe complications of COVID-19: a retrospective cohort study in primary care in the Netherlands

Ron M C Herings ,[1,2] Karin M A Swart,[3] Bernard A M van der Zeijst,[4] Amber A van der Heijden,[5] Koos van der Velden,[6] Eric G Hiddink,[7] Martijn W Heymans,[1] Reinier A R Herings,[8] Hein P J van Hout ,[5] Joline W J Beulens,[1] Giel Nijpels,[5] Petra J M Elders[5]

To cite: Herings RMC, Swart KMA, van der Zeijst BAM, *et al*. Development and validation of an algorithm to estimate the risk of severe complications of COVID-19: a retrospective cohort study in primary care in the Netherlands. *BMJ Open* 2021;**11**:e050059. doi:10.1136/bmjopen-2021-050059

► Prepublication history and additional supplemental material for this paper are available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2021-050059).

Received 04 March 2021
Accepted 06 December 2021

For numbered affiliations see end of article.

Correspondence to
Dr Ron M C Herings;
ron.herings@pd-groep.nl

ABSTRACT

Objective To develop an algorithm (sCOVID) to predict the risk of severe complications of COVID-19 in a community-dwelling population to optimise vaccination scenarios.

Design Population-based cohort study.

Setting 264 Dutch general practices contributing to the NL-COVID database.

Participants 6074 people aged 0–99 diagnosed with COVID-19.

Main outcomes Severe complications (hospitalisation, institutionalisation, death). The algorithm was developed from a training data set comprising 70% of the patients and validated in the remaining 30%. Potential predictor variables included age, sex, chronic comorbidity score (CCS) based on risk factors for COVID-19 complications, obesity, neighbourhood deprivation score (NDS), first or second COVID-19 wave and confirmation test. Six population vaccination scenarios were explored: (1) random (*naive*), (2) random for persons above 60 years (*60plus*), (3) oldest patients first in age band of 5 years (*oldest first*), (4) target population of the annual influenza vaccination programme (*influenza*), (5) those 25–65 years of age first (*worker*), and (6) risk based using the prediction algorithm (*sCOVID*).

Results Severe complications were reported in 243 (4.8%) people with 59 (20.3%) nursing home admissions, 181 (62.2%) hospitalisations and 51 (17.5%) deaths. The algorithm included age, sex, CCS, NDS, wave and confirmation test (c-statistic=0.91, 95% CI 0.88 to 0.94) in the validation set. Applied to different vaccination scenarios, the proportion of people needed to be vaccinated to reach a 50% reduction of severe complications was 67.5%, 50.0%, 26.1%, 16.0%, 10.0% and 8.4% for the *worker, naive, influenza, 60plus, oldest first* and *sCOVID* scenarios, respectively.

Conclusion The sCOVID algorithm performed well to predict the risk of severe complications of COVID-19 in the first and second waves of COVID-19 infections in this Dutch population. The regression estimates can and need to be adjusted for future predictions. The algorithm can be applied to identify persons with highest risks from data in the electronic health records of general practitioners (GPs).

Strengths and limitations of this study

- This large population-based cohort study used electronic health record data of n=6074 patients with COVID-19 in n=264 Dutch general practitioner (GP) practices.
- The routine electronic data of each patient with COVID-19 were enriched with a brief COVID-19 information and communication technology (ICT) system-linked questionnaire filled in by the GP.
- Least absolute shrinkage and selection operator (LASSO) regression was used for prediction modelling in a train data set and a validation data set, including data up to January 2021.
- Although the LASSO regression accounted for shrinkage of coefficients, the split sample method did not allow to further study model optimism.
- The data collection relies on a GP registration and may overpresent patients with manifest complaints.

INTRODUCTION

In the Netherlands, as in many other countries, the SARS-CoV-2 outbreak had severe consequences from March 2020 onwards. The fast spread of the infection and the unexpected severe complications required, in the absence of treatment, hospitalisation for many days in intensive care units (ICU), thereby occupying all available ICU beds in the Dutch hospitals. This urged the Dutch government to install social distancing measures including a lockdown. Although the number of hospitalisations dropped fast in the summer, a sudden increase started in August 2020 leading to a second lockdown

on 15 December 2020 and a curfew on 23 January 2021. The still limited capacity of available ICU beds, the unpredictable course of the COVID-19 infections, the limited knowledge on how these infections spread among the population, the absence of proper treatments, and the in-time and location unsuspected flare-ups of infections paralysed the Dutch care system and economy.

To prioritise high-risk individuals for vaccination or shielding from corona infections, or to start treatment in primary care as soon as possible, accurate identification of patients at risk for severe COVID-19 is of utmost importance. This requires living, accurate risk prediction algorithms that are easy to apply in general practice as suggested by Clift *et al*.[1] Initially, prediction algorithms for mortality or progression to severe disease were mainly developed for hospitalised patients.[2–4] In the mean time, several prediction algorithms for patients infected with COVID-19 in the general population have been developed.[1 5 6] Although the performance of these algorithms is fairly good, they have to deal with bias due to country-specific policy measures that change in time. This is in part because these studies were conducted based on the first wave of the infections, when testing was scarce and policy measures were still in its infancy.

By now, vaccines have become available and vaccination campaigns are ongoing but the shortage of vaccines limits the outroll of these campaigns.[7] Efforts to prioritise risk groups for vaccination are ongoing, focusing on populations with the highest risk of COVID-19 complications.[8] The development of our algorithm was aimed to provide predictions for subpopulations at risk for severe COVID-19 infections leading to hospitalisation, institutionalisation or death. The prediction algorithms are based on data of the Dutch NL-COVID database, containing nationwide geodemographical and medical data.[9] Building on this algorithm, we estimated the effectiveness of six different scenarios for vaccination of high-risk persons in order to prevent severe COVID-19 complications.

