Development and Establishment of Initial Validity Evidence for a Novel Tool for Assessing Trainee Admission Notes

Danielle E. Weber, MD, MEd[1,2], *Justin D. Held, MD*[1], *Roman A. Jandarov, PhD*[1], *Matthew Kelleher, MD, MEd*[1,2], *Ben Kinnear, MD, MEd*[1,2], *Dana Sall, MD, MEd*[1], *and Jennifer K. O'Toole, MD, MEd*[1,2]

[1]Department of Internal Medicine, University of Cincinnati College of Medicine, University of Cincinnati Medical Center, Cincinnati, OH, USA; [2]Department of Pediatrics, Cincinnati Children's Hospital Medical Center, University of Cincinnati College of Medicine, Cincinnati, OH, USA.

BACKGROUND: Documentation is a key component of practice, yet few curricula have been published to teach trainees proper note construction. Additionally, a gold standard for assessing note quality does not exist, and no documentation assessment tools integrate with established competency-based frameworks.

OBJECTIVE: To develop and establish initial validity evidence for a novel tool that assesses key components of trainee admission notes and maps to the Accreditation Council for Graduate Medical Education (ACGME) milestone framework.

DESIGN: Using an iterative, consensus building process we developed the Admission Note Assessment Tool (ANAT). Pilot testing was performed with both the supervising attending and study team raters not involved in care of the patients. The finalized tool was piloted with attendings from other institutions.

PARTICIPANTS: Local experts participated in tool development and pilot testing. Additional attending physicians participated in pilot testing.

MAIN MEASURES: Content, response process, and internal structure validity evidence was gathered using Messick's framework. Inter-rater reliability was assessed using percent agreement.

KEY RESULTS: The final tool consists of 16 checklist items and two global assessment items. Pilot testing demonstrated rater agreement of 72% to 100% for checklist items and 63% to 70% for global assessment items. Note assessment required an average of 12.3 min (SD 3.7). The study generated validity evidence in the domains of content, response process, and internal structure for use of the tool in rating admission notes.

CONCLUSIONS: The ANAT assesses individual components of a note, incorporates billing criteria, targets note "bloat," allows for narrative feedback, and provides global assessments mapped to the ACGME milestone framework. The ANAT can be used to assess admission notes by any attending and at any time after note completion with minimal rater training. The ANAT allows programs to implement routine note assessment for multiple functions with the use of a single tool.

KEY WORDS: documentation; electronic health records; assessment; evaluation; medical education.

Prior Presentations *none.*

Received September 19, 2019
Accepted January 12, 2020
Published online January 28, 2020

Abbreviations

ANAT	Admission note assessment tool
CC	Chief complaint
HPI	History of present illness
ROS	Review of systems
PFSH	Past/family/social history
DD	Diagnostic data
A&P	Assessment and plan.

J Gen Intern Med 35(4):1078–83
DOI: 10.1007/s11606-020-05669-6

INTRODUCTION

Clinical notes are the means by which physicians document and communicate important information regarding the care of their patients.[1, 2] Appropriate documentation is a necessary component of practice,[3–5] with notes being used for patient care, medical education, billing, quality improvement, and legal proceedings.[1, 2] Since the advent of electronic health records (EHRs), physician notes have become more legible and accessible,[2, 6] but these advances have come at the cost of increased note length due to note "bloat," or "clutter," and increased errors created by "cut and paste" or "copy forward" practices.[6–8] Faculty and trainees recognize these pitfalls[9–11] while also perceiving these functions as helpful for efficiency.[11] Despite this, there are few curricula available to teach trainees how to construct their notes appropriately[6, 12] and a gold standard for assessing note quality does not exist.[12]

