Applying a cumulative ordinal regression model to infer possible biomarkers associated with response to PD-1/PD-L1 inhibition in Merkel cell carcinoma

Summary
Merkel cell carcinoma a is type of neuroendocrine skin cancer that in some cases can be treated with anti-PD-1/PD-L1 antibodies that act as immune checkpoint inhibitors and therefore enhance immune response against tumor cells. In an effort to identify biomarkers that distinguish treatment responders from non-responders, data of 114 patients had been collected and analyzed using a cumulative ordinal regression model. Conditioned on the model and the observed data, there is moderate statistical evidence that absence of immunosuppression, usage of anti-PD-1 antibodies (as opposed to anti-PD-L1 antibodies), and limited spread of the main tumor are associated with a higher probability of responding to the treatment.

Contents

1. Dataset

The dataset had been collected from 4 patients with Merkel cell carcinoma. All patients were either treated with anti-PD-1 or anti-PD-L1 antibodies and treatment response was classified into progressive disease (PD), stable disease (SD), partial response (PR) and complete response (CR). In total, there are 19 different predictors:

- gender (categorical: male, female)
- primary localisation (categorical: head + neck, occult, extremities, trunk)
- immunosuppression (binary)
- tumor PD-L1 expression (binary)
- MCPyV+ status (binary)
- prior chemotherapy (binary)
- prior radiotherapy (binary)
- checkpoint inhibition (categorical: PD-1, PD-L1)
- metastasic stage (categorical: M0, M1a, M1b/M1c)
- ≥ 2 organs involved (binary)
- elevated LDH levels (binary)
- elevated CRP levels (binary)
- neutrophil count at therapy start (numeric)
- lymphocyte count at therapy start (numeric)
- neutrophil/lymphocyte ratio (NLR) ≥ 4 (binary)
- ECOG performance status ≥ 1 (binary)
- age ≥ 70 years (binary)
- year of therapy start (ordered categorical)
- participation in a clinical trial (binary)

2. Introduction

To analyze these data, we fit a Bayesian model. This has several advantages:

A unique feature of Bayesian statistics is that it allows to describe model parameters with probability distributions. This means that instead of point estimates (with more or less reliable standard deviations) we obtain

a distribution of parameter values that is consistent with the observed data. In this way, it is possible to quantify the uncertainty of the estimates.

Additionally, Bayesian statistics provides an accessible way to test models: By comparing data generated under the model's assumptions to the actually observed data, it is possible to identify important aspects of the dataset that the model fails to capture, and subsequently improve the model until it is consistent with the observed data.

The dataset also contains missing values, and Bayesian statistics allows to incorporate data points where some of the inputs are missing in such a way that the uncertainty of the missing values directly translates into uncertainty of the estimates.

3. Model description

The treatment response (progressive disease, stable disease, partial response or complete response) is on an ordinal scale, which means that the different levels have an inherent order, e.g a partial response is clearly better than stable disease.

Unfortunately, statistical models that are applied to this kind of data often do not adequately account for this order: In practice, a common approach is to encode the categories as increasing integer values (1, 2, 3, …) and to apply a linear regression model. While this preserves the order, it assumes equal distances between the outcomes, which means that the distance between progressive- and stable disease is identical to the distance between partial response and complete response. Other common approaches involve multinomial models (which ignore the order) or fitting logistic regression models after arbitrarily binarizing the response.

All these approaches share the common problem that they do not make efficient use of the available data and might lead to over- or underestimated effect sizes [1]. For instance when binning the data into two categories (responder and non-responder), patients with a partial response have the same influence on the overall estimates as patients with a complete response. The grouping is often also arbitrary, for example in terms of survival time, our data shows that patients with stable disease are actually more similar to partial- and complete responders than to patients with progressive disease (data not shown).

3.1 Cumulative ordinal regression model

To circumvent the aforementioned issues, we apply a cumulative ordinal regression model, which adequately takes the order between the categories into account.

The main idea of the cumulative regression model is to regard the tendency of a patient to respond to the treatment as a latent (= unobserved) variable that is determined by the patient characteristics. By convention, this latent variable is usually assumed to be distributed according to a logistic distribution with a scale of 1 and a mean that is determined by the predictor values for that patient.

