Augmenting Predictive Models in Forensic Psychiatry with Cultural Consensus Theory

Don van den Bergh[1], Erwin Schuringa[2], and Eric-Jan Wagenmakers[1]

[1]Department of Psychological Methods, University of Amsterdam
[2]Forensic Psychiatric Center Dr. S. van Mesdag

Contents

Abstract

1 Introduction

The mental health and forensic risk factors of patients in forensic psychiatric hospitals is regularly monitored with methods such as Routine Outcome Monitoring (de Beurs et al., 2011). A staff member (e.g., a clinician or psychiatrist), henceforth a *rater*, scores a patient on variety of criteria, such as problematic behavior (e.g., hostility) or protective behavior (e.g., coping skills). These scores are used to track the mental state of patients over time, to measure the effectiveness of treatment, and as a risk indicator for violent outbursts by patients.

Typically, multiple raters score each patient on different items. Standard practice is to average the scores across raters and use the averages to inform decisions. However, this is

suboptimal for multiple reasons (van den Bergh et al., 2020). For example, taking the average implies that raters are interchangeable and patients are independent, rather than taking into account raters' bias or patients' offense.

An improvement over averaging the scores is to use Cultural Consensus Theory (CCT; Batchelder & Anders, 2012; Batchelder & Romney, 1988; Romney et al., 1986) to construct an appropriate model for the scores that accounts for the hierarchical structure among patients, raters, and items. In previous work, we developed such a model based on the Latent Truth Rater model (LTM; Anders & Batchelder, 2015), and demonstrated that, in theory, this model predicted better than the average and several machine learning alternatives. However, due to the lack of an empirical dataset, we could not demonstrate whether the theoretical claims hold up in practice.

Here, we apply the CCT-based model to data of a Dutch maximum-security forensic psychiatric center and use its inferences to predict whether or not a patient becomes violent. First, we briefly introduce the LTM model used and discuss two changes made compared to van den Bergh et al. (2020). Afterward, we use the LTM to augment a logistic regression. We use the augmented model to predict violent outbursts in patients and compare the predictive performance to that of frequently used machine learning models. We find that our LTM approach outperforms all other methods, albeit by a small margin. Next, we interpret the fitted model and ... Finally, we discuss some practical limitations of the data set at hand and how future monitoring of patients could be adjusted to maximize the added benefit of our CCT-based approach.

2 Mesdag Data

DB: Edwin, het zou fijn zijn als jij naar deze sectie kan kijken!

2.1 Method of collection

The data were collected in the Dutch maximum-security Forensic Psychiatric Center Dr. S. van Mesdag in the period XXX to XXX. The individual records were retrospectively merged into a single data set.

DB: April 2010 t/m juli 2016 gok ik gebaseerd op Schuringa et al. 2016.

2.2 IFTE

The data were collected using a Routine Outcome Monitoring instrument called the the Instrument for Forensic Treatment Evaluation (IFTE). The IFTE consists of 22 items, of which 14 items are criminogenic need indicators of the Dutch risk assessment instrument HKT-R (Spreen et al., 2013), five items were designed in consultation with psychologists and psychiatrists, and three items are based on the Atascadero Skills Profile (Vess, 2001). The 22 items can be grouped into three factors, Protective behaviors, Problematic behaviors, and Resocialization Skills. The individual items are shown in Table 2. All items are scored on a 18 point scale.

2.3 Covariates

In addition to the IFTE, several other background variables about the patients are considered for predicting violent behavior. These are age (0-30, 31-40, 41-50, or 55+), treatment duration (0-2 years, 2-4 years, 4-6 years, or 6+ years), diagnosis (Autism spectrum disorder, Personality disorder, Personality disorder B, Schizophrenia or other psychotic disorders, As 1 overig), offence (Manslaughter, Aggravated assault, Murder, Arson, Violent property crime, Moderately violent property crime, Sexual offence), violent behavior before measurement 1, and violent behavior in between measurements 1 and 2.

DB: Check de juiste engelse vertalingen.

2.4 Descriptives

Figure 1 shows a histogram of the raw scores across all IFTE items, raters, and patients. It is clear that some scores are given more often than others, such as 1, 2, 6, 10, and 14. Furthermore, it appears that, with the exception of 1, even scores are given more often than odd scores.