METHODS

Design

This cohort study was performed by using data from an extensive and representative general practice population database in the Netherlands.

Data sources

Data were obtained from general practitioner (GP) practices who reported information of the diagnoses and comorbidities of patients suffering from COVID-19 in the NL-COVID database. This database was set up in April 2020 as a collaborative initiative of general practitioners, public health specialists, virologists, epidemiologists, data scientists, data specialists, privacy specialists and information and communication technology (ICT) companies providing electronic health records (EHR). Together, the ICT companies cover about 95% of all GP practices in the Netherlands. GPs were asked to complete a brief questionnaire protocol for patients suffering from COVID-19 in their ICT systems. From a total of 264 practices (~5% of all Dutch GP practices), both questionnaire data and EHR with information regarding selected comorbidities were included in this study. Data until 21 January 2021 were used. Vaccines were not yet available during the study period.

The selected comorbidities (online supplemental appendix A) were those indicated by the National Institute for Public Health and the Environment to be relevant for the prognosis of severe outcomes of COVID-19 infections.[10] The following information was collected on a daily basis: a diagnosis of COVID-19 and whether the diagnosis was confirmed with a PCR test, the severity of the infection defined as treated at home, treated in a hospital or special care institution, or death from COVID-19. Updates of the patient's status were recorded using the same form. For this paper, we used the last status report. In addition, age, gender, body mass index (BMI), a chronic comorbidity score (CCS) and postal code were collected from the electronic registries of the GP. The neighbourhood deprivation score (NDS) was based on the quartile distribution of relative wealth of the neighbourhood as derived by Statistics Netherlands.[11] There were no missing data in the NL-COVID database: questionnaire data were complete and the registration of comorbidities in the EHR was considered to be complete as well.

Participants

A cohort study was performed among patients suffering from COVID-19 symptoms registered in the NL-COVID database certified by their GP.

Primary outcome

The primary outcome was the occurrence of severe complicated COVID-19 disease defined as hospitalisation, institutionalisation or death, as collected by the questionnaire from patients' GP.

Predictors

Predictors included age and sex, the NDS, BMI ≥30 kg/m^2, the period of registration (before or after August 2020) as first of the second wave, whether the diagnosis was confirmed with a PCR test or CT scan and a CCS. The CCS was based on the chronic diseases identified as predictors for complications of COVID-19 infection by the National Institute for Public Health and the Environment.[10] The comorbidities were mapped to the international classification of primary care (ICPC) coding system used in Dutch GP practices and subsequently grouped into nine disease clusters (online supplemental appendix A). A patient was scored in each of these respective disease clusters and assigned a point per cluster. For example, a patient suffering from epilepsy and diabetes scored a point in the category neurological diseases and a point for diabetes yielding a CCS of 2. The absence of registration was considered to be the absence of the disease/

Table 1 Characteristics of the cohort of patients with COVID-19

	Characteristic	Complicated (%) (n=291)	Home (%) (n=5783)
Outcome	Nursing home	59 (20.3)	–
	Hospitalisation	181 (62.2)	–
	Deceased	51 (17.5)	–
Age group (years)	<40	17 (5.8)	2186 (37.8)
	40–49	16 (5.5)	985 (17.0)
	50–59	34 (11.7)	1204 (20.8)
	60–69	43 (14.8)	758 (13.1)
	70–79	66 (22.7)	436 (7.5)
	80–89	79 (27.1)	178 (3.1)
	90+	36 (12.4)	36 (0.6)
Sex	Woman	144 (49.5)	3352 (58.0)
	Man	147 (50.5)	2431 (42.0)
Timing	Period before August	197 (67.7)	2123 (36.7)
	Period starting in August	94 (32.3)	3660 (63.3)
Province	South Holland	92 (31.6)	2102 (36.3)
	North Brabant	82 (28.2)	2495 (43.1)
	North Holland	67 (23)	821 (14.2)
	Limburg	33 (11.3)	227 (3.9)
	Other	17 (5.8)	138 (2.4)
Tested	Negative or not	49 (16.8)	1025 (17.7)
	Positive	242 (83.2)	4758 (82.3)
NDS	Middle	85 (29.2)	2241 (38.8)
	Low	150 (51.5)	1900 (32.9)
	High	56 (19.2)	1642 (28.4)
Obesity	BMI ≥30 kg/m^2	50 (17.2)	611 (10.6)
Comorbidity	Cancer	84 (28.9)	507 (8.8)
	Cardiovascular disease	188 (64.6)	1405 (24.3)
	Diabetes	64 (22.0)	440 (7.6)
	Heart valve disease	21 (7.2)	87 (1.5)
	Immunosuppressants	41 (14.1)	421 (7.3)
	Kidney disease	67 (23.0)	280 (4.8)
	Liver disease	7 (2.4)	89 (1.5)
	Lung disease	64 (22.0)	864 (14.9)
	Neurological disease	42 (14.8)	187 (3.2)
Indication	Influenza vaccination	225 (77.3)	2392 (42.3)
CCS	0	60 (20.6)	3250 (56.2)
	1	49 (16.8)	1430 (24.7)
	2	74 (25.4)	655 (11.3)
	3	65 (22.3)	297 (5.1)
	4	31 (10.7)	113 (2.0)
	5+	12 (4.1)	38 (0.7)

BMI, body mass index; CCS, chronic comorbidity score; NDS, neighbourhood deprivation score.

