Covid-19 risk factors: statistical learning from German healthcare claims data

Roland Jucknewitz, Oliver Weidinger and Anja Schramm
AOK Bayern, Regensburg, Germany

ABSTRACT

Background: Precise individual risk quantification of severe courses of Covid-19 is needed to prioritize protective measures and to assess population risks in a phase of increased immunization. So far, results for the German population are lacking. Furthermore, existing studies pre-specify comorbidity risks by broad categories rather than deriving them from the data using statistical learning techniques.

Methods: Risk factors for severe, critical and lethal courses of Covid-19 are identified from a large German claims dataset covering more than 4 million individuals. To avoid prior grouping and pre-selection of risk factors, fine-grained hierarchical information from medical classification systems for diagnoses, pharmaceuticals and procedures are used, resulting in more than 33,000 covariates. These are processed using a LASSO approach.

Results: We identify relevant risk factors, among which hypertensive diseases, heart disease and the corresponding medications are most relevant at population level. Prior use of diuretics is the strongest single medical predictor for severe course (e.g. Torasemide, odds ratio (OR) 1.801), but also for a critical course (OR 2.304) and death (OR 2.523). To assess risk profiles at the individual level, our approach sums up many such factors and has better predictive ability than using pre-specified morbidity groups (AUC for predicting critical course 0.875 versus AUC \( \leq 0.865 \)).

Conclusions: The proposed method can help to identify risk factors and assess risk at the individual level for other infectious diseases. The results can be used by administrative data holders to guide protective policies, while a risk index can be applied in clinical studies with a narrower focus.

KEYWORDS

SARS-CoV-2
prioritization
prediction
routine data
machine learning

ARTICLE HISTORY

Received 12 April 2021
Revised 12 August 2021
Accepted 13 September 2021

CONTACT

Roland Jucknewitz
roland.jucknewitz@by.aok.de
AOK Bayern, Regensburg, Germany
**Introduction**

In early 2020, Covid-19 has rapidly become a public health emergency of international concern. Besides general containment approaches, such as social distancing or wearing community masks, protecting those at highest risk for severe illness is critical to prevent rising death tolls and overloaded healthcare systems. Such measures include intensified shielding, issuing of protecting masks (FFP2), and prioritization of pharmaceutical interventions during the phase of limited availability. To be effective in these measures, there is an urgent need for a precise quantification and ranking of groups who are at increased risk of severe Covid-19. Quantification of individual risks is still important when immunization of the population proceeds, since adequate policies could then be backed by information about the risks of those not immunized so far.

Several authors have studied the risk factors for severe outcomes like death in an overall population [1,2], in the population of Covid-19-positive [3] or in those hospitalized with Covid-19 [4,5]. These studies assess a broad range of risk factors, especially comorbidity conditions, and some also consider risk prediction to rank the most vulnerable individuals [6]. Beyond these, numerous primary studies and meta-analyses focus on single-risk factors such as diabetes [7], HIV infection [8] or pregnancy [9]; for an overview see the according systematic reviews (e.g. [10]) and official communications (https://www.cdc.gov/coronavirus/2019-ncov/science/science-briefs/underlying-evidence-table.html).

Wherever they are examined in the Covid-19 literature, (co-)morbidity risk factors are identified by a few broad risk groups such as coronary heart disease or diabetes mellitus. Usually, about a dozen such groups, are specified a priori, e.g. by referring to risk groups from other infectious respiratory diseases like influenza, or from groups defined for other purposes, such as the Charlson or Elixhauser comorbidity groups or the corresponding indices. Clift et al. [6] in contrast, with their focus on prediction, have considered an unusually large number of 38 selected risk factors.

In this study, we try to identify risk factors from a German population of a statutory health insurance without pre-specifying such groups. The risk index, which can be constructed from the factors, allows the public health system to prioritize protective measures, while research studying single risk factors or a limited set thereof can use it to statistically account for the most relevant comorbidities within a single variable.