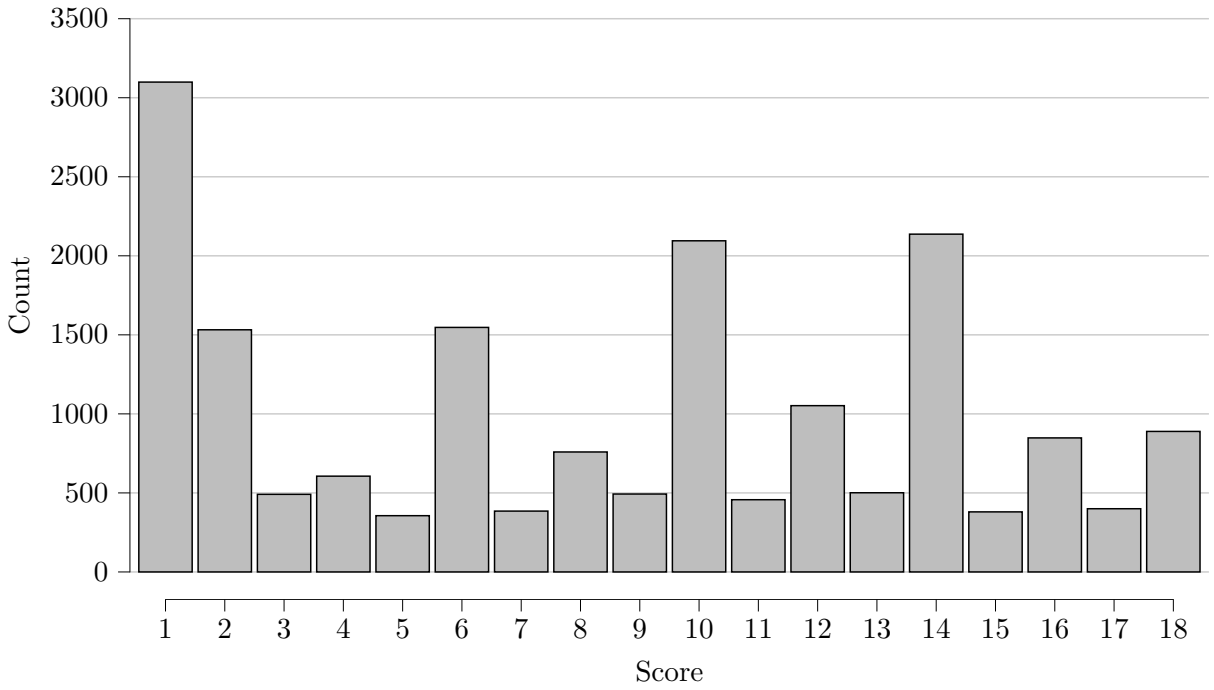


Figure 1: Histogram of observed scores across all patients, items, raters, and time points.

Not all raters rated every patient, in contrast, Figure 2 shows that the rater by patient matrix is rather sparse.

3 Cultural Consensus Theory

Cultural Consensus Theory, sometimes called "test theory without an answer key" (Batchelder & Romney, 1988), is a method to discover the "true answer" for items from the consensus among the responses. For example, suppose a patient is scored by multiple raters on aggressiveness. Multiple scores are obtained that need to be aggregated to arrive at a single score for this patient. The naive solution is to average these scores. However, as shown in Figure 3 averaging may lead to severely biased estimates. The average score disregards all additional information that is available. It ignores the individual differences between raters, for example, this assumes that all psychiatrists score aggressiveness in the same way, and it ignores group differences among raters, for example, there is no difference in scores by psychiatrists as opposed to clinicians, or other staff member. In addition, the average ignores any additional information about the patient at hand, such as offence and diagnosis.

Cultural consensus theory provides a model-based framework for pooling information from multiple raters to form a consensus (Anders et al., 2014). There exist a variety of CCT models, each applicable to different types of data. For example, the General Condorcet model (Batchelder & Romney, 1986) applies to dichotomous data, the Continuous Response model (Anders et al., 2014) is suited for continuous data, and the Latent Truth Rater model (Anders

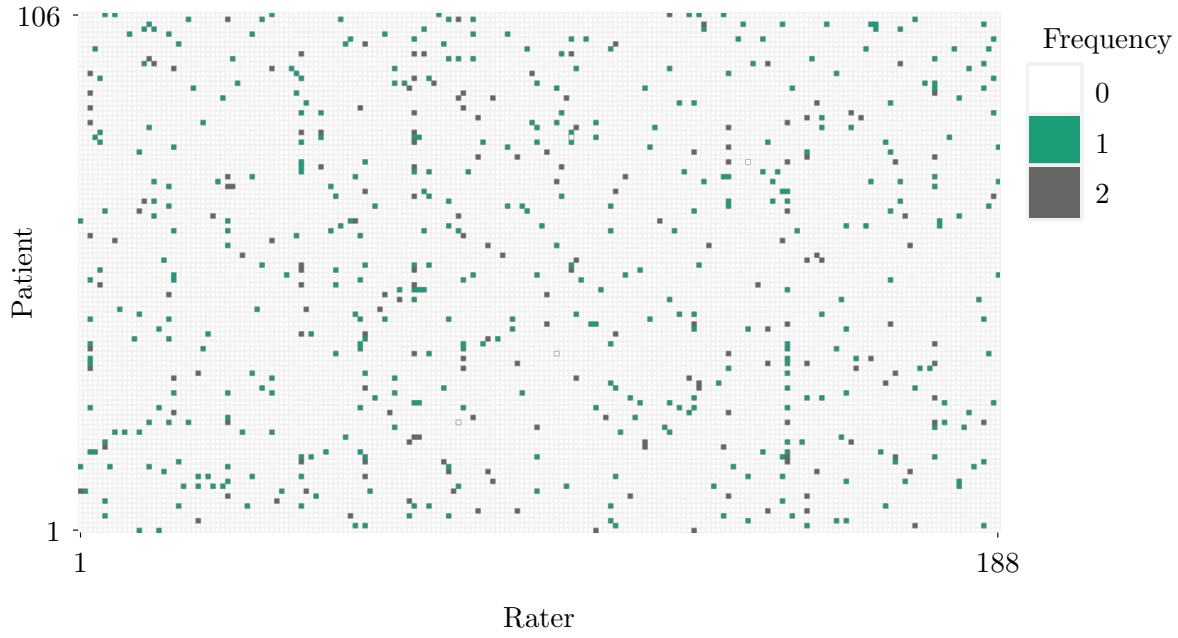


Figure 2: Heatmap of observed scores of rater (x-axis) against patients (y-axis).

