coding based on established codebook, identification of overarching themes and narratives); GenAI platform(s) and large language model(s) used; types of GenAI prompts used to facilitate analyses; processes and measures/metrics to evaluate the accuracy and quality of GenAI's results; outcomes and methods used to increase GenAI's accuracy and/or quality. DISCUSSION/SIGNIFICANCE OF IMPACT: Qualitative research is essential to understand patient perspectives and increase treatment effectiveness and accessibility for all. To actualize GenAI's potential to facilitate rapid and large-scale qualitative analysis, we first need to understand its strengths, weaknesses, and best practices for maximizing GenAI's accuracy and quality.

75

Validation of clinical vignettes for a translational study evaluating ChatGPT's diagnostic support in adolescent and young adults' mental health

Samuel Colón De La Rosa[1], Karen G. Martínez González[1], Abiel Roche Lima[1], Claudia P. Amaya Ardila[1], Yamil Ortiz Ortiz[2], Yaniris Colón Rivera[2], Sebastián Burgos Fernández[2], Alondra García Rosa[2] and Mairim Figueroa Fernández[2]
[1]University of Puerto Rico, Medical Science Campus and [2]University of Puerto Rico, Río Piedras Campus

OBJECTIVES/GOALS: To validate 20 culturally tailored clinical vignettes for a translational study assessing ChatGPT's diagnostic support in adolescent and young adult mental health. The validation ensures methodological, ethical, and cultural rigor to examine ChatGPT's potential in collaborative clinical decision-making. METHODS/STUDY POPULATION: Twelve licensed clinical psychologists evaluated 50 culturally adapted vignettes using a nine-item instrument assessing diagnostic agreement (Fleiss κ), content validity (I-CVI, S-CVI/Ave), comorbidity sensitivity, difficulty, ethical sensitivity, and perceived risk. Qualitative feedback refined clarity and realism. Vignettes meeting $\kappa \geq 0.60$, S-CVI ≥ 0.80, and ethics ≥ 3 were retained, ensuring one easy, two moderate, and one difficult case per condition. The resulting 20 validated cases will serve as standardized stimuli to assess ChatGPT's diagnostic accuracy and its ability to collaborate effectively with clinical psychologists. RESULTS/ANTICIPATED RESULTS: High inter-rater reliability ($\kappa \geq 0.60$) and strong content validity (S-CVI ≥ 0.85) are expected. Balanced difficulty and accurate risk detection will confirm calibration. Qualitative feedback will enhance realism and cultural appropriateness. Validated vignettes will enable rigorous assessment of ChatGPT's diagnostic accuracy and clinical decision-support potential within a translational research framework. Vignette cases will be presented to three groups: 1) ChatGPT app, 2) Clinical mental health care providers, and 3) Clinical mental health care providers collaborating with ChatGPT. Each group will provide a diagnosis and confidence level to compare accuracy, sensitivity, and collaborative value across human-AI and human-only conditions. DISCUSSION/SIGNIFICANCE OF IMPACT: This validation provides a rigorous foundation for evaluating ChatGPT's translational role in mental health diagnostics. Psychometrically and ethically sound vignettes strengthen evidence-based integration of AI tools into precision mental health care and promote responsible human-AI collaboration.

76

Benchmarking automated segmentation tools for white matter hyperintensities quantification to inform stroke and dementia risk stratification

Amir Ali Satani, Michael Hughes and David Kent
Tufts University

OBJECTIVES/GOALS: To benchmark the extent to which variability in automated WMH quantification influences stroke and dementia risk stratification by characterizing differences in volume estimates and spatial concordance across three segmentation tools (BIANCA, SAMSEG, and nnU-Net). METHODS/STUDY POPULATION: This retrospective, cross-sectional study will analyze de-identified clinical brain MRI scans from patients aged ≥50 years, including those with and without prior stroke, obtained from institutional imaging repositories. Eligible scans include co-registered T1-weighted and FLAIR sequences. Scans with poor image quality and scans from patients with multiple sclerosis, other demyelinating disorders, or brain tumors will be excluded. WMHs will be segmented using three tools: BIANCA (supervised k-NN classifier), SAMSEG (unsupervised Bayesian model), and nnU-Net (self-configuring deep learning). Agreement in WMH burden will be assessed with intraclass correlation coefficients (ICC[3,1]) and spatial concordance with Dice similarity coefficients (DSC). Bland-Altman analyses will evaluate bias. RESULTS/ANTICIPATED RESULTS: We will analyze 600 brain MRI scans. All three tools are expected to detect similar WMH spatial distributions, though total WMH volume estimates may differ. nnU-Net may yield higher WMH burden estimates than BIANCA or SAMSEG. WMH burden will be neuroradiologist-adjudicated to establish the gold standard, against which concordance for each tool and inter-tool agreement will be assessed using ICC[3,1] and DSC. We anticipate moderate-to-high agreement, but tool-specific variability could influence reproducibility and interpretation of WMH burden for future WMH-related research studies. This study will quantify the extent of variability among tools in clinical brain MRI scans and evaluate whether these differences could meaningfully alter future stroke risk classification thresholds. DISCUSSION/SIGNIFICANCE OF IMPACT: This study will benchmark automated segmentation tools for WMH-related research by standardizing WMH quantification to improve reproducibility and reliability for stroke and dementia risk stratification. Establishing tool-specific benchmarks may also guide integration of automated WMH measures into clinical decision-support systems.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 19 Nov 2026 at 18:42:15, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/cts.2026.10301