Uncovering clinical risk factors and prediction of severe COVID-19: A machine learning approach based on UK Biobank data

Kenneth C.Y. Wong[1], Hon-Cheong SO[1-7^]

[1]School of Biomedical Sciences, The Chinese University of Hong Kong, Shatin, Hong Kong

[2]CUHK Shenzhen Research Institute, Shenzhen, China

[3]KIZ-CUHK Joint Laboratory of Bioresources and Molecular Research of Common Diseases, Kunming Institute of Zoology and The Chinese University of Hong Kong, China

[4]Department of Psychiatry, The Chinese University of Hong Kong, Shatin, Hong Kong

[5]Margaret K.L. Cheung Research Centre for Management of Parkinsonism, The Chinese University of Hong Kong, Shatin, Hong Kong

[6]Brain and Mind Institute, The Chinese University of Hong Kong, Shatin, Hong Kong

7Hong Kong Branch of the Chinese Academy of Sciences Center for Excellence in Animal Evolution and Genetics, The Chinese University of Hong Kong, Hong Kong SAR, China

^*Corresponding author*

Correspondence to: Hon-Cheong So, Lo Kwee-Seong Integrated Biomedical Sciences Building, The Chinese University of Hong Kong, Shatin, Hong Kong. Tel: +852 3943 9255; E-mail: hcso@cuhk.edu.hk

Submitted on 18 Sep 2020

Abstract

Background: COVID-19 is a major public health concern. Given the extent of the pandemic, it is urgent to identify risk factors associated with severe disease. Accurate prediction of those at risk of developing severe infections is also important clinically.

Methods: Based on the UK Biobank (UKBB data), we built machine learning(ML) models to predict the risk of developing severe or fatal infections, and to evaluate the major risk factors involved. We first restricted the analysis to infected subjects, then performed analysis at a population level, considering those with no known infections as controls. Hospitalization was used as a proxy for severity. Totally 93 clinical variables (collected prior to the COVID-19 outbreak) covering demographic variables, comorbidities, blood measurements (e.g. hematological/liver and renal function/metabolic parameters etc.), anthropometric measures and other risk factors (e.g. smoking/drinking habits) were included as predictors. XGboost (gradient boosted trees) was used for prediction and predictive performance was assessed by cross-validation. Variable importance was quantified by Shapley values and accuracy gain. Shapley dependency and interaction plots were used to evaluate the pattern of relationship between risk factors and outcomes.

Results: A total of 1191 severe and 358 fatal cases were identified. For the analysis among infected individuals (N=1747), our prediction model achieved AUCs of 0.668 and 0.712 for severe and fatal infections respectively. Since only pre-diagnostic clinical data were available, the main objective of this analysis was to identify baseline risk factors. The top five contributing factors for severity were age, waist-hip ratio(WHR), HbA1c, number of drugs taken(cnt_tx) and gamma-glutamyl transferase levels. For prediction of mortality, the top features were age, systolic blood pressure, waist circumference (WC), urea and WHR.

In subsequent analyses involving the whole UKBB population (*N* for controls=489987), the corresponding AUCs for severity and fatality were 0.669 and 0.749. The same top five risk factors were identified for both outcomes, namely age, cnt_tx, WC, WHR and cystatin C. We also uncovered other features of potential relevance, including testosterone, IGF-1 levels, red cell distribution width (RDW) and lymphocyte percentage.

Conclusions: We identified a number of baseline clinical risk factors for severe/fatal infection by an ML approach. For example, age, central obesity, impaired renal function, multi-comorbidities and cardiometabolic abnormalities may predispose to poorer outcomes. The presented prediction models may be useful at a population level to help identify those susceptible to developing severe/fatal infections, hence facilitating targeted prevention strategies. Further replications in independent cohorts are required to verify our findings.

Introduction

Coronavirus Disease 2019 (COVID-19) has resulted in a pandemic affecting more than a hundred countries worldwide [1-3]. More than 30 million confirmed cases and 900,000 fatalities have been reported worldwide as at 17th Sep 2020 (https://coronavirus.jhu.edu/map.html), while a large number of mild or asymptomatic cases may remain undetected. Given the extent of the pandemic, it is urgent to identify risk factors that may be associated with severe disease, and to gain deeper understanding its pathophysiology. Accurate prediction of those at risk of developing severe infections is also clinically very important.

Machine learning (ML) approaches are powerful tools to predict disease outcomes and have been increasingly applied in biomedical research. An important advantage is that ML methods can capture complex, non-linear and even interactions between variables, hence leading to better predictive power in many circumstances. In view of the COVID-19 pandemic, many ML models have been developed for diagnostic or prognostic purposes. A recent review nicely summarized many of these models[4].

Here we made of the UK Biobank (UKBB) data to build ML models to predict the severity and fatality from COVID-19, and evaluate the contributing risk factors. While predictive performance is the main concern in most previous studies, we argue that ML models can also provide important insight into individual contributing factors, and the pattern of complex relationship between risk factors and the outcome, which may not be captured by conventional linear models. We note that in the UKBB, clinical data were collected years before the outbreak of infection in 2020, which is a limitation. Ideally, the predictors should be measured at the time when the model is intended to be applied (e.g. at admission). However, we believe building ML models with previously collected clinical data is useful for several reasons.

First of all, this approach may facilitate the identification of potential causal risk factors. As the predictors are collected prior to the outbreak, there is no concern of reverse causality. In practice, infection itself will lead to changes in many clinical parameters (e.g. glucose, inflammatory markers, liver/renal functions etc.); hence it is often difficult to tell the direction of effect in cross-sectional observational studies. While many have studied risk factors on COVID-19 susceptibility or severity in the UKBB[5-7] or other cohorts (e.g. see[4,8-11]), most relied on conventional linear models. As such, non-linear effects and interactions between variables may be missed. We hypothesize that this study will identify general or 'baseline' risk factors or laboratory measurements that may be predictive of outcome.

Secondly, the UKBB is a huge population-based sample (*N* close to 500,000) that enables ML models to be developed *at the general population level*. For example, at a population level, who may be more susceptible to developing severe or fatal infections? This may have implications for prioritizing individuals for specific prevention (e.g. vaccination) strategies and diagnostic testing under limited resources. Importantly, there is a lack of such population-level prediction models, and this study may fill the gap.

In this study we performed four sets of analysis. In the first two sets of analysis, we built ML models to predict severity and mortality of COVID-19 within those who are tested positive for the virus. In the other two analyses, we aimed to predict severity and mortality of COVID-19 at the population level, considering subjects not known to be infected as 'controls'. We employed XGboost, a state-of-the-art ML tool for prediction, and identified how different risk factors and their interactions impact on disease severity.

Methods

UK Biobank data

For details of UK Biobank data please refer to ref[12]. The UK Biobank is a large-scale prospective cohort comprising over 500,000 subjects aged 40–69 years when they were recruited in 2006–2010. The present analysis was conducted under project number 28732.

COVID-19 phenotypes

COVID-19 outcome data were downloaded from data portal provided by the UKBB. Details of data release are provided at http://biobank.ndph.ox.ac.uk/showcase/exinfo.cgi?src=COVID19. Briefly, COVID test data was extracted on 25 Jul 2020, which covers test specimens taken from 16 March 2020 to 19 July 2020. The data also included an indicator on whether the patient was an inpatient when the specimen was taken. We consider inpatient (hospitalization) status as a proxy for severity, as more sophisticated indicators of severity were not available. Data on mortality and cause of mortality were also extracted (as at 31 Jul 2020). Death cause indicated by U07.1 was considered to be a fatal case with laboratory-confirmed COVID-19. Please also refer to http://biobank.ndph.ox.ac.uk/showcase/exinfo.cgi?src=COVID19_tests on relevant details.

We considered a case as 'severe COVID-19' if the subject is an inpatient and/or if the cause of mortality is U07.1. For a very small number of subjects (*N*=21) whose cause of mortality was U07.1 but test result(s) was negative, they were excluded from subsequent analysis.

Sets of analysis

Four sets of analysis were performed. The first two sets were restricted to test-positive cases (total N=1747), with 'severe COVID-19' (N=1191) and death (N=358) due to COVID-19 as outcomes. Since only pre-diagnostic clinical data were available, the main objective of this analysis was to identify baseline risk factors for severe/fatal illness among the infected. We then performed another two sets of analysis with the same outcomes, but the 'unaffected' group was composed of the general population (N=489987) who were *not* diagnosed to have COVID-19 or tested negative. The four sets of analysis were also referred to as cohorts A to D as shown in Table 1.

Data analysis

We extracted a total of 93 clinical variables of potential relevance. Among these variables, 15 were categorical and 78 were quantitative traits. The missing rates were <20%. We chose a relatively large set of predictors as the ML model (XGboost) is able to deal with a large number of predictors without overfitting. Another reason is that this study also aims to explore a wider variety of potential (new) risk factors for the disease. The full list of variables being analyzed is shown in Table S1. Briefly, we included basic demographic variables (e.g. age at recruitment, sex, ethnic group), comorbidities (e.g. heart diseases, type 2 diabetes, hypertension, asthma/COPD, cancer etc.), indicators of general health (number of medications taken, number of illnesses etc.), blood measurements (hematological measures, liver and renal function measures, metabolic parameters such as lipid levels, HbA1c etc.), anthropometric measures (e.g. waist circumference, waist-hip ratio, body fat percentage, body mass index[BMI] etc.) and lifestyle risk factors (e.g. smoking, drinking habits etc.).

Missing values of remaining features were imputed by two different approaches according to their data type. For categorical features, missing values were replaced by UK Biobank data coding representing "Prefer not to answer". Following Twala et al.[13], missing values were treated as a separate category, which has been shown to be a reasonably good approach for decision-tree models. One-hot encoding was then applied to categorical features. For the continuous features, missing values were replaced by the column mean (more complex methods are computationally expensive in view of the large sample size).

