**Sample size analysis for prediction**

We calculated the minimum sample size for clinical prediction models (Riley, Snell et al. 2019, Riley, Snell et al. 2019). To predict quantitative outcomes, such as the PainImpact, the regression model was used in the calculation as the standard. Since the PainImpact range is 8-50, we assumed the mean is about 29, and the standard deviation is about 20. The minimum sample size was calculated to satisfy all four recommended criteria:

1. Small overfitting is defined by an expected shrinkage of predictor effects by 10% or less.
2. Small absolute difference of 0.05 in the model's apparent and adjusted R-squared value.
3. Precise estimation of the residual standard deviation with a multiplicative margin of error (MMOE) less than 1.1.
4. Precise estimation of the average outcome value within 95% confidence interval.

Table 1 gives the minimum sample sizes calculated based on the coefficient of determination (R2) and the number of parameters to be used in the predictive model.

Table 1: Minimum sample sizes over $R^2$ values and the numbers of parameters to be used in predictive regression model for quantitative outcomes.

| | Number of parameters | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $R^2$ | 10 | 15 | 20 | 25 | 30 | 35 | 40 | 45 | 50 |
| 0.4 | 244 | 249 | 312 | 400 | 488 | 576 | 664 | 751 | 839 |
| 0.5 | 244 | 249 | 254 | 294 | 359 | 424 | 488 | 553 | 618 |
| 0.6 | 244 | 249 | 254 | 259 | 271 | 320 | 369 | 418 | 466 |
| 0.7 | 244 | 249 | 254 | 259 | 264 | 269 | 280 | 317 | 354 |
| 0.8 | 244 | 249 | 254 | 259 | 264 | 269 | 274 | 279 | 284 |
| 0.9 | 244 | 249 | 254 | 259 | 264 | 269 | 274 | 279 | 284 |

We used the logistic regression model as standard to predict binary outcomes, such as the TreatmentResponse. The prevalence of response in the data was assumed to be 50%. The minimum sample size was calculated to satisfy three criteria:

1. Small overfitting defined by an expected shrinkage of predictor effects by 15% or less
2. Small absolute difference of 10% in the model's apparent and adjusted Nagelkerke's R-squared value
3. Precise estimation (within +/- 10%) of the average outcome risk in the cohort of the study

Table 2 gives the minimum sample sizes calculated based on the area under the ROC curve (AUC) and the number of parameters to be used in the predictive model.

Table 2: Minimum sample sizes over AUCs and the numbers of parameters to be used in predictive logistic regression model for binary outcomes.

| | Number of parameters | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| AUC | 10 | 15 | 20 | 25 | 30 | 35 | 40 | 45 | 50 |
| 0.70 | 436 | 653 | 871 | 1089 | 1306 | 1524 | 1742 | 1959 | 2177 |
| 0.75 | 272 | 408 | 544 | 680 | 815 | 951 | 1087 | 1223 | 1359 |
| 0.80 | 182 | 273 | 364 | 455 | 546 | 637 | 728 | 819 | 910 |
| 0.85 | 128 | 191 | 255 | 318 | 382 | 445 | 509 | 572 | 636 |
| 0.90 | 95 | 142 | 189 | 236 | 284 | 331 | 378 | 425 | 473 |
| 0.95 | 84 | 126 | 169 | 211 | 253 | 295 | 337 | 379 | 421 |

The above minimum sample size calculations satisfy some general model fitting and prediction requirements. Beyond that, we moved further and applied simulations to study the sample size and prediction accuracy based

on hypothetical models and parameters for pain and treatment-related outcomes according to well-recognized literature studies. Following the proposed data collection plan, we assumed that patients were randomized to acupuncture, MBSR, and control group, each containing 1/3 of the sample. We considered the influences of factors:

- "Basic" predictors (Baker, Buchanan et al. 2008, Witt, Schützler et al. 2011): acupuncture and MBSR treatments, heterogenous racial groups (assuming three clusters of similar responses to the treatments), gender, age, education level (high/low), long duration of pain (yes/no), baseline pain score, presence of certain concomitant diseases (yes/no).
- Extra predictors and modifiers related to omics- and biopsychosocial PainMarkers. We hypothesized two types of effects: 1) 10 predictors contributed to the outcome through main effects, and 2) 8 modifiers (4 for each of acupuncture and MBSR) that contributed through both main and interaction effects.