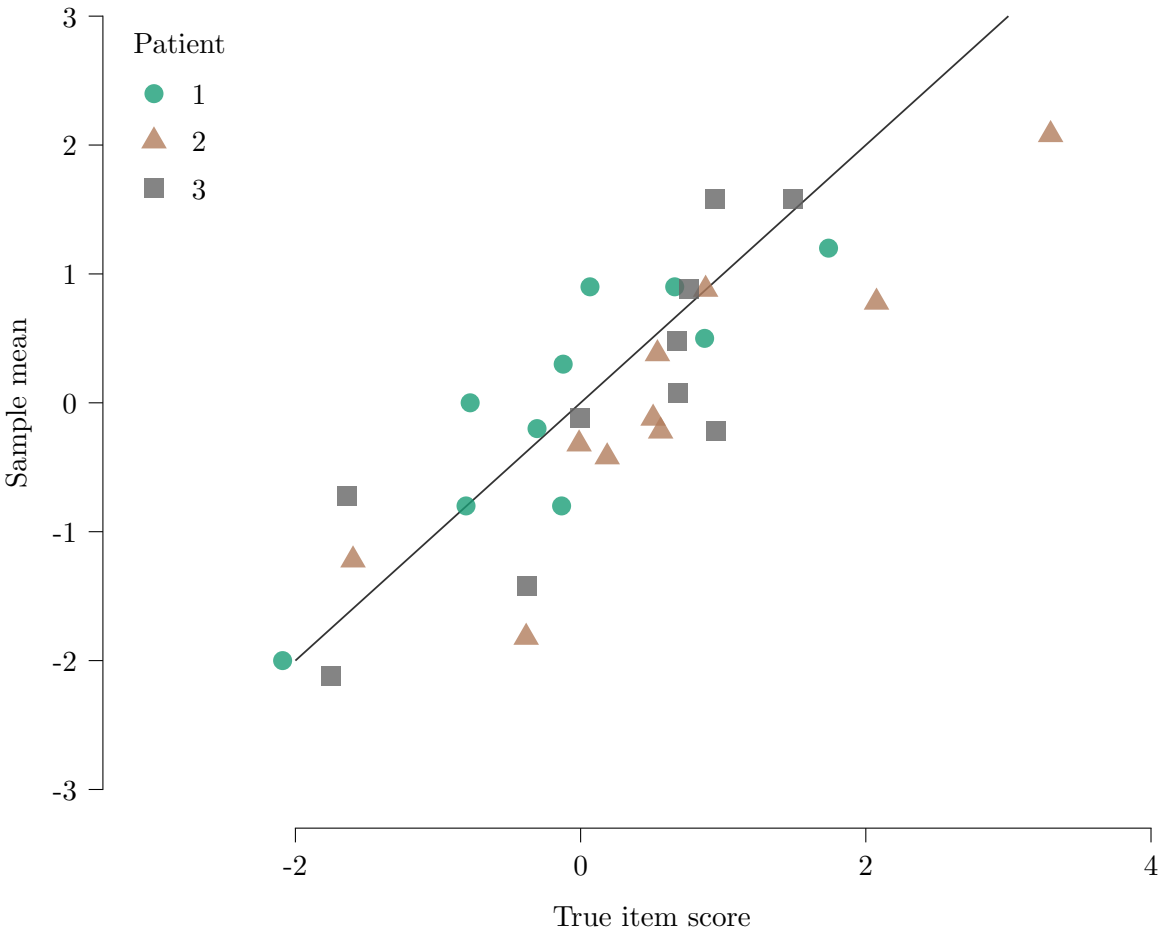
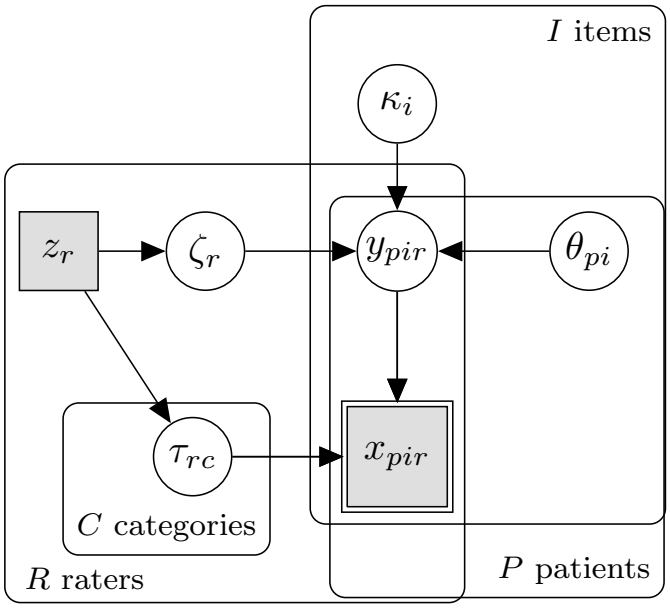


Figure 3: True item scores (x-axis) versus the sample mean across raters (y-axis).

& Batchelder, 2015) is suited for ordinal data. As the ratings on the are ordinal, we use the Latent Truth Rater model to analyze the Mesdag data.