XGboost prediction model

XGboost with gradient-boosted trees was employed for building prediction models. Analysis was performed by the R package 'xgboost'. We employed a 5-fold (nested) cross-validation strategy to develop and test the model. To avoid overoptimistic results due to choosing the best set of hyper-parameters based on test performance, the test sets were *not* involved in the tuning of hyper-parameters.

In each iteration, we divided the data into 5 folds, among which 1/5 was reserved for testing only. For the remaining 4/5 of the data, we further sampled 4/5 for training and 1/5 for hyper-parameter tuning. The best prediction model was applied to the test set. The process was repeated five times. A grid-search procedure was used to search for the best combination of hyper-parameters (e.g. tree depth, learning rate, regularization parameters for L1/L2 penalty etc.). The full range of hyper-parameters chosen or for grid-search is given in supplementary Table S2.

Identifying and quantifying the effects of important predictors

Identification the main contributing factors to severe/fatal infection is crucial to understanding of the model and uncovering new risk factors potentially linked to the severity of infection.

We employed three sets of indices to assess variable importance. The first index is known as 'gain'. Intuitively, it is the improvement in prediction accuracy as a result of splitting based on the studied feature. The second index is the Shapley value (ShapVal)[14,15], which is a measure based on game theory to assess the contribution of each feature. Shapley values also enable local explanation of the model as they could be

computed for each observation. Intuitively, the Shapley value of the *i*-th feature (for subject *k*) is the contribution of this feature to prediction of outcome for the subject considers all possible orderings of the features as the contribution may differ when variables enter the prediction algorithm in different orders. The third set of index is Shapley *interaction* value[15], which computes the difference in Shapley value of feature *i* with and without another feature *j*. The gain and Shapley values were averaged across 5 folds.

Results

An overview of the sample sizes in each set of analysis is presented in Table 1.

Prediction performance

We performed 5-fold CV and the average AUC under the ROC curve is given in Table 1. We observed better predictive performances in cohort B (fatal cases vs outpatient cases) and D (fatal cases vs population with no known infection) when fatalities from COVID-19 were modeled. The corresponding mean AUC were 0.712 and 0.749 respectively. The mean AUC for cohort A (hospitalized/fatal cases vs mild cases) was 0.668 and AUC for cohort C (hospitalized/fatal cases vs population with no known infection) was 0.669.

Important contributing variables identified

The Shapley dependence plots (ranked by mean absolute ShapVal) and variable importance plots (ranked by gain) of the top 15 features are shown in Figures 1-12.

Cohort A (hospitalized/fatal cases vs outpatient cases)

For this set of analysis, the top 5 contributing features by ShapVal included age at recruitment, waist-hip ratio (WHR), HbA1c, number of medications received and gamma-glutamyl transferase (GGT). Higher levels of these risk factors generally lead to higher severity of disease among the infected. Interestingly, Shapley dependence plots revealed potential *non-linear* effects of risk factors on the outcome. For example, age at recruitment of 50 (age at diagnosis ~60-64) or above was associated with a marked increased risk of severe/fatal infection. (As recruitment was performed in 2006-2010, the age at diagnosis is ~10-14 years added to the recruitment age). For waist-hip ratio, levels of ~0.9 or higher appeared to be associated with marked increase in risks. Elevated risks were also observed true for HbA1c > ~40 mmol/mol and number of drugs received >= ~5. Impaired renal function (raised cystatin C and urea) were also linked to worse outcomes. For the effect of other features please also refer to Figure 1. We note that at more extreme levels of the variables, the observations are sparse so the trend (e.g. decreased risk) shown by the loess curve may not be reliable (this also applies to other cohorts). Variable importance based on gain revealed similar patterns of important features (Figure 2).

With regards to interaction between variable, most the top pairs of interacting variable involves age. For example, GGT interacts with age in affecting the risk of severe disease. As shown in Figure 3, younger individuals were observed to have more extreme ShapVal at similar ranges of GGT, suggesting the effect of

GGT on disease severity is more marked among younger subjects. The effects of HbA1c or WHR also seemed to be larger among younger subjects, but the reverse was observed for testosterone.

Cohort B (fatal cases vs outpatient cases)

The results are presented in Figures 4-6. The top 5 contributing variables by ShapVal included age, systolic blood pressure (SBP), waist circumference (WC), urea and WHR. Again certain non-linear 'threshold' effects appeared to be present for many top-ranked features. For age at recruitment, the risk for mortality was more marked beyond ~50, but unlike the case for cohort A, the risk continues to rise beyond this age up to ~70, indicating continuously escalating mortality with age among the elderly. SBP greater than ~140 mmHg was also associated with higher risks, with another rise at ~160 mmHg. Similarly, higher WC, WHR and urea were associated with elevated risk of mortality, but the effects were non-linear and showed threshold effects. Other possible associations included increased red cell distribution width (RDW), reduced percentage of lymphocytes and reduced IGF-1 levels with increased risk of mortality. Variable importance based on gain yielded similar results, but glucose level was also ranked among the top 5. As for interaction between the variables (Figure 6), the impact of waist circumference on the outcome was more prominent among younger individuals. On the other hand, the impact of SBP appeared more marked among the elderly.

Cohort C (hospitalized/fatal cases vs *population with no known infection)*

The results are presented in Figures 7-9. Based on ShapVal, age was the top contributing factor, similar to before. Among the top 10 variables, 3 were related to obesity (WC, WHR and BMI) but WC/WHR were ranked higher, suggesting central obesity may be a stronger predictor for severe disease than BMI alone. Two are related to multi-comorbidities in general (cnt_treatment/cnt_noncancer). Increased cystatin C and HbA1c as well as smoking status were also associated with higher susceptibility to severe infections. Smoking status was not ranked among the top in our previous analyses (A and B) that are restricted to infected individuals. The variable importance plot based on accuracy gain is shown in Figure 8. The top variables in general agree with the rankings from ShapVal. Compared to cohorts A or B, while age was still the most important contributing factor, the distance between age and other risk factors was less marked. Interaction plot (Figure 9) shows WHR and WC may interact with age, with elderly individuals showing more prominent effects from changes in WHR/WC.

Cohort D (fatal cases vs *population with no known infection)*

Relevant results are shown in Figure 10-12. Age was the top feature, followed by WC, number of drugs taken, WHR and cystatin C based on ShapVal. The top five features were in fact identical to those identified in cohort C. Other top features included testosterone, sex (male), mean platelet volume, RDW and total protein level. The features ranked by gain are shown in Figure 11. There top 5 features ranked by ShapVal were among the top 7 ranked by gain, indicating high consistency. Note that the genetic principal components were not directly interpretable, but were included here to adjust for possible confounding due to population substructures or more subtle ethnic differences. The other top-ranked features by gain included lymphocyte count, IGF-1 and HbA1c.

Shapley interaction analysis suggested the top interacting pairs involved age and some of top contributing features.

Discussions

In this study we performed four sets of analysis, predicting severe or fatal COVID-19 cases among the affected individuals or in the population. We also identified risk factors for increased severity or mortality from infection, which may shed light on disease mechanisms.

Prediction of severity/mortality

Regarding the predictive performance, the models predicted mortality (AUC ~71% to 75%) better than severity of disease. As discussed earlier, in the absence of better alternatives, hospitalization (test performed as inpatient) was used as a proxy for severity. However, reasons or criteria for hospitalization may vary across individuals or different hospitals, and some tests may be performed in in-patients for surveillance or due to other confirmed/suspected cases in the ward. On the other hand, mortality from infection is a more objective outcome.

A number of studies focused on prediction of severity/mortality of disease (corresponding to our prediction in cohorts A and B) were available and reviewed in ref [4]. For model A (prediction of severity among infected), the AUC is ~67%, which is moderate but not as good as many previous ML models for severity prediction[4]. The AUC for prediction of mortality is higher (~71%), but we noted many studies have reported much higher predictive power from clinical symptoms, blood biochemistry on admission and imaging features[4]. We understand that without access to the above features, predictive performance may be inferior. Here we are not aimed at deriving a highly accurate prediction model; the main purpose is to identify general or 'baseline' risk factors for severe disease, thereby gaining insight into disease pathophysiology. However, we also showed that such clinical features or blood measurements, even when collected much earlier in time, may still be predictive of outcomes and hence may be incorporated into existing prediction algorithms.

For cohorts C and D, the general population (with no known infection) was also included. Compared to cohorts A and B, the identified risk factors may also increase the overall susceptibility to infection. The AUC for cohort C (severe/fatal disease) is ~67% but is much higher when mortality is considered as the outcome (AUC~75%). To our knowledge, there are very few predictive models built at a *general population level* to identify susceptible individuals. DeCaprio et al.[16] proposed an ML model to assess the vulnerability to COVID-19 in the population. However, due to limited data, no actual COVID-19 patients were included and 'proxy' outcomes were used instead. Models were built from mainly demographic and comorbidity data to predict hospitalization due to acute respiratory distress syndrome, pneumonia, influenza, acute bronchitis and other respiratory tract infections. In view of limitations of previous works and lack of models applicable to the population, our prediction models may be practically useful. If further validated, the presented models may help prioritize individuals for specific prevention strategies (e.g. vaccination) or viral testing (e.g. in a screening program).

A few other studies have studied risk factors (especially comorbidities) for COVID-19 infection in the UKBB. For example, Atkins et al. [5] studied elderly subjects (age>65) in UKBB, and found that hypertension, history of falls, coronary heart disease and type 2 diabetes and asthma as the top comorbidities among those hospitalized cases. The analysis was restricted to the elderly population however. In a more recent work, McQueenie et al. [6] studied multi-comorbidity and polypharmacy on the risk of developing the disease. The main risk factors were having >=2 long-term conditions, cardiometabolic disorders and polypharmacy were associated with heightened risk of infection. Among individuals with multi-comorbidities, severe obesity and impaired renal function may lead to increased risks. Another study of primary care patients in the UK revealed deprivation, males, older age, ethnicity (being black) and chronic renal disease were associated with higher risks of being tested positive. Another large-scale British primary care study of more than 17 million subjects revealed similar risk factors as above [17]. There is also a relatively large literature on the study of risk factors associated with severe or fatal disease[8-11,18-21]. Some commonly reported risk factors included age, sex, obesity, diabetes, hypertension, cardiometabolic and respiratory disorders.

