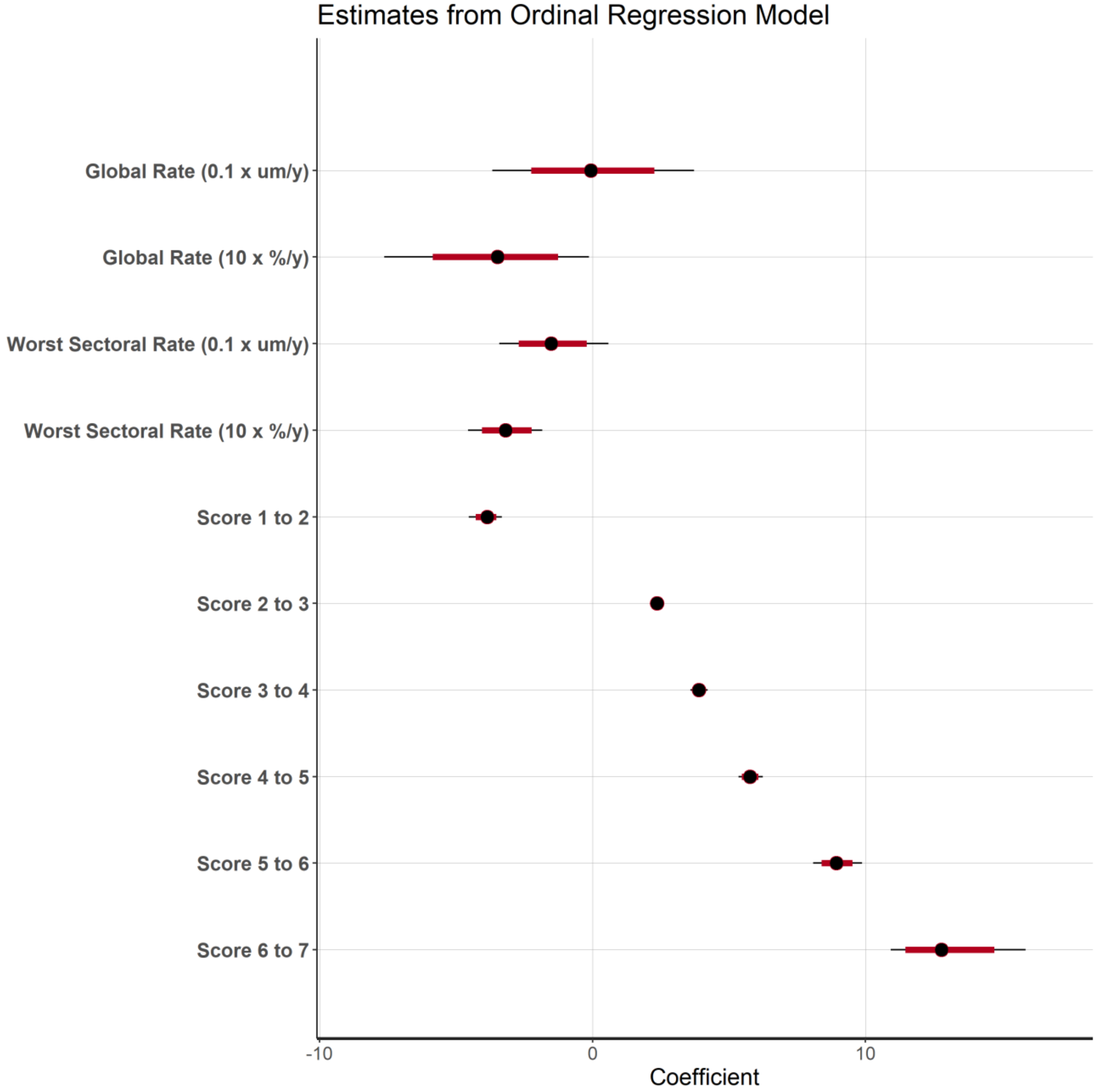
Supplemental Figure 1



The primary analysis in the paper uses as its outcome measure the mean score from 11 clinicians, each of whom assigned integer scores between 1 (improvement) and 7 (very rapid progression) to each series. This implicitly assumes that the scores can be considered as existing on a continuous linear scale, whereby the difference between levels is constant. To test this assumption, an ordinal regression model was formed, to predict the scores of individual clinicians, using four predictors; the rate of global change of retinal nerve fiber layer thickness (RNFLT) expressed in µm/y and in %/y, and the rate in the fastest changing sector expressed in µm/y and in %/y. The ordinal regression model asssumes that there exists a continuous latent variable expressing clinician opinions, and the assigned scores 1-7 represent seven ordered but unknown values of that latent variable. The primary (linear) analysis is then implicitly assuming that these seven unknown values are evenly-spaced, and could be estimated wth equal precision for a given sample size.

The four predictors were first transformed to be of similar magnitude, as indicated on the figure. A clinician-level random effect was incorporated to allow for different mean scores and variability for each clinican. To constrain model parameters and facilitate convergence, a weakly informative *N*(0, 10) Bayesian prior was used for each predictor. The model was then fit using the rstan package (Stan Development Team (2020). “RStan: the R interface to Stan.” R package version 2.21.2, http://mc-stan.org/). The figure shows the point esimate of each coefficient, with thick horizontal error bars showing the 80% credible interval, and thin horizontal bars showing the 95% credible interval (the Bayesian equivalent of a 95% confidence interval for that parameter). If a linear model perfectly fit the data, then the transitions between consecutive levels (shown above as “Score 1 to 2”, “Score 2 to 3” etc.) would be evenly spaced, and estimated with equal uncertainty (i.e. no evidence of heteroscedasticity).

The differences in the widths of the credible intervals in the figure can be explained by the fact that only one score of 7 and fifteen scores of 6 were given by clinicians (compared to 806 scores of 2, and 242 scores of 3). The sample sizes for estimating the transitions from score 5 to 6 and from score 6 to 7 are much smaller than for estimating the transition from score 2 to 3. Thus the assumption of homoscedasticity appears reasonable. The assumption that the scores represent evenly-spaced values of the latent variable was mostly supported, but the difference between scores 1 (“improved”) and 2 (“stable”) appeared to be greater than the difference between other consecutive levels. This would be consistent with clincians being reluctant to say that a series had improved, thus requiring stronger evidence before assigning a score of 1. Thus, the primary linear analysis was repeated excluding series for which at least one clinician assigned a score of 1. This did not materially affect any results or conclusions, and so the primary analysis assuming a continuous linear scale was retained for greater statistical power and easier interpretation.