3.1 The Latent Truth Rater Model

The Latent Truth Rater Model (LTM) is a CCT model for ordinal data (Anders & Batchelder, 2015). Previously, we extended the LTM to handle data from multiple patients (van den Bergh et al., 2020) and Figure 4 shows the LTM for multiple patients.



$$
\begin{aligned}
x_{pir} &= \begin{cases} 1 & \text{if } y_{pir} \leq \tau_{r1} \\ c & \text{if } \tau_{r,c-1} < y_{pir} \leq \tau_{rc} \\ C & \text{if } y_{pir} > \tau_{r,C-1} \end{cases} \\
y_{pir} &\sim \text{Logistic}\left(\theta_{pi},\, \kappa_i/\zeta_r\right) \\
\tau_{rc} &\sim \text{Normal}\left(0,\, 1\right) \\
\theta_{pi} &\sim \text{Normal}\left(\mu_\theta,\, \sigma^2_\theta\right) \\
\kappa_i &\sim \text{Gamma}\left(\mu^2_\kappa/\sigma^2_\kappa,\, \mu_\kappa/\sigma^2_\kappa\right) \\
\zeta_r &\sim \text{Gamma}\left(\mu^2_{\zeta_r}/\sigma^2_{\zeta_r},\, \mu_{\zeta_r}/\sigma^2_{\zeta_r}\right)
\end{aligned}
$$

Figure 4: Graphical model of the Latent Truth Rater Model for multiple patients. Note that the thresholds $\boldsymbol{\tau}_r$ are constrained to be ordered, that is, for all raters we have $\tau_{r1} \leq \cdots \leq \tau_{rc} \leq \cdots \leq \tau_{r,C-1}$.

Here, x_{pir} is the observed score given to patient p on item i by rater r. This score is assumed to be deterministically generated from a continuous latent appraisal y_{pir} that is discretized to an ordinal scale by the thresholds τ_{rc}. In particular, we have that

$$
x_{pir} = \begin{cases} 1 & \text{if } y_{pir} \leq \tau_{r1} \\ c & \text{if } \tau_{r,c-1} < y_{pir} \leq \tau_{rc} \\ C & \text{if } y_{pir} > \tau_{r,C-1} \end{cases}
$$

Since the appraisal score is latent, the deterministic function above implies the following probabilistic model over the observed scores:

$$
P(x_{pir} \mid y_{pir}, \boldsymbol{\tau}_r) = \begin{cases} 1 - F\left(y_{pir} - \tau_{r1}\right) & \text{if } x_{pir} = 1, \\ F\left(y_{pir} - \tau_{r,c-1}\right) - F\left(y_{pir} - \tau_{rc}\right) & \text{if } 1 < x_{pir} < C, \\ F\left(y_{pir} - \tau_{r,C-1}\right) & \text{if } x_{pir} = C. \end{cases}
$$

where $F\left(\right)$ is the logistic cumulative distribution function.[1]

Next, we explain how the latent appraisals and thresholds come about. The appraisals are draws from a logistic distribution with location θ_{pi}, the true score for patient p on item i. The scale of the logistic distribution is the ratio of the item difficulty κ_i to the rater competence ζ_r. A higher item difficulty means that the appraisals are more noisy, which leads to a more

[1]Note that this choice is arbitrary and that it is possibly to use any continuous cumulative distribution function.

dispersed probability distribution over possible scores. Conversely, a higher rater competence means that the appraisals are less noisy, which leads to a more concentrated distribution over the outcomes. There are $C-1$ ordered thresholds for each rater, which are assigned a standard normal prior for identification purposes.

There are two differences in the model specification above compared to our previous work (van den Bergh et al., 2020). First, we previously modeled the thresholds using two rater specific parameters. However, in simulations we noticed that these two parameters provide too little flexibility when the ordinal scale consists of 18 categories and has a multimodal distribution (see Figure 1), as in the Mesdag data. Therefore we decided to model the thresholds individually. This complicates interpreting the differences between the thresholds across raters, however, that is also not the goal of this paper. Second, we previously allowed the item difficulty parameter to vary across patients, which captures that some items may be more difficulty or easy to assess for some patients (e.g., some patients may cooperate more than others). This parameter is mainly informed by the number or raters and there needs to be a sufficient amount of raters that score each patient for a reliable estimate. However, when simulating data with a ratio of raters to patients similar to that in the data at hand that we noticed that there are simply too few observations to reliably estimate the deviations in item difficulty across patients. Therefore, we only vary item difficulty across items and not across patients.

3.2 Augmenting Logistic Regression with the LTM

In a next step, we use logistic regression for to predict violent behavior, where we use the results from the LTM as additional predictors. We do so in a fully Bayesian approach, that is, we constructed a joint model for the violent behavior and then patient ratings.[2] Figure 5 shows a graphical model of the logistic regression combined with the LTM. The latent truth for each patient on each item is seen as a covariate in the logistic regression model.

3.3 Implementation

We estimated the parameters of the LTM and the combined LTM – Logistic regression using a Bayesian approach. To explore the posterior distributions of the model parameters we used Stan (Carpenter et al., 2017). Rather than Markov chain Monte Carlo (MCMC) we used variational inference, as variational inference was computationally fast while providing similar results in terms of parameter retrieval and model predictions (Kucukelbir et al., 2017). All analyses were done using R (R Core Team, 2022) and Stan models were run using the R package `cmdstanr` (Gabry & Češnovar, 2022). Code for the analyses is available in the online appendix at https://github.com/vandenman/CCT-Logistic.