Some groups have created note templates to improve progress note[7, 11] or discharge summary quality.[12] These interventions have shown modest effect by reducing "clutter"[7] or decreasing note length,[11] however interventions based solely on compliance with institution-specific note templates limit generalizability. Others have developed assessment tools to improve notes. QNOTE assesses outpatient notes,[1, 2] while PDQI-9 assesses inpatient notes.[6, 8, 13] Both QNOTE and PDQI-9 are based on subjective adjectives as assessment items, such as "concise" or "up-to-date", which do not provide concrete, actionable feedback to learners. PDQI-9 also requires assessors to be familiar with the patient or perform significant chart review. The RED checklist[13] assesses inpatient progress notes with four global measures of quality (truthful, reasoned, updated, and succinct) via open-ended questions and assesses individual note components via a checklist, but requires assessors to also review the previous progress note. Additionally, none of the tools integrate with competency-based frameworks such as the Accreditation Council for Graduate Medical Education (ACGME) milestones. While these tools have furthered our understanding of assessing trainee documentation, a tool that simultaneously assesses the multiple functions of a note, targets note "bloat," assesses overall clarity, and maps to other educational frameworks is needed.

The purpose of this study was to develop and create initial validity evidence for a single tool that assesses admission note quality and serves multiple functions, including assessment of key individual components of a note and provides global assessments mapped to the ACGME milestone framework.

METHODS

Setting

We conducted our study from 2017 to 2018 at the University of Cincinnati Medical Center. Our Internal Medicine residency program has approximately 92 categorical and preliminary residents each year. All documentation is entered into the electronic health record (Epic Hyperspace; Epic Systems, Verona, Wisconsin). This study was approved by the University of Cincinnati Institutional Review Board.

Assessment Tool Development

An initial draft of the Admission Note Assessment Tool (ANAT) was developed by two authors (DW and JH). The study team consisted of two internal medicine hospitalists (JH and DS) and four internal medicine-pediatrics hospitalists (DW, MK, BK, and JO). Members of this group have content expertise in tool development, learner assessment, and billable documentation. Tool development continued with the goal that the ANAT (Fig. 1) would help accomplish the following objectives:

1. Ensure proper documentation for billing
2. Decrease note "bloat"
3. Provide global assessments mapped to the ACGME milestone framework

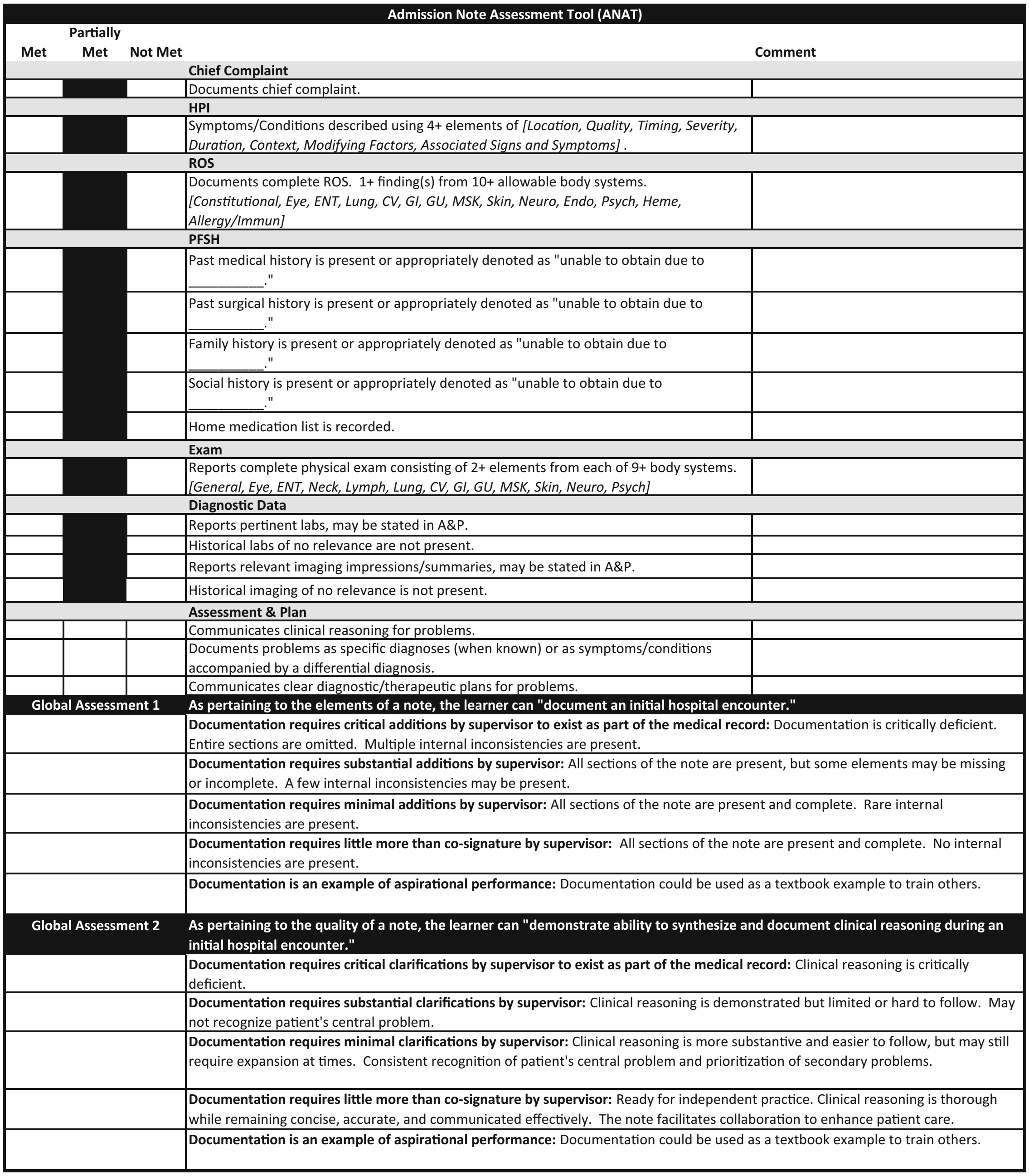
Validity evidence was sought utilizing Messick's validity framework.[14]