For the quantitative outcome (such as PainImpact), the simulations were based on a linear mixed-effect model to account for the heterogeneity of patients and predictive factors. We assumed that the binary predictors had a prevalence of around 50% and the continuous variables were standardized. According to the literature, we assumed that the "basic" predictors accounted for 45% of total variation (Baker, Buchanan et al. 2008). All predictors involve 37 coefficients in the true model, which explains up to 97% variation. The predictions were carried out based on the 5-fold cross-validation procedure. Prediction accuracy was measured by 1) the correlation coefficient between the predicted and the observed outcomes, 2) the relative mean-squared predictive error (RMSPE, i.e., the ratio between the mean-squared predictive error and the observed variance of responses). Large correlation and small RMSPE indicate accurate prediction. Table 3 presents how the correlation and RMSPE (averaged over 100 simulations) depend on the sample size and the percentage of predictors included into the prediction model (proportional to both main and interaction effects). It shows that the prediction accuracy is stable when sample size is equal or larger than 240. After this number, the increase of sample size is not cost-efficient to increase prediction capability. Instead, the innovative statistical and machine-learning approaches proposed in Projects 2 and 3 are critical to revealing more effective predictors and improving the predictive power.

Table 3: Prediction accuracy (correlation / RMSPE) over sample sizes and the percentages of non-basic predictors included in the predictive regression model.

| | Percentage of non-basic predictors included | | | | |
|---|---|---|---|---|---|
| Sample size | 0% | 25% | 50% | 75% | 100% |
| 60 | 0.45/0.97 | 0.55/0.92 | 0.64/0.85 | 0.76/0.63 | 0.94/0.15 |
| 120 | 0.48/0.82 | 0.61/0.69 | 0.73/0.51 | 0.86/0.29 | 0.98/0.05 |
| 180 | 0.51/0.77 | 0.63/0.64 | 0.75/0.46 | 0.88/0.25 | 0.98/0.04 |
| 240 | 0.52/0.75 | 0.64/0.60 | 0.77/0.42 | 0.88/0.23 | 0.98/0.03 |
| 300 | 0.52/0.74 | 0.65/0.59 | 0.77/0.42 | 0.89/0.22 | 0.98/0.03 |
| 360 | 0.53/0.73 | 0.65/0.58 | 0.78/0.40 | 0.89/0.21 | 0.99/0.03 |

For the binary outcome (such as TreatmentResponse), the simulations were based on a generalized linear mixed-effect model to account for the heterogeneity of patients and predictors. The predictors are the same as above, except the ORs of the basic predictors were set from 0.77 to 4.9 according to literature (Witt, Schützler et al. 2011), and the intercept was set so that the prevalence of the binary outcome is about 0.5. Predictions were based on the 5-fold cross-validation procedure. Prediction accuracy was measured by the AUC. Table 4 presents how the AUC (averaged over 100 simulations) depends on the sample size and the percentage of predictors included into the prediction model (proportional to both main and interaction effects). Again, to increase AUC, finding more predictors is more critical than getting more sample size.

Table 4: Prediction accuracy (AUC) over sample sizes and the percentages of non-basic predictors included in the predictive regression model.