The main difference between the present work and previous epidemiological studies is that we employed a machine learning approach which is able to capture also non-linear and more complex interactive effects. As shown in our Shapley dependence plots, the models were able to reveal non-linear effects in a data-driven manner. We also included a larger number of blood measurements to shed light on potential new risk factors and the mechanisms underlying the disease.

Highlights of potential risk factors

For limit of space, we shall only highlight the top five to ten risk factor ranked by Shapley values here. Across the four cohorts, age and cardiometabolic risk factors predominate the top risk factors. For the population-based cohorts (C and D), the same five risk factors (age, number of medications received, waist circumference, WHR and cystatin C) were top-ranked. In fact, age and waist circumference was ranked among top 5 across all four cohorts, while number of medications of taken was in top 5 across three cohorts. Of note, cystatin C or urea (reflecting renal function) was among the top 5 across three cohorts, and top 10 in all cohorts. HbA1c is a top-10 risk factor across two cohorts.

Notably, obesity has been observed to be a major risk factor for susceptibility or severity of infection in the UK Biobank [7,22], and in many other studies[23,24]. The observation that waist circumference/WHR were highly ranked suggests that *central* obesity is a major risk factor and may be a better predictor of severity than BMI alone.

Among other comorbidities, another major risk factor we identified was impaired renal function (IRF), as reflected by elevated risks with raised urea and cystatin C. Several studies also suggested IRF increases risk of mortality [21,25,26], although it is probably not as widely recognized as cardiometabolic disorders as a main risk factor. Since COVID-19 itself may lead to renal failure, our findings specifically suggest that *underlying* or baseline IRF is an important risk factor. The high ranking of cystatin C also indicates this measure may better

reflect renal function than urea or creatinine (which were also included in our analysis) [27,28], and may serve as a superior predictor for COVID-19 severity. HbA1c reflects glycemic control, and it is unsurprising that it showed up as a risk factor for severity or mortality. Diabetes has been shown to raise the risk and severity of infection[29,30].

Other potential risk factors briefly highlighted below were less reported; as most were listed only once or twice among the top 10 list, and their Shapley values were close to other risk factors, further replications are required. For example, testosterone was top-ranked by XGboost, with higher levels associated with increased risk. Studies have suggested elevated or reduced testosterone levels may be both associated with a more severe clinical course [31]. Higher IGF-1 appeared to be associated with greater mortality in cohort B. A recent preprint using simpler logistic models (adjusted for age, sex, ethnicity and lifestyle factors) and a slightly smaller sample from UKBB also revealed increased mortality among those with high IGF-1, but here we employed a more complex ML model with adjustment for a larger range of risk factors and comorbidities. We also found a few hematological indices that may be potential risk factors. High red cell distribution width (RDW) was associated with mortality in our study and was also identified in a recent meta-analysis of three studies as a risk factor [32]. Low lymphocyte percentage was a top-10 risk factor in cohort B, which may be related to immune functioning and response to infections. Lymphopenia has been reported as a main hematological finding in those with severe illness [33,34]. Most previous studies considered hematological indices at admission or during hospitalization. Slightly surprisingly, this study suggested that high RDW or reduced lymphocyte percentage *prior to the diagnosis* may also be predictive of worse outcomes.

Some of other risk factors reported in other studies were not highly ranked or identified in this study. For example, smoking, diagnosis of coronary artery disease (CAD) or type 2 diabetes (DM), cancers, asthma/COPD were not consistently highly ranked by ShapVal or gain. One possibility is that other associated risk factors (e.g. obesity, IRF, high BP, raised HbA1c, multi-comorbidities as reflected by polypharmacy) may have largely accounted for the risks conferred by the above diagnoses. Also, it is possible that some patients only develop the above diseases after the last assessment and some 'misclassification' may be present, although this may not be only reason as other measurements are also subject to a similar kind of bias.

Limitations

Some limitations have been discussed above, for example the use of hospitalization as a proxy for severity, and that the predictors were recorded prior to the pandemic. We briefly discuss other limitations here. The UK biobank is a very large-scale study with detailed phenotypic data, but still the number of COVID-19 cases is not large (N ~1700), with a smaller number of fatal cases. Also, UKBB is not entirely representative of the UK population, as participants tend to be healthier and wealthier overall[35]. Also, it remains to be studied whether the findings are generalizable to other populations. Symptom measures and lung imaging features were not available. Despite adjusting for a rich set of predictors and that all were recorded prior to the outbreak, causality cannot be confirmed from the current study, as there is risk of residual confounding by unknown factors. In

cohorts C and D, the population with no known infection was regarded as controls. It is expected that some may become infected in the future, and some may have been infected but not tested; however, the chance of missing cases of severe infection is probably not high. Since the UKBB represents a relatively healthy population with low rate of severe COVID-19 cases so far (~0.2%), we expect the use of 'unscreened' controls is unlikely to result in substantial bias.

Regarding the ML model, XGboost is a state-of-the-art method that has been consistently shown to the best or one of the best ML methods in supervised learning tasks/competitions [36] (especially for tasks not involving computer vision or natural language processing). Nevertheless, other ML methods may still be useful or may uncover novel risk factors. Assessing variable importance is a long-standing problem in ML; here we mainly employed Shapley values which is both computationally fast and was shown to have good theoretical properties[14,15].

Conclusions

In conclusion, we identified a number of baseline risk factors for severe/fatal infection by an ML approach. The Shapley plots revealed possible non-linear and 'threshold' effects on risks of severity. To summarize, age, central obesity, impaired renal function, multi-comorbidities, cardiometabolic abnormalities may predispose to poor outcomes, among other risk factors. The prediction models (for cohorts C/D) may be useful at a population level to identify those susceptible to developing severe/fatal infections, facilitating targeted prevention strategies. Further replication and validation in independent cohorts are required to confirm our findings.

Acknowledgements

This work was supported partially by the Lo Kwee Seong Biomedical Research Fund from The Chinese University of Hong Kong. We thank Prof. Pak Sham for support on data access and analyses.

Supplementary Tables are available at

https://drive.google.com/file/d/1LC0wKiguslS0dpnR3Pr8rqaEE4OD_JTe/view?usp=sharing

Conflicts of interest

The authors declare no conflict of interest.

Table 1 The four sets of analysis performed and predictive performances

Model	Group 1	Group 2	*N* (Group 1)	*N* (Group 2)	AUC (%)
A	Hospitalized or fatal cases	Non-hospitalized cases	1191	556	66.8
B	Fatal cases	All other COVID-19 cases	358	1389	71.2
C	Hospitalized or fatal cases	UKBB subjects without a COVID-19 Dx or tested -ve	1191	489987	66.9
D	Fatal cases	UKBB subjects without a COVID-19 Dx or tested -ve	358	489987	74.9

Dx, diagnosis. *N*, sample size. AUC was taken from the average from 5 folds of cross-validation.

References

1. Li, Q. *et al.* Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia. *N Engl J Med* (2020).
2. Novel-Coronavirus-Pneumonia-Emergency-Response-Epidemiology-Team. [The epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (COVID-19) in China]. *Zhonghua Liu Xing Bing Xue Za Zhi* **41**, 145-151 (2020).
3. Guan, W.-j. *et al.* Clinical Characteristics of Coronavirus Disease 2019 in China. *New England Journal of Medicine* (2020).
4. Wynants, L. *et al.* Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *BMJ* **369**, m1328 (2020).
5. Atkins, J.L. *et al.* Preexisting Comorbidities Predicting COVID-19 and Mortality in the UK Biobank Community Cohort. *The Journals of Gerontology: Series A* (2020).
6. McQueenie, R. *et al.* Multimorbidity, polypharmacy, and COVID-19 infection within the UK Biobank cohort. *PLOS ONE* **15**, e0238091 (2020).
7. Yates, T., Razieh, C., Zaccardi, F., Davies, M.J. & Khunti, K. Obesity and risk of COVID-19: analysis of UK biobank. *Primary care diabetes* **14**, 566-567 (2020).
8. Rod, J.E., Oviedo-Trespalacios, O. & Cortes-Ramirez, J. A brief-review of the risk factors for covid-19 severity. *Rev Saude Publica* **54**, 60 (2020).
9. Romero Starke, K. *et al.* The Age-Related Risk of Severe Outcomes Due to COVID-19 Infection: A Rapid Review, Meta-Analysis, and Meta-Regression. *Int J Environ Res Public Health* **17**(2020).
10. Wingert, A. *et al.* Risk factors for severe outcomes of COVID-19: a rapid review. *medRxiv*, 2020.08.27.20183434 (2020).
11. Wolff, D., Nee, S., Hickey, N.S. & Marschollek, M. Risk factors for Covid-19 severity and fatality: a structured literature review. *Infection* (2020).