The study group revised the initial draft of the ANAT based on discussion and consensus building[15] around optimal tool content and format to meet the tool's objectives. The ANAT then underwent further iterative revisions as follows. Each study group member used the tool individually to evaluate an admission note. The group then discussed how each study group member interpreted and applied the assessment items in their rating of the note via think alouds. The ANAT was then revised and the process was repeated until there was agreement that the tool accomplished the above stated goals, had utility,[16] and was easy to use. Notes were taken throughout this process and the group's experiences were used to create a rater training manual.

At the start of the project, two commonly identified issues were the lack of adequate review of systems (ROS) and physical exam (PE) documentation needed for appropriate evaluation and management (E&M) billing.[17] Since the majority of our patients are significantly complex, billing requirements for a level three E&M encounter became the benchmark for assessing the elements of the note. We found that agreeing on the quality of certain elements (e.g. completeness of history of present illness) was challenging given the subjective nature of assessing quality. Therefore, the scope of many items in the ANAT was narrowed to focus on billing criteria and scored as "met" or "not met" in a checklist format. Similarly, to help decrease note "bloat," we included items aimed at reducing irrelevant historical labs or imaging. A separate area was created for narrative feedback related to each item. The incorporation of narrative feedback throughout the tool became an important focus to allow specific, actionable feedback related to more subjective or nuanced aspects of documentation not captured in the checklist, and to inform the global assessment ratings.

Within the assessment and plan (A&P) items our goal was to assess clinical reasoning, but we found differing opinions amongst our group regarding what we considered adequate. Therefore, we decided to assess the presence or absence of clinical reasoning in the A&P items and focus on narrative feedback, while assessing the overall adequacy of clinical reasoning in one of the global assessment items. It was felt that by scoring these items on a three-point scale (i.e. "met", "partially met", and "not met") we could still provide some discrimination between trainees.