**Methods**

We conduct a retrospective cohort study using observational claims data from AOK Bayern, a regional statutory health insurer with a market share of approximately 40% in the German region of Bavaria. The study population is selected from all 4,6 mio. individuals insured by AOK Bayern on 1/3/2020. We exclude individuals with a loss of follow-up due to a change of insurance before 1/12/2020, persons who were not enrolled during the entire year 2019 and non-German residents (identified by postal address or by an according label in the risk adjustment data). Thus, our population-at-study includes $N = 4,101,745$ individuals. The population is not restricted to patients tested Covid-19-positive or to hospitalized patients, so that the cohort is designed to measure an overall population risk that includes both the infection risk and the risk given an infection. For this population, we collect co-morbidities and other characteristics in a baseline period and measure the Covid-19 outcomes in a follow-up period beginning 1/3/2020. The data are drawn from different reimbursed healthcare services such as ambulatory or hospital treatments and drug prescriptions, as well as socio-economic factors, all of which are available with high data quality.

As study outcomes, we consider severe, critical and fatal clinical courses of Covid-19. A severe Covid-19 case (Y1) is defined by an admission to hospital with confirmed Covid-19 diagnosis (ICD1U07.1) and either documented pneumonia (ICD J12.- to J18.-) or mechanical ventilation. A critical course (Y2) restricts this set to persons who received intensive care or mechanical ventilation, who had a documented sepsis (ICD R65.0, R65.1 or R57.2) or who died within hospital (identified by reason of discharge). Intensive care is identified if the first, last or longest stay at hospital is an intensive care unit (ICU), at least one diagnosis is coded by an ICU or by the OPS keys 8-980.-, 8-98 D.-, 8-98 F.-, 8-712.0, 8-721.1, 8-721.2 or 8-721.3. In-hospital-death (Y3) further restricts this set to those who died within hospital. We consider critical course (Y2) as primary outcome, while we use Y1 and Y3 in additional analyses to assess the sensitivity with respect to outcome. Outcome data are collected from all hospital admissions on 1/3/2020 or later on for which final invoices were available on 22/11/2020. We use data that became available later until 25/1/2021 for a prediction validation set in section ‘Predictive performance’.

To identify risk factors, we use the following predictors: age and gender in 40 groups (age is categorized...
by 5%-percentile), nationality (the 10 most frequent nationalities and ‘other’), nursing home residence, insurance and occupational status (compulsorily insured employee, voluntarily insured employee, compulsorily insured pensioners, voluntarily insured pensioners, roughly short-term unemployed (‘Arbeitslosengeld 1’), roughly long-term unemployed (‘Arbeitslosengeld 2’), welfare recipient, self-employed, dependently insured spouse, dependently insured child), and labour income by quintile groups. We approximate the regional infection risk by the 7-day Covid-19-incidence from Robert Koch Institute, measured on the date of admission to hospital for individuals with the considered outcome. For all other individuals, the reference date is randomly assigned from a distribution that matches the distribution of the admissions to hospital for the aforementioned Covid-19 cases. Hence, regional biases due to regional disparities of infection rates and morbidities are mitigated. Whenever nationality is missing (in 27 cases), we set it to German as the most probable category. No other characteristics are missing in our dataset.

With regard to the risk associated with morbidity and medical treatments, we use data on outpatient and inpatient diagnoses according to the ICD, prescriptions of drugs from ready-made medicines or mixtures according to the German Anatomical Therapeutic Chemical (ATC)-Classification, as well as medical interventions according to the German procedural classification (Operationen- und Prozedurenschlüssel or OPS); for information about the classification systems see BfArM. These data are collected for the baseline period from 1/1/2019 until 29/2/2020. As an exception, ambulatory diagnoses are only used until 31/12/2019, since these data merely refer to a quarter as a whole.

We aim to avoid a loss of medical information due to prior grouping, and nonetheless to structure the information where it is found appropriate. We, therefore, retain data on the granular level of full digit ICD and ATC, and 5-digit OPS codes, and add the hierarchical structure of the classification systems of ICD, ATC and OPS into the information set (see Stausberg and Hagn [11] who suggests using ICD structure on chapter and group levels). Hence, if a high level code such as the 5-digit-ICD I25.22 (‘Old myocardial infarction, more than one year past’, hierarchical level 5) is coded, we activate an according indicator variable, but also assign a value of 1 to indicators for the lower levels of the hierarchy, i.e. the 4-digit-code I25.2 ‘Old myocardial infarction’ (level 4), the 3-digit-code I25 ‘Chronic ischaemic heart disease’ (level 3), the group I20-I25 ‘Ischaemic heart diseases’ (level 2) and Chapter IX ‘Diseases of the circulatory system’ (level 1). The same logic holds for ATC codes, where anatomical main group, therapeutic, pharmacological and chemical subgroup as well as chemical substance are the five hierarchical levels. For OPS keys, where groups, 3-, 4- and 5-digit codes are the second to the fifth level analogously to ICD, we use only codes from Chapters 5 (operations), 6 (medications) and 8 (non-operative therapeutic interventions) and drop the chapter code. This total enriched set of medical information amounts to a number of 33,362 binary predictor variables, from which risk factors are identified.