4 Predictive Performance

Here we compare the predictive performance of the Logistic regression augmented with the LTM (LR-LTM) to several reasonable alternatives and three baseline models. As a first baseline model we use an intercept only logistic regression (LR-Intercept), which quantifies if we can outperform the base-prevalence of violence. The second baseline model we use is a logistic regression model with all covariates but not the IFTE items (LR-No IFTE). As a third baseline model, we use an logistic regression model with only prior violence as predictors (LR-Violence). As more reasonable alternatives, we consider logistic regression (LR), random forest, and boosted regression trees (GBM) which have access to all history of violence, patient covariates, and IFTE scores. These four alternative have in common that they are designed for purely rectangular

[2] An alternative is to fit the LTM separately and then use e.g., the posterior means in a second step in a logistic regression. While this would be more computationally efficient, it would also ignore the uncertainty in the analysis of the patient ratings.

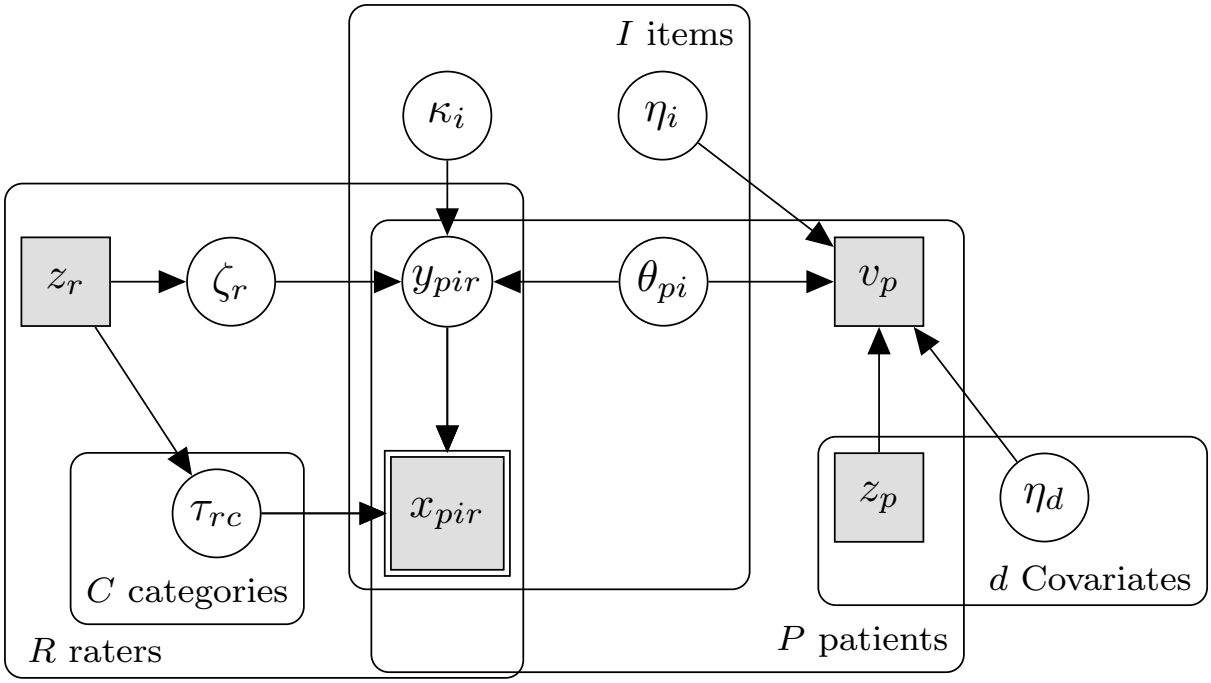


Figure 5: Graphical model of Logistic Regression Augmented with the Latent Truth Rater Model.

data. That is, each row of the data set contains one outcome (aggressive behavior or not) and a number of predictors. However, the raw data of the IFTE contains repeated observations, since patients were rated multiple times by different raters. To accommodate this we average across different raters to obtain a single scores for each item and time point. By comparing these models we aim to answer the following three questions: (1) Can predictive models outperform the base-prevalence of violent behavior? (2) Do models with the IFTE scores perform better than the baseline models without? (3) Does the LR-LTM perform better than the models that naively average across raters?

To examine predictive performance we used 10-fold stratified cross validation. Each fold consisted of 8 nonviolent observations and 2 or 3 violent observations. To quantify model performance, we used two metrics, prediction accuracy and the Brier score. Prediction accuracy is defined as the fraction of correct predictions. We converted model probabilities for violent or non-violent behavior into binary predictions by comparing with 0.5. In other words, if a model prediction for observation i, $\hat{y}_i \in [0, 1]$, is larger than 0.5 then the predicted label is violent, otherwise it is non-violent. The Brier score is defined as $N^{-1} \sum_{i=1}^{N} (\hat{y}_i - y_i)^2$, i.e., the mean squared error between the observed labels, $y \in \{0, 1\}$ and the model predictions.

Table 1 shows the prediction accuracy and the Brier score averaged over the 10-folds. The LR-LTM has the best classification of violence and the lowest Brier score, and the LR-Violence performs second-best. The difference in classification performance is $0.894 - 0.866 = 0.028$, which for a data set of 104 patients implies that the LR-LTM makes more accurate predictions than the LR-Violence for about 3 patients. Possibly striking is the poor performance of the LR. The standard logistic regression model clearly suffers from overfitting, as indicated by the high training performance but poor test performance. This result makes sense as the standard logistic regression model does not do anything special to combat overfitting, unlike the machine learning alternatives or Bayesian logistic regression used by the LR-LTM.