12. Sudlow, C. *et al.* UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Medicine* **12**, e1001779 (2015).
13. Twala, B., Jones, M. & Hand, D.J. Good methods for coping with missing data in decision trees. *Pattern Recognition Letters* **29**, 950-956 (2008).
14. Lundberg, S.M. & Lee, S.-I. A unified approach to interpreting model predictions. in *Advances in neural information processing systems* 4765-4774 (2017).
15. Lundberg, S.M. *et al.* From Local Explanations to Global Understanding with Explainable AI for Trees. *Nat Mach Intell* **2**, 56-67 (2020).
16. DeCaprio, D. *et al.* Building a COVID-19 Vulnerability Index. *medRxiv*, 2020.03.16.20036723 (2020).
17. Williamson, E.J. *et al.* Factors associated with COVID-19-related death using OpenSAFELY. *Nature* **584**, 430-436 (2020).
18. Noor, F.M. & Islam, M.M. Prevalence and Associated Risk Factors of Mortality Among COVID-19 Patients: A Meta-Analysis. *J Community Health* (2020).
19. Rahman, A. & Sathi, N.J. Risk Factors of the Severity of COVID-19: a Meta-Analysis. *medRxiv*, 2020.04.30.20086744 (2020).
20. Zhou, Y., Chi, J., Lv, W. & Wang, Y. Obesity and diabetes as high-risk factors for severe coronavirus disease 2019 (Covid-19). *Diabetes/Metabolism Research and Reviews* **n/a**, e3377 (2020).
21. Harrison, S.L., Fazio-Eynullayeva, E., Lane, D.A., Underhill, P. & Lip, G.Y.H. Comorbidities associated with mortality in 31,461 adults with COVID-19 in the United States: A federated electronic medical record analysis. *PLoS Med* **17**, e1003321 (2020).
22. Hamer, M., Gale, C.R., Kivimäki, M. & Batty, G.D. Overweight, obesity, and risk of hospitalization for COVID-19: A community-based cohort study of adults in the United Kingdom. *Proceedings of the National Academy of Sciences* **117**, 21011-21013 (2020).
23. Popkin, B.M. *et al.* Individuals with obesity and COVID-19: A global perspective on the epidemiology and biological relationships. *Obes Rev* (2020).
24. Tamara, A. & Tahapary, D.L. Obesity as a predictor for a poor prognosis of COVID-19: A systematic review. *Diabetes Metab Syndr* **14**, 655-659 (2020).
25. Di Castelnuovo, A. *et al.* Common cardiovascular risk factors and in-hospital mortality in 3,894 patients with COVID-19: survival analysis and machine learning-based findings from the multicentre Italian CORIST Study. *Nutr Metab Cardiovasc Dis* (2020).
26. Uribarri, A. *et al.* Impact of renal function on admission in COVID-19 patients: an analysis of the international HOPE COVID-19 (Health Outcome Predictive Evaluation for COVID 19) Registry. *J Nephrol* **33**, 737-745 (2020).
27. Hojs, R., Bevc, S., Ekart, R., Gorenjak, M. & Puklavec, L. Serum cystatin C as an endogenous marker of renal function in patients with mild to moderate impairment of kidney function. *Nephrology Dialysis Transplantation* **21**, 1855-1862 (2006).
28. Shlipak, M.G., Mattes, M.D. & Peralta, C.A. Update on Cystatin C: Incorporation Into Clinical Practice. *American Journal of Kidney Diseases* **62**, 595-603 (2013).

29. Gupta, R., Hussain, A. & Misra, A. Diabetes and COVID-19: evidence, current status and unanswered research questions. *Eur J Clin Nutr* **74**, 864-870 (2020).
30. Apicella, M. *et al.* COVID-19 in people with diabetes: understanding the reasons for worse outcomes. *Lancet Diabetes Endocrinol* **8**, 782-792 (2020).
31. Giagulli, V.A. *et al.* Worse progression of COVID-19 in men: Is testosterone a key factor? *Andrology*, 10.1111/andr.12836 (2020).
32. Lippi, G., Henry, B.M. & Sanchis-Gomar, F. Red Blood Cell Distribution Is a Significant Predictor of Severe Illness in Coronavirus Disease 2019. *Acta Haematologica* (2020).
33. Terpos, E. *et al.* Hematological findings and complications of COVID-19. *American Journal of Hematology* **95**, 834-847 (2020).
34. Tan, L. *et al.* Lymphopenia predicts disease severity of COVID-19: a descriptive and predictive study. *Signal Transduction and Targeted Therapy* **5**, 33 (2020).
35. Fry, A. *et al.* Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *American Journal of Epidemiology* **186**, 1026-1034 (2017).
36. Nielsen, D. Tree Boosting With XGBoost. *Master thesis, Norwegian University of Science and Technology* (2016).

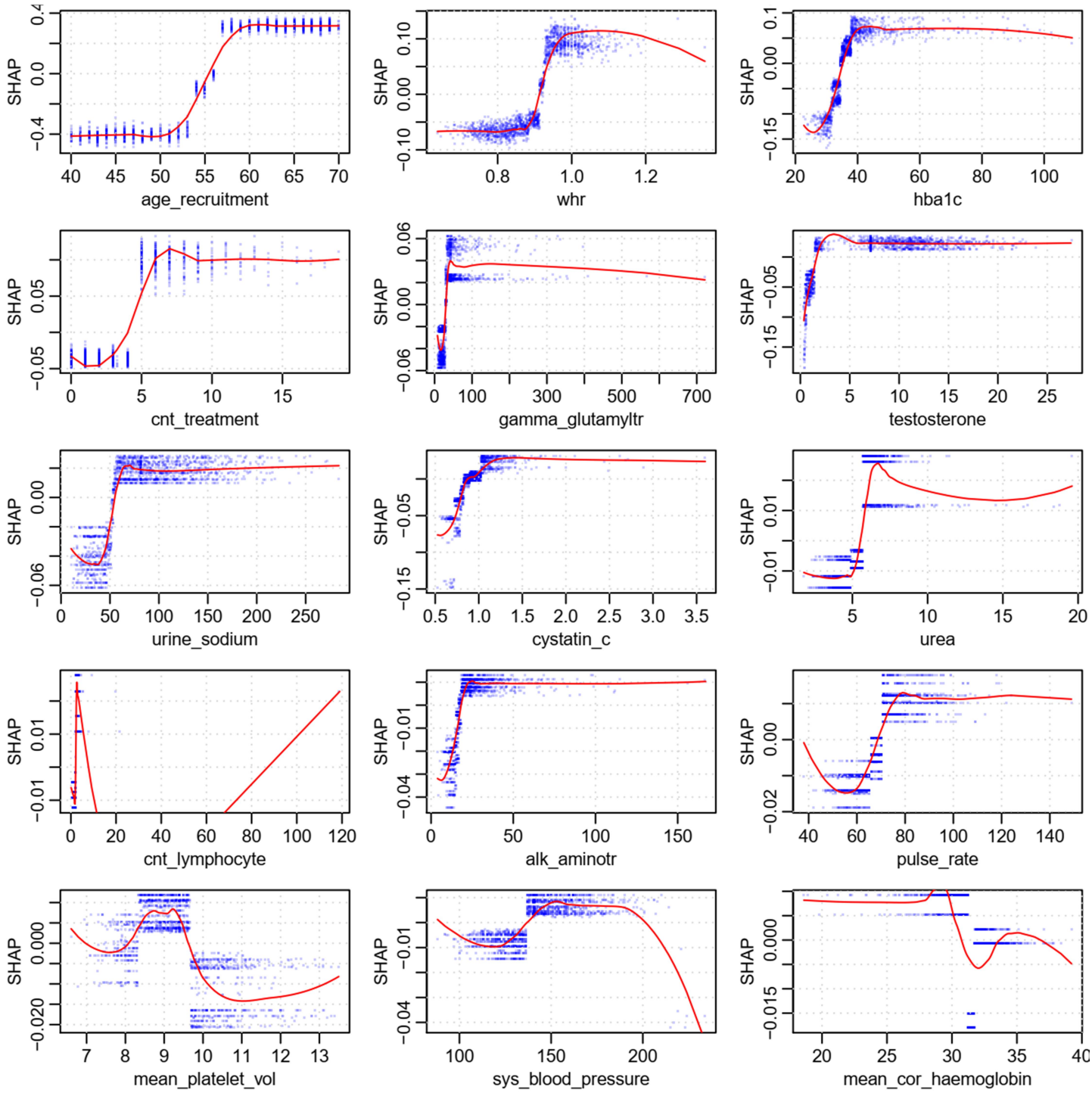

Figure 1 Shapley dependence plot for cohort A (hospitalized/fatal cases vs outpatient cases)

Features are ranked by mean absolute ShapVal (the 1st row corresponds to features ranked 1st, 2nd and 3rd; the 2nd row corresponds to features ranked 4th, 5th and 6th and so on).

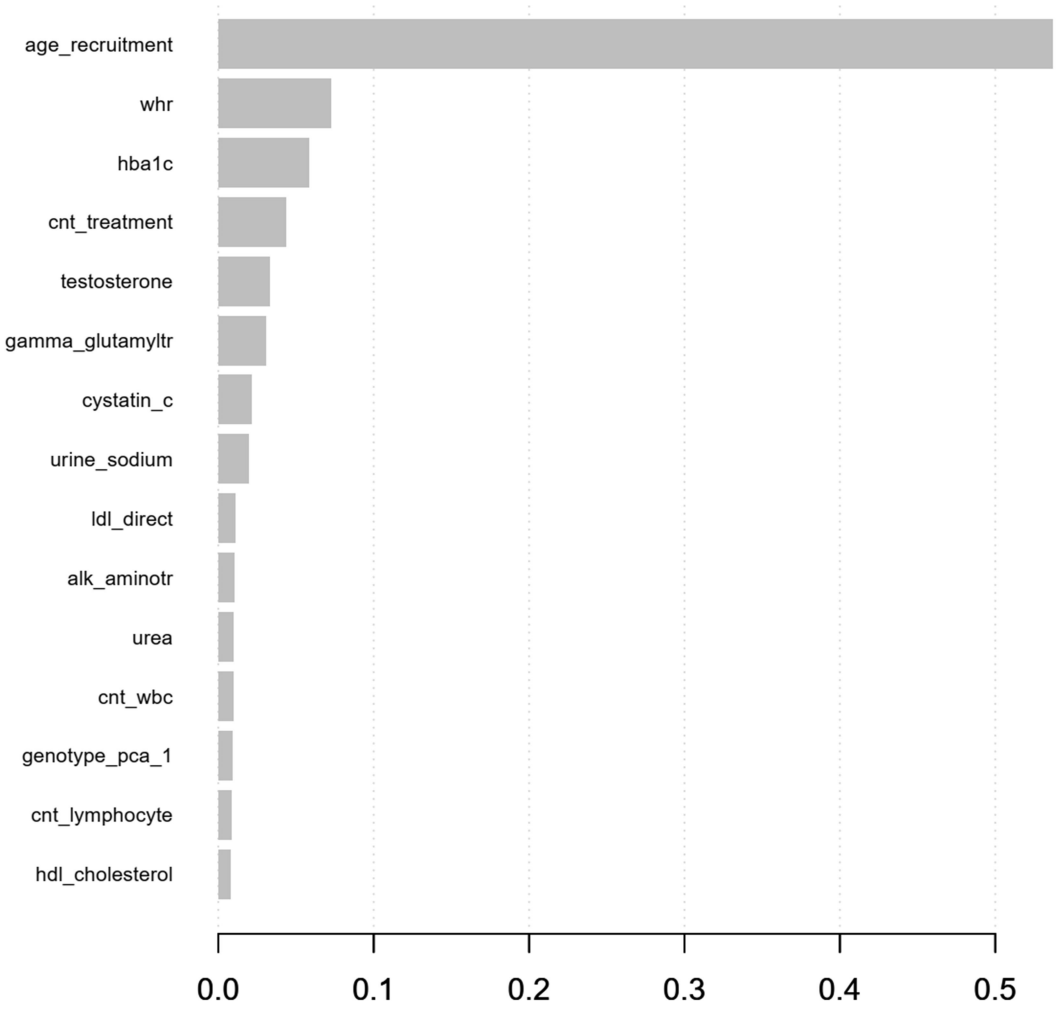

Figure 2 Variable importance ranked by gain for cohort A (top 15 variables shown)