Table 2 Characteristics of training and test cohort of patients with COVID-19

Predictors	Sampling	Training (70% random)		Test (30% random)	
		Complicated (%) (n=197)	Home (%) (n=4054)	Complicated (%) (n=94)	Home (%) (n=1729)
Age group (years)	<40	15 (7.6)	1514 (37.3)	2 (2.1)	672 (38.9)
	40–49	14 (7.1)	682 (16.8)	2 (2.1)	303 (17.5)
	50–59	21 (10.7)	850 (21.0)	13 (13.8)	354 (20.5)
	60–69	33 (16.8)	539 (13.3)	10 (10.6)	219 (12.7)
	70–79	42 (21.3)	321 (7.9)	24 (25.5)	115 (6.7)
	80–89	49 (24.9)	122 (3.0)	30 (31.9)	56 (3.2)
	90+	23 (11.7)	26 (0.6)	13 (13.8)	10 (0.6)
Sex	Man	87 (44.2)	1711 (42.2)	57 (60.6)	720 (41.6)
	Woman	110 (55.8)	2343 (57.8)	37 (39.4)	1009 (58.4)
Timing	Before August 2020	59 (29.9)	2587 (63.8)	35 (37.2)	1073 (62.1)
	After August 2020	138 (70.1)	1467 (36.2)	59 (62.8)	656 (37.9)
Tested	Negative or not	32 (16.2)	698 (17.2)	17 (18.1)	327 (18.9)
	Positive	165 (83.8)	3356 (82.8)	77 (81.9)	1402 (81.1)
NDS	Middle	58 (29.4)	1568 (38.7)	27 (28.7)	673 (38.9)
	Low	102 (51.8)	1346 (33.2)	48 (51.1)	554 (32.0)
	High	37 (18.8)	1140 (28.1)	19 (20.2)	502 (29.0)
Overweight	BMI <30 kg/m^2	164 (83.2)	3621 (89.3)	77 (81.9)	1551 (89.7)
	BMI ≥30 kg/m^2	33 (16.8)	433 (10.7)	17 (18.1)	178 (10.3)
CCS	0	42 (21.3)	2273 (56.1)	18 (19.1)	977 (56.5)
	1	40 (20.3)	1010 (24.9)	9 (9.6)	420 (24.3)
	2	52 (26.4)	471 (11.6)	22 (23.4)	184 (10.6)
	3	39 (19.8)	198 (4.9)	26 (27.7)	99 (5.7)
	4	19 (9.6)	73 (1.8)	12 (12.8)	40 (2.3)
	5+	5 (2.5)	29 (0.7)	7 (7.4)	9 (0.5)

BMI, body mass index; CCS, chronic comorbidity score; NDS, neighbourhood deprivation score.

condition. This also applied to BMI, that is, if no record of BMI or obesity was observed then it was assumed that the patient had a normal weight.

Risk mitigation scenarios

The prediction models yield a probability that a patient with COVID-19 develops a severe complication. In a single normalised Dutch GP practice (n=2090 patients), the summarised sCOVID-predicted probability is 85. It is assumed that if all patients would be infected with COVID-19, an expected 85 patients would develop severe COVID-19 complications. We further assumed that this probability can be reset to (almost) zero by vaccination, or by shielding patients from contact with others. By vaccination or shielding of the 10 highest ranked patients, ranging from a probability of 0.64 to 0.45, 85 minus 5=80 patients were expected to develop severe COVID-19 complications, a decrease of 100*(1-80/85) of 5.9%.

For 300 randomly selected, fully anonymised GP practices from the STIZON Database Network, including data from 1.2 million inhabitants, the predicted number of patients developing severe COVID-19 complications was estimated as the summarised sCOVID probabilities as Base (B). Depending on the vaccination coverage and the policy who to vaccinate (scenario), the number of patients developing severe complications can be estimated for different vaccination scenarios.

The impact of the vaccination strategy can be followed in time by division of the summarised probabilities Pt divided by B as 100% times Pt/B yielding the percentage expected decrease in severe complications at a given percentage of the population vaccinated. The vaccination coverage needed for a 50% decrease of hospitalisation was defined as VC50 as a measure of the efficiency of a particular hypothetical vaccination or shielding scenario. We explored and compared six different hypothetical vaccination scenarios. A first scenario was defined as a *naive* scenario, a scenario in the absence of any policy, that is,

inhabitants are randomly vaccinated. A second scenario was defined as a *plus60* scenario where all inhabitants, 60 years of age or older, are randomly vaccinated, followed by random vaccination of those under 60 years of age (also random). A third scenario (*oldest first*) prioritised vaccination from the oldest down from 100 to 60 years of age in age band of 5 years. Within the respective age bands, allocation is random. A fourth scenario was defined as the *influenza* scenario. Here, patients with an indication for influenza vaccination are prioritised for vaccination. A fifth scenario (*worker*) prioritised random vaccination of inhabitants 25–65 years of age. The sixth and last scenario was based on the sCOVID risk-ranking algorithm, the *sCOVID* scenario. Here, we start vaccination based on the absolute risk ranking, the patient with the highest risk first, followed by the second patient in line, etc.

Statistical analyses

Least absolute shrinkage and selection operator (LASSO) regression analysis was used to select predictors in the model and to estimate and shrink regression coefficients. Tenfold cross-validation was used to estimate the optimal shrinkage factor (λ) used in the LASSO regression, such that the sum of the squared residuals was minimised. Age was included as quadratic function. The final regression formula allowed calculation of predicted probabilities for each registered patient at their GP. We randomly allocated 70% of the patients in a training data set to develop the model. The other 30% of the patients were allocated into a validation data set. We assessed the model performance in terms of discrimination and calibration in the validation set. Discrimination was assessed using the c-statistic. The c-statistic indicates the extent to which the model can distinguish between a patient with and without the outcome and varies between 0.5 and 1. Calibration was assessed using calibration plots showing the predicted risk against the observed frequency of the study population's outcome using 10 risk groups. Goodness of fit was assessed with the Brier Score to quantify the difference between the observed and fitted probabilities ranging from 0 to 1, with a score of 0 representing the best model.[12] With an outcome proportion of 0.05 and eight candidate predictors at least 891 patients would be needed.[13] R V.4.0.2, GLMNET package (4.0-2), was used for statistical analyses and constructing figures. We adhered to the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis statement.[14]

Patient and public involvement

Patients were not involved in the design and conduct of the study. General practitioners were consulted to reflect on their ideas about different vaccination scenarios and their practicality.