Figure 6 show the Receiver Operating Characteristic (ROC) curve and area under the curve (AUC) averaged across cross-validation runs for all methods except the LR-Intercept.[3] For each

[3]The ROC for the intercept only model is by definition the identity function with an area under the curve of

Table 1: Predictive performance of violent behavior. The values are the average of 10 cross validations; the standard deviation is shown in parentheses.

	Classification		Brier score	
Method	Train	Test	Train	Test
LR-LTM	0.980 (0.017)	0.894 (0.074)	0.043 (0.005)	0.093 (0.034)
LR-Violence	0.865 (0.007)	0.866 (0.063)	0.100 (0.003)	0.107 (0.030)
LR-No IFTE	0.948 (0.034)	0.837 (0.105)	0.032 (0.020)	0.142 (0.109)
Random forest	1.000 (0.000)	0.835 (0.084)	0.032 (0.002)	0.114 (0.033)
GBM	0.867 (0.009)	0.828 (0.073)	0.095 (0.002)	0.126 (0.020)
LR	1.000 (0.000)	0.735 (0.144)	0.000 (0.000)	0.257 (0.147)
LR-Intercept	0.769 (0.004)	0.771 (0.038)	0.177 (0.002)	0.177 (0.020)

cross-validation run, we computed the true positive rate and false positive rate with the same set of thresholds run and afterward we averaged these. The AUC was obtained by averaging the AUCs of each individual cross-validation, rather than computing the AUC for the averaged ROC curve. In line with the previous results, the LR-LTM performs best and the LR-Violence method performs second best.

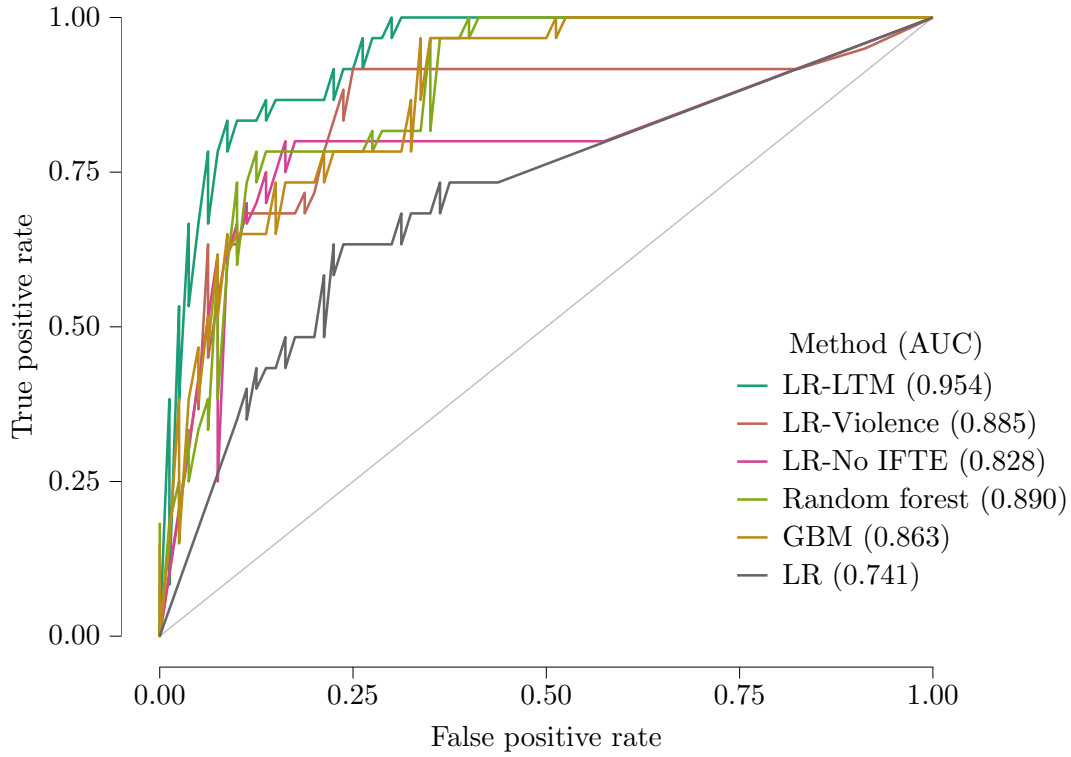


Figure 6: ROC curves for the considered methods. The legend shows the area under the curve in parentheses.

To summarize, it is evident that all models, except for logistic regression, outperform the intercept-only baseline model. Furthermore, the results show that the LTM augmented logistic regression model performs best, albeit by a small margin. The logistic regression model with only violence (LR-Violence) outperformed the two machine learning alternatives that naively

0.5.

averaged the scores from the IFTE (GBM and Random forest). This indicates that the IFTE has added value for prediction, but only if it is analyzed properly.