The logistic distribution with a scale of 1 closely resembles a normal distribution with a standard deviation of 1.6. In fact, the exact choice of distribution does not matter in practice and other choices are possible, e.g. using a standard normal distribution instead would lead to a class of models called probit models (whereas the cumulative regression model using the logistic distribution is a generalization of the logistic regression model for ordered responses with more than two categories).

The probabilities of the four possible response categories are then determined by three thresholds that are also estimated.

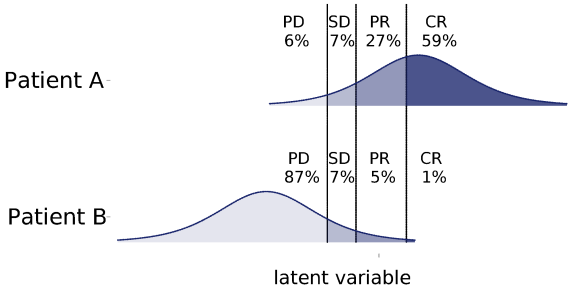

Figure 1. Distributions of the inferred latent variable for two patients in the dataset. PD: progressive disease, SD: stable disease, PR: partial response, CR: complete response.

Figure 1 shows the distribution of the latent variable for two patients in the dataset along with the 3 estimated thresholds (vertical lines). Patient A has predictor values that favor treatment response, whereas Patient B has predictor values that do not favor treatment response. The probability for each of the response categories is given by the area under the curve that is enclosed by the corresponding thresholds. Please note that, as we fit the model in a Bayesian fashion, neither the location of

the latent variables nor the tresholds are fixed values (as shown in Fig. 1), but follow some distribution of values that are consistent with the observed data.

In formula notation, the model can be written as:

$$\begin{aligned}
\mu_i &= \beta_{\text{age}} \cdot x_{\text{i, age}} + \beta_{\text{LDH}} \cdot x_{\text{i, LDH}} \dots \\
p(\text{PD})_i &= \int_{-\infty}^{\tau_1} \text{logistic_distribution}(\mu_i, 1) dx \\
p(\text{SD})_i &= \int_{\tau_1}^{\tau_2} \text{logistic_distribution}(\mu_i, 1) dx \\
p(\text{PR})_i &= \int_{\tau_2}^{\tau_3} \text{logistic_distribution}(\mu_i, 1) dx \\
p(\text{CR})_i &= \int_{\tau_3}^{\infty} \text{logistic_distribution}(\mu_i, 1) dx
\end{aligned} \tag{1}$$

,

where

- μ_i is the location of the latent variable for patient i,
- β_{age} is the β coefficient for the predictor age,
- $x_{\text{i, age}}$ is the indicator variable of patient i for age (in this case, 0 if patient i's age is $\geq$ 70 years, 1 otherwise),
- $p(\text{PD})_i$ is the probability of a progressive disease response
- τ_1, τ_2, τ_3 are the three estimated tresholds.

The β coefficients of the predictors are of main interest in this analysis, as they give information on whether or not a predictor is associated with a higher probability of responding to the treatment. A more comprehensive explanation of ordinal regression models that is also accessible without a background in statistics is given in [2].

3.2 Model fitting

Fitting the model to the dataset was done with the R software package 'brms' [3], which utilizes 'Stan' [4] in the background. Student t priors with 7 degrees of freedom and a standard deviation of 1 were chosen as weakly informative priors for the β coefficients. This is in line with the Stan prior choice recommendations [1]. The t distribution has a similar shape as the normal distribution, but with higher density in the tail areas. In this way, we rule out unreasonably large parameter values (e.g. anything larger than 10-15 for coefficients on the log-odds scale), but the model is still flexible enough to allow for values that might make sense. Model fit is performed numerically by Markov chain Monte Carlo. In total, 2000 samples from 4 different Markov chains were generated. We use the split-$\hat{R}$ diagnostic [5, 6] to identify possible Markov chain convergence issues. All parameters satisfied $\hat{R} < 1.01$, the effective sample size N_{eff} [7] exceeded 1000 in all cases.