RESULTS

Overall study population

A total of 264 GP practices (~5% of all Dutch GP practices) reported 6074 patients with a diagnosis of COVID-19 in the period 10 April 2020 until 21 January 2021. Severe complications were reported for 291 (4.7%) patients of whom 59 (20.3%) were treated in a nursing home, 181 (62.2%) were hospitalised and 51 (17.5%) died. Training and test model included 4251 and 1823 persons, respectively.

Baseline characteristics

The characteristics of patients with COVID-19 recorded in the first and second-time periods differed in age, baseline risk, frequency of testing and region. The percentage of people developing severe complications dropped from 8.5% in the first period in the Spring 2020 to 2.5% in the second-time period in the Autumn 2020 which was reflected in institutionalisation, hospitalisation and death. In the first wave, infected patients were from older age groups, whereas relatively more adolescents, 12–19 years of age, were reported in the second wave. The proportion of patients recorded with a positive COVID-19 test increased from 63% in the first wave to 95% in the second wave. The general characteristics are presented in table 1. Most of the patients with severe complications suffered from cardiovascular conditions (64.6%) and other chronic conditions such as diabetes, neurological diseases (ie, dementia, Parkinson's disease) and lung disease. Almost 80% of patients with severe complications suffered from at least one chronic disease. More than 62% had multiple chronic conditions. The characteristics of the training and validation set are shown in table 2.

Predictor variables

The predictor variables in the final COVID-19 models included age, sex, positive test result, period (first or second wave), NDS, obesity and the CCS (table 3). The strongest predictors included age, NDS, the time period, a positive PCR test and male sex. Obesity was eliminated by the LASSO predictor selection. The final model showed a very good calibration and fit. Figure 1 illustrates the receiver operating characteristic curve from the validation set with a c-index of 0.91 (95% CI 0.88 to 0.94). The

Table 3 Selected predictors of least absolute shinkage and selection operation (LASSO) regression coefficients

Predictors	LASSO coefficients
Intercept	−5.1765361941
Age (squared)	0.0005142466
Male	0.1133382856
After July 2020	−1.3845037172
Positive COVID-19 test	0.6423757757
Obese (BMI >30 kg/m^2)	–
Low NDS	0.4400574635
High NDS	–
CCS	0.1186100356

BMI, Body Mass Index; CCS, chronic comorbidity score; NDS, neighbourhood deprivation score.

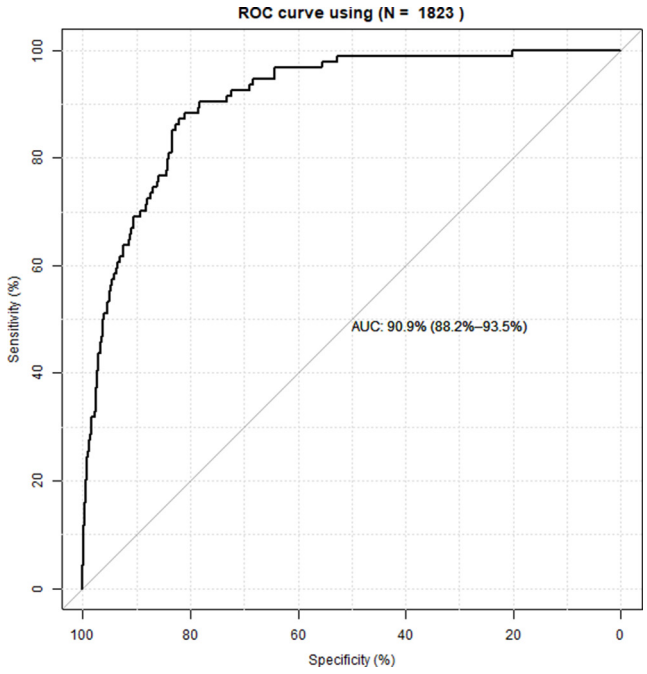

Figure 1 Receiver operating characteristic (ROC) curve of sCOVID based on the validation samples. The c-index was 0.91. AUC, area under the curve.

model yielded a good calibration (Brier Score=0.034) (figure 2).

Risk prediction in practice

Examples of individual risk ranking

The risk of developing severe complications for a 60-year-old man, with a positive PCR test, living in a neighbourhood with a low NDS and who suffers from diabetes,

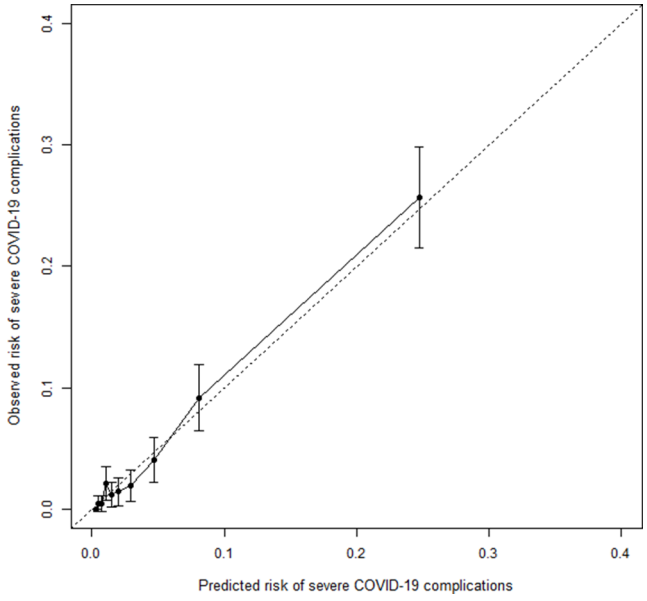

Figure 2 Calibration plot of least absolute shrinkage and selection operation (LASSO) regression of severe COVID-19 complications predicted by the sCOVID algorithm versus the observed complications. The Brier Score was 0.034. The calibration intercept was 0.09. The calibration slope was 1.04.