With the development of this tool we created two global assessment items: 1) As pertaining to the elements of a note, the learner can "document an initial hospital encounter" and 2) As pertaining to the quality of a note, the learner can "demonstrate ability to synthesize and document clinical reasoning during an initial hospital encounter". We created a behaviorally-anchored five-point rating scale based on the amount of clarification/editing needed by a supervisor (e.g. "documentation requires substantial clarifications by supervisor"). In order to integrate ANAT ratings into our program of assessment, and to examine relationship to other variables of trainee performance in the future, a key aspect of development included mapping these global assessment items to ACGME sub-competencies. We mapped item one to interpersonal and communication skills (ICS)-3 ("appropriate utilization and completion of health records") and professionalism (PROF)-4 ("exhibits integrity and ethical behavior in professional conduct"), and item two to patient care (PC)-1 ["gathers and synthesizes essential and accurate information to define each patient's clinical problem(s)"] and ICS-2 ["communicates effectively in interprofessional teams (e.g. peers, consultants, nursing, ancillary professionals and other support personnel)"].[4]



Admission Note Assessment Tool (ANAT)				
Met	Partially Met	Not Met		Comment
			Chief Complaint	
	■		Documents chief complaint.	
			HPI	
	■		Symptoms/Conditions described using 4+ elements of *[Location, Quality, Timing, Severity, Duration, Context, Modifying Factors, Associated Signs and Symptoms]* .	
			ROS	
	■		Documents complete ROS. 1+ finding(s) from 10+ allowable body systems. *[Constitutional, Eye, ENT, Lung, CV, GI, GU, MSK, Skin, Neuro, Endo, Psych, Heme, Allergy/Immun]*	
			PFSH	
	■		Past medical history is present or appropriately denoted as "unable to obtain due to ___________."	
	■		Past surgical history is present or appropriately denoted as "unable to obtain due to ___________."	
	■		Family history is present or appropriately denoted as "unable to obtain due to ___________."	
	■		Social history is present or appropriately denoted as "unable to obtain due to ___________."	
	■		Home medication list is recorded.	
			Exam	
	■		Reports complete physical exam consisting of 2+ elements from each of 9+ body systems. *[General, Eye, ENT, Neck, Lymph, Lung, CV, GI, GU, MSK, Skin, Neuro, Psych]*	
			Diagnostic Data	
	■		Reports pertinent labs, may be stated in A&P.	
	■		Historical labs of no relevance are not present.	
	■		Reports relevant imaging impressions/summaries, may be stated in A&P.	
	■		Historical imaging of no relevance is not present.	
			Assessment & Plan	
			Communicates clinical reasoning for problems.	
			Documents problems as specific diagnoses (when known) or as symptoms/conditions accompanied by a differential diagnosis.	
			Communicates clear diagnostic/therapeutic plans for problems.	
Global Assessment 1			**As pertaining to the elements of a note, the learner can "document an initial hospital encounter."**	
			Documentation requires critical additions by supervisor to exist as part of the medical record: Documentation is critically deficient. Entire sections are omitted. Multiple internal inconsistencies are present.	
			Documentation requires substantial additions by supervisor: All sections of the note are present, but some elements may be missing or incomplete. A few internal inconsistencies may be present.	
			Documentation requires minimal additions by supervisor: All sections of the note are present and complete. Rare internal inconsistencies are present.	
			Documentation requires little more than co-signature by supervisor: All sections of the note are present and complete. No internal inconsistencies are present.	
			Documentation is an example of aspirational performance: Documentation could be used as a textbook example to train others.	
Global Assessment 2			**As pertaining to the quality of a note, the learner can "demonstrate ability to synthesize and document clinical reasoning during an initial hospital encounter."**	
			Documentation requires critical clarifications by supervisor to exist as part of the medical record: Clinical reasoning is critically deficient.	
			Documentation requires substantial clarifications by supervisor: Clinical reasoning is demonstrated but limited or hard to follow. May not recognize patient's central problem.	
			Documentation requires minimal clarifications by supervisor: Clinical reasoning is more substantive and easier to follow, but may still require expansion at times. Consistent recognition of patient's central problem and prioritization of secondary problems.	
			Documentation requires little more than co-signature by supervisor: Ready for independent practice. Clinical reasoning is thorough while remaining concise, accurate, and communicated effectively. The note facilitates collaboration to enhance patient care.	
			Documentation is an example of aspirational performance: Documentation could be used as a textbook example to train others.	

