**Supplementary Information**

**Supplementary statistical methods**

*Data preparation*

We extracted 122,250 COVID-19 cases with a diagnosis date between 1 Feb 2020 and 1 Jul 2020. Diagnosis was confirmed either by documented ICD10 codes in the OPTUM*®* diagnosis table any one of U07.1, U07.2, U07.3 or a combination of one of J12.89/J20.8/J40/J22/J98.8/J80 plus B97.29 on the same encounter or, alternatively, via positive test result for PCR laboratory tests (Supplementary Table S1). Survival status was derived from the OPTUM*®* patient table. Survival time was computed as the number of days between the date of COVID-19 diagnosis and last documented clinical activity (vitals, labs, medication, encounter).

Our analysis focused on predisposition. We kept all laboratory and vital data from a time window between 12 months prior to COVID-19 diagnosis to 1 month prior to COVID-19 diagnosis. In case of multiple measurements taken in the time window, the median value per patient was used. Medical history (MH), as documented by ICD codes, was included for the entire patient history available 1 month or more before the COVID-19 diagnosis. We defined indicator variables of whether a patient had or had not a prior history of a particular disease entity. We investigated all disease entities that a part of the Charlson comorbidity index (Quan et al. 2011), the AHRQ (Agency for Healthcare Research and Quality 2018), Arrhythmia from the former Elixhauser definition (Elixhauser et al. 1998), as well as a series of entities previously implicated in COVID-19 outcomes, including COPD, hypertension/hypotension, asthma and heart failure (HF). A detailed description of assignment of ICD codes to disease entities can be found in (Supplementary Table S2). Prior immunizations were handled analogously.

*Quality control, transformation and handling of missing data*

In the first step, we kept all lab and vital parameters, which were available for at least 1,000 patients, leaving 249 variables for primary analysis. We set implausible measurements to missing (e.g. negative values, percentages greater than 100, “0” if not biologically sensible, etc.). For continuous variables, observations more than +/-3 interquartile ranges from the median were set to missing, unless more extreme values were deemed to be plausible according to the judgement of an experienced physician. We performed log-transformation of variables when the Shapiro-Wilk test (Shapiro and Wilk 1965) showed less significant deviation from normality on the log-scale than it did on the original scale. We confirmed the decision on transformation via visual inspection of the density plots. The accordingly processed and quality-controlled data were used for univariate association analysis (cf. below).

For multivariate analysis, we additionally removed all patients from the analysis, for which less than 10 different variables were available, leaving 55,757 patients for analysis. Variables available for less than 10,000 patients were mean-imputed, resulting in a remaining missing rate of 12.4% in total. Those remaining missing values were imputed using the missForest R-package (Stekhoven and Buhlmann 2012). The package iteratively imputes missing values, and is suited for mixed-type data. It averages over unpruned trees, using the built-in out-of-bag error estimates of random forest.

We computed pairwise Pearson correlation coefficients (“VII. Note on Regression and Inheritance in the Case of Two Parents” 1895) of variables, for the purpose of sanity checking of variables via to be expected correlations, and to investigate potential collinearity.

*Association analysis*

As primary analysis, we performed survival analysis, in particular Cox regression (Enderlein 1987), to analyze time-to-death from the date of COVID-19 diagnosis. Univariate analysis was conducted using age, sex, ethnicity, rate, insurance status, and US region/division as covariate parameters for adjustment. We applied a Bonferroni-correction with the number of variables (m=250) to account for multiple testing and judged variable association to be significant if it met an 𝜶-level of 𝜶=0.05/m=2\*10-4. We used the R survival package (T. Therneau and Grambsch 2000) and in particular the survfit() function for survival times and the coxph() function for cox regression. Median survival times were computed using the methodology as described in (Kalbfleisch and Prentice 2002) and visualized as Kaplan-Meier plots (Kaplan and Meier 1958). For our main model, we computed Martingale residual (T. M. Therneau, Grambsch, and Fleming 1990) and visualized them as suggested in (Hosmer, Lemeshow, and May 2008).An interpolation curve was fitted with the R lowess() function (Cleveland 1979). We tested if the Cox proportional hazards (PH) assumption (Grambsch and Therneau 1994) was fulfilled and computed Schoenfeld residuals (Schoenfeld 1982).

In order to allow comparison of hazard ratios (HRs) between different variables, we report the 2-standard-deviations hazard ratio “HR\_2SD”. It is computed as HR\_2SD=HR^{2\*SD}, where SD is the standard deviation of the respective variable.

The association analysis between a priori measurements of hypotensive DBP and a priori median albumin levels (n=45,819 patients) was performed using logistic regression. To be consistent with the Cox regression the results are shown as 2-standard-deviations odds ratio “OR\_2SD”.

*Model development*

As secondary analysis, multivariable modelling was performed. We pursued two approaches in parallel. First, we performed a backward selection procedure on the Cox regression model of all eligible variables. We iteratively removed the variable with least impact on model performance, until all remaining parameters were significant at α1=0.05/250=2\*10-4 (Bonferroni-correction). By construction, the procedure controls the family-wise error rate at α=0.05. In parallel, we derived a regularized Lasso model (Simon et al. 2011). We fitted a L1 (Lasso) regularized Cox-Proportional Hazards Model using glmnet version 3.02 (Simon et al. 2011), with the concordance index (C-index) (Steck, Krishnapuram, and Dehing-Oberije 2008) as the performance measure. The regularization parameter λ was optimized using ten-fold cross-validation. We selected λ such that we extracted the most regularized model with a C-index within one standard error of the best performing model.

To compare model performance between the two approaches, we used the C-index and and receiver operating characteristics area under the curve (ROC-AUC) for censored data, implemented in the R package timeROC (Blanche, Dartigues, and Jacqmin-Gadda 2013). ROC-AUC was computed for different time points (5, 10, 20, 30, 60s-days survival).