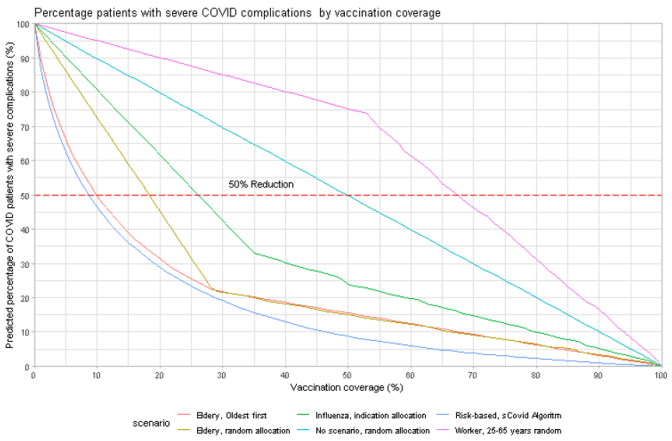

Figure 3 Effect of vaccination on the reduction of severe COVID-19 complications in different scenarios. Vaccination coverages needed for a 50% reduction of the burden of disease can be estimated at the intersection of the graphs with the 50% reduction line.

hypertension and kidney failure can be estimated. His comorbidities comprise three different classes (online supplemental appendix A). Summarising the coefficients (Cf) from the column LASSO regression in table 3, the equation yields as total score of Cf(intercept)+60*Cf (age times age)+Cf(man)+Cf (after July 2020)+Cf (positive COVID-19 test)+Cf(low NDS)+3*Cf(CCS)=−3.456. His risk to develop severe complications is subsequently calculated as $100 * (\exp(-3.456)) / (1 + \exp(-3.456)) = 4.1\%$. The risk equals that of a 73-year-old woman without any chronic condition living in a neighbourhood with a high socioeconomic status.

Practice risk ranking

The results of six prioritising scenario analyses were obtained by deploying the different algorithms to 300 randomly selected, fully anonymised GP practices, including data from 1.2 million inhabitants. The results for the six scenarios are plotted in figure 3 and summarised in table 4. A reduction of 50% of the patients with severe complications was observed already with a vaccination coverage of 8% if all high-risk persons according to the sCOVID algorithm are vaccinated first. This scenario was superior to all other scenarios with vaccination scheme in which the oldest are consecutively vaccinated in age band of 5 years, being second best. The worst scenario was the *worker* scenario prioritising patients 25–65 years of age, followed by the naive scenario where patients are randomly vaccinated.

DISCUSSION

Using data from the NL-COVID database, an algorithm was developed to predict the probability of patients developing severe complications once infected with COVID-19 using EHR from general practices. This sCOVID algorithm, which can be deployed in all Dutch GP practices, showed a very good performance in terms of discrimination (c-index: 0.91) and calibration and can be used

Table 4 Vaccination coverage needed for 50% risk reduction by selected scenarios

Scenario	Strategy	Vaccination coverage for 50% risk reduction (%)	Needed vaccinations in a normalised size general practitioner (GP) practice* (Netherlands)
sCOVID based	Highest risk first, based on ranking from highest to lowest risk	8.4	175 (1.2 million)
Oldest first	Stepwise in 5 years group from 100 to 65, random in age band	10.0	210 (1.7 million)
60plus	60 plus first, random	18.1	380 (3.1 million)
Influenza	Those with an indication for influenza vaccination first, random	26.1	545 (4.4 million)
Naive	None, random, everybody	50.0	1050 (8.5 million)
Worker	Prioritise inhabitants 25–65 years of age	67.5	1414 (11.5 million)

*Normalised GP practice=2095 patients (www.lhv.nl).

to rank the most susceptible patients for prioritisation of vaccinations. Our vaccination scenarios showed that ranking and vaccinating patients based on their complication risk (sCOVID scenario) would be the most efficient vaccination scenario to reduce hospitalisation and deaths. The second-best scenario was to vaccinate the oldest people first in consecutive order. With shortage of vaccines, the most vulnerable patients and not the oldest patients are prioritised.

Comparison with other studies

The sCOVID risk models yield a high discrimination rate (c-statistic=0.91). Calibration plots show a good fit in all risk categories, although the lowest risks were most challenging to estimate due to the limited numbers of patients developing severe complications. These results are similar to those of other prediction algorithms. Two earlier studies developed prediction algorithms for hospitalisation and/or death due to COVID-19 infection and showed similar prognostic performance.[1 5]

The major predictors, selected by the LASSO procedures, were higher age, male gender, the number of chronic comorbidities but also a positive test result and neighbourhood deprivation status. These selected predictors resemble the predictors reported in earlier studies by Clift *et al*, Jehi *et al* and Williamson *et al*.[1 5 6] The most obvious differences were the summary score of comorbidities (CCS) compared with separate conditions and inclusion of symptoms and laboratory measures for the study by Jehi *et al*. Most predictors found in this study relate to poor health and a complex of comorbidities. More than 60% of the patients with severe complications suffered from more than one chronic condition against less than 20% of those without comorbidities. Therefore, we preferred to include a chronic disease summary score to come to a more comprehensive and practical algorithm. Moreover, from a clinical perspective, our sample size was relatively small and would exclude rare but clinically relevant outcomes.