Fig. 1 Admission Note Assessment Tool (ANAT) The ANAT consists of 16 discrete checklist items: the first 13 items are scored as "met" or "not met" and the last three checklist items are scored as "met", "partially met", or "not met". The two global assessment items are scored on a five-point behaviorally-based scale mapped to ACGME sub-competencies. Global assessment item one is mapped to ICS-3 ("appropriate utilization and completion of health records") and PROF-4 ("exhibits integrity and ethical behavior in professional conduct"), while global assessment item two is mapped to PC-1 ["gathers and synthesizes essential and accurate information to define each patient's clinical problem(s)"] and ICS-2 ["communicates effectively in interprofessional teams (e.g. peers, consultants, nursing, ancillary professionals and other support personnel)"].

Assessment Tool Piloting

Raters were trained prior to each pilot. Rater training consisted of a one-hour training session with the principal investigator (DW) in-person or via conference call. Each component of the tool was explained, and raters were instructed on proper use of the tool through situational examples, the rater training manual, and simulated review of a sample note.

To determine if ANAT could be used by any assessor without firsthand knowledge of the patient, chart review, or review of other notes, pilot testing was performed comparing admission note ratings by the supervising attending to note ratings by study team raters. For this comparison to be legitimate, notes needed to be reviewed in close proximity to the supervising attending's time on service, to ensure memory of the patient and limit recall bias. This was a crucial step to determine whether firsthand knowledge of the patient is necessary for note assessment. The number of notes available was limited by the number of patients seen and thus determined the number of additional study team raters needed to power reliability calculations. In pilot one, a total of 28 notes were assessed: 18 notes were assessed by the supervising attending and two study team raters; and an additional 10 notes were assessed by the same two study team raters only. Results of ratings by the supervising attending and study team raters (i.e. first 18 notes) were then compared to ratings by the study team raters alone (i.e. all 28 notes).

After reviewing the results from pilot one, the tool was refined using the same iterative process described previously. Feedback from the supervising attending was also incorporated into the discussion. A second round of pilot testing was performed, again using a supervising attending and study team raters. A new supervising attending was used, again for note review to be completed in close proximity to the supervising attending's time on service. Based on power calculations, additional study team raters were used to decrease the number of notes needed for review. In pilot two, a total of 15 notes were assessed: 13 notes were assessed by the supervising attending and four study team raters; and an additional two notes were assessed by the same four study team raters only. Results of ratings by the supervising attending and study team raters were again compared to ratings by the study team raters alone. During pilot two, raters recorded time spent on each note assessment. After reviewing the results from pilot two and discussing the group's experience using the tool, the group felt no further revisions were needed.

A final pilot was performed with one study team rater and three attending physicians from other institutions using the finalized tool to assess feasibility of using the tool at a different institution with similar rater training. A total of 18 notes were reviewed by all four raters. Results of ratings by the study team rater and the three attending physicians from other institutions were compared to ratings by the three attending physicians from other institutions alone.

Statistical Analysis

For interrater reliability calculations we assigned scores between 0 and 1 for each individual item on the ANAT. For all items, ratings were scored as follows: "not met" as 0, "met" as 1; when applicable, "partially met" as 0.5. The two global assessment items were rated on a five-point scale. As discussed by de Vet et al.,[18] agreement parameters should be used for an instrument developed for evaluative purposes where only measurement error of the instrument itself matters and not the variability between subject matters. Thus, average percent agreement was used to measure interrater reliability. All data was analyzed using R, version 3.3.3.[19]

RESULTS

The ANAT (Fig. 1) includes 16 checklist items that are key "elements of the note" and two global assessment items.

The results of pilot two testing with five raters, consisting of the supervising attending and four study team raters, can be found in Table 1. Rater agreement ranged from 86% to 100% for thirteen of the items. The three A&P items had rater agreement ranging from 72% to 81%. Rater agreement on the two global assessment items was 69% and 68% respectively. Results of ratings by the supervising attending and study team raters, compared to ratings by the study team raters alone, can be seen in Table 1. Results from pilot testing with raters from other institutions can be seen in Table 2.

Overall, raters in pilot two took an average of 12.3 min to complete the note assessment (SD 3.7). Supervising attendings took an average of 11.2 min (SD 1.9) while study team raters took an average of 12.6 min (SD 3.9).