[1] https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations

4. Model results

Figure 2 shows marginal posterior distributions of the estimated β coefficients. Values larger than 0 denote that these predictors favor a response to the treatment, whereas values less than 0 favor treatment non-response. The width of the distribution gives an impression of the uncertainty of the estimate: A distribution tightly concentrated around some value means that the dataset allows for a precise estimate of that parameter, while a broader distribution means that the data is consistent with a wide range of parameter values.

Please note that all these estimates are conditioned on the model and the observed data, which means that they are not a statement about the general population of patients with Merkel cell carcinoma.

Most of estimates include 0, which means that the absence of association between that predictor and the treatment response is a reasonable explanation for the observed data. The widths of the distributions are also broad, so while no effect is a possible explanation, it could also be quite large.

Notable exceptions are the predictors immunosuppression and organs involved, where most of the probability mass is located at values less than 0 (denoting they are associated with a decreased probability of treatment response); and the use of an anti-PD-1 antibody, where most of the probability mass is located at values greater than 0 (denoting it is associated with an increased probability of treatment response), as compared to checkpoint inhibition with an anti-PD-L1 antibody.

4.1 Average Predictive Comparisons

As with logistic regression models, the β coefficients of cumulative ordinal regression models are in units of log-odds, which means that a value of 1 of the corresponding predictor increases the expected log-odds of the next higher response category by 1.

This has the disadvantage that it is difficult to have an intuition about the effect size, i.e. whether an increase in the log-odds by 1 corresponds to a large, moderate or small change. To circumvent this issue, we show average predictive comparisons in addition to the regression coefficients.

Briefly, average predictive comparisons are calculated as the expected change in the response associated with a unit difference in one of the inputs. A technical description of average predictive comparisons is given in [8]. In this analysis, it allows us to reinterpret the regression coefficients (which are in units of log-odds) into a summary that is on the probability scale.

Figure 2 shows average predictive comparisons for each of the model inputs. They were calculated with respect to having at least a partial response to the treatment, e.g. for immunosuppression, values between -0% and -40% denote that comparing a patient with immunosuppression to an otherwise identical patient without immunosuppression, the patient with immunosuppression has (on average) a 0% to 40% lower probability of having at least a partial response to treatment.

As a result of the limited sample size, the uncertainty around the estimates is rather large. It also shows that just because a predictor includes 0%, it should not be confused with having no association with the treatment response, because the data is still consistent with large effect sizes in either direction.

5. Imputing missing values

The dataset contains missing values in some of the predictors. Standard practice is usually to delete them, either by row-wise exclusion (removing all samples that contain any missing value), or by removing the predictors that contain missing values.

With only 114 patients, removing all samples that contain missing values would mean to remove import information.

As a more sensible approach, multiple imputation with the R package mice [9] was used instead. In multiple imputation, several imputed versions of the dataset are created where the missing values are replaced with plausible values. The imputed datasets are identical for the non-missing entries, but differ in the imputed values. The uncertainty about the missing values is reflected in the degree of variation between the datasets.

To translate these different datasets into as single estimate, we simply fit the model independently on each dataset and combine the posterior samples of each model fit. In this way, the uncertainty of the missing values propagates directly into uncertainty of the estimates.

6. Model testing

A useful way to test Bayesian models is called posterior predictive check. In posterior predictive checks, the inferred parameter estimates are used to sample an arbitrary number of new datasets that are generated under the model's assumptions. By comparing these datasets to the actually observed dataset, it is possible to identify aspects of the data that the model fails to capture.

One possible way to perform a posterior predictive check for the model described here is to compare the observed proportion of the different treatment response categories to the proportion of treatment response categories expected under the model's assumptions. Figure 3 shows a histogram of the proportion of patients with progressive disease, stable disease, partial response and complete response in the generated datasets with the actually observed proportions highlighted in blue. The observed proportions lie directly in the center of what is expected by the model.

Another form of posterior predictive check focuses on individual predictions instead. For each patient, the expected probability that a given patient has at least a partial response to the antibody treatment is calculated. If the model produces reasonable estimates, we expect that patients with a higher estimated probability really do respond more frequently to the treatment than patients with a lower estimated probability.