Complexity of modelling

Estimating the risk of severe COVID-19 complications is permanently subject to changing policy measures and interventions to shield high-risk people by vaccinations.[7 8] The time biases caused by these measures and interventions are complex and difficult to unravel. First analyses confirmed suggestions from Clift *et al* that these time biases are indeed present,[1] showing an age and sex-adjusted three to five times lower complication rate compared with the patients in the first wave. Estimates, needed to predict hospitalisation and/or death, therefore need permanent recalibration of the prediction algorithms. Such recalibration is necessary to monitor the effect of policy intervention on managing care capacity. The infrastructure of the NL-COVID database permits the recalibration on a regional and daily basis.

Strengths and limitations

The NL-COVID database also has limitations and strengths. First, we have substantial under-reporting of positive cases since our 264 registration practices consisting of about 5% of all GP practices only reported 0.7% of the registered cases. This is explained by several factors: first, practices enrolled into the programme over time and some practices only joined the programme and the end of 2021. Second, COVID-19 testing was done by the regional health authorities whereas the administrations of the regional health authorities were not linked with the GP administration. Therefore, our registration relies on whether the patient contacted the GP and whether the GP registered the patient. This makes it likely that we have a selection bias towards the more severe disease manifestations of the COVID-19 infection. Also, our prediction partly relied on the judgement of the GP whether a patient was COVID-19 positive (in case of lacking test results). It should therefore be stressed that absolute risk estimates of severe complications should be interpreted with care only by healthcare professionals for prioritising strategies. A weakness of the sCOVID scenario is that we did not perform an external validation and that the model was not retrained in a random sample of the general population. The large number of GP practices that came from all over the country and the good testing

characteristics of the validation set makes it likely that the accuracy of the scenarios is adequate. For the comparison of the different scenarios this has no importance since they were compared in the same sample.

A first strength was the coverage and representativity of the practices most strongly confronted with the pandemic. The first wave of COVID-19 hit hard in the southern part of the country and most participating practices were situated here. Second, we used training and validation samples to estimate the accuracy of the algorithms. Third, by law and regulation, almost every Dutch citizen has a designated GP and therefore we were able to study the general population. Fourth, this study is the first to demonstrate the potential impact and efficiency of more and less targeted vaccination scenarios. Fifth, the prediction algorithm can be adapted, updated and validated on a daily basis and learn from new insights and policy measures.

Practical implications

Our study showed that within the framework of privacy regulations, COVID-19 infections and consequences can be monitored fast, efficiently and safely on a very detailed local level and on a day-to-day basis using country-wide data from currently available ICT systems in GP practice. The costs of such a database are relatively low. Insights can be generated that help GPs and involved regional and local health authorities to shield the patients from infection and to reduce hospitalisation and death very efficiently in a selected group of persons with the highest risks. Moreover, such database may demonstrate and underpin the effectiveness and efficiency of policy measures to plan and manage care facilities. Second, the prediction accuracy could be improved with flexible access to the complete GP patient dossier and linkage to hospital admission under strict compliance with the general data protection regulation to adapt and improve the algorithms if new insights become available. The vaccination scenarios did not fully address the complexity of the real world. For instance, the scenarios assumed that vaccination is always effective, and did not consider the time needed to vaccinate the population. Furthermore, implementation requires embedding in guidelines and acceptance by general practitioners to be used in current practice and need to be weighed against social and political measures. Therefore, the vaccination coverages needed for a 50% reduction may be underestimated. However, the scenarios showed that in case of remaining shortage of vaccines, vaccination based on the sCOVID scenario performs best with a consecutive age-based scenario as second best. Hybrid scenarios that do not follow the risk of COVID-19 complications have worse performances, for example, the influenza scenario in which age and influenza risk are combined. Currently, vaccination in the Netherlands is performed from a practical perspective based on factors not only related to COVID-19 risk complications. This makes it likely that more efficient scenarios are thinkable.

In conclusion, the sCOVID algorithm has been developed to predict which patients are at high risk to develop severe complications due to COVID-19 and showed a good model performance. In remaining shortage of vaccines, prioritising vaccination of patients based on sCOVID risk complications is the most efficient way to reduce hospitalisations, institutionalisations and death.

Author affiliations
[1]Department of Epidemiology and Data Science, Amsterdam UMC - Location VUmc, Amsterdam Public Health, Amsterdam Cardiovascular Science, Amsterdam, The Netherlands
[2]Stichting Informatievoorziening voor Zorg en Onderzoek (STIZON), Utrecht, The Netherlands
[3]PHARMO Institute for Drug Outcomes Research, Utrecht, The Netherlands
[4]Department of Medical Microbiology, Leiden Universitair Medisch Centrum, Leiden, The Netherlands
[5]Department of General Practice, Amsterdam UMC - Locatie VUmc, Amsterdam Public Health, Amsterdam, The Netherlands
[6]Department of Primary and Community Care, Academic Collaborative Center AMPHI, Integrated Health Policy, Radboud University Medical Center, Nijmegen, The Netherlands
[7]Stichting Health Base, Houten, The Netherlands
[8]Julius Center for Health Science and Primary Care, UMC Utrecht, Utrecht, The Netherlands

Acknowledgements The authors like to thank the unconditional support of Guus Vaassen and Johan Ruiter (Medworq), Marjoleine van der Zwan, Piet-Hein Knoop, Arjan den Ouden (PharmaPartners), Mark van Vliet, Chris Tromp (Health Base), Eric Grosveld, Meefa Hogenes (ExpertDoc), and Frank Carlebur (ZONH), Ernst de Graag, Michiel Meulendijk (STIZON), Theo Peters (CGM) and more than 450 Dutch general practitioners who contribute to the COVID database in time, cash or in kind to fight COVID-19.

