Development and Validation of a Prediction Model to Estimate the Hurley Score in Patients With Hidradenitis Suppurativa

Alves P; Palucci O; Gopalakrishnan C; Boussios C | OM1, Inc., Boston, MA, USA

Background

- The Hurley staging system classifies hidradenitis suppurativa (HS) severity and is critical for guiding treatment decisions and evaluating patient outcomes.
- However, documentation of the Hurley Score is inconsistent in real-world data (RWD) sources, limiting the size of available HS patient cohorts for research.

Objectives

To develop and validate a prediction model to estimate the Hurley Score (eHurley) from unstructured clinical notes in the electronic medical record (EMR).

Methods

- We used the OM1 PremiOM™ HS, a linked claims and EMR dataset on HS patients in the US. Physician-recorded Hurley Scores and predictive features were extracted from clinical notes.
- Candidate tokens were prioritized by clinical relevance and absolute difference in Hurley Scores between notes with and without the token, yielding 85 baseline predictors for the estimation model.
- A logistic regression binary classifier (HS Stage 1 vs. HS Stage ≥2) was trained on an 80/20 patient-level train/test split. Model discrimination was assessed using the area under the receiver operating characteristic curve (AUROC).

Results

- Hurley Scores were successfully extracted from 3,259 patients across 5,117 encounters, which served as the labeled training (2,608 patients; 4,097 encounters) and test (651 patients; 1,020 encounters) sets.
- The overall cohort had a mean (SD) age of 36.68 (14.86) years and was predominantly female (79.93%) and White (51.46%); most patients resided in the South (66.80%) and had commercial insurance (49.98%) (Figure 2). The most frequently recorded baseline comorbidities were obesity (42.74%), back pain (28.26%), hypertension (25.10%), acne vulgaris (24.33%) and anxiety (22.80%) (Figure 3).
- The logistic regression model achieved an AUROC of 0.84 on the training set and 0.81 on the test set (Figures 4 and 5).
- The strongest predictors spanned use of medications such as adalimumab, apremilast, metformin, lisinopril, trimethoprim, explicit severity and lesion descriptors (severe, active lesion, drainage, subcutaneous involvement), anatomic site terms, and associated symptoms and comorbidities (joint pain, aches, anemia, weight gain).
- The predicted probability distribution of the estimation set closely mirrored that of the test set, supporting the validity of scores generated.
- The eHurley model was subsequently applied to generate scores for approximately 15,000 additional patients across 57,000 encounters, representing a greater than four-fold increase in patients and a greater than ten-fold increase in available encounters.

Conclusions

- This study developed and validated a prediction model to estimate Hurley Scores from clinical notes with good performance.
- Application of the model to real-world data sets may allow estimated Hurley Scores to be used for research purposes.
- Potential use cases include filling in missingness in the patient journey, more comprehensively measuring disease severity, improving understanding of treatment response, and for risk stratification.

FIGURE 1 Development of estimation models

Applying machine learning to unstructured clinical notes to produce research-ready Hurley Stage data.

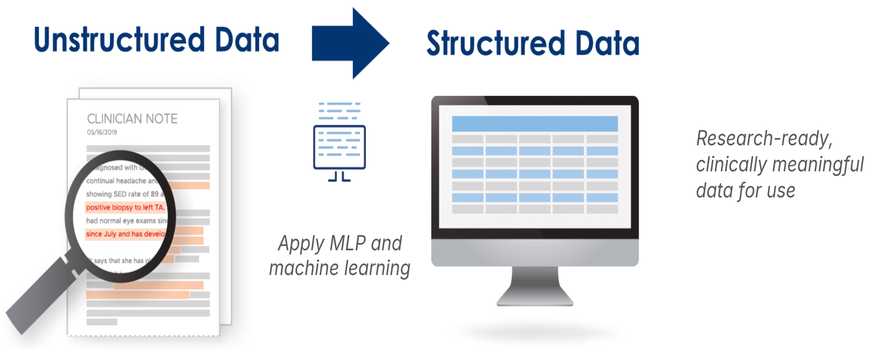
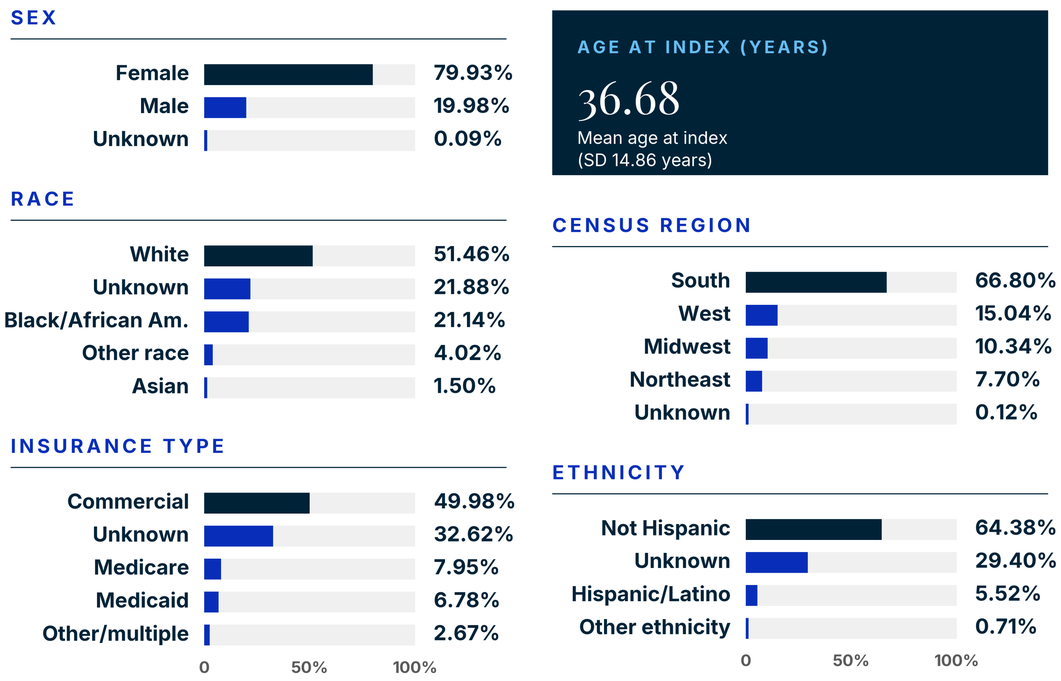


FIGURE 2 Baseline demographics

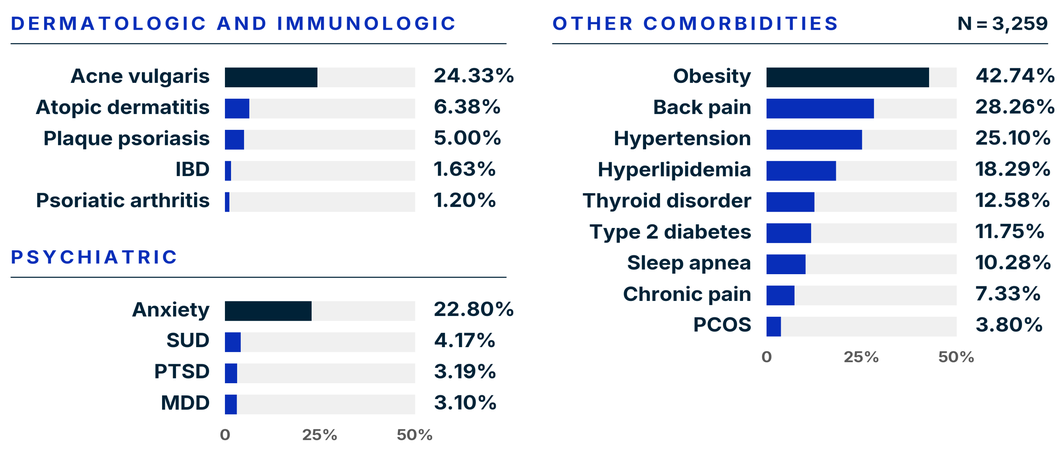
Demographic composition of the overall eHurley cohort, N = 3,259 patients.



Sex, race, ethnicity, census region and insurance type as recorded in the source electronic medical records. "Unknown" combines unknown and not-reported responses and is shown for transparency. Percentages are of the overall cohort (N = 3,259) and may not sum to 100% because of rounding.

FIGURE 3 Baseline clinical characteristics

Comorbidities recorded during the baseline period, N = 3,259 patients.



Comorbidities identified from diagnosis codes recorded during the baseline period; categories are not mutually exclusive, so percentages do not sum to 100%. Bars are drawn on a 0–50% scale. Percentages are of the overall cohort (N = 3,259).
IBD, inflammatory bowel disease (Crohn's disease or ulcerative colitis); MDD, major depressive disorder; PCOS, polycystic ovary syndrome; PTSD, post-traumatic stress disorder; SUD, substance use disorder.

FIGURE 4 ROC curve

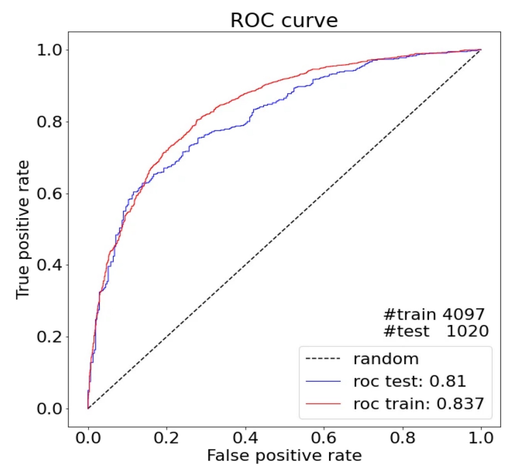


FIGURE 5 Confusion matrix

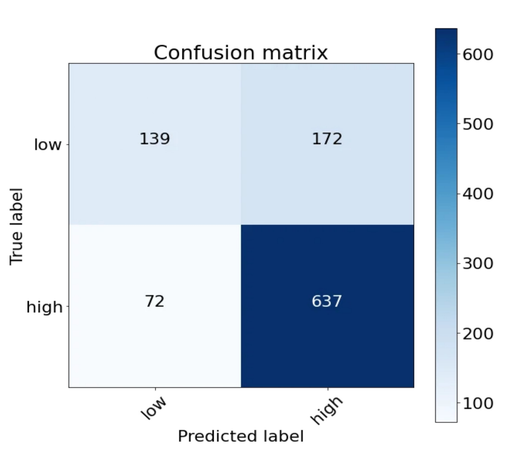


Figure 4: receiver operating characteristic curves for the training (n = 4,097 encounters) and test (n = 1,020 encounters) sets.
Figure 5: classification of HS Stage 1 ("low") versus HS Stage ≥2 ("high") in the test set at the default probability threshold.

Presented at ISPOR Europe 2026, 8–11 November 2026, Vienna, Austria