| | percentage of non-basic predictors included | | | | |
|---|---|---|---|---|---|
| Sample Size | 0% | 25% | 50% | 75% | 100% |
| 60 | 0.50 | 0.58 | 0.64 | 0.69 | 0.73 |
| 120 | 0.51 | 0.61 | 0.69 | 0.72 | 0.80 |
| 180 | 0.51 | 0.63 | 0.71 | 0.76 | 0.83 |
| 240 | 0.51 | 0.64 | 0.73 | 0.78 | 0.85 |
| 300 | 0.51 | 0.65 | 0.74 | 0.80 | 0.87 |
| 360 | 0.51 | 0.65 | 0.75 | 0.82 | 0.87 |
| 420 | 0.51 | 0.66 | 0.76 | 0.83 | 0.88 |
| 480 | 0.51 | 0.66 | 0.76 | 0.84 | 0.89 |
| 540 | 0.51 | 0.67 | 0.77 | 0.84 | 0.90 |
| 600 | 0.51 | 0.67 | 0.77 | 0.85 | 0.90 |
| 660 | 0.52 | 0.67 | 0.78 | 0.86 | 0.91 |
| 720 | 0.52 | 0.68 | 0.78 | 0.86 | 0.92 |
| 780 | 0.52 | 0.68 | 0.78 | 0.86 | 0.92 |
| 840 | 0.52 | 0.68 | 0.79 | 0.87 | 0.93 |
| 900 | 0.52 | 0.68 | 0.79 | 0.87 | 0.93 |
| 960 | 0.52 | 0.68 | 0.79 | 0.87 | 0.93 |
| 1020 | 0.52 | 0.69 | 0.79 | 0.87 | 0.94 |
| 1080 | 0.52 | 0.69 | 0.79 | 0.88 | 0.94 |
| 1140 | 0.52 | 0.69 | 0.79 | 0.88 | 0.94 |
| 1200 | 0.52 | 0.69 | 0.80 | 0.88 | 0.94 |

Based on Tables 3 and 4, we could make two conclusions/interpretations.

- Prediction accuracy highly depends on the percentage of variations explained by the predictors. Therefore, the proposed research of identifying more valid predictors is particularly critical to improving clinic prediction on pain impact and responses.
- At a given set of predictors, sample increase does not significantly improve prediction accuracy (especially after sample size reaches a certain “adequate” level, roughly around 300 in our simulation). Such sample size–accuracy relationship observation is consistent with literature studies (van Smeden, Moons et al. 2019).

A few further comments:

- This report intends to address all four hypotheses by considering both omics-based and biopsychosocial factors in Projects 2 and 3 as generic predictors with main and interaction effects. Further results on 95% C.I. (empirical) for the accuracy measures in Tables 3 and 4 are also available if needed.
- It remains to be determined how to deliver the sample size study in the proposal.
- In general, predicting binary responses requires a larger sample size than predicting quantitative responses. Also, the more parameters used in the predictive model, the larger the sample size is needed. Binary responses could require a sample size at a thousand-level if many predictors are involved and if we demand very high accuracy and low variation in prediction. A smaller sample size (e.g., around 300) is still justifiable if the corresponding level of predictive accuracy and variation is acceptable.
- The predictive models are based on linear statistical models. Some literature (van der Ploeg, Austin et al. 2014) claims that machine/deep learning approaches are more data-hungry because they routinely include many parameters and categorize continuous predictors (more categories mean more

parameters). Therefore, further justifications for the sample size from both machine-learning methodology and literature perspectives could be helpful in the proposal writing.

## References


Baker, T. A., et al. (2008). "Factors influencing chronic pain intensity in older black women: Examining depression, locus of control, and physical health." Journal of women's health **17**(5): 869--878.

Riley, R. D., et al. (2019). "Minimum sample size for developing a multivariable prediction model: Part I--Continuous outcomes." Statistics in Medicine **38**(7): 1262--1275.

Riley, R. D., et al. (2019). "Minimum sample size for developing a multivariable prediction model: PART II-binary and time-to-event outcomes." Statistics in Medicine **38**(7): 1276--1296.

van der Ploeg, T., et al. (2014). "Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints." BMC medical research methodology **14**(1): 1--13.

van Smeden, M., et al. (2019). "Sample size for binary logistic prediction models: beyond events per variable criteria." Statistical methods in medical research **28**(8): 2455--2474.

Witt, C. M., et al. (2011). "Patient characteristics and variation in treatment outcomes: which patients benefit most from acupuncture for chronic pain?" The Clinical journal of pain **27**(6): 550--555.