We use a logistic lasso regression to predict the binary outcomes by a selected set of predictors and shrinkage. To favour the selection on low-level rather than high-level codes where appropriate, we apply a differential shrinkage factor that corresponds to the hierarchical level of the medical codes, while first level codes are shrunk with the same factor as demographic and socio-economic variables. No shrinkage is applied to the regional incidence, since this variable is measured at a different (non-binary) scale than all other variables and bias is avoided in this variable. Five-fold cross validation is applied to maximize the AUC with respect to the shrinkage parameter. For each medical category (or code), we obtain its estimated contribution to Covid-19 risk at logistic scale by adding the corresponding parameters of all hierarchical levels.

To assess the predictive performance of the chosen approach, we compare the described method of full information to prior approaches of pre-specified Covid-19 risk groups. To be precise, we reconstruct medical groups of Williamson et al. [2], McKeigue et al. [1], Schröder et al. [12] and Clift et al. [6], and also apply the Charlson Morbidity groups used by Harrison et al. [3]. Furthermore, we quantify the possible improvement that nonlinear machine learning techniques can have over the linear lasso, and add tree boosting with full quantitative information to the comparison. As a first evaluation, the predictive performance is assessed in the same cross validation folds as mentioned above, so that variable selection and parameter estimation does not use the fold to be predicted beyond the univariate maximization of the shrinkage parameter. A second evaluation setup is constructed from outcome data that became available after all model specification and estimation steps where finished. We compare the AUC, the expected information for discrimination $\Lambda$, and predictive likelihoods for statistical comparisons.
We construct a risk index which is suitable to adjust for comorbidities when investigating the effect of specific variables in greater detail. To this end, we compute the logistic link value on the cross-validation folds. No outcome data from the same folds are used to compute the index, and, hence, it can be consistently used on the right side of a regression explaining the outcome. The contributions of risk factors to be investigated in greater detail can be easily cancelled by setting their coefficients in the lasso model of each fold to zero.

**Estimation results**

An overview over population characteristics by outcome groups is provided in Table 1. Among the 4,101,745 individuals in the study population, 1611 are characterized by the primary outcome critical Covid-19 (Y2), with admissions to hospital occurring between 1/3/2020 and 10/11/2020, while the broader outcome (Y1) is documented for 3,731 and the hospital mortality outcome (Y3) is documented for 931 individuals. Age, gender, multimorbidity (measured as number of distinct ICD groups) and polymedication (number of distinct pharmaceutical subgroups) differ substantially between the groups, with the more severe outcomes dominated by a greater percentage of male, older and more multimorbid and polymedicated patients.

Estimating the baseline logistic lasso regression for the primary outcome Y2 over the full population, we obtain an optimal shrinkage parameter $\lambda = 3.4882 \times 10^{-7}$ that yields an in-sample AUC of 0.890. All coefficients are available online as Supplementary Table S1. In this setup, 4497 coefficients are nonzero, among them 3292 ICD, 792 ATC and 413 OPS codes. While all of the 1st level (e.g. ICD chapters) and 85.5% of 2nd level codes (e.g. ICD groups) are nonzero, only 10.5% of 5th-level codes differ from zero. This shows the necessity to subtly differentiate some, but not all of the major morbidity groups. Due to the hierarchical structure, each ICD and ATC code has a nonzero coefficient when aggregated over hierarchical levels.

The most severe single medical risk factors for critical course Y2 (by magnitude of the odds ratios and applying a minimum group size of 1000) are three diuretics in the ATC class C03CA (e.g. Torasemid, OR 2.3) and the ICD I98.2 (oesophageal varices without bleeding in diseases classified elsewhere, OR 2.2); see Table 2. Overall, diagnoses related to the circulatory system and pharmaceuticals for the cardiovascular system are dominant in the medical factors with the highest impact and account for 97 of the 100 top codes. A similar pattern is found documented for 3,731 and the hospital mortality outcome (Y3) is documented for 931 individuals. Age, gender, multimorbidity (measured as number of distinct ICD groups) and polymedication (number of distinct pharmaceutical subgroups) differ substantially between the groups, with the more severe outcomes dominated by a greater percentage of male, older and more multimorbid and polymedicated patients.