Contributors RMCH and EGH were involved in the development of the database. RMCH, KMAS, BAMvdZ, AAvdH, KvdV, HvH, JWJB, GN and PJME contributed to the development of the research question and study design. RMCH and RARH conducted the statistical analyses. MWH was involved in advanced statistical aspects. RMCH, KMAS, BAMvdZ, AAvdH, KvdV, EGH, MWH, RARH, HPJvH, JWJB, GN and PJME contributed to the interpretation of the results. RMCH wrote the first draft of the manuscript. RMCH, KMAS, BAMvdZ, AAvdH, KvdV, EGH, MWH, RARH, HPJvH, JWJB, GN and PJME contributed to the critical revision of the manuscript for important intellectual content and approved the final version of the manuscript. RMCH is the guarantor of the study.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests RMCH is the director of STIZON, an organisation that is the processor of the NL-COVID data on behalf of the participating general practitioners. EGH is an employee of Health Base, an independent multidisciplinary foundation active in developing content for medical and pharmaceutical decision support systems, which was also applied for the NL-COVID database. KMAS is an employee of the PHARMO Institute for Drug Outcomes Research. This independent research institute performs financially supported pharmacoepidemiological studies for the government, healthcare authorities and pharmaceutical companies.

Patient consent for publication Not required.

Ethics approval Patients and GPs were asked to consent with data sharing regarding sending data to the NL-COVID database and extract information on the four-digit postal code level in an anonymised format for public decision-making. The procedure was approved and tested for compliance with the General Data Protection Regulation by the Institutional Review Board of 'Stichting Informatievoorziening voor Zorg en Onderzoek' (STIZON, ID 10042020).

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request. The data will be available upon reasonable request (ron.herings@pd-groep.nl).

ORCID iDs
Ron M C Herings http://orcid.org/0000-0001-9316-2161
Hein P J van Hout http://orcid.org/0000-0002-2495-4808

REFERENCES

1 Clift AK, Coupland CAC, Keogh RH, *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.
2 Hu C, Liu Z, Jiang Y. Early prediction of mortality risk among severe COVID-19 patients using machine learning. *medRxiv* 2020.
3 Wynants L, Van Calster B, Collins GS, *et al*. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *BMJ* 2020;369:m1328.
4 Guo Y, Liu Y, Lu J. Development and validation of an early warning score (EWAS) for predicting clinical deterioration in patients with coronavirus disease 2019. *medRxiv* 2020.
5 Jehi L, Ji X, Milinovich A, *et al*. Development and validation of a model for individualized prediction of hospitalization risk in 4,536 patients with COVID-19. *PLoS One* 2020;15:e0237419.
6 Williamson EJ, Walker AJ, Bhaskaran K, *et al*. Factors associated with COVID-19-related death using OpenSAFELY. *Nature* 2020;584:430–6.
7 WHO. *22Nd who regulatory update on COVID-19*. World Health Organization, 2020.
8 Gezondheidsraad. *Strategieën voor COVID-19-vaccinatie*. Den Haag: Gezondheidsraad, 2020.
9 COVID-19 Datacoalitie, 2021. Available: www.covid-data.nl
10 RIVM, 2020. Available: https://www.rivm.nl/coronavirus-covid-19/risicogroepen [Accessed 20 April 2020].
11 CBS. *Sociaaleconomische status van huishoudens in Nederland den Haag*. CBS, 2020. www.cbs.nl
12 Steyerberg EW, Vickers AJ, Cook NR, *et al*. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* 2010;21:128–38.
13 Riley RD, Snell KI, Ensor J, *et al*. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. *Stat Med* 2019;38:1276–96.
14 Collins GS, Reitsma JB, Altman DG, *et al*. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ* 2015;350:g7594.

Appendix A: Characteristics and risk factors selected for the development of the sCOVID prediction model

Type of data	Characteristic	ICPC/ Diagnostic code/ ATC
Demographic	Age in years (continuous)	
	Sex	
	NDS (Area postal level)	
	Postal code (4 – digits)	
Lifestyle		
Obesity	Obesity (BMI ≥ 30 kg/m2)	T82
	BMI ≥ 30 kg/m2 in the last 2 years	1272
COVID-19 (Information extracted from questionnaire)		
Current Status	Clinical suspicion (not tested)	
	COVID-19 test positive (any)	
Severity	Hospitalisation	
	Institutionalisation	
	Deceased	
Time period	First wave (March - August 2020)	
	Second wave (September-December 2020)	
Chronic Diseases (CMS: score 0- 9)*		
Cardiovascular	Ischaemic heart disease with angina	K74
	Unstable angina	K74.01
	Stable angina	K74.02
	Acute myocardial infarction	K75
	Ischaemic heart disease without angina	K76
	Coronary sclerosis	K76.01
	Old heart attack	K76.02
	Heart failure	K77
	Acute heart failure	K77.01
	Chronic heaft failure	K77.02
	Atrial fibrillation/flutter	K78
	Transient cerebral ischaemia	K89
	Stroke/cerebrovascular accident	K90
	Subarachnoid hemorrhage	K90.01
	Intracerebral haemorrhage	K90.02
	Cerebral Infarction	K90.03
	Cerebrovascular disease	K91
	Hypertension uncomplicated	K86
	Hypertension complicated	K87
	Paroxysmal tachycardia	K79
	Supraventricular tachycardia	K79.01