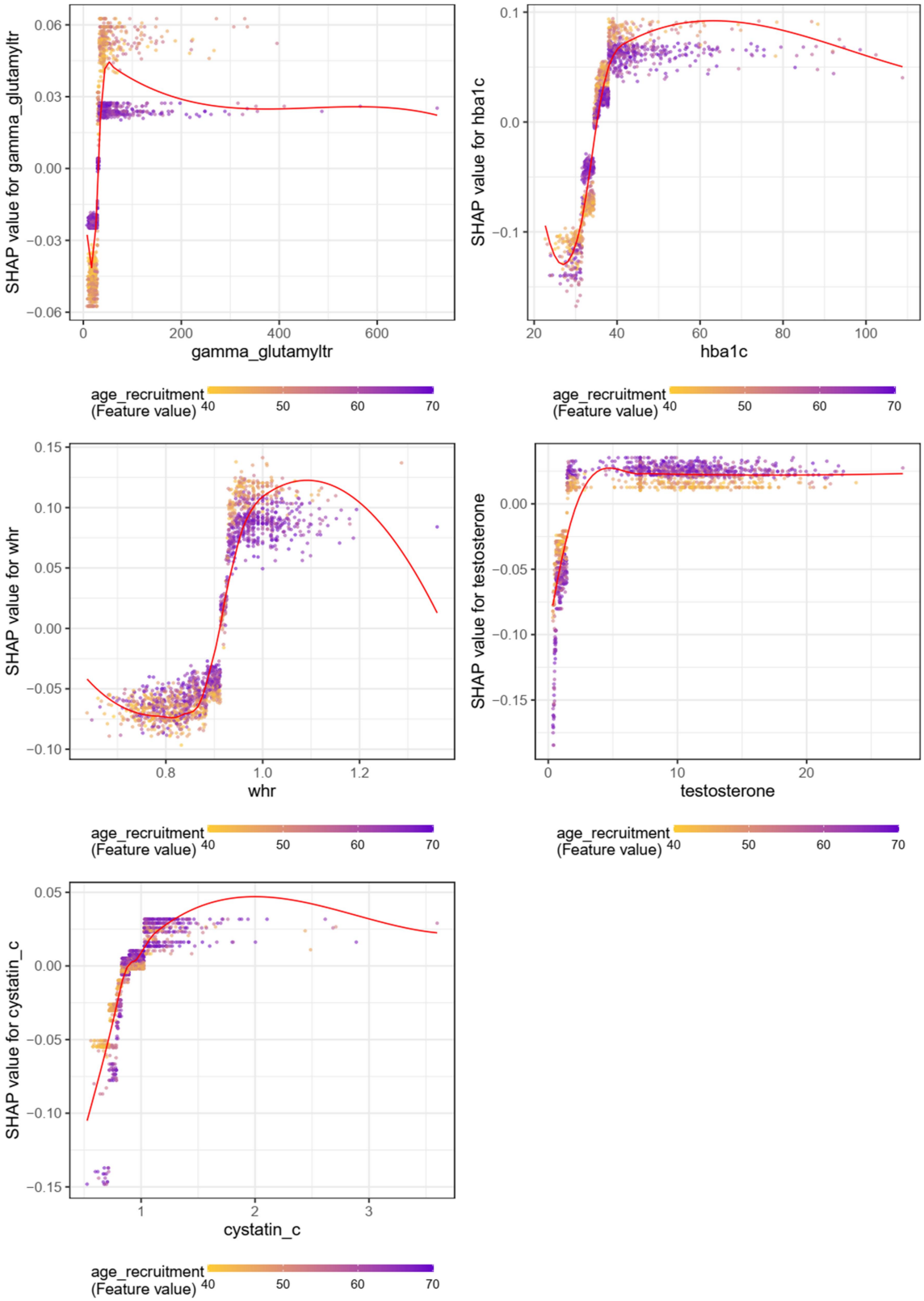

Figure 3 Shapley dependence plot with color coding based on an interacting variable (cohort A)

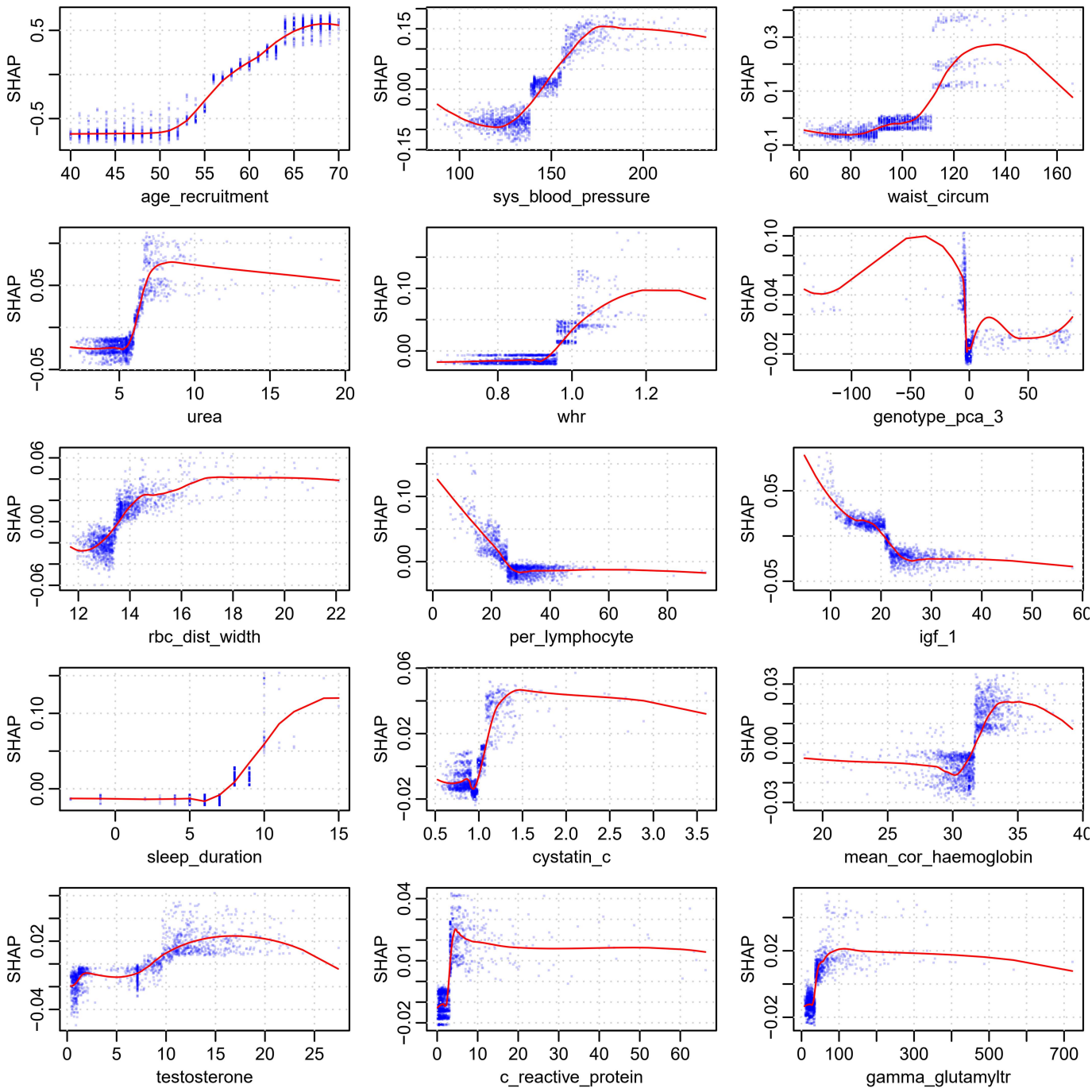

Figure 4 Shapley dependence plot for cohort B (fatal cases vs outpatient cases)