Figure 4 shows a so-called calibration plot. All 114 patients were sorted according to their expected probability of having at least a partial response to the treatment and placed into 7 distinct bins. For each bin, the mean probability of having a partial response is plotted against the observed proportion of patients in that bin which at least partially respond to the treatment. As each uncertainty interval around the observed proportion touches the diagonal line, this type of posterior predictive check shows again no large discrepancies between expected and observed data.

In conclusion, the posterior predictive checks show that the clinical data is consistent with data expected under the model's assumptions.

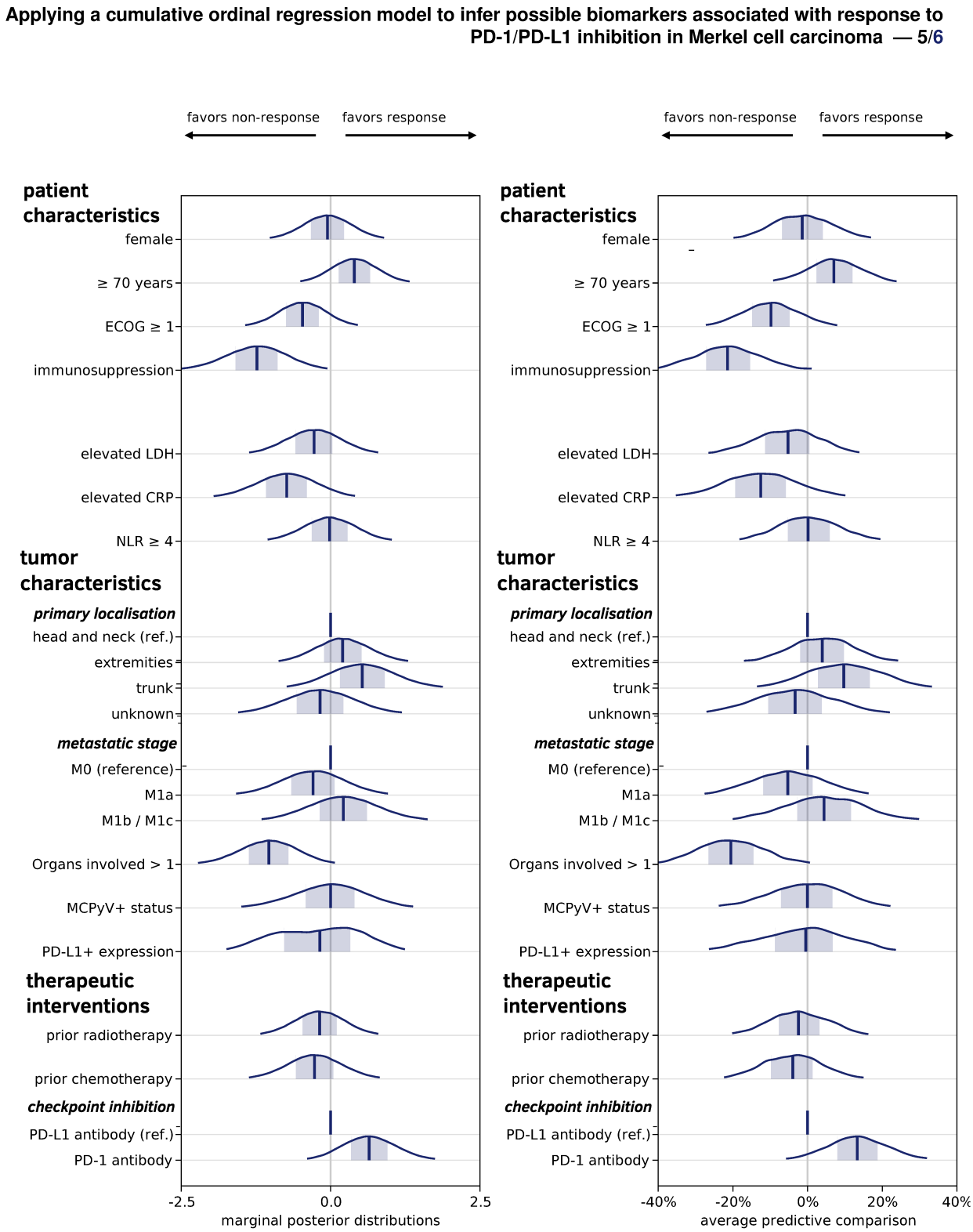

Figure 2. Marginal posterior distributions of β coefficients (left) and average predictive comparisons of the expected probability of having at least a partial response to the treatment (right). The regression coefficients of the full model have been projected onto a probability scale.

