Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Journal of Biomedical Informatics 130 (2022) 104086

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Original Research

Automated interpretable discovery of heterogeneous treatment effectiveness: A COVID-19 case study

Benjamin J. Lengerich [a,*], Mark E. Nunnally [b], Yin Aphinyanaphongs [c], Caleb Ellington [d], Rich Caruana [e]

[a] Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 32 Vassar Street, Cambridge 02139, MA, USA
[b] Departments of Anesthesiology, Perioperative Care and Pain Medicine, Neurology, Surgery and Medicine, NYU Langone Health, 560 1st Avenue, New York 10016, NY, USA
[c] Department of Population Health, NYU Langone Health, 227 East 30th Street, New York 10016, NY, USA
[d] Computational Biology Department, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh 15213, PA, USA
[e] Microsoft Research, 14820 NE 36th Street, Redmond 02139, WA, USA

ARTICLE INFO

Keywords:
COVID-19
Heterogeneous Treatment Effects
Personalized Medicine
Interpretable Machine Learning

ABSTRACT

Testing multiple treatments for heterogeneous (varying) effectiveness with respect to many underlying risk factors requires many pairwise tests; we would like to instead automatically discover and visualize patient archetypes and predictors of treatment effectiveness using multitask machine learning. In this paper, we present a method to estimate these heterogeneous treatment effects with an interpretable hierarchical framework that uses additive models to visualize expected treatment benefits as a function of patient factors (identifying personalized treatment benefits) and concurrent treatments (identifying combinatorial treatment benefits). This method achieves state-of-the-art predictive power for COVID-19 in-hospital mortality and interpretable identification of heterogeneous treatment benefits. We first validate this method on the large public MIMIC-IV dataset of ICU patients to test recovery of heterogeneous treatment effects. Next we apply this method to a proprietary dataset of over 3000 patients hospitalized for COVID-19, and find evidence of heterogeneous treatment effectiveness predicted largely by indicators of inflammation and thrombosis risk: patients with few indicators of thrombosis risk benefit most from treatments against inflammation, while patients with few indicators of inflammation risk benefit most from treatments against thrombosis. This approach provides an automated methodology to discover heterogeneous and individualized effectiveness of treatments.

1. Introduction

Medical treatments can have different effectiveness for different patients based on a wide variety of factors including patient histories, comorbidities, and concurrent treatments; for this reason, a wide variety of statistical tools [1–4] have been developed to robustly estimate heterogeneous treatment effects (HTE) and individualized treatment effects (ITE) as functions of patient features. However, independently testing multiple treatments for heterogeneous effects with respect to many underlying patient risk factors requires large numbers of pairwise tests. To avoid reducing statistical power, we would like an automated method which uses multi-task learning to simultaneously discover heterogeneous treatments effects of many treatments and risk factors.

In this paper, we focus on the S-learner strategy [5] of estimating the conditional average treatment effect (CATE). This strategy first seeks $\widehat{\mu}_t(\boldsymbol{x}) = \mathbb{E}(Y_i|T_i = t, X_i = \boldsymbol{x})$, where X_i are patient factors and T_i are treatments; the heterogeneous treatment effects are then provided as $\widehat{\tau}(\boldsymbol{x}) = \mu_1(\boldsymbol{x}) - \mu_0(\boldsymbol{x})$. While this strategy is straightforward, there are statistical challenges with modeling interactions between treatments and risk factors. To overcome this challenge, we propose to use multitask learning to share statistical power between treatments to identify patient representations which predict the effectiveness of several related medications. This can be interpreted as an S-learner where T_i is a vector of treatments for patient i rather than a scalar indicator of a single treatment. Our proposed framework (Fig. 1) uses additive models to estimate and visualize expected treatment benefits as a function of patient factors (identifying personalized treatment benefits) and concurrent treatments (identifying interactive treatment benefits). This

* Corresponding author.
E-mail address: blengeri@mit.edu (B.J. Lengerich).

https://doi.org/10.1016/j.jbi.2022.104086
Received 23 October 2021; Received in revised form 25 April 2022; Accepted 26 April 2022
Available online 30 April 2022

framework has three main benefits: (1) it is a multi-task learning method which shares statistical power between related treatments, (2) it inherits the interpretability of additive models, allowing us to visualize the impacts of risk factors, and (3) it naturally extends to continuous-valued, dosed treatments. These advantages compare favorably against prior works which select linear regression models based on subgroup identification [6–8], and are thus limited to clustering patients, or regression trees [9,10] which do not provide interpretable maps linking risk factors to treatment benefits.

We first validate the proposed method on the MIMIC-IV dataset [11] to test recovery of HTEs. Next we examine a dataset of over 3000 patients hospitalized for COVID-19, and find evidence of HTEs influenced largely by indicators of inflammation and thrombosis risk: patients with few indicators of thrombosis risk benefit most from treatments against inflammation, while patients with few indicators of inflammation risk benefit most from treatments against thrombosis. This approach provides an automated methodology to discover heterogeneous and individualized effectiveness of treatments, which can be followed up by targeted statistical tests and clinical studies.

2. Materials and Methods

Cohort. Our dataset consists of 11080 hospitalized patients who had lab-confirmed cases of COVID-19 from March 2020 to August 2020. To filter out patients who were hospitalized for reasons other than COVID-19, we excluded patients who have indicators of (1) pregnancy: outpatient prenatal vitamins, in-patient oxytocics, folic acid preparations; or (2) scheduled surgery: urinary tract radiopaque diagnostics, laxatives, general anesthetics, antiemetic/antivertigo agents, or antiparasitics. We also require that the patients have recorded temperature, age, BMI, and admission date. Finally, we remove patients who died within six hours of admission. As predictors of patient risk, we include pre-admission features (demographics, comorbidities, and outpatient medications), and features measured on admission (vitals and initial in-patient lab tests). We exclude any measurement taken within 24 h of the patient mortality. The full list of features and more details regarding the cohort are provided in Section S1.