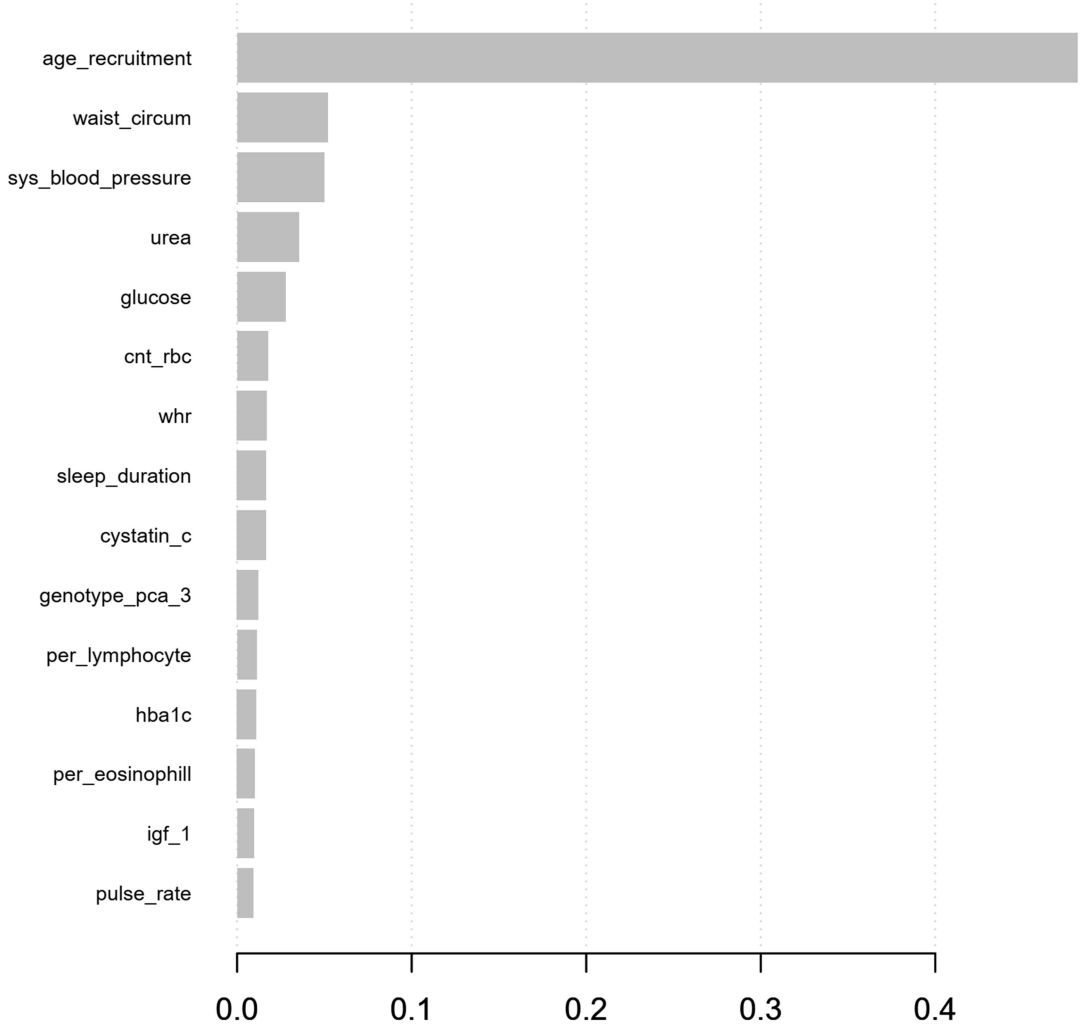

Figure 5 Variable importance ranked by gain for cohort B (top 15 variables shown)

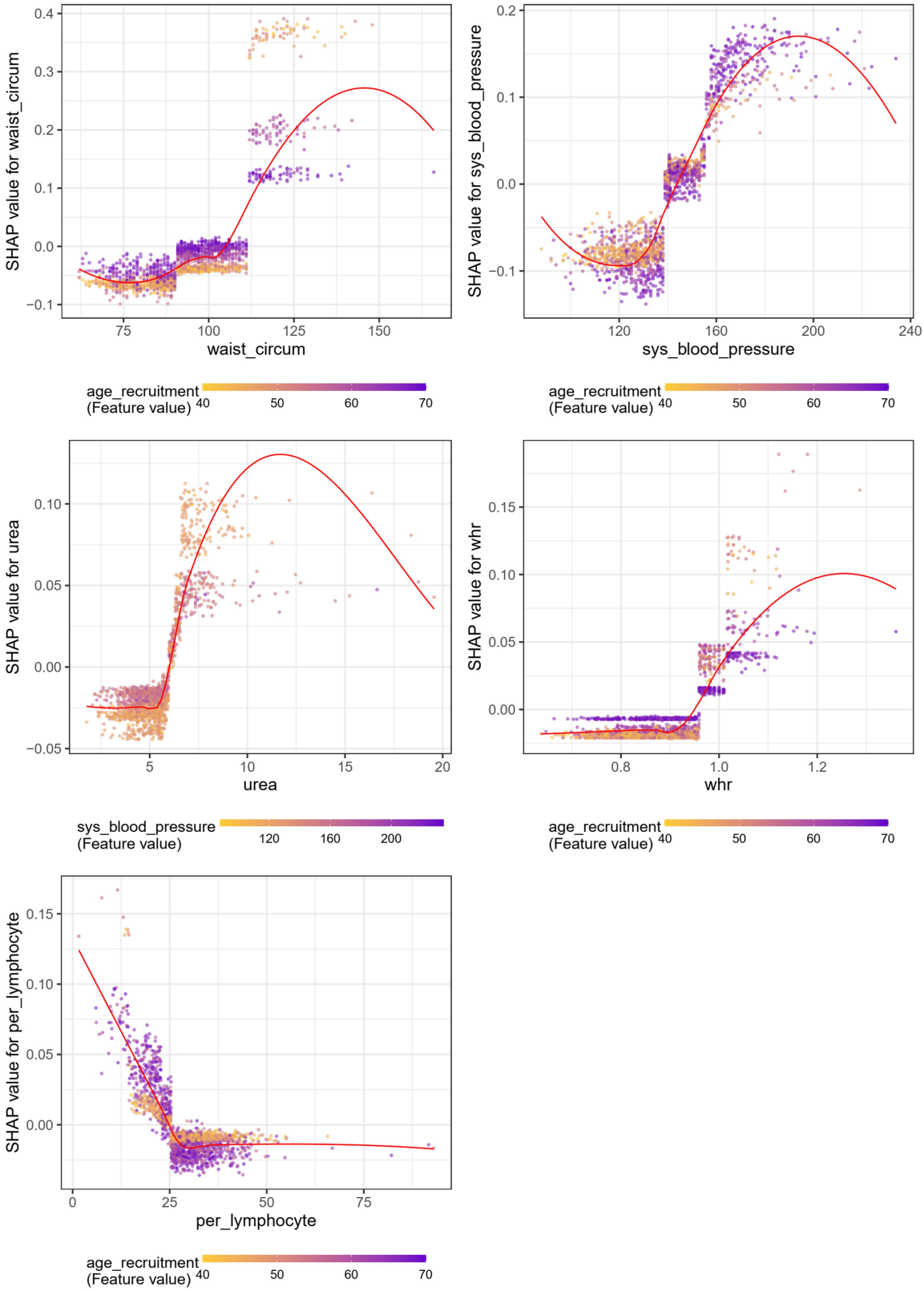

Figure 6 Shapley dependence plot with color coding based on an interacting variable (cohort B)

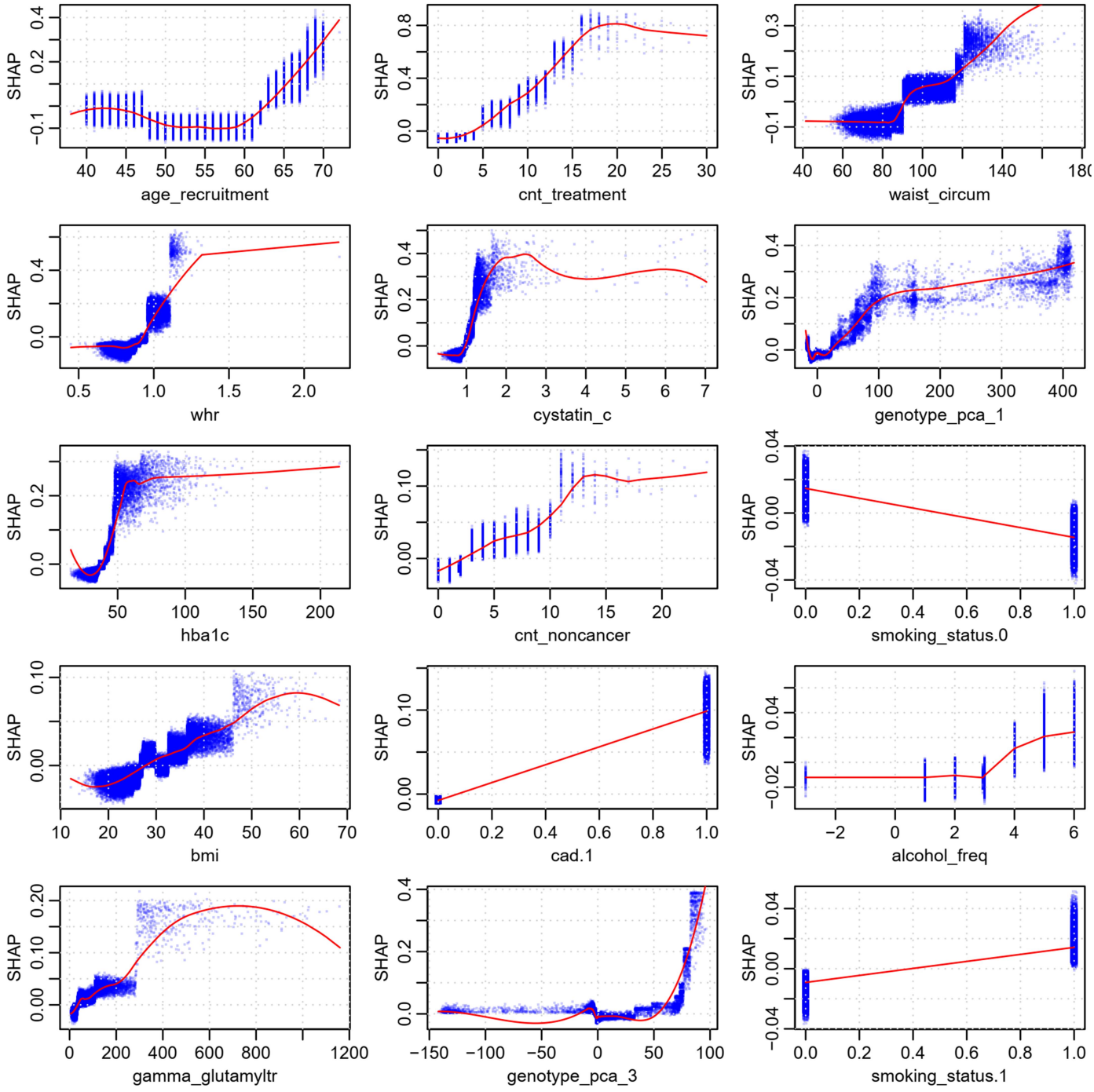

Figure 7 Shapley dependence plot for cohort C *(hospitalized/fatal cases* vs *population with no known infection)*