For the final model, we checked if recently suggested requirements for developing multivariable prediction models were fulfilled (Riley et al. 2019).

**Supplementary statistical results**

The plot of the martingale residuals of the multivariate model (Supplementary Figure S1) and the red fitting line, which increases very slightly and only at the end, suggest that linearity on the log hazard ratio scale is a reasonable model assumption. Also the Schoenfeld residuals (Supplementary Figure S2) are distributed symmetrically, and do not suggest relevant systematic deviations from the proportional hazards assumption.

**Supplementary Figure S1**: Martingale residuals of the combined multivariate model with lowess fitting line



**Supplementary Figure S2**: Schoenfeld residuals of the combined multivariate model



**Supplementary references**

Agency for Healthcare Research and Quality. 2018. “Elixhauser Comorbidity Software for ICD-10-CM Healthcare Cost and Utilization Project.” Agency for Journal of Statistical Software 23 Healthcare Research and Quality, Rockville, MD, USA. https://www.hcup-us.ahrq.gov/toolssoftware/comorbidityicd10/comorbidity\_icd10.jsp.

Blanche, Paul, Jean-François Dartigues, and Hélène Jacqmin-Gadda. 2013. “Estimating and Comparing Time-Dependent Areas under Receiver Operating Characteristic Curves for Censored Event Times with Competing Risks.” *Statistics in Medicine* 32 (30): 5381–97. https://doi.org/10.1002/sim.5958.

Cleveland, William S. 1979. “Robust Locally Weighted Regression and Smoothing Scatterplots” 74 (368): 829–36.

Elixhauser, A, C Steiner, DR Harris, and RM Coffey. 1998. “Comorbidity Measures for Use with Administrative Data.” *Medical Care* 36 (1): 8–27.

Enderlein, G. 1987. “Cox, D. R.; Oakes, D.: Analysis of Survival Data. Chapman and Hall, London – New York 1984, 201 S., £ 12,–.” *Biometrical Journal* 29 (1): 114–114. https://doi.org/10.1002/bimj.4710290119.

Grambsch, Patricia M., and Terry M. Therneau. 1994. “Proportional Hazards Tests and Diagnostics Based on Weighted Residuals.” *Biometrika* 81 (3): 515–26. https://doi.org/10.1093/biomet/81.3.515.

Hosmer, David W., Stanley Lemeshow, and Susanne May. 2008. *Applied Survival Analysis: Regression Modeling of Time-to-Event Data*. 2nd ed.

Kalbfleisch, John D., and Ross L. Prentice. 2002. *The Statistical Analysis of Failure Time Data: Kalbfleisch/The Statistical*. Wiley Series in Probability and Statistics. Hoboken, NJ, USA: John Wiley & Sons, Inc. https://doi.org/10.1002/9781118032985.

Kaplan, EL, and P Meier. 1958. “Nonparametric Estimation from Incomplete Observations.” *Journal of the American Statistical Association* 53: 457–81.

Quan, H., B. Li, C. M. Couris, K. Fushimi, P. Graham, P. Hider, J.-M. Januel, and V. Sundararajan. 2011. “Updating and Validating the Charlson Comorbidity Index and Score for Risk Adjustment in Hospital Discharge Abstracts Using Data From 6 Countries.” *American Journal of Epidemiology* 173 (6): 676–82. https://doi.org/10.1093/aje/kwq433.

Riley, Richard D, Kym IE Snell, Joie Ensor, Danielle L Burke, Frank E Harrell Jr, Karel GM Moons, and Gary S Collins. 2019. “Minimum Sample Size for Developing a Multivariable Prediction Model: PART II - Binary and Time-to-Event Outcomes.” *Statistics in Medicine* 38 (7): 1276–96. https://doi.org/10.1002/sim.7992.

Schoenfeld, David. 1982. “Partial Residuals for the Proportional Hazards Regression Model.” *Biometrika* 69 (1): 239–41. https://doi.org/10.1093/biomet/69.1.239.

Shapiro, S. S., and M. B. Wilk. 1965. “An Analysis of Variance Test for Normality (Complete Samples).” *Biometrika* 52 (3/4): 591. https://doi.org/10.2307/2333709.

Simon, Noah, Jerome Friedman, Trevor Hastie, and Rob Tibshirani. 2011. “Regularization Paths for Cox’s Proportional Hazards Model via Coordinate Descent.” *Journal of Statistical Software* 39 (5): 1–13. https://doi.org/10.18637/jss.v039.i05.

Steck, H, B Krishnapuram, and C Dehing-Oberije. 2008. “On Ranking in Survival Analysis: Bounds on the Concordance Index.” In Platt JC, Koller D, Singer Y, Roweis ST (eds): Advances in Neural Information Processing Systems 20, Curran Associates, Inc.

Stekhoven, D. J., and P. Buhlmann. 2012. “MissForest--Non-Parametric Missing Value Imputation for Mixed-Type Data.” *Bioinformatics* 28 (1): 112–18. https://doi.org/10.1093/bioinformatics/btr597.

Therneau, T. M., P. M. Grambsch, and T. R. Fleming. 1990. “Martingale-Based Residuals for Survival Models.” *Biometrika* 77 (1): 147–60. https://doi.org/10.1093/biomet/77.1.147.

Therneau, TM, and PM Grambsch. 2000. *Modeling Survival Data: Extending the Cox Model*. New York: Springer-Verlag.

“VII. Note on Regression and Inheritance in the Case of Two Parents.” 1895. *Proceedings of the Royal Society of London* 58 (347–352): 240–42. https://doi.org/10.1098/rspl.1895.0041.