Treatment and Outcome Measure Construction. To ensure a proper linking of lab values, treatments, and outcomes, we consider only treatments that are administered within 24 h of the initial lab test and at least 24 h before mortality. Our outcome is in-hospital mortality; the cohort had an average mortality rate of 18.1%; with a peak rate over 25% in March 2020 that decreased to less than 5% by August 2020 (Figure S2).

2.1. Methods

Our goal in this study is to decompose the patient mortality risk into underlying risk factors and treatment benefits:

$$\text{logit}(\mathbb{P}(\text{Mortality})) = f(X_1) + \langle g(X_1, X_2), X_2 \rangle \quad (1)$$

where X_1 are features observed at or prior to hospital admission, and X_2 are in-patient treatments. The function $g(X_1, X_2)$ estimates the vector of expected treatment benefits for each patient; the inner product $\langle g(X_1, X_2), X_2 \rangle$ converts this potential benefit into an estimated scalar benefit under the observed treatments X_2. The estimated treatment benefits may vary with X_2 (in-patient medications); g thus learns to identify combination therapies which have super-additive effectiveness as well as personalized benefits which vary with patient characteristics X_1.

This model decomposes patient mortality risk into a background risk and a treatment benefit. By formulating g as a multivariate function with a vector output, we share statistical power between multiple treatments and improve estimation of patient representations and treatment responses. This framework allows extensibility; for example we can encourage information sharing between multiple risk factors and treatments by constraining the estimated set of treatment benefits to be a low-rank matrix summarizing archetypal patients and benefits.

There are two main challenges in estimation: estimating f the background mortality risk model and estimating g the personalized treatment benefit model. We choose to use generalized additive models for both f and g so that both functions are highly interpretable. We call this model a Contextual GAM (CGAM). `Python` code for this model is available at github.com/blengerich/ContextualGAM.

Background Mortality Risk Model. We use generalized additive models (GAMs) to model patient mortality risk. GAMs [12] are a version of logistic regression that are able to model more complex effects: while logistic regression summarizes the influence of each feature with a single coefficient, GAMs estimate the influence of a feature as a graph. This means that GAMs naturally accommodate non-linear and non-monotonic effects, which improves both model accuracy and interpretation. Non-linear effects are particularly important when features have multiple regions of high or low risk (e.g., both hyperthermia and hypothermia are associated with high risk). We use tree-based GAMs [13] implemented in the Python `Interpret` package [14]. These tree-based "Explainable Boosting Machine" (EBM) GAMs are invariant to all monotonic feature transforms, so log-transforms of lab values are not

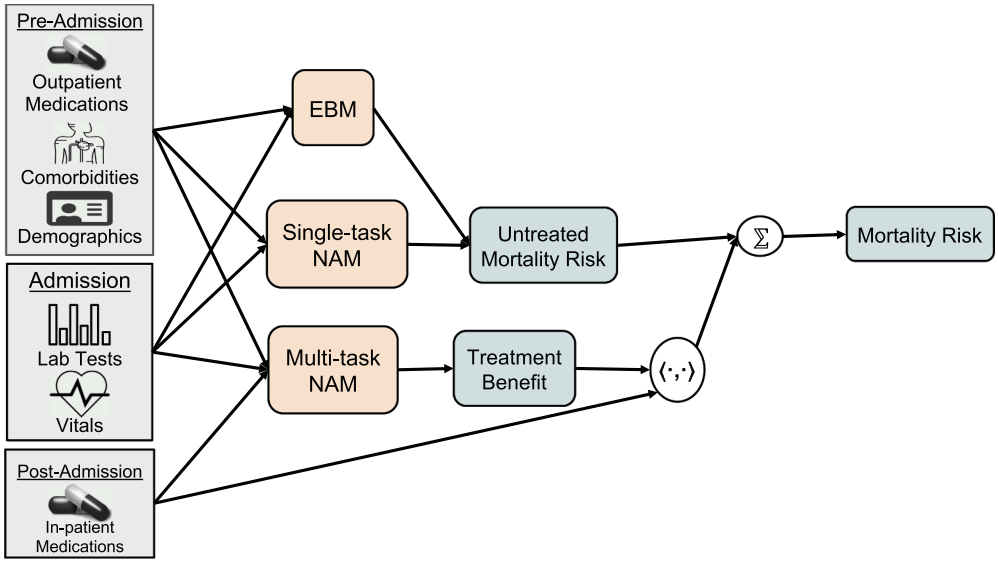

Fig. 1. Architecture to estimate latent personalized treatment benefits. Gray boxes indicate observed data, orange boxes are learned models, and blue boxes are latent variables. We first train an Explainable Boosting Machine (EBM) to predict mortality risk from pre-treatment features to generate baseline mortality risk. This model, based on gradient-boosted trees, captures discontinuities in risk, but is not differentiable and so must be trained in isolation from the rest of the architecture. After the EBM is trained, we train the single-task Neural Additive Model (NAM) and the multi-task NAM in parallel to decompose mortality risk into untreated mortality risk and personalized treatment benefits. This framework provides state-of-art mortality risk prediction and decomposes the risk into interpretable additive factors contributing to underlying risk and treatment benefits.

necessary.

Estimating Personalized Treatment Benefits. To estimate g, the map between patient characteristics and expected treatment benefits, we use a neural additive model (NAM) [15]. The NAM is a differentiable (gradient-based) additive model which can adapt as the estimated latent treatment benefits are updated during training. While the tree-based EBM of background mortality risk (f) includes discontinuous risk curves that often result from treatment protocols and heuristics for clinical decision making which the NAM struggles to capture, it is not differentiable and thus must be pre-trained and frozen before training g. For this reason, we estimate a third function $h(X_1)$ as a NAM to adjust the EBM f by removing effects which were initially estimated as effects in f but are better captured as effects in g. This model architecture is shown in Fig. 1.

