Nicolas Lescano

Review of article by Shah et al.

This is a short, embedded QI pilot at Stanford testing an Epic-integrated ambient scribe (Nuance DAX Copilot) over roughly three months. Forty-eight physicians participated; 38 contributed paired pre/post surveys and 46 completed post-only items. The authors report sizeable *perceived* gains: physician task load −24.4/100 and burnout −1.94/10, with a moderate SUS bump of +10.9/100. Post-survey sentiment was warm (65% endorsed better documentation efficiency, 52% better quality, 98% ease of use) and the median perceived time change was +20 minutes saved per half-day, albeit with a very wide range (−90 to +90).

I appreciate the sociotechnical posture. Capture ran through Haiku; drafts flowed into SmartSections (HPI/PE/Results/A&P) and were edited/attested inside Epic. Training and office hours were provided. This is the right direction—meet the workflow where it actually lives instead of demoing in a vacuum.

Where I struggle is with what the study can prove. The outcomes are all surveys. There are no audit-log time series (time-in-notes, after-hours, turnaround) to corroborate “saves time,” no draft→final edit-distance or note-length comparisons to detect bloat, and no linkage to safety/quality (corrections, addenda, near-misses, etc.). The manuscript itself frames this as a first step and calls for ROI and quality/safety work before scale, which, to me, highlights the gap more than it closes it.

Generalizability is also thin. This is a volunteer sample, iPhone-only, with rolling onboarding and a heterogeneous mix of settings and specialties. Sensible for a pilot, difficult for inference with n=38 pairs. The most interesting signal may be the heterogeneity itself: even with 98% reporting usability, roughly a third did **not** experience efficiency gains, and some perceived *more* time (consistent with the median-with-wide-range finding). Without week-by-week logs, attributing this to “learning curve” or “shifting from writing to editing” reads like a just-so story. Usability measures *interaction ease*; productivity in clinical documentation is dominated by verification liability and fit (templates, acoustics, section mapping, trust cues). Pleasant tools can still cost time when verification is heavy.

What would move me from “promising” to “recommend”? A short list: (1) audit-log time series by week since activation showing medians crossing baseline (note time, after-hours, turnaround, addenda); (2) draft→final edit-distance and note-length against baseline to rule out bloat; (3) safety/quality linkage (corrections, near-misses, SOAP completeness) tagged to AI-draft use; (4) heterogeneity analysis by specialty/visit type/acoustics to design for non-responders; and (5) ROI that distinguishes time *reclaimed* from time merely *displaced* to verification or inbox. Until then, I read these results as encouraging signals (burden/burnout perceptions move the right way) but not yet as evidence of structural relief. Proceed, but measure mercilessly!