	Ventricular tachycardia	K79.02
	Congenital anomaly cardiovascular	K73
	Atrial septal defect	K73.01
	Ventricular septal defect	K73.02
	Heart disease other	K84
	Wolff-Parkinson-White (WPW) syndrome	K84.01
	Atrioventricular block	K84.02
	Cardiomyopathy	K84.03
	Long QT syndrome	K84.07
	Atherosclerosis/peripheral vascular disease	K92
	Intermittent claudication	K92.01
	Raynaud syndrome	K92.02
	Buerger's disease	K92.03
	Cardiac arrhythmia NOS	K80
	Supraventricular extrasystoles	K80.01
	Ventricular extrasystoles	K80.02
	Sick sinus syndrome	K80.03
	Pulmonary heart disease	K82
Diabetes	Diabetes non-insulin dependent	T90
	Diabetes mellitus type 1	T90.01
	Diabetes mellitus type 2	T90.02
Neurological	Dementia	P70
	Parkinsonism	N87
	Parkinson's disease	N87.01
	Epilepsy	N88
	Multiple sclerosis	N86
	Neurological disease other	N99
	ALS	N99.01
	Myasthenia	N99.02
	Other neuron muscle diseases	N99.03
Heart valve	Rheumatic fever/heart disease	K71
	Rheumatic and heart disease	K71.02
	Heart valve disease NOS	K83
	Stenosis of aorta	K83.01
	Mitral insufficiency	K83.02
Liver	Liver disease NOS	D97
	Cirrhosis	D97.04
	Liver steatosis (< 5 years)	D97.05
	Viral hepatitis	D72
	Viral hepatitis A (< 6 month)	D72.01
	Viral hepatitis B (< 6 month)	D72.02
	Viral hepatitis C (< 6 month)	D72.03
	Carrier Hepatitis B	D72.04
	Carrier Hepatitis C	D72.05

Lung	Chronic obstructive pulmonary disease	R95
	Astma	R96
	Allergic astma	R96.02
	Chronic bronchitis	R91
	Chronic bronchitis	R91.01
	Bronchiectasis	R91.02
	Pulmonary embolism	K93
	Tuberculosis	R70
	Pleurisy/pleural effusion	R82
	Congenital anomaly respiratory	R89
	Respiratory disease other	R99
	Pneumoconiosis	R99.06
	Cystic fibrosis	T99.10
	GOLD classification (<24 month)	2209
	Nr of exacerbations (<12 month)	3549
	Prednisolone (<24 month)	H02AB06
	Prednisone (<24 month)	H02AB07
Cancer	Malignant neoplasm breast female	X76
	Malignant Adenocarcinoma	X76.01
	Malignant neoplasm male genital other	Y78
	Malignant carcinoma penis	Y78.01
	Malignant carcinoma testis	Y78.02
	Malignant carcinoma breast	Y78.03
	Malignant neoplasm nervous system	N74
	Malignant neoplasm of bladder	U76
	Malignant neoplasm thyroid	T71
	Malignant neoplasm pancreas	D76
	Malignant neoplasm stomach	D74
	Malignant neoplasm related to pregnancy	W72
	Malignant neoplasm urinary tract other	U77
	Malignant neoplasm of kidney	U75
	Malignant neoplasm colon/rectum	D75
	Malignancy NOS	A79
	Malignant neoplasm cervix	X75
	Malignant neoplasm bronchus/lung	R84
	Malignant neoplasm blood other	B74
	Multiple myeloma	B74.01
	Malignant neoplasm genital female other	X77
	Endometrial carcinoma	X77.01
	Malignant ovary carcinoma	X77.02
	Malignant neoplasm of skin	S77
	Basal cell carcinoma	S77.01
	Spinocellular carcinoma	S77.02
	Malignant melanoma	S77.03

	Kaposi's sarcoma	S77.04
	Malignant neoplasm respiratory other	R85
	Benign neoplasm respiratory	R86
	Hodgkin Lymphoma	B72
	Hodgkin Lymphoma	B72.01
	Non-hodgkin Lymphoma	B72.02
	Leukaemia	B73
Kidney	Congenital anomaly urinary tract	U85
	Polycystic kidney disease	U85.01
	Glomerulonephritis/nephrosis	U88
	Urinary disease other	U99
	Chronic kidney disease	U99.01
	Renal hypoplasia	U99.02
	Hydronephrosis	U99.03
	Creatin clearance (CKD-EPI) <30 (< 12 months)	3583
	Creatin clearance (MDRD) <30 (< 12 months) or	1919
	Creatin clearance (COCKCROFT) <30 (< 12 months) or	1918
	Creatin clearance (< 12 months)	524
Immune system	HIV-infection/AIDS	B90
	Seropositive without symptoms	B90.01
	HIV	B90.02
	Immunodeficiencies	T99.01
	Prednison/prednisolon (< 6 months)	H02A*
	Oncolytics (< 6 months)	L01*
	Immunosuppressants (< 6 months)	L04A*
	TNF-Alpha (< 6 months)	L04AB01
	Interleukine inhibititors (< 6 months)	L04AC
Indication for influenza vaccination (score 0-1)		
	Diabetes	
	Cardiovascular disease	
	Lung disease	
	Kidney disease	
	Immunosuppressive disease/immunosuppressants	
	Heart valve disease	
	Cirrhosis	D97.04
	Liver disease NOS	D97
	Congenital anomaly NOS/multiple	A90
	Down syndrome	A90.01
	Ruptured spleen traumatic	B76
	Hereditary hemolytic anemia	B78
	Sikkelcell anemia	B78.02
	Congenital anomaly musculoskeletal	L82
	Acquired deformity of the spine	L85

Scoliosis	L85.01
Endocrine/metabolic/nutritional disease other	T99
Cushing syndrome	T99.08
Addison syndrome	T99.09
Adrenal insufficiency	T99.12
Adreno-genital syndrome	T99.13

**A patient with dementia, epilepsy and diabetes generates a score of 3*