Estimating the baseline logistic lasso regression for the primary outcome Y2 over the full population, we obtain an optimal shrinkage parameter $\lambda = 3.4882 \times 10^{-7}$ that yields an in-sample AUC of 0.890. All coefficients are available online as Supplementary Table S1. In this setup, 4497 coefficients are nonzero, among them 3292 ICD, 792 ATC and 413 OPS codes. While all of the 1st level (e.g. ICD chapters) and 85.5% of 2nd level codes (e.g. ICD groups) are nonzero, only 10.5% of 5th-level codes differ from zero. This shows the necessity to subtly differentiate some, but not all of the major morbidity groups. Due to the hierarchical structure, each ICD and ATC code has a nonzero coefficient when aggregated over hierarchical levels.

**Table 1.** Descriptive statistics of selected independent variables, by outcome status.

| No outcome | Y1 severe | Y2 critical | Y3 death |
|------------|-----------|-------------|----------|
| N          | 4,098,014 | 3731        | 1611     | 931      |
| Age        | 45.5      | 69.9        | 73.9     | 79.8     |
| Gender (male) | 48.4%     | 54.8%       | 59.2%    | 56.1%    |
| Multimorbidity | 11.3      | 22.2        | 24.0     | 26.5     |
| Polymedication | 3.5       | 7.7         | 8.5      | 9.4      |
| Nursing home | 1.1%      | 13.7%       | 17.3%    | 25.1%    |
| Regional incidence | 56.8      | 81.2        | 74.4     | 82.3     |
| Nationality: German | 81.0%     | 81.5%       | 86.5%    | 93.1%    |
| Nationality: other | 19.0%     | 18.5%       | 13.5%    | 6.9%     |
| Wage: na/not working | 57.6%     | 79.6%       | 87.7%    | 96.9%    |
| Wage: 1st quintile | 8.5%      | 2.9%        | 1.6%     | 0.9%     |
| Wage: 2nd quintile | 8.5%      | 4.5%        | 2.4%     | 0.4%     |
| Wage: 3rd quintile | 8.5%      | 4.8%        | 2.9%     | 0.3%     |
| Wage: 4th quintile | 8.5%      | 3.7%        | 2.0%     | 0.4%     |
| Wage: 5th quintile | 8.5%      | 4.5%        | 3.4%     | 1.1%     |

**Table 2.** Adjusted log odds ratios (logOR) of single clinical predictors, ranked by the logOR of outcome Y2 (column ‘Rank’).

| Rank | Code | Diagnosis or drug name | Y1 severe | Y2 critical | Y3 death |
|------|------|------------------------|-----------|-------------|----------|
| 1    | C03CA04 | Torasemide | 0.589 | 0.835 | 0.925 |
| 4    | I98.2 | Oesophageal varices without bleeding in diseases classified elsewhere | 0.582 | 0.776 | 0.371 |
| 12   | C01BD01 | Amiodarone | 0.447 | 0.730 | 0.582 |
| 17   | I05.0 | Mitral stenosis | 0.387 | 0.644 | 0.600 |
| 19   | C02CA06 | Urapidil | 0.508 | 0.627 | 0.561 |
| 20   | I15.10 | Hypertension secondary to other renal disorders | 0.349 | 0.623 | 0.487 |
| 21   | C05AD04 | Cinchocaine | 0.447 | 0.607 | 0.553 |
| 22   | E10.30 | Type 1 diabetes mellitus with ophthalmic complications | 0.471 | 0.598 | 0.291 |
| 25   | C07CB23 | Atenolol and chlorothalidone | 0.409 | 0.587 | 0.554 |
| 27   | C09DB01 | Valsartan and amlodipine | 0.440 | 0.561 | 0.543 |
| 29   | I67.10 | Cerebral aneurysm, nonruptured | 0.335 | 0.557 | 0.341 |
| 32   | A10AE05 | Insulin detemir | 0.332 | 0.547 | 0.503 |
| 33   | I61.0 | Intracerebral haemorrhage in hemisphere, subcortical | 0.338 | 0.544 | 0.338 |
| 35   | I46.0 | Cardiac arrest with successful resuscitation | 0.315 | 0.538 | 0.530 |
| 36   | I70.24 | Atherosclerosis of arteries of extremities | 0.434 | 0.537 | 0.601 |
| 40   | I21.9 | Acute myocardial infarction, unspecified | 0.333 | 0.536 | 0.453 |
| 42   | C08CA08 | Nifedipine | 0.351 | 0.534 | 0.554 |
| 52   | C10AA04 | Fluvoxamine | 0.320 | 0.525 | 0.540 |
| 58   | A02AH01 | Sodium hydrogen carbonate | 0.368 | 0.522 | 0.399 |
| 77   | I65.3 | Occlusion and stenosis of multiple and bilateral precerebral arteries | 0.377 | 0.516 | 0.386 |