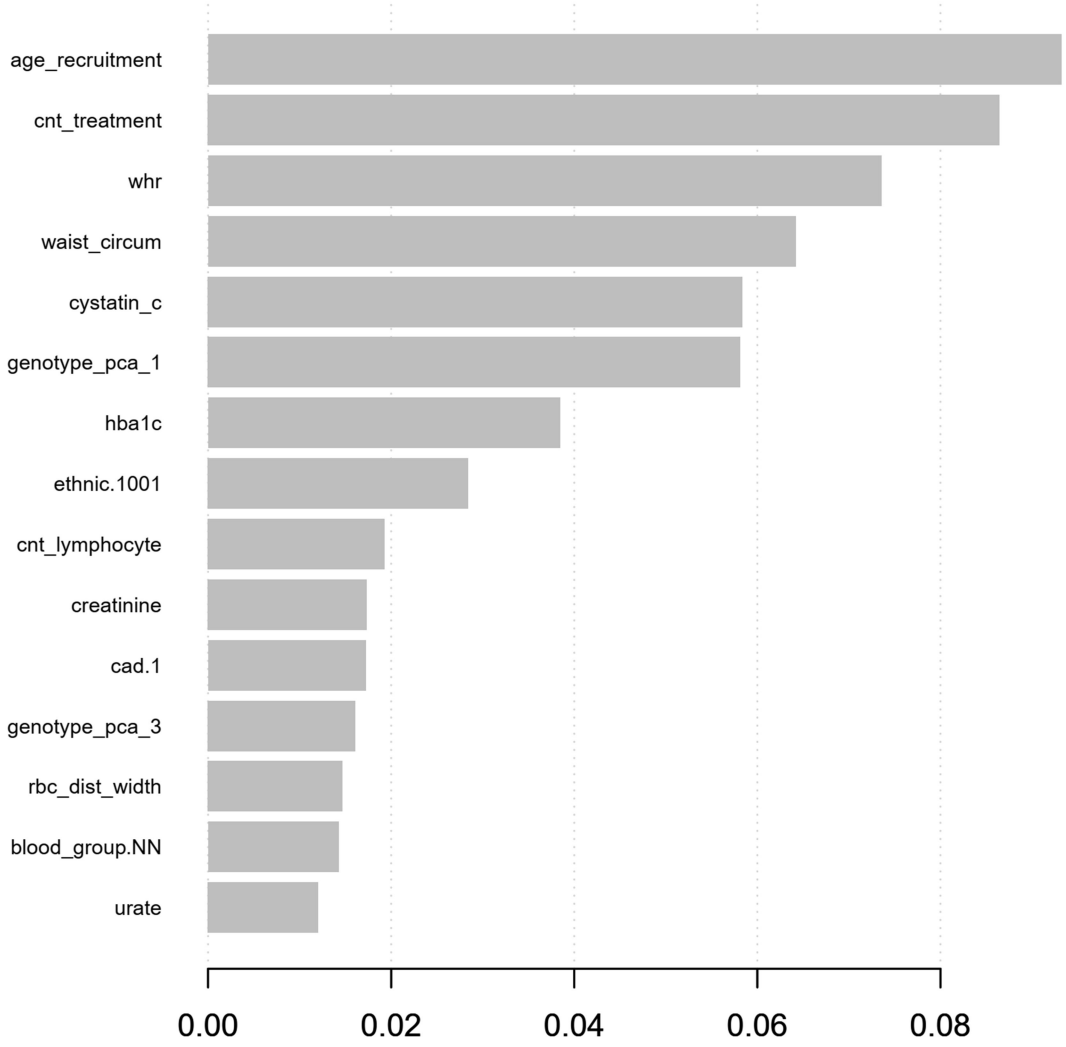

Figure 8 Variable importance ranked by gain for cohort C (top 15 variables shown)

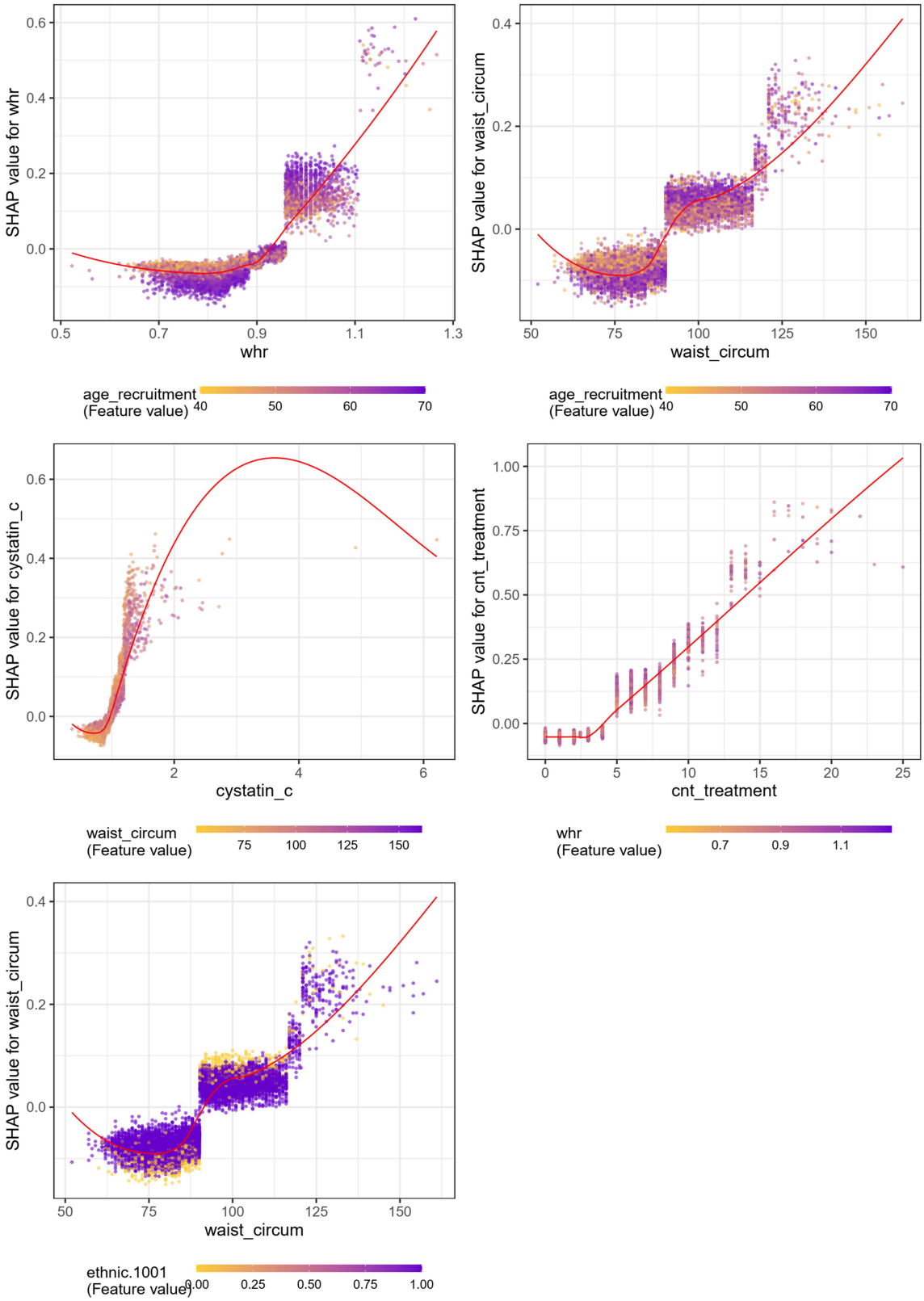

Figure 9 Shapley dependence plot with color coding based on an interacting variable (cohort C)