5 Interpretation of the LTM

Figure 7 shows the posterior means and 95% credible intervals for the LTM fitted to the complete data. Prior history of violence has the largest coefficients, whereas the other coefficients all appear close to zero. The influence of all items in the IFTE at both time points appears to be

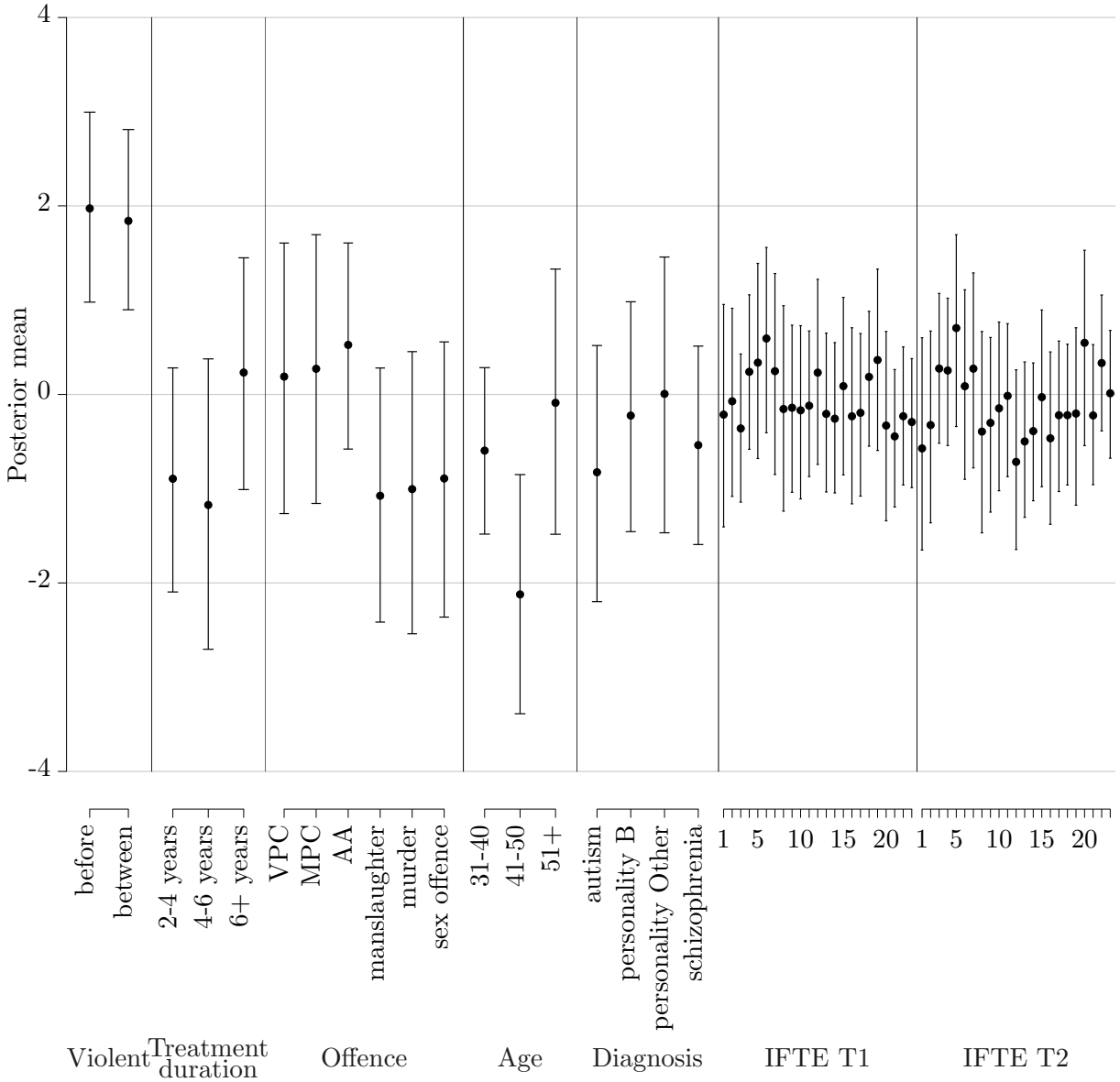


Figure 7: Posterior means and 95% credible intervals for the coefficients of the logistic regression model (η_d and η_i in Figure 5). The reference categories are treatment duration 0-2 years, diagnosis As 1 overig, and offence arson. The abbreviations VPC, MPC, and AA stand for violent property crime, moderately violent property crime, and aggravated assault.

close to zero. However, unlike the other coefficients the IFTE coefficients should be interpreted as random effects as they are multiplied with the estimated latent truth θ_{pi} which varies across patients.

6 Discussion

Here we applied a Cultural Consensus Theory model to scores of patients in a Forensic Psychiatric Center. We used this CCT model to augment the predictive performance of a logistic regression model and showed that our CCT-infused logistic regression model outpredicted all other candidate models.

An alternative approach to Hybrid is probably most predictive.

6.1 Suggestions for Future Data Collection

A limitation of the CCT approach applied here is that the use of unconstrained thresholds makes the model parameters hard to interpret. However, this modeling decision was mandated by the structure in the data, which, due to the large number of response categories, showed patterns that could not be described by a simpler function for the thresholds (e.g., a preference for even scores).

Furthermore, the frequency of measurements is key to the usefulness of the IFTE scores. The scores' predictive value likely decreases with time. Rather than scoring patients every 6 months, as in the data at hand, it would make more sense to rate them every few weeks. Although it may be practically difficult to score patients regularly, there are opportunities to use self-reports for this purpose (Bousardt et al., 2016; Tuente et al., 2021).

6.2 Limitations

A key limitation of our approach here is that patients' violent behavior is not a discrete event as we modeled it, but rather a continuous process where the risk of violent behavior changes over time. This limits the direct applicability of our approach in practice. To accommodate time series data, the model could be be extended, for example, by adding autoregressive components.