Only the first diagnosis or pharmaceutical substance of each level-3-group is shown. Example: The use of torasemide increases the risk of outcome Y2 approximately by a factor of $\exp(0.835) = 2.305$, since for rare outcomes odds ratios approximates relative risks.
for the other outcomes. Several non-medical factors (not shown in Table 2) like nursing home (OR 3.0) have a relevant impact. Categorical groups of age and gender are not leading the list of risk factors, but the effects are interdependent with socio-economic characteristics like pension status.

Considering aggregated (2nd level) groups only, we obtain aggregated log odds ratios by averaging over all fine-grained codes and weighting by their prevalence. The importance of the group in the population (Pop) is obtained by multiplying the log odds ratios by group size and ranking over this excess relative risk. The top 20 groups with respect to the critical outcome Y2 are shown in Table 3. Hypertensive diseases (ICD I10-I15), other forms of heart disease (ICDs I30-I52) and agents acting on the renin–angiotensin system (ATC C09) add most to the population-level risk, followed by metabolic disorders (E70–E90), which is the first group not related to the circulatory system.

The effects are strongly correlated between the outcomes, as can be seen in the top groups in Table 3, and in Supplementary Figure S1 which shows the scatter plot of log odds ratios corresponding to each predictor. The overall risk score on logistic scale follows a right-skewed distribution, see Supplementary Figure S2. From the distribution of scores over age groups younger and older than 80, it becomes evident that even if age is a significant risk factor, morbidity-based criteria often outweigh a higher age, and 41% of the top-5% scorers (with score at probability scale greater than 0.00151) are neither 80 years old nor reside in a nursing home.

Table 3. Adjusted log odds ratios (logOR) and ranks in population importance (Pop) of aggregated medical risk factors at aggregation level 2.

| Code     | Name                                      | Outcome Y1 severe | Outcome Y2 critical | Outcome Y3 death |
|----------|-------------------------------------------|-------------------|---------------------|------------------|
| I10-I15  | Hypertensive diseases                      | 0.307             | 0.385               | 0.350            |
| I30-I52  | Other forms of heart disease               | 0.319             | 0.440               | 0.430            |
| C09      | Agents acting on the renin-angiotensin system | 0.363             | 0.511               | 0.538            |
| E70-E90  | Metabolic disorders                        | 0.283             | 0.261               | 0.135            |
| C03      | Diuretics                                 | 0.561             | 0.787               | 0.867            |
| E10-E14  | Diabetes mellitus                          | 0.299             | 0.299               | 0.177            |
| C07      | Beta blocking agents                       | 0.339             | 0.522               | 0.556            |
| I20-I25  | Ischaemic heart diseases                   | 0.307             | 0.421               | 0.369            |
| A02      | Drugs for acid related disorders           | 0.280             | 0.377               | 0.351            |
| C10      | Lipid modifying agents                     | 0.316             | 0.496               | 0.536            |
| I80-I89  | Diseases of veins, lymphatic vessels and lymph nodes, not elsewhere classified | 0.317             | 0.400               | 0.329            |
| A10      | Drugs used in diabetes                     | 0.287             | 0.418               | 0.399            |
| E00-E07  | Disorders of thyroid gland                 | 0.245             | 0.194               | 0.092            |
| E65-E68  | Obesity and other hyperalimentation        | 0.321             | 0.282               | 0.108            |
| C08      | Calcium channel blockers                   | 0.343             | 0.516               | 0.529            |
| I70-I79  | Diseases of arteries, arterioles and capillaries | 0.289             | 0.430               | 0.414            |
| I60-I69  | Cerebrovascular diseases                   | 0.327             | 0.427               | 0.332            |
| B01      | Antithrombotic agents                      | 0.231             | 0.240               | 0.410            |
| N17-N19  | Renal failure                             | 0.245             | 0.274               | 0.426            |
| N02      | Analgesics                                | 0.161             | 0.115               | 0.221            |