Using Messick's validity framework we gathered content, response process, and internal structure validity evidence for the ANAT. The ANAT was developed by faculty members with content expertise in tool development, learner assessment, and billable documentation, evidence for the content validity of our tool. Our think-alouds during consensus building, standardized rater training, and comprehensive rater training manual created from this process generated evidence of response process validity. The finding that scores from the supervising attending were comparable to scores from the study team raters (Table 1) demonstrated internal structure validity.

DISCUSSION

The ANAT had high agreement for simple and objective items (e.g. chief complaint) while it had lower agreement for more complex items, such as ROS and PE, and more subjective items like the A&P items and global assessment items. Less agreement on subjective items is similar to other published tools. For example, the RED checklist had lower agreement

Table 1 Results of Pilot Testing Comparing a Supervising Attending Rater and Study Team Raters

ANAT item	% agreement, 5 raters*	% agreement, 4 raters †
CC	100%	100%
HPI	91%	92%
ROS	86%	85%
PFSH 1	100%	100%
PFSH 2	100%	100%
PFSH 3	100%	100%
PFSH 4	97%	97%
PFSH 5	96%	95%
Exam	93%	97%
DD 1	90%	88%
DD 2	96%	95%
DD 3	97%	97%
DD 4	99%	98%
A&P 1	81%	83%
A&P 2	72%	72%
A&P 3	79%	83%
Global Assessment 1	69%	70%
Global Assessment 2	68%	63%

Supervising attending rater compared to study team raters
†*Study team raters only*

Table 2 Results of Pilot Testing Comparing a Study Team Rater and Outside Institution Raters

ANAT Item	% agreement, 4 raters*	% agreement, 3 raters†
CC	100%	100%
HPI	99%	98%
ROS	92%	94%
PFSH 1	99%	98%
PFSH 2	99%	98%
PFSH 3	96%	94%
PFSH 4	96%	96%
PFSH 5	96%	96%
Exam	83%	83%
DD 1	76%	70%
DD 2	57%	72%
DD 3	90%	93%
DD 4	100%	100%
A&P 1	82%	87%
A&P 2	76%	80%
A&P 3	83%	91%
Global Assessment 1	65%	65%
Global Assessment 2	67%	70%

*Study team rater compared to outside institution raters
†Outside institution raters only

for their A&P items with 71% agreement for the item "active problems are accompanied by clinical reasoning," and as low as 51% agreement for the item "problems are associated with brief, clear plans."[13]

Lower agreement for the global assessment items is also expected given the nuances assessors bring to global assessment. Our behaviorally-anchored scale includes guidance for rating the global assessment items and faculty were trained to rate based on amount of clarification/editing *needed*, not their own documentation attestation practices. However, just like variation seen in actual practice, faculty bring their own interpretations to assessment of performance.[20, 21] Further, the global assessment items are rated on a five-point scale, making perfect agreement less likely.

Results were similar between the supervising attending and study team raters across all items, indicating that use of the ANAT without prior knowledge of the patient is expected to yield similar results to use of the ANAT by the supervising attending. Average time to complete the note assessment was 12 min. These findings improve feasibility of implementing a system for routine assessment of trainee notes, as assessments are not limited to the supervising attending, can be completed quickly, and can be completed at any time after note completion. Results were similar between raters from other institutions and a study team rater, suggesting that the ANAT can be implemented at other institutions with minimal rater training.

Unlike other published tools, the ANAT serves multiple functions. None of the published tools incorporate billing criteria, while ANAT does. Additionally, items in ANAT assess incorporation of irrelevant historical labs and imaging. These items target note "bloat" and other pitfalls of EHR shortcuts without depending on adoption of a particular note template, unlike published interventions. The ANAT focuses on assessing documentation behaviors not limited to one EHR or one institution, suggesting the ANAT can be easily adopted by others. Maximizing assessment opportunities to provide additional feedback on practical skills, such as billing criteria and appropriate use of the EHR, provides trainees with important education without additional faculty effort. PDQI-9 and the RED checklist require assessors to review information other than the individual note assessed, while the ANAT only requires review of the note being assessed, further maximizing faculty effort.