The predictive comparison is possibly incomplete as likely overfitting

6.3 Conclusion

In sum, we applied the LTM introduced by Anders and Batchelder (2015) and adapted previously by us in van den Bergh et al. (2020) to data of patients in a Forensic Psychiatric Center. We showed that including the IFTE items slightly improves predictive performance, but only if the scores from different raters are analyzed properly and not when the scores of different raters are averaged.

References

Anders, R., & Batchelder, W. H. (2015). Cultural consensus theory for the ordinal data case. *Psychometrika*, *80*(1), 151–181.

Anders, R., Oravecz, Z., & Batchelder, W. H. (2014). Cultural consensus theory for continuous responses: A latent appraisal model for information pooling. *Journal of Mathematical Psychology*, *61*, 1–13.

Batchelder, W. H., & Anders, R. (2012). Cultural consensus theory: Comparing different concepts of cultural truth. *Journal of Mathematical Psychology*, *56*(5), 316–332.

Batchelder, W. H., & Romney, A. K. (1986). The statistical analysis of a general Condorcet model for dichotomous choice situations. In G. O. (B. Grofman (Ed.), *Information pooling and group decision making: Proceedings of the second University of California Irvine conference on political economy* (pp. 103–112). JAI Press Greenwich, CN.

Batchelder, W. H., & Romney, A. K. (1988). Test theory without an answer key. *Psychometrika*, *53*(1), 71–92.

Bousardt, A. M., Hoogendoorn, A. W., Noorthoorn, E. O., Hummelen, J. W., & Nijman, H. L. (2016). Predicting inpatient aggression by self-reported impulsivity in forensic psychiatric patients. *Criminal behaviour and mental health*, *26*(3), 161–173.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*, 1–32. https://doi.org/10.18637/jss.v076.i01

de Beurs, E., den Hollander-Gijsman, M. E., van Rood, Y. R., van der Wee, N. J. A., Giltay, E. J., van Noorden, M. S., van der Lem, R., van Fenema, E., & Zitman, F. G. (2011). Routine outcome monitoring in the Netherlands: Practical experiences with a web-based strategy for the assessment of treatment outcome in clinical practice. *Clinical Psychology & Psychotherapy*, *18*(1), 1–12.

Gabry, J., & Češnovar, R. (2022). *Cmdstanr: R interface to 'cmdstan'* [https://mc-stan.org/cmdstanr/, https://discourse.mc-stan.org].

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., & Blei, D. M. (2017). Automatic differentiation variational inference. *Journal of machine learning research*.

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Romney, A. K., Weller, S. C., & Batchelder, W. H. (1986). Culture as consensus: A theory of culture and informant accuracy. *American Anthropologist*, *88*(2), 313–338.

Schuringa, E., Spreen, M., & Bogaerts, S. (2014). Inter-rater and test-retest reliability, internal consistency, and factorial structure of the instrument for forensic treatment evaluation. *Journal of Forensic Psychology Practice*, *14*(2), 127–144.

Spreen, M., Brand, E., Ter Horst, P., & Bogaerts, S. (2013). Handleiding HKT-R. *Historische Klinische Toekomst–Revisie*.

Tuente, S. K., Bogaerts, S., & Veling, W. (2021). Mapping aggressive behavior of forensic psychiatric inpatients with self-report and structured staff-monitoring. *Psychiatry research*, *301*, 113983.

van den Bergh, D., Bogaerts, S., Spreen, M., Flohr, R., Vandekerckhove, J., Batchelder, W. H., & Wagenmakers, E.-J. (2020). Cultural consensus theory for the evaluation of patients' mental health scores in forensic psychiatric hospitals. *Journal of Mathematical Psychology*, *98*, 102383.

Vess, J. (2001). Development and implementation of a functional skills measure for forensic psychiatric inpatients. *The Journal of Forensic Psychiatry*, *12*(3), 592–609.

A IFTE items

Table 2: Overview of the 22 IFTE Items, the factor on which they load, and origin of the question. Adapted from Table 1 of Schuringa et al. (2014).

Item description	Factor
Does the patient show problem insight?	Protective behaviors
Does the patient cooperate with your treatment?	Protective behaviors
Does the patient admit and take responsibility for the crime(s)?	Protective behaviors
Does the patient show adequate coping skills?	Protective behaviors
Does the patient have balanced daytime activities?	Resocialization Skills
Does the patient show sufficient labor skills?	Resocialization Skills
Does the patient show sufficient common social skills?	Resocialization Skills
Does the patient show sufficient skills to take care of oneself?	Resocialization Skills
Does the patient show sufficient financial skills?	Resocialization Skills
Does the patient show impulsive behavior?	Problematic behavior
Does the patient show antisocial behavior?	Problematic behavior
Does the patient show hostile behavior?	Problematic behavior
Does the patient show sexual deviant behavior?	Problematic behavior
Does the patient show manipulative behavior?	Problematic behavior
Does the patient comply with the rules and conditions of the center and/or the treatment?	Problematic behavior
Does the patient have antisocial associates?	Problematic behavior
Does the patient use his medication in a consistent and adequate manner?	Protective behaviors
Does the patient have psychotic symptoms?	Problematic behavior
Does the patient show skills to prevent drug and alcohol use?	Protective behaviors
Does the patient use any drug or alcohol?	Problematic behavior
Does the patient show skills to prevent physical aggressive behavior?	Protective behaviors
Does the patient show skills to prevent sexual deviant behavior?	Protective behaviors