Method Validation with MIMIC-IV. Before applying CGAMs to estimate HTEs for the novel COVID-19 disease, we first validate recovery of HTEs from realistic observational data. We use the MIMIC-IV dataset [11,16], which provides de-identified critical care data for over 50,000 patients admitted to intensive care units (ICU) at the Beth Israel Deaconess Medical Center (BIDMC), and records detailed lab values, treatments, and outcomes for all patients. In this analysis, we use all 53,150 patients admitted to the ICU in MIMIC-IV, spanning 69,211 hospital admission events and 76,540 ICU stays. Our outcome is in-hospital mortality, and we use demographics, comorbidities, and in-hospital medications analogous to the procedure used for the COVID-19 dataset.

First, we perform semi-synthetic experiments by simulating HTEs and adding these effects to the observed patient outcomes. We generate a simulated covariate $X_{sim} \sim \text{Bern}(p)$ and use this simulated covariate to modulate an HTE: $g(X_{sim}) = \mu X_{sim}$. In our experiments, p is an experimental setting giving the probability of treatment while μ is an experimental setting giving the HTE strength. Finally, we re-sample observed mortality rates adjusted by the new additive HTE. This procedure retains the distribution of patient risk factors, covariates, and treatments, allowing us to test the recovery of a known HTE in a realistic setting. For these simulated HTEs, we evaluate the expected ℓ_2 error of the estimated HTE: $\mathbb{E}_{X_{sim}}[\| g(X_{sim}) - \widehat{g}(X_{sim})\|^2]$, and compare the CGAM against a baseline EBM trained to recover heterogeneous interaction effects. The CGAM accurately accurately recovers simulated HTEs from the MIMIC-IV dataset (Fig. 2), with benefit especially for HTEs which are strong and treatments which are frequent.

Beyond these semi-synthetic experiments, we examine the HTEs suggested by the CGAM on the original MIMIC-IV data. The CGAM identifies HTEs governed by patient age and hemoglobin levels (Section S4), which are plausible the known treatment effectiveness and real-world behavior captured in the MIMIC-IV dataset.

3. Results

Our approach decomposes mortality risk into underlying risk, homogeneous treatment effects, and heterogeneous treatment effects. The risk model is accurate, achieving an AUROC of 0.912 ± 0.001 and an F1-score of 0.598 ± 0.002 on held-out patients, outperforming a logistic regression model which achieves an AUROC of 0.859 ± 0.001 and F1-score of 0.455 ± 0.002 (confidence intervals provided by bootstrap resampling). These scores are state-of-the-art, outperforming published models which integrate CT scans with biological and clinical variables [17].

3.1. Risk Factors

The most important features to the background risk model are (in decreasing order): Neutrophil–Lymphocyte Ratio (NLR), Temperature, Age, Procalcitonin and C-Reactive Protein. While elevated NLR is a strong predictor of mortality, none of these risk factors solely determine mortality, and in many patients with low NLR levels nevertheless have large probabilities of mortality (Figure S1).

Comorbidities and Pre-Admission Medications. The baseline mortality risks associated with comorbidities and pre-admission out-patient medications suggest mechanisms of mortality involving inflammation and/or thromboses (Fig. 3). The most protective association is valve replacement, for which patients are often prescribed long-term anti-coagulants. The second-most protective association is platelet aggregation inhibitors (low-dose aspirin), which is an anti-coagulant. The most deleterious associations are congestive heart failure, hypertension, and myocardial infarction.

Demographics and Vital Signs. Elderly patient age and elevated temperature on hospital admission are strong predictors of mortality, while patient body mass index (BMI) is not a strong risk factor for hospitalized patients (Fig. 4).

Lab Tests. The effects of 12 lab tests are shown in Fig. 4. The largest effect is for Neutrophil/ Lymphocyte Ratio (NLR), a measure of inflammation and an indicator of COVID-19 severity [18,19]. Procalcitonin and C-reactive protein also increase likelihood of mortality. Also of note are serum calcium, for which no elevated level is associated with increased risk, and serum creatinine, which shows a step-function drop in risk at 4 mg/dL, a common threshold for dialysis decisions. Finally,

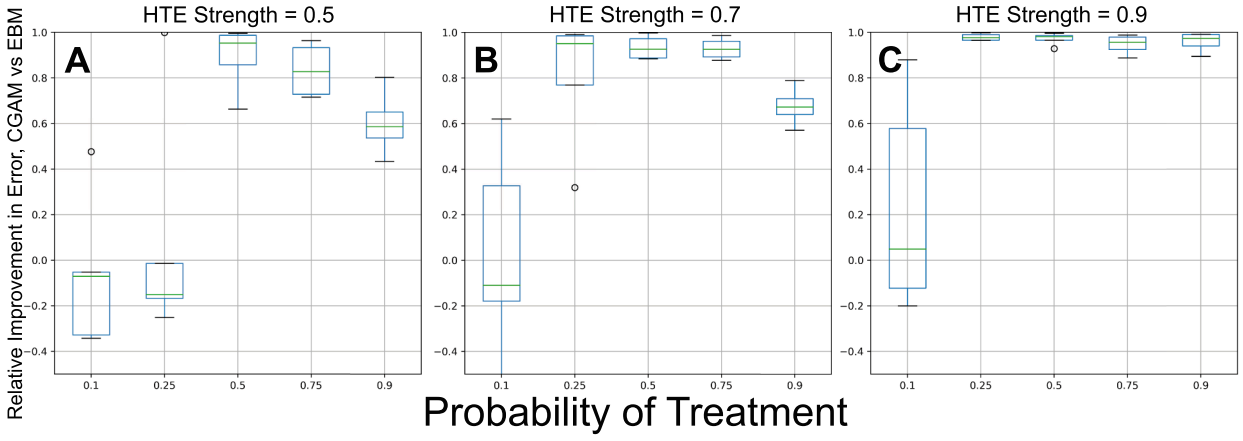

Fig. 2. Recovery of simulated heterogeneous treatment effects (HTEs) added to the MIMIC-IV dataset shows that CGAM more accurately recovers HTEs when the HTEs are strong and the treatment is frequently prescribed. In each pane, we plot the relative improvement in ℓ_2 error of the HTE estimated by the CGAM compared against the HTE estimated by a baseline method (EBM with interactions). Positive values indicate the CGAM performs better than the baseline, while negative values indicate the baseline performs better than the CGAM. Each box-and-whiskers plot represents the distribution of recovery errors over 10 experimental runs. We vary two experimental conditions in the simulation: probability of treatment p (indexed along the horizontal axis), and HTE strength μ (indexed by panes).