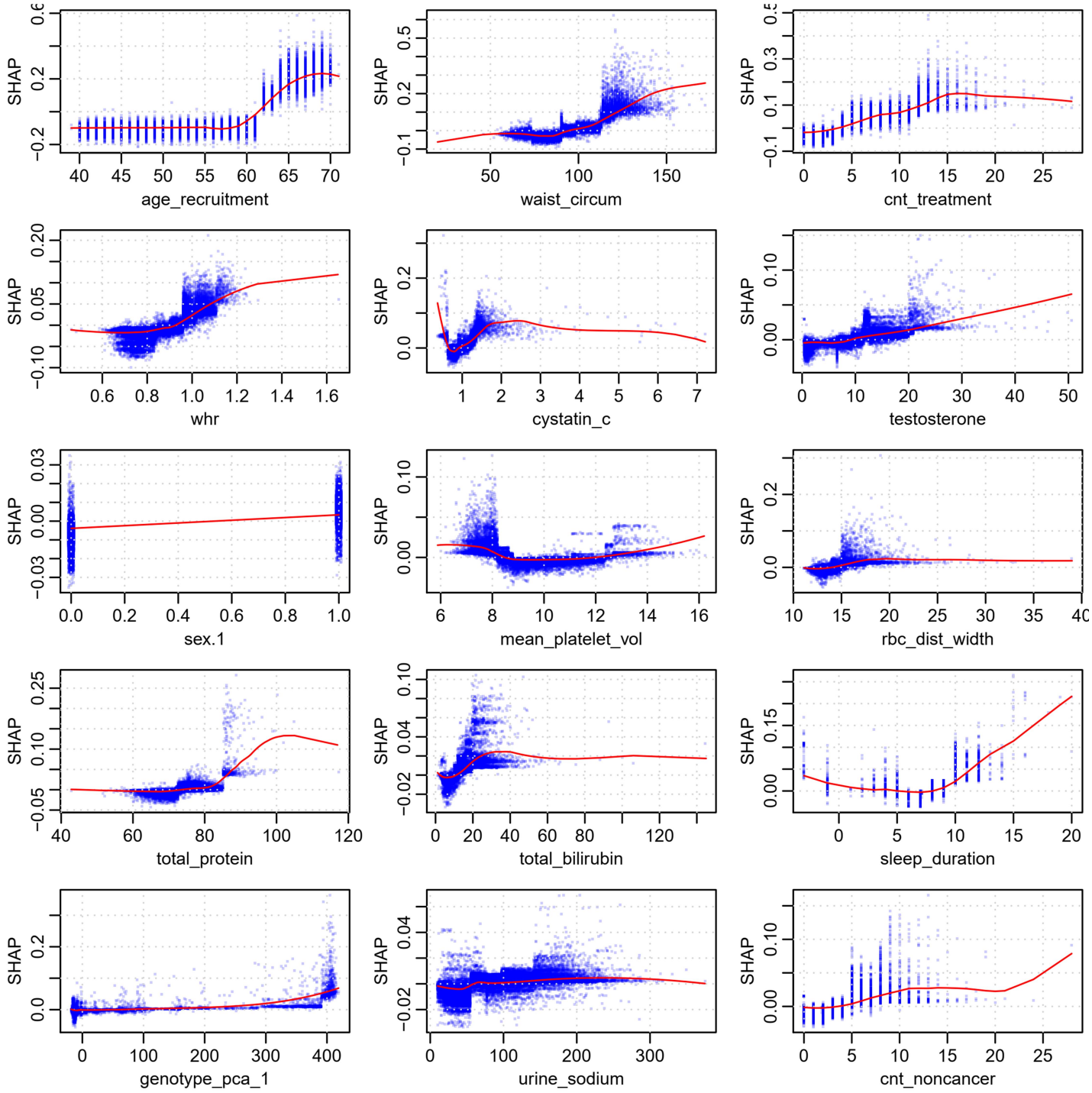

Figure 10 Shapley dependence plot for cohort D *(fatal cases vs population with no known infection)*

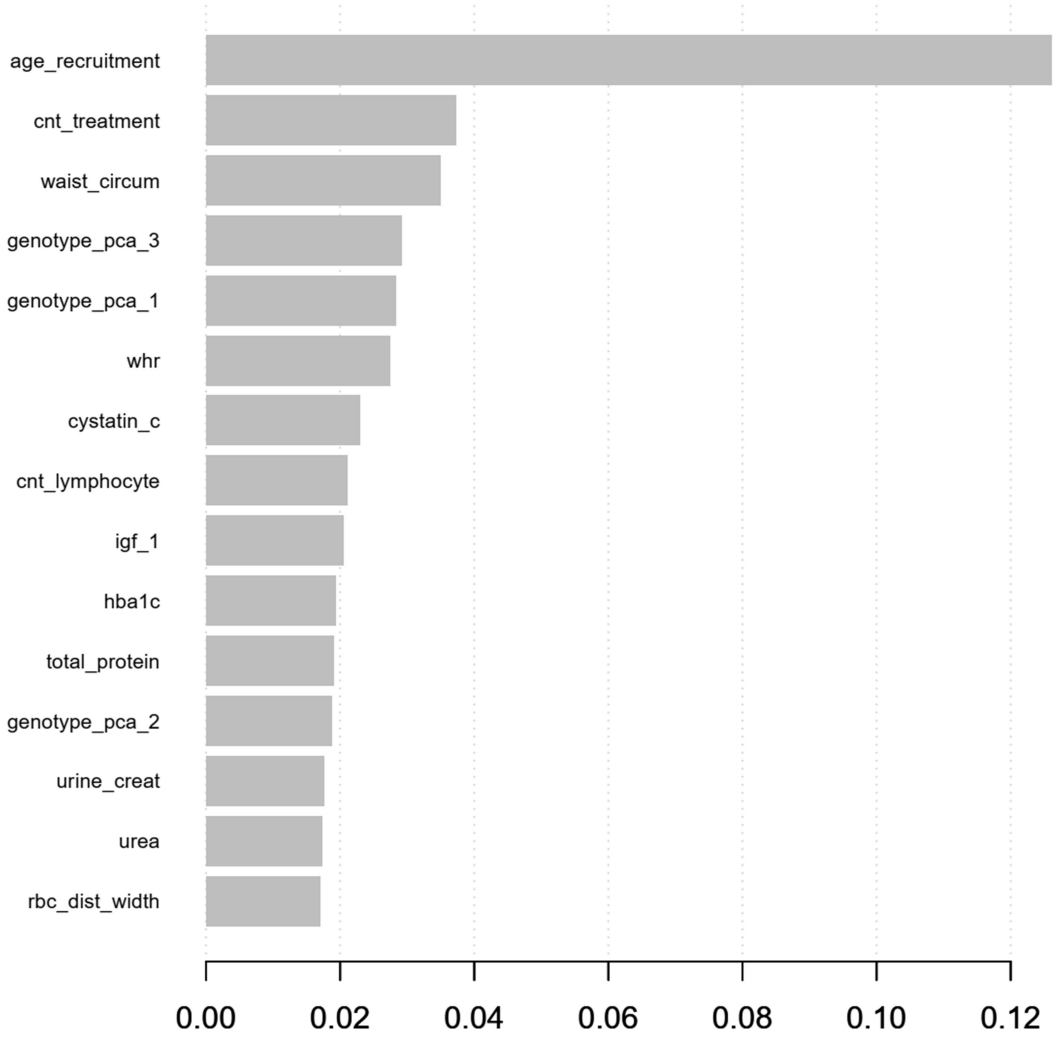

Figure 11 Variable importance ranked by gain for cohort D (top 15 variables shown)

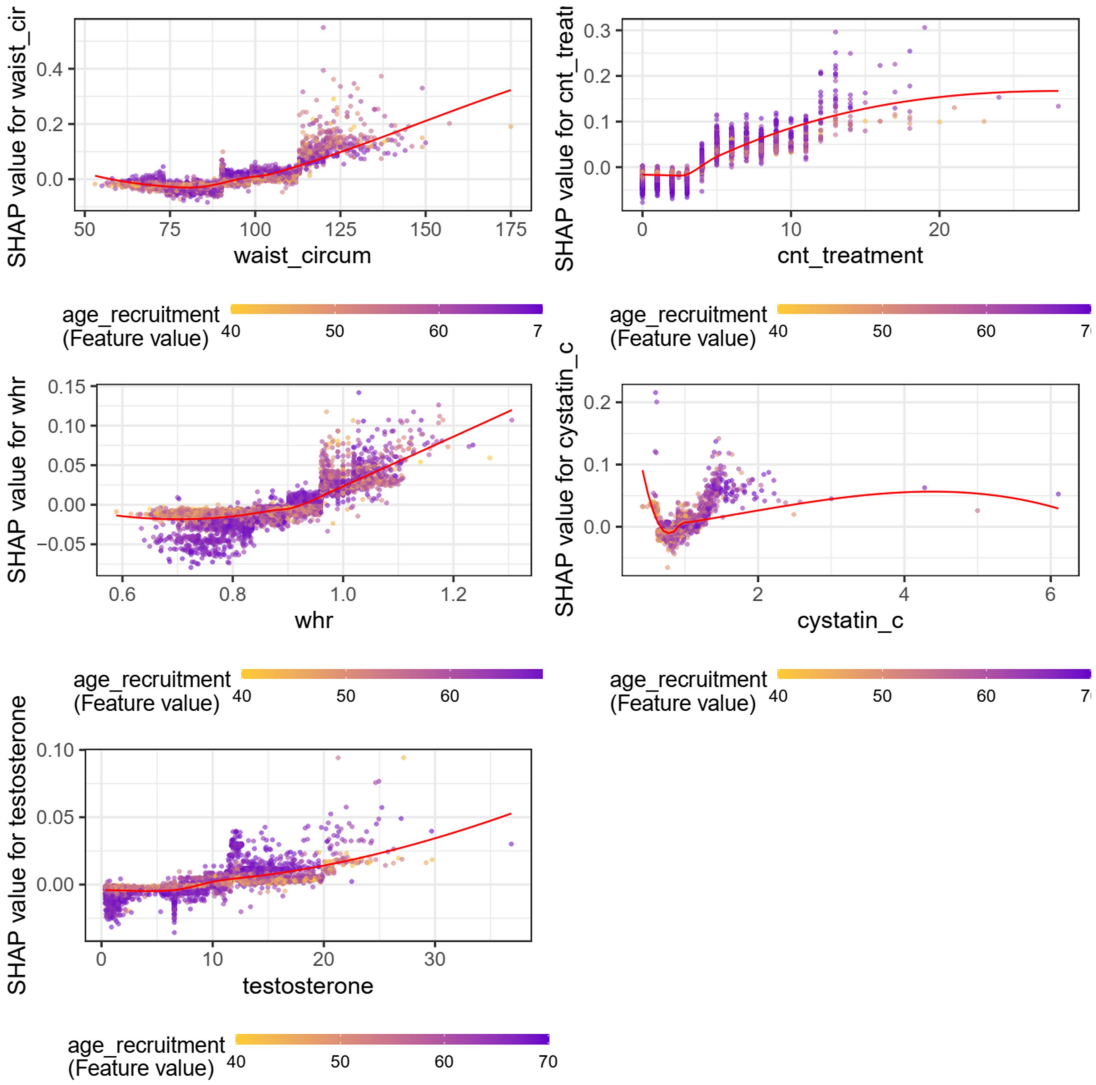

Figure 12 Shapley dependence plot with color coding based on an interacting variable (cohort D)