Applying a cumulative ordinal regression model to infer possible biomarkers associated with response to PD-1/PD-L1 inhibition in Merkel cell carcinoma — 6/6

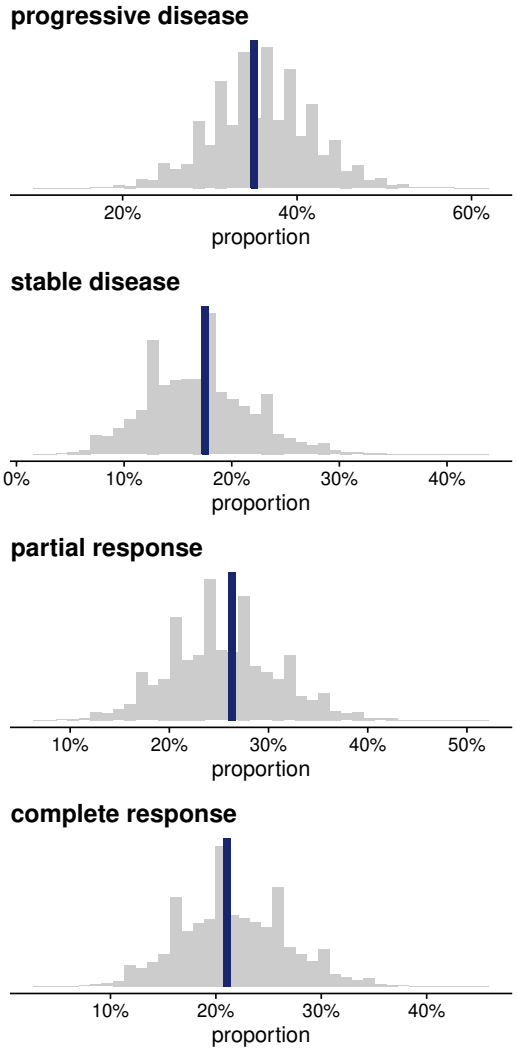

Figure 3. Expected proportion of each response category under the model (histogram) vs. observed proportion (blue line).

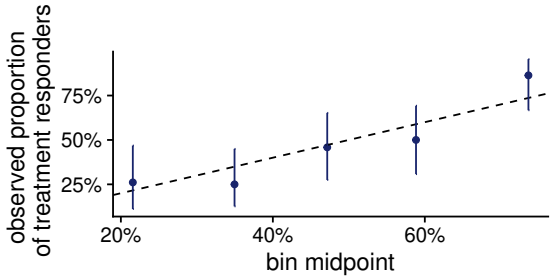

Figure 4. Calibration plot of proportion of patients with at least a partial response to the treatment.

References

[1] Torrin M. Liddell and John K. Kruschke. Analyzing ordinal data: Support for a Bayesian approach. *SSRN Electronic Journal*, 2015.

[2] Paul-Christian Bürkner and Matti Vuorre. Ordinal regression models in psychology: A tutorial. 2:77–101, 2019.

[3] Paul-Christian Bürkner. brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1), 2017.

[4] Bob Carpenter, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1), 2017.

[5] Andrew Gelman and Donald B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472, nov 1992.

[6] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. CRC press, 2013.

[7] Stan Development Team. *Stan Modeling Language Users Guide and Reference Manual, 2.23. https://mc-stan.org*, 2021.

[8] Andrew Gelman and Iain Pardoe. 2. Average Predictive Comparisons for models with nonlinearity, interactions, and variance components. *Sociological Methodology*, 37(1):23–51, aug 2007.

[9] Stef van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3), 2011.