A risk index, computed from the coefficients in Table S1, can be used to adjust for a large set of confounders within a single variable. To illustrate this, we focus on the risk effect of age and gender. To construct the index, we compute predictions of the baseline model at logistic scale, and – in concordance to the question at hand – leave out age and gender groups in the predictions by setting their coefficients to zero. Using this index on the right-hand side of a regression that explains the outcome would result in a bias, since the same outcome enters also on the right-hand side of the regression through the coefficients of the index. Therefore, cross validation is applied so that for each data fold the model to construct the index is estimated only from data outside this fold.

To study age and gender effects we estimate a generalized additive model where age effects are modelled by smooth cubic splines separated by gender, and the risk index enters the model linearly. Age and gender effects from the full model are compared to unconditional age and gender effects without adjustment for other confounders such as comorbidities. To keep the effects comparable, we account for socio-demographic groups of pensioners and co-insured children even in the unconditional model. The risk profiles are shown in Figure 1. It is evident that the rise in risk that is due to older age and being male is highly overstated by the unconditional profile in the lower panel, while the sharp rise from age 65 on is not visible in the upper panel, where the comorbidity risks are held constant. Still, age and gender remain statistically highly significant and are
– considered individually – among the most important risk factor even in the conditional model. Other risk factors or morbidities can be studied similarly, by canceling related variables when computing the risk score and running individual regression with a more detailed specification of the risk factors at question.

Predictive performance

We evaluate the predictive performance of the described high-dimensional regression approach and compare it to several benchmarks. To that end, we obtain predictions on five-fold cross validated samples, where outcome data of the same fold have neither been used to select the variables nor to compute the predictions. Additionally, we assess the predictive performance in an a priori unseen holdout sample of recent data which became available after the models have been estimated.

First, we assess whether there are gains in using widely unstructured medical data in our baseline model as compared to using age and gender alone. Additionally, we use only age, gender and the described socio-demographic variables together with risk groups defined in the prior literature. We adjust all predictions at logistic scale to fit the overall prevalence of the outcome, since with one method (boosting, see below) the scale is shifted by differential weights on outcomes and controls.

The results of this predictive comparison can be seen in Table 4. Consider the middle panel corresponding to the primary outcome Y2 and the left three numeric columns corresponding to cross-validation first. All three measures lead to the same ranking of models, where the baseline model outperforms the group-based approaches. While the area under the curve (AUC, aka C-statistic) is a widely used metric, we compute also the expected weight of evidence $\Lambda$ promoted by McKeigue [13] which was used in McKeigue et al. [1]. A doubling of $\Lambda$ (‘adding one bit of information’) is understood as the quantity of information that halves the hypothesis space. Hence, adding socio-demographic and medical information in the baseline model adds a substantial amount of information to the age and gender model. The predictive log likelihoods show a high statistical evidence in favour of the baseline model as compared to the smaller group-based models. E.g. the log likelihood difference of the baseline and the McKeigue et al. [1] model is about 160. It is notable, however, that the groups by McKeigue et al. [1] and Williamson et al. [2] outperform the baseline model in the cross validation framework for outcome Y1, and also for some performance measures with outcome Y3, see the upper left and lower left panels of Table 4.

In the last line of each panel, we add tree boosting as a nonlinear approach which is considered the state of the art in machine learning using tabular data. We use the XGBoost library, version 1.0.0.2, in R, where tree depth is set to 3, the shrinkage parameter is 0.01, the
The number of iterations is determined by cross validation and the scale_pos_weight parameter is used to weight the outcome group equally to the much larger group of controls in the overall population. Each panel of Table 4 shows that the nonlinear approach improves over the baseline model with a slightly larger AUC and a statistically relevant log likelihood difference (of 89 in the middle left panel). The ROC curve in Figure 2 shows that...
the baseline model performs similarly to boosting for the highest-risk groups at the left, while persons with low-risk profile are ranked slightly worse. The steep slope for the highest-risk groups at the left illustrates the potential to avoid a high proportion of critical outcomes by properly prioritized protective measures of only a small number of effectively protected individuals.