The ANAT evaluates individual components of notes via a checklist and provides global assessment via the two global assessment items. This structure addresses the need to provide both types of assessment for clinical documentation. A checklist is well suited to assess discrete, concrete items such as components of a note and billing criteria, also allowing direct feedback to specific portions of a note. Whereas, global assessment is needed for more complex skills not well captured in a checklist, such as the learner's ability to synthesize information and communicate their thought process. The only other assessment tool that evaluates individual components of notes and evaluates global measures is the RED checklist developed by Bierman et al.[13] Unlike the RED checklist, our global assessment items utilize a behaviorally-based numerical scale and maps these items to the sub-competencies for Internal Medicine within the ACGME milestone framework, an important part of ANAT's construct.[22, 23] In the future, any resident who receives a level one will get flagged by our clinical competency committee for further review as part of our standard review processes. This will allow us to track and intervene on critical deficiencies in documentation.[24]

The ANAT provides space for narrative comments allowing for specific, directed feedback not otherwise captured in the checklist ratings. Incorporation of narrative feedback is highlighted throughout the rater training process to ensure assessment is not limited to checklist criteria, to allow more subjective measures of quality to be assessed via narrative feedback, and to instruct assessors on aspects of the note that should inform the global assessment ratings. The emphasis on narrative comments throughout the ANAT has the ability to generate rich formative feedback aimed at concrete, actionable improvement, which is less of a focus with other published tools.

This study is limited by being conducted at a single institution within a single program, although pilot testing included attendings from multiple institutions. Another limitation is the lack of comparison of ANAT results to results using other published tools or other standards of documentation practice. Additionally, the ANAT incorporates billing criteria in the assessment items that may not be applicable in the future if billing criteria is modified. However, a change in billing criteria would require minimal tool modification which would not impact the tool's underlying assessment construct. Importantly, we do not yet know if use of the ANAT will affect documentation behaviors amongst learners. Further, we have not yet established evidence of relationship to other variables or consequence validity, although including global assessment items mapped to the ACGME milestone framework is an important step to being able to establish this validity evidence in the future.

CONCLUSIONS

We present initial validity evidence for the ANAT that uniquely serves multiple functions by incorporating billing criteria, targeting note "bloat," assessing individual note elements, and utilizing global assessment items mapped to the ACGME milestone framework while simultaneously providing an opportunity for narrative feedback. Routine use of the ANAT as a part of trainee assessment is feasible given that any attending can complete the evaluation with minimal training and minimal time required, with overall high agreement expected. Further testing of the ANAT should be undertaken to continue to build validity evidence and to see how the tool performs as part of routine assessment. Next steps should include development of a documentation curriculum that can be implemented along with the ANAT, assessment of documentation behaviors after implementation of ANAT, and spread to other institutions.

Acknowledgements:

Contributors: *Lindsay Huber, MD[1].*
Adam Rose, MD, MSc[1].
Nick Favela, MD[2].
Ryan Munyon, MD[3].
Bo Stapler, MD[4].
1. University of Cincinnati College of Medicine, Department of Internal Medicine, University of Cincinnati Medical Center, Cincinnati, OH.
2. Department of Internal Medicine and Department of Pediatrics, Maury Regional Medical Center, Columbia, TN.
3. Pennsylvania State College of Medicine, Department of Internal Medicine, Penn State Health Milton S. Hershey Medical Center, Hershey, PA.
4. Department of Internal Medicine and Department of Pediatrics, St. Vincent Healthcare, Billings, MT.

Corresponding Author: *Danielle E. Weber, MD, MEd; Department of Internal Medicine University of Cincinnati College of Medicine, University of Cincinnati Medical Center, 231 Albert Sabin Way, ML 0535, Cincinnati, OH 45267, USA (e-mail: weber2de@ucmail.uc.edu).*

Funders *none.*

Compliance with Ethical Standards:

Conflict of Interest: *Dr. Jennifer K. O'Toole is a paid consultant for the I-PASS Patient Safety Institute. She has no