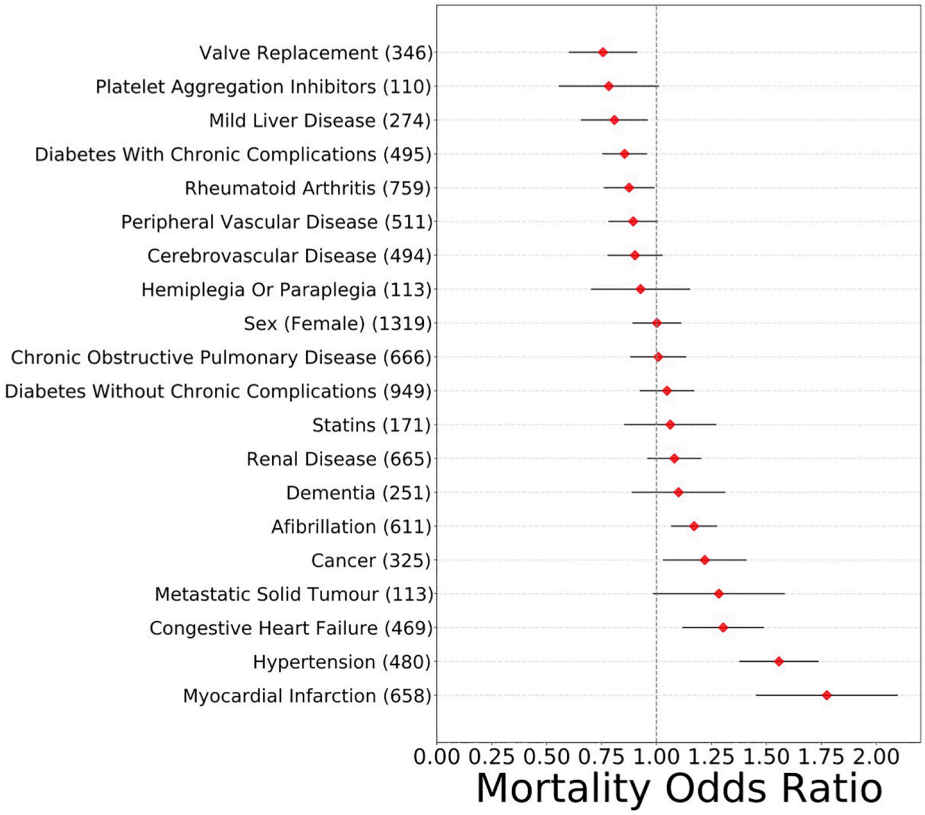

Fig. 3. Mortality risk conferred by pre-admission comorbidities and medications, with 95% confidence intervals. For each treatment, the number in parentheses indicates the sample size.

hematocrit exhibits opposite effects when estimated in isolation or after correcting for other factors – marginalization would associate a decrease in hematocrit with increased mortality risk, while the multivariable GAM identifies that increased hematocrit is associated with increased mortality risk.

3.2. Homogeneous Treatment Effects

For homogeneous average treatment effects (ATEs), we are interested in the adjusted risk difference (ARD) $\sum_i Y_i T_i - \sum_i \mathbb{P}(\text{mortality}|X_i)$ and the adjusted risk ratio (ARR) $\sum_i Y_i T_i / \sum_i \mathbb{P}(\text{mortality}|X_i)$, where T_i is the treatment indicator for patient i and $\mathbb{P}(\text{mortality}|X_i) = \widehat{f}(X_i)$ is given by the mortality risk model. For both measures, smaller values indicate protective associations while larger values indicate harm (a treatment with no effect would have ARD $= 0$ and ARR $= 1$).

The majority of treatments are not significantly associated with homogeneous treatment benefits after correcting for patient risk factors (Fig. 5). There is some evidence of a benefit of thyroid hormones, concording with observed dysregulation of thyroid hormones in COVID-19 patients [20]. In addition, there is some evidence of a protective effect of NSAIDs (defined as either Ibuprofen or Ketorolac) in a small sample size of 101 treated patients. We also see a possibility of a beneficial effect of direct factor Xa inhibitors; however, the confidence intervals are very wide and it is difficult to distinguish any protective effect from association with administration prior to patient discharge.

3.3. Heterogeneous Treatment Effects

Next, we turn to the HTEs of 6 treatments: Anti-coagulants (Heparin), NSAIDs, Azithromycin, HCQ, Zinc replacement, and Glucocorticoids (GCs). All 6 treatments are associated with diminished effectiveness in older patients (Figure S3). In addition, 3 of these treatments are associated with HTEs governed by NLR, a marker of inflammation and severe COVID-19 [18,19] (Fig. 6). The statistical benefit of anti-coagulants, NSAIDs, and Azithromycin decrease with increased NLR, while the effectiveness of HCQ, Zinc, and GCs do not show statistically significant heterogeneity. We examine the association between GCs, NLR, and mortality in detail in the discussion.

Treatment-specific HTEs associated with comorbidities and concurrent treatments are summarized in Table 1; for each effect, we report the increased benefit associated with the interactive factor: $\widehat{\mu}_t(x_1) - \widehat{\mu}_t(x_0)$ where X_1 indicates that the interactive factor is present while X_0 indicates that the interactive factor is not present. All of these benefit increases should be considered as an additive effect in coordination with the homogeneous treatment effects summarized in Fig. 5. These results suggest GCs are associated with decreased benefit in combination therapy with Azithromycin, while NSAIDs are associated with increased benefit for patients with a history of alcoholism, a pro-inflammatory condition [21]. Zinc replacement is associated with diminished benefit for patients with congestive heart failure; this result would be expected if protective effects of zinc were due to a mediation of thromboses [22] because patients with congestive heart failure are routinely prescribed anticoagulants and platelet aggregation inhibitors as out-patient medications. Similarly, anti-coagulants, which have been shown to have prevent mortality in COVID-19 patients with a likelihood of thromboses [23], appear to be less protective for patients with congestive heart failure, while increasing in effectiveness for patients with histories of substance abuse. Azithromycin is associated with decreased benefit for patients with history of opioid usage and congestive failure, with a mildly positive interaction with HCQ. Finally, HCQ is associated with beneift increase in patients who are at reduced risk of negative side effects including arrythmias (e.g., patients without a history of afibrillation or congestive heart failure); in addition, HCQ is associated with increased benefit in combination with Azithromycin and for