A second evaluation of predictive accuracy uses only data on outcomes that have become available after the estimation of the models and consists of all hospital stays for which invoices have entered the database between 23/11/2020 and 25/01/2021. We thus assess the stability of the models over time and with respect to the larger time lag between outcomes and input data (where still only data until 28/2/2020 are used). We estimate the models with the whole population and with outcome data available until 22/11/2020. The test set consists of all individuals who have not incurred a considered outcome in the previously available dataset. The results in the three right columns of Table 4 show that the ranking of methods are similar to that of the cross validation study. The baseline model is consistently better than group-based approaches, however, while the superiority of the tree boosting approach is somewhat more pronounced.

We conclude that fine-grained medical information is able to improve predictions of critical Covid-19 outcomes in an overall population, while using predeter-

Discussion and limitations

We developed a statistical learning approach which is successful in identifying risk factors and competitive with respect to predictive performance.

The benefits of our approach in individual risk ranking or self-assessment can be utilized by web tools similar to that of the QCOVID project (digital.nhs.uk/coronavirus/risk-assessment) or by data holders like national healthcare agencies or insurances, using the publicly available coefficients of Supplementary Table S1. But we see potential also in the quantification of population risks. In situations in which the immunization status is present in the data, the potential utilization of healthcare capacities can be assessed in different phases of the pandemic by aggregating individual risks.

Importantly, due to the ceteris paribus notion of (logistic) regression coefficients, identifying a risk factor as having a relatively high coefficient does not mean that the corresponding individuals are also at highest risk from a ranking point of view. The overall effect of having a certain health condition is substantially affected by complementary medication and frequent co-morbidities. Hence, while the regression-based odds ratio of the oldest male age group (AGG20) is higher than the odds ratio for end-stage kidney disease (ICD N18.5), the morbidity risk including possible further conditions like dialysis of the latter group dominates that of AGG20 (OR 17.6 versus 4.6), and also the overall risk (including the age effects) of individuals with end-stage kidney disease exceeds that of AGG20. We, therefore, emphasize that a ranking of risks should be in an additive manner accounting for all risk factors of the individuals rather than looking at the risk factors with the highest odds ratios alone. The full risk profiles of certain individuals are obtained by summing up the appropriate coefficients of Supplementary Table S1.

Our identified main risk factors are very fine-grained and specific, which makes them hard to compare to previous results that uses a more aggregate approach. While chronic heart disease, for example, is a single risk factor (OR 1.17) without further specification in [2], with our approach the excess risk depends on the type of heart disease and its medication. In many cases, it is helpful to view medications like diuretics and morbidity codes like oesophageal varices as indicative for the severity of an underlying morbid condition rather than as a causal individual risk factor.

From the perspective of clinical research on Covid-19 risk factors, studies differ with respect to the co-morbidities that are controlled for when specific conditions are investigated. In small studies it is possible to account for only a handful such factors. In such situations, using an externally constructed index as the one obtained from the coefficients in Supplementary Table S1 appears beneficial, at least if sufficient information on co-morbidities, medications and prior operations are collected within the study. It can help to gain more comparable insights in the risk of several conditions.

For a valid interpretation of the results, a few limitations of the study are in order. First, we have studied risk in a population of individuals insured by AOK Bayern. While composed of nearly all groups of the Bavarian population (besides civil servants and a small group without any health insurance), it is not representative for the overall population due to differential selection of public and private health insurers.
Second, studying risk in an overall population might not give appropriate answers to the question who is at highest risk when infected with Covid-19, as shortly stated in the introduction. Differences between the results of these two concepts occur when considered risk groups have different risks of getting infected because of e.g. different protective measures between these groups. Another possible source of bias is heterogeneous regional Covid-incident when also the risk factors are not evenly dispersed regionally. We tried to account for this by including the regional incidence as a predictor.

Third, the risk groups and coefficients identified by the statistical learning procedures are not always easy to interpret from a medical point of view. One reason is that diagnoses and treatments (drugs, operations) are introduced as separate variables. Their effect might be cumulated (e.g. adding the effects of diabetes diagnoses and of antidiabetic drugs), but in many cases, the treatments are not all specific for certain medical conditions. Thus, when studying medical risk groups in greater detail, it might be necessary to build the corresponding risk variables from a subject-specific perspective, while other confounding risks can be accounted for using our approach as outlined in the discussion of age and gender risks above.

