Supplementary statistical methods

Rather than evaluating a limited set of covariates, this study leveraged machine learning methods that can learn from the data to identify predictors of particular target response variables [10]. This study was designed to evaluate the extensive baseline patient profile data collected in PSoHO and identify variables that may predict who would or would not achieve complete skin clearance at different points in time. A second objective was to further examine these variables to quantify the likelihood of different outcomes.

Descriptive analyses

To understand the extent of missingness of treatment outcomes for PSoHO patients, descriptive analyses summarised the available data. Descriptive analyses used the Fisher's exact test or Chi-square for categorical variables, and F-test for continuous variables.

Missing data imputation

Missing values in the baseline data were imputed using the random forest missing data algorithm (impute.rfsrc function in R package randomForestSRC) [S1, 14]. Imputation was completed for each treatment group separately without using the outcomes.

Identifying candidate predictors via modelling

The three modelling strategies used to identify potential predictors of complete skin clearance were logistic regression (LR), penalized logistic regression (PLR) using lasso penalty, and gradient boosting [14, S2–S5]. LR was implemented using forward selection. PLR used sparse group lasso [S6]. Finally, XGBoost, a decision-tree-ensembles machine learning algorithm, was used for the gradient boosting modelling [S7]. SHapley Additive exPlanations (SHAP) was used for XGBoost to identify the variable importance. Figure S1 shows the standard SHAP value plots for each of the three PASI100 outcomes.

Data splits for modelling

For the final logistic model, 20% of data was selected randomly as the hold-out sample. The above 3 modelling strategies were executed on the remaining 80% of data 1000 times using the 70%:30% training:validation random splits. The performance of each modelling strategy and the predictors were identified on the validation data.

Identifying candidate predictors via Treatment-Specific Subgroup Detection Tool (TSDT)

The Treatment-Specific Subgroup Detection Tool (TSDT) is a tree-based tool that was developed to identify subgroups with a superior response relative to the overall sample[S8, 13]. TSDT was implemented on the full data set with no hold out sample.

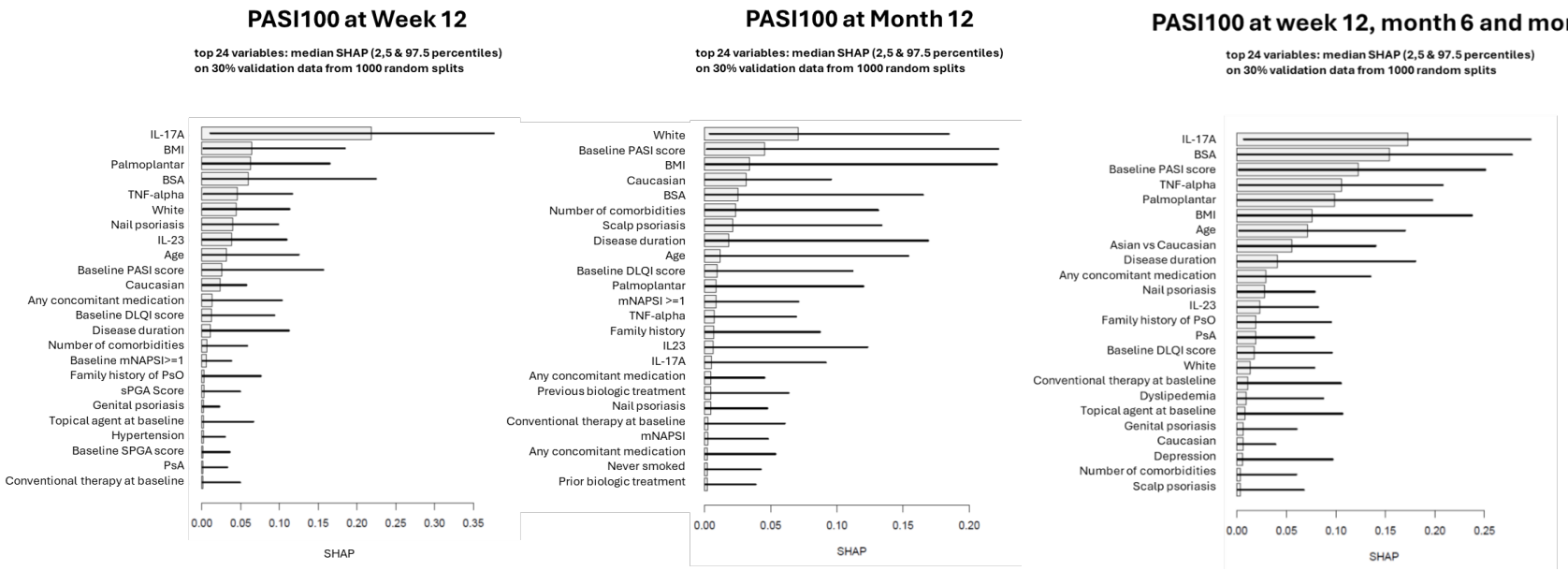
Final logistic regression analysis

For the final model choice, the most important 14 variables selected by TSDT and LR, XGBoost and penalized logistic regression (PLR) modeling on random splits were added to the 12 variables selected based on the literature for clinical relevance and completeness. Of these variables, 3 treatment variables were included that compared outcomes for patients treated with IL-17A biologics versus TNF-α, IL-23 and IL-12/23 biologics. Table S1 lists the 26 predictor variables and whether they were identified using the modelling strategies or from the literature. These 26 variables were then separately used in the final logistic regression model within the three PASI100 outcomes to calculate model estimates on the 20% hold-out sample.

All analyses were conducted using SAS 9.4 (SAS Institute Inc., Cary, NC, USA) and R Version 4.0.3 (R Foundation for Statistical Computing, Vienna, Austria).

References

S1. Fast Unified Random Forests for Survival, Regression, and Classification [Available from: https://cran.r-project.org/web/packages/randomForestSRC/index.html.
S2. Chen T, Guestrin C, editors. Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining; 2016.
S3. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. Journal of statistical software. 2010;33(1):1.
S4. Simon N, Friedman J, Hastie T. A blockwise descent algorithm for group-penalized multiresponse and multinomial regression. arXiv preprint arXiv:13116529. 2013.
S5. Simon N, Friedman J, Hastie T, Tibshirani R. Regularization paths for Cox's proportional hazards model via coordinate descent. Journal of statistical software. 2011;39(5):1.
S6. Sparse Group Lasso [Available from: https://cran.r-project.org/web//packages/sparsegl/index.html.
S7. Extreme gradient boosting [Available from: https://cran.r-project.org/web/packages/xgboost/index.html.
S8. Battioui C, Shen L, Ruberg SJ, editors. A resampling-based ensemble tree method to identify patient subgroups with enhanced treatment effect. Proceedings of 2014 joint statistical meetings; 2014: American Statistical Association Alexandria, VA, USA.



(a) **(b)** **(c)**

Figure S1: Standard SHAP value plots for each of the three PASI100 outcomes: (a) PASI100 at week 12, (b) PASI100 at month 12, and (c) PASI100 at week 12 and maintained at month 6 and month 12. Each plot provides the top 24 variables based on median SHAP (with 95% rCI) on 30% validation data from 1000 random splits.

Abbreviations: BMI = Body Mass Index; BSA = Body Surface Area; DLQI = Dermatology Life Quality Index; IL = interleukin; mNAPSI = modified Nail Psoriasis Severity Index; PASI = Psoriasis Area and Severity Index; PsA = psoriatic arthritis; PsO = psoriasis; sPGA = static Physician's Global Assessment. SHAP = SHapley Additive exPlanations; TNF = tumor necrosis factor.