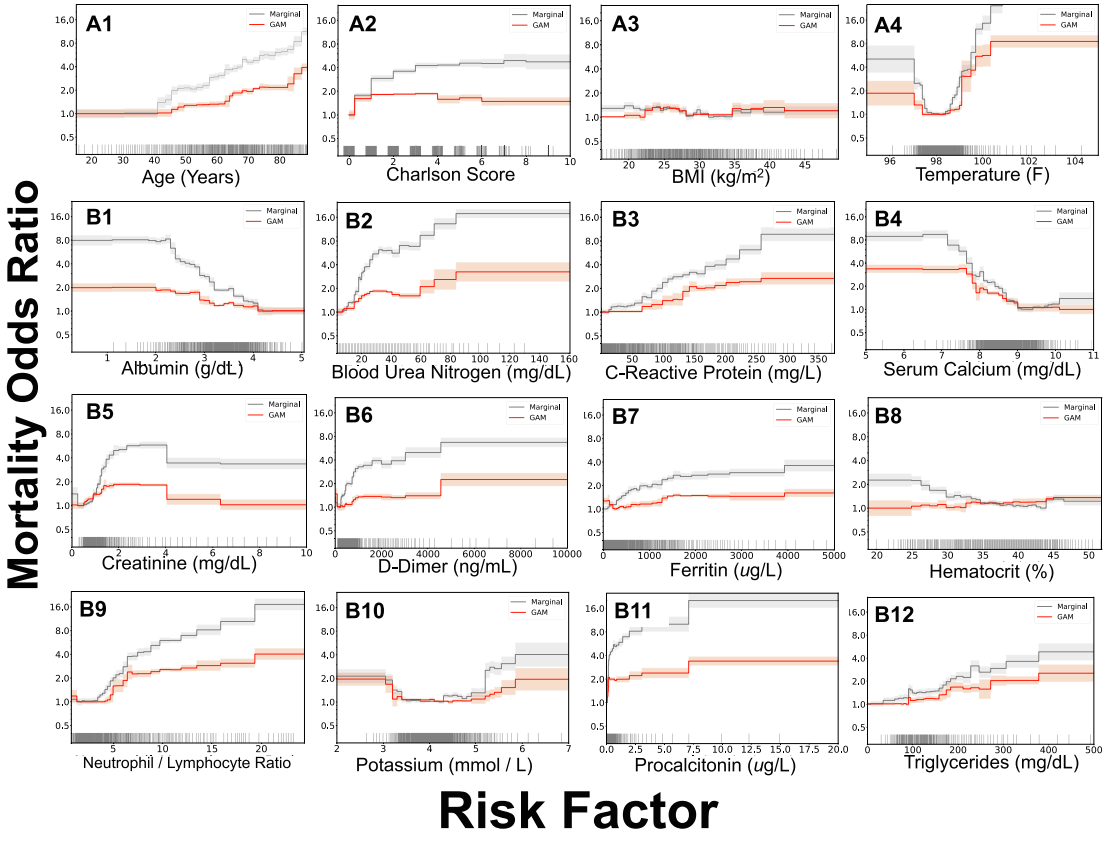

Fig. 4. Baseline mortality risk associated with a variety of risk factors. In each pane, we show both the effect estimated by univariable marginalization (gray), and the multivariable GAM which corrects for other risk factors (red). Each black tick mark along the horizontal indicates 10 patients, with noise added to visualize data density.**(A)** Baseline mortality risk conferred by patient age, BMI, Charlson score, and temperature. **(B)** Baseline mortality risk conferred by demographics, vitals, and lab tests.

patients with hallucinogen usage. These mild HTEs contrast against the lack of statistical evidence for any homogeneous treatment effectiveness for either Azithromycin or HCQ (Fig. 5).

4. Discussion

This method of estimating personalized effectiveness of multiple treatments is best considered as a tool for hypothesis generation to be followed up by targeted statistical tests. Here we examine a case study of manual exploration of the effectiveness of GCs modulated by NLR, which was suggested by the heterogeneous treatment effect model.

Case Study: NLR and GCs. GCs have been shown to improve outcomes of patients with severe cases of COVID-19 [24]; however, criteria for prescribing GCs are currently limited. GC prescription is highly correlated with later admission dates and mortality is lower at later dates for a number of reasons. To correct for this confounding, we use the mortality risk model to correct for all risk which can be attributed to factors other than GCs (a full discussion of methodology is provided in Sec. S3) to isolate the impact of GCs.

Corresponding sample sizes and ARRs for GC treatment, stratified by patient NLR value into three groups, are given in Table 2. Of these three ARRs, we observe statistically significant evidence of GC effects only for patients with NLR 6–25. We hypothesize that for patients who are not at risk of severe inflammation, GCs have little effect; for patients who are admitted with extremely high NLR, GCs may be insufficient. Finally, while elevated NLR is a strong predictor of mortality, it is not the only risk factor (Figure S1). These results suggest that GCs may have limited benefit to patients who are at high risk without having an elevated NLR and other treatments may be required.

Limitations. As with all analyses of observational data, this approach has several limitations. Firstly, while the optimization allocates effect sizes to the most statistically reliable indicators, we do not use any side information (such as treatment mechanism of action or time-series data) to perform causal inference. In addition, in this study we have considered only binary indicators for each treatment, choosing to assume that providers are following dosage protocols to standardize care. Finally, while additive models are interpretable and accurate, they are still susceptible to statistical biases [25] which may cause different model classes to recover different effects from a single dataset. Further works should investigate the potential for other classes of additive models to corroborate or dispute these findings.