Fourth, we have used country-specific medical codes, and the risk coefficients might not be representative beyond the coding environment of the German statutory health insurance system. For other setups, it might be necessary to use other coding systems and a re-estimation of the coefficients, but the overall methodology should still be applicable.

Finally, the aim of the study was to consider fine-grained medical information and to assess their informative content, but the baseline model studied here is not designed to reach maximum predictive performance, as results with the gradient boosting benchmark show. Superior predictive performance might be obtained by combining risk groups and fine-grained information in a nonlinear machine learning algorithm. For practical purposes (e.g. prioritization), however, predictive performance might not be the only target. Ethical issues (leaving out nationality or other socio-economic factors) or arguments for a simpler more practicable model (grouping on lower levels and linear score models) could outweigh small boosts in predictive accuracy in most practical situations.

Conclusion
We have identified and quantified risk factors for several Covid-19 outcomes in the overall population of AOK Bayern insured. For all outcomes of severe, critical and lethal course of Covid-19, hypertensive and other heart diseases and their medication appear most relevant in the population, followed by metabolic disorders and diabetes mellitus. From a clinical perspective, we find that it is important to divide such large groups further, to get an accurate view of individual risks. The methodology and our published coefficients may also be of interest for researchers who aim to adjust for confounders in smaller clinical studies of Covid-19 outcomes.

Ethical approval
The data used in this analysis were completely anonymized. Therefore, an approval of an ethics committee or another respective institution was not required. Because of anonymization, a written consent was not required as well.

Disclosure statement
They have no conflicts of interest to declare.

Notes
1. International Statistical Classification of Diseases, German Modification or ICD-10-GM.
2. https://www.dimdi.de/dynamic/en/classifications/

References
[1] McKeigue PM, Weir A, Bishop J, et al. Rapid epidemiological analysis of comorbidities and treatments as risk factors for COVID-19 in Scotland (REACT-SCOT): a population-based case-control study. PLoS Med. 2020;17(10):e1003374.
[2] Williamson EJ, Walker AJ, Bhaskaran K, et al. Factors associated with COVID-19-related death using OpenSAFELY. Nature. 2020;584(7821):430–436.
[3] Harrison SL, Fazio-Eynullayeva E, Lane DA, et al. Comorbidities associated with mortality in 31,461 adults with COVID-19 in the United States: a federated electronic medical record analysis. PLoS Med. 2020;17(9):e1003321.
[4] Karagiannidis C, Mostert C, Hentschker C, et al. Case characteristics, resource use, and outcomes of 10 021 patients with COVID-19 admitted to 920 German hospitals: an observational study. Lancet Respir Med. 2020;8(9):853–862.
[5] Knight SR, Ho A, Pius R, et al. Risk stratification of patients admitted to hospital with covid-19 using the ISARIC WHO clinical characterisation protocol: development and validation of the 4C mortality score. BMJ. 2020;370:m3339.
[6] Clift AK, Coupland CA, 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(8265):m3731.

[7] Barron E, Bakhai C, Kar P, et al. Associations of type 1 and type 2 diabetes with COVID-19-related mortality in England: a whole-population study. Lancet Diabetes Endocrinol. 2020;8(10):813–822.

[8] Bhaskaran K, Rentsch CT, MacKenna B, et al. HIV infection and COVID-19 death: a population-based cohort analysis of UK primary care data and linked national death registrations within the OpenSAFELY platform. Lancet HIV. 2021;8(1):e24–e32.

[9] Villar J, Ariff S, Gunier RB, et al. Maternal and neonatal morbidity and mortality among pregnant women with and without COVID-19 infection. JAMA Pediatr. 2021;175(8):817–826.

[10] Treskova-Schwarzbach M, Haas L, Reda S, et al. Pre-existing health conditions and severe COVID-19 outcomes: umbrella review and meta-analysis of global evidence. BMC Med. 2021;19:212.

[11] Stausberg J, Hagn S. New morbidity and comorbidity scores based on the structure of the ICD-10. PLoS One. 2015;10(12):e0143365.

[12] Schröder H, Brückner G, Schüssel K, et al. 2020. Monitor: Vorerkrankungen mit erhöhtem Risiko für schwere COVID-19-Verläufe. WldOmonitor. https://www.wido.de/news-events/aktuelles/2020/monitor-covid-19/.

[13] McKeigue P. Quantifying performance of a diagnostic test as the expected information for discrimination: relation to the C-statistic. Stat Methods Med Res. 2019;28(6):1841–1851.