content; while this varied by region, ranging from 41.4% in the Midwest to 60.3% in the Northeast, the differences did not reach statistical significance [$\chi^2(3,N=283)=6.86$, p=0.076]. No significant associations were found between wellness content and program size (p=0.304), age (p=0.387), or length (p=0.807). Descriptive content analysis revealed substantial heterogeneity in how programs portray wellness to applicants. Common domains among programs providing descriptive information included mental health resources and access to counseling (30.1%), social and community activities (28.1%), retreats (26.7%), and presence of a formal wellness infrastructure or committee(s) (24.0%) (Table 1).

Conclusions: Wellness content on EM residency websites was inconsistently represented, with substantial variation in how support was described. The lack of program characteristic-based differences suggests that wellness communication is unsystematic and may not reflect broader institutional priorities. Clearer and more consistent wellness information could help programs better convey their culture and enable applicants to identify environments aligned with their needs.

48 Optimizing EM Residency Application Reviews a Comparative Study: Faculty versus AI

Robert Steele, Taylor Route, Michael Macias, Lauren Donnelly

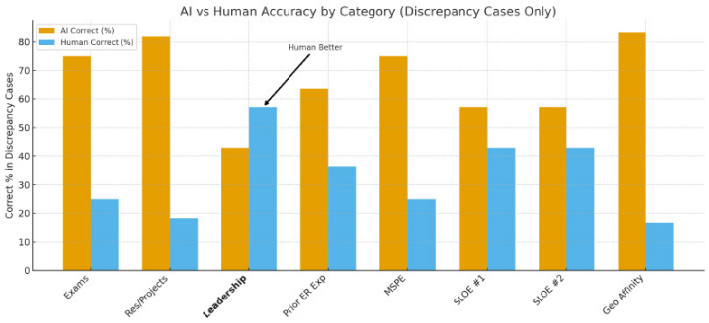
Background: Residency application volumes have increased, straining faculty time for holistic medical student application review. Artificial intelligence (AI)–assisted screening may be a solution, but can it really review student applications?

Objectives: To compare agreement and accuracy between an AI-based application scoring tool versus trained faculty application reviewers for EM residency interview screening.

Methods: We developed an AI algorithm using the cloud-based platform Airtable to generate scores in 10 predefined domains (exams, research/special projects/community service, leadership/distinction, prior EM experience, MSPE, two SLOE/letter domains, geographic affinity, exceptional bonus, and red flag). Thirty-four applicants from one recruitment cycle were independently scored by both AI and a faculty reviewer. For each domain we calculated percent exact agreement between AI and faculty. All discrepant scores were adjudicated by a trained independent blinded reviewer, who determined which score (Faculty/AI) was most accurate. We calculated the proportion of discrepant cases using McNemar's test.

Results: Across 340 domain-level ratings, AI and faculty agreed exactly on 252 (74.1%). Agreement by domain ranged from 38.2% for leadership/distinction to 94.1% for red flag scores. Among 88 discrepant ratings, the AI score was adjudicated correct in 58 (65.9%) and the faculty score in 30 (34.1%) (p=0.004). Patterns were similar across most individual domains.

Conclusions: An AI-based application scoring tool demonstrated substantial agreement with faculty ratings and was more likely than the individual faculty reviewer to match adjudicated scores when disagreement occurred. AI-assisted scoring may be a feasible adjunct for initial residency application screening, potentially reducing faculty workload while preserving decision quality.



49 Content Validity Index (CVI) as a Tool for Instrument Development: A Methodological Case Study in Neonatal Lumbar Puncture and Umbilical Vein Catheterization for Emergency Medicine Simulation

Brendan Freeman, Kathryn Zabinski, Darya Ryndych

Background: The Content Validity Index (CVI) is a structured framework for evaluating instrument content. Although well-described in nursing literature, its use is limited elsewhere. Simulation-based procedural tools, particularly for neonatal lumbar puncture (LP) and umbilical vein catheterization (UVC), often lack documented content validity, creating uncertainty about instrument quality.

Objectives: To apply CVI methodology to develop and refine task-specific checklists and global rating scales for neonatal LP and UVC for emergency physicians. We hypothesized that iterative CVI analysis with expert raters would improve item- and scale-level validity metrics.

Methods: Design: Cross-sectional, survey-based instrument validation study using CVI methodology.

Setting: Electronic ratings from experts in neonatal and emergency medicine simulation.

Participants: Six expert raters recruited by purposive sampling completed two rounds of item relevance ratings. Measures: Item-Level CVI (I-CVI) and Scale-Level CVI/Average (S-CVI/Ave) were calculated. Items below accepted thresholds (I-CVI < 0.78) were revised or removed. Descriptive statistics were used for CVI calculations.

Results: Several items showed low I-CVI values initially