Generalizable Impact and Comparisons to Related Approaches. We have proposed and studied an automated method to identify and interpret HTEs with additive models. The method, CGAM, uses additive models to decompose mortality likelihood into risk factors and personalized treatment benefits. Estimating HTEs from observational data has been a long-standing goal or clinical informatics, and many methods have been proposed, including models which seek to estimate effects of multiple treatments [26], multiple outcomes [2], or nonparametric effect models [27]. The CGAM framework combines the advantages of these frameworks, simultaneously modeling multiple treatments, multiple outcomes, and a nonparametric additive model which has both large representational capacity (using deep learning as the basis of the additive models) and an intelligible map from observed patient covariates to an estimate of personalized treatment benefits (retaining the interpretability of additive models). This compares against prior works which select linear regression models based on subgroup identification [6–8], and are thus limited to clustering patients, or regression trees [9,10] which do not provide interpretable maps linking risk factors to treatment benefits.

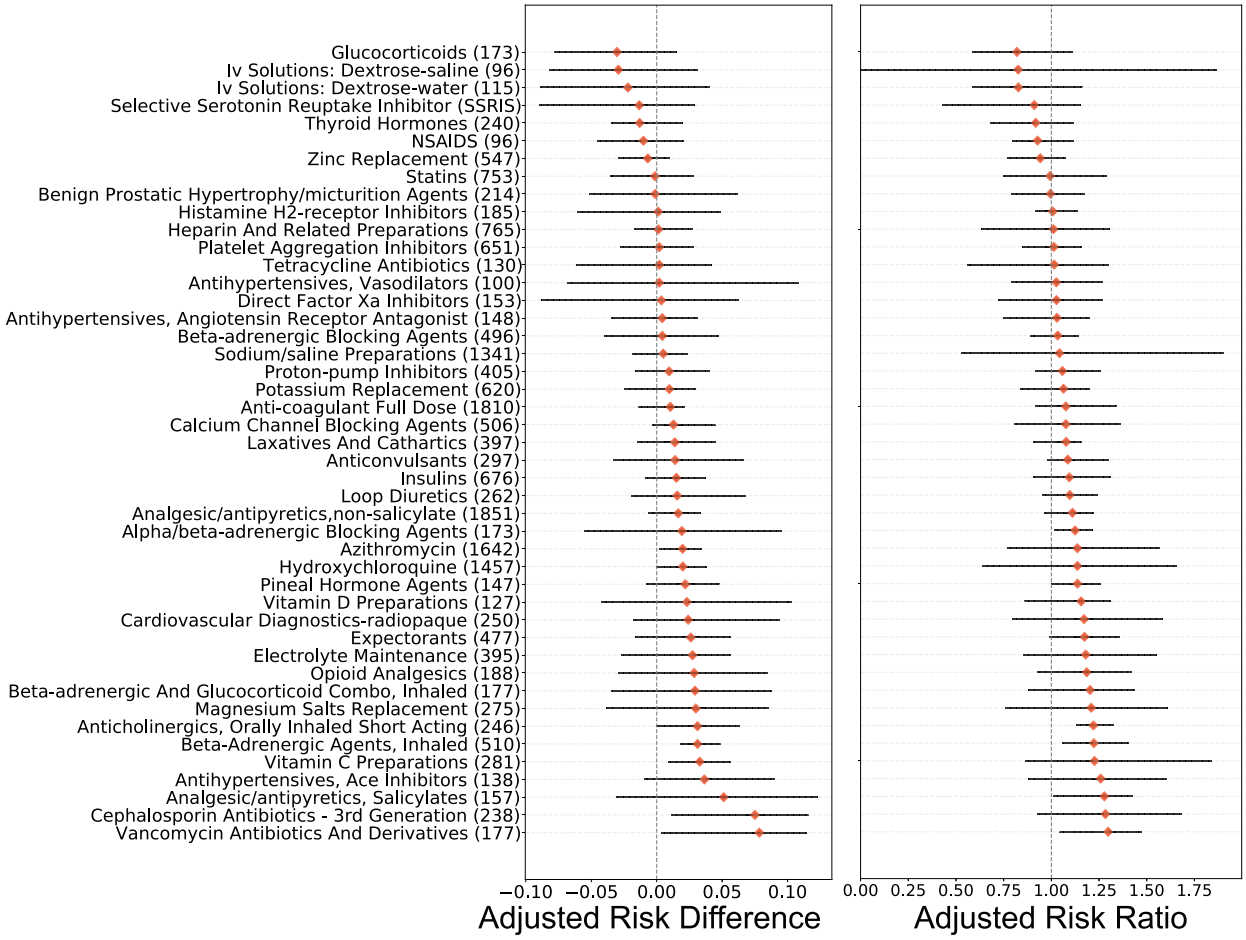

Fig. 5. Homogeneous treatment effects of in-patient medications, calculated as the adjusted risk difference (left) and the adjusted risk ratio (right). For each treatment, the number in parentheses is the number of patients treated with this medication in the dataset.

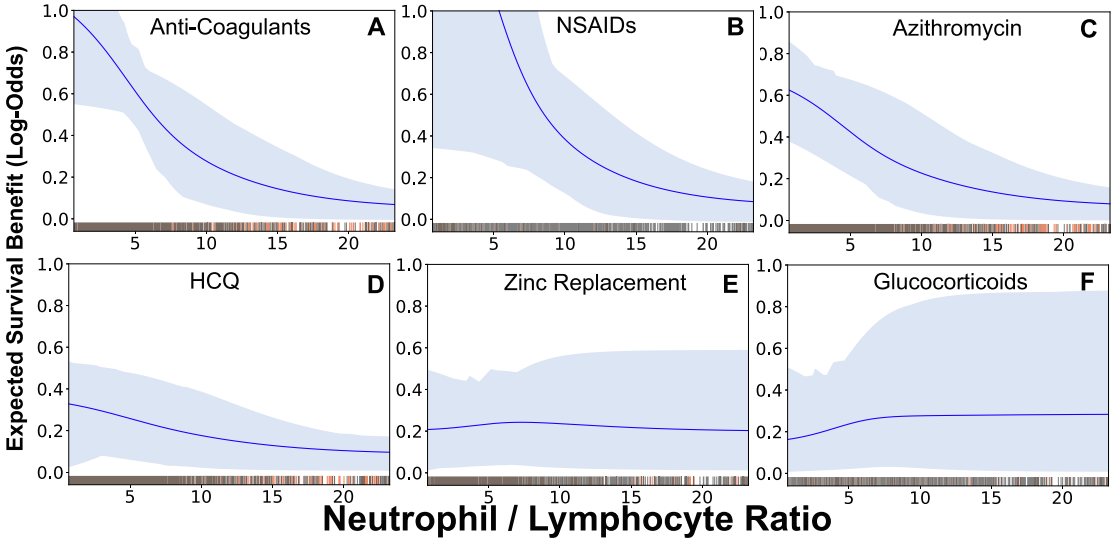

Fig. 6. HTEs with respect to NLR. In each pane, we plot the estimated treatment benefit as a function of patient NLR, with shaded regions indicating 95% bootstrap confidence intervals. Red tick marks along the horizontal axis denote treated patients, while black tick marks denote untreated patients. NLR is a biomarker of inflammation and severe COVID-19; most of these treatments have lower effectiveness for higher NLR values. In contrast, zinc replacement and GCs appear to have to have stable or increasing effectiveness for larger values of NLR, consistent with an anti-inflammatory mechanism of action.

Table 1
Heterogeneous Treatments Effects. We report the additive increase in survival probability (log-odds) of all HTEs in which the 95% confidence interval does not overlap a zero effect. For each effect, we report the increased benefit associated with the interactive factor: $\widehat{\mu}_t(x_1) - \widehat{\mu}_t(x_0)$ where X_1 indicates that the interactive factor is present while X_0 indicates that the interactive factor is not present.

Treatment	Interactive Factor	Benefit Increase (95% CI)
GCs	Azithromycin	−0.28 (-0.54, −0.02)
NSAIDs	History of Alcoholism	0.29 (0.05, 0.53)
Zinc Replacement	Congestive heart failure	−0.26 (-0.47, −0.05)
Anti-Coagulant	Congestive heart failure	−0.17 (-0.32, −0.02)
	History of Hallucinogen Usage	0.44 (0.09, 0.80)
	History of Alcoholism	0.82 (0.22, 1.43)
Azithromycin	History of Opioid Usage	−0.50 (-0.98, −0.03)
	Congestive heart failure	−0.15 (-0.26, −0.04)
	HCQ	0.10 (0.02, 0.18)
HCQ	Congestive heart failure	−0.14 (-0.25, −0.03)
	Myocardial infarction	−0.09 (-0.17, −0.01)
	History of Alcoholism	0.17 (0.01, 0.33)
	Azithromycin	0.24 (0.06, 0.42)
	History of Hallucinogen Usage	0.39 (0.04, 0.73)

Table 2
ARRs for patients treated with GCs compared to patients not treated with GCs, calculated by patient NLR. ARRs smaller than 1 indicate reduced mortality for patients treated with GCs, i.e., a beneficial effect. P-values are calculated by bootstrap resampling of the test set.

NLR Values	N GC	N Control	ARR	95% CI	P-Value (Uncorr.)
0–6	79	1728	0.99	0.68–1.31	0.975
6–25	99	1079	0.74	0.49–0.99	**0.042**
>25	15	116	0.90	0.58–1.22	0.549

All of these methods are based on the goal of estimating a CATE and assuming that treatment assignments are sufficiently randomized after some set of observed covariates have been conditioned on [28]. Unfortunately, this "ignorability" assumption is rarely satisfied in practice – clinicians use all tools available to guide treatment decisions, including recorded and unrecorded patient factors and can apply deterministic treatment protocols after categorizing patients in subtypes. In these cases, ignorability is not satisfied and intelligibility of the learned model is critical in order to audit any interactions between treatment protocols and HTE estimates. In future work, we are interested in automatically identifying and correcting for hidden confounding from latent variables (e.g. adapting the frameworks of [29,30] to use the CGAM model). Finally, CGAMs are extensible frameworks: we are interested in future work to model dose-specific effects of continuously-dosed treatments, regularization to encourage similarity between patient types, or low-rank patient subtype representations[31] to learn archetypes of patients and treatment responses.

5. Conclusions

In this paper, we have proposed a method to estimate heterogeneous (varying) effectiveness of medical treatments by training additive models to estimate personalized treatment benefits and share statistical power between many treatments. We applied this method to mortality risk models of COVID-19 patients and uncovered evidence supporting two pathways of mortality: inflammation and thrombosis. We see that many treatments appear to have heterogeneous effectiveness; in particular, anti-inflammation treatments tend to be more effective for patients with lower likelihood of thromboses, while anti-coagulation treatments tend to be more effective for patients with lower likelihood of inflammatory attacks. We also see some evidence consistent with super-additive effectiveness of combinatorial treatments.

5.1. Funding

No funding was received for this work.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.jbi.2022.104086.

References

[1] C. Lee, N. Mastronarde, M. van der Schaar, Estimation of individual treatment effect in latent confounder models via adversarial learning, arXiv preprint arXiv:1811.08943 (2018).
[2] A.M. Alaa, M. van der Schaar, Bayesian inference of individualized treatment effects using multi-task gaussian processes, arXiv preprint arXiv:1704.02801 (2017).
[3] I. Bica, J. Jordon, M. van der Schaar, Estimating the effects of continuous-valued interventions using generative adversarial networks, arXiv preprint arXiv:2002.12326 (2020).
[4] I. Bica, A. Alaa, M. Van Der Schaar, Time series deconfounder: Estimating treatment effects over time in the presence of hidden confounders, in: International Conference on Machine Learning, PMLR, 2020, pp. 884–895.
[5] S.R. Künzel, J.S. Sekhon, P.J. Bickel, B. Yu, Metalearners for estimating heterogeneous treatment effects using machine learning, Proc. Natl. Acad. Sci. 116 (10) (2019) 4156–4165.
[6] K. Imai, M. Ratkovic, Estimating treatment effect heterogeneity in randomized program evaluation, The Annals of Applied Statistics 7 (1) (2013) 443–470.
[7] L. Tian, A.A. Alizadeh, A.J. Gentles, R. Tibshirani, A simple method for estimating interactions between a treatment and a large number of covariates, Journal of the American Statistical Association 109 (508) (2014) 1517–1532.
[8] H.I. Weisberg, V.P. Pontes, Post hoc subgroups in clinical trials: Anathema or analytics? Clinical trials 12 (4) (2015) 357–364.
[9] X. Su, C.-L. Tsai, H. Wang, D.M. Nickerson, B. Li, Subgroup analysis via recursive partitioning, Journal of Machine Learning Research 10 (2) (2009).
[10] S. Athey, G. Imbens, Recursive partitioning for heterogeneous causal effects, Proc. Nat. Acad. Sci. 113 (27) (2016) 7353–7360.
[11] A. Johnson, L. Bulgarelli, T. Pollard, S. Horng, L. Celi, R. Mark, Mimic-iv (version 1.0, PhysioNet (2021).
[12] T.J. Hastie, R.J. Tibshirani, Generalized additive models, Vol. 43, CRC Press, 1990.
[13] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, N. Elhadad, Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission, in: Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, 2015, pp. 1721–1730.
[14] H. Nori, S. Jenkins, P. Koch, R. Caruana, Interpretml: A unified framework for machine learning interpretability, arXiv preprint arXiv:1909.09223 (2019).
[15] R. Agarwal, N. Frosst, X. Zhang, R. Caruana, G.E. Hinton, Neural additive models: Interpretable machine learning with neural nets, arXiv preprint arXiv:2004.13912 (2020).
[16] P. PhysioBank, Physionet: components of a new research resource for complex physiologic signals, Circulation 101 (23) (2000) e215–e220.
[17] N. Lassau, S. Ammari, E. Chouzenoux, H. Gortais, P. Herent, M. Devilder, S. Soliman, O. Meyrignac, M.-P. Talabard, J.-P. Lamarque, et al., Integrating deep learning ct-scan model, biological and clinical variables to predict severity of covid-19 patients, Nature communications 12 (1) (2021) 1–11.
[18] F.A. Lagunas-Rangel, Neutrophil-to-lymphocyte ratio and lymphocyte-to-c-reactive protein ratio in patients with severe coronavirus disease 2019 (covid-19): a meta-analysis, Journal of medical virology 92 (10) (2020) 1733–1734.
[19] S. Jimeno, P.S. Ventura, J.M. Castellano, S.I. García-Adasme, M. Miranda, P. Touza, I. Lllana, A. López-Escobar, Prognostic implications of neutrophil-lymphocyte ratio in covid-19, Eur. J. Clin. Invest. 51 (1) (2021) e13404.
[20] A. Lania, M.T. Sandri, M. Cellini, M. Mirani, E. Lavezzi, G. Mazziotti, Thyrotoxicosis in patients with covid-19: the thyrcov study, European journal of endocrinology 183 (4) (2020) 381–387.
[21] E. González-Reimers, F. Santolaria-Fernández, M.C. Martín-González, C. M. Fernández-Rodríguez, G. Quintero-Platt, Alcoholism: a systemic proinflammatory condition, World Journal of Gastroenterology: WJG 20 (40) (2014) 14660.
[22] T.T. Vu, J.C. Fredenburgh, J.I. Weitz, Zinc: an important cofactor in haemostasis and thrombosis, Thrombosis and haemostasis 109 (03) (2013) 421–430.
[23] R.L. Flumignan, J.D. de Sá Tinôco, P.I. Pascoal, L.L. Areias, M.S. Cossi, M.I. Fernandes, I.K. Costa, L. Souza, C.F. Matar, B. Tendal, et al., Prophylactic anticoagulants for people hospitalised with covid-19, Cochrane Database of Systematic Reviews (10) (2020).

[24] P. Horby, W. Lim, J. Emberson, M. Mafham, J. Bell, L. Linsell, N. Staplin, C. Brightling, A. Ustianowski, E. Elmahi, et al., Dexamethasone in hospitalized patients with covid-19—preliminary report, N. Engl. J. Med. (2020).
[25] C.-H. Chang, S. Tan, B. Lengerich, A. Goldenberg, R. Caruana, How interpretable and trustworthy are gams?, in: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, 2021, pp. 95–105.
[26] Y. Zhang, A. Bellot, M. van der Schaar, Learning overlapping representations for the estimation of individualized treatment effects, in: S. Chiappa, R. Calandra (Eds.), Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics, Vol. 108 of Proceedings of Machine Learning Research, PMLR, 2020, pp. 1005–1014. URL: https://proceedings.mlr.press/v108/zhang20c.html.
[27] S. Horii, Heterogeneous treatment effect estimation based on a partially linear nonparametric bayes model, arXiv preprint arXiv:2201.12016 (2022).
[28] P.R. Rosenbaum, D.B. Rubin, The central role of the propensity score in observational studies for causal effects, Biometrika 70 (1) (1983) 41–55.
[29] I. Bica, A. Alaa, M. Van Der Schaar, Time series deconfounder: Estimating treatment effects over time in the presence of hidden confounders, in: H.D. III, A. Singh (Eds.), Proceedings of the 37th International Conference on Machine Learning, Vol. 119 of Proceedings of Machine Learning Research, PMLR, 2020, pp. 884–895. URL: https://proceedings.mlr.press/v119/bica20a.html.
[30] C. Louizos, U. Shalit, J.M. Mooij, D. Sontag, R. Zemel, M. Welling, Causal effect inference with deep latent-variable models, Adv. Neural Inform. Process. Syst. 30 (2017).
[31] B.J. Lengerich, M. Al-Shedivat, A. Alavi, J. Williams, S. Labbaki, E.P. Xing, Discriminative subtyping of lung cancers from histopathology images via contextual deep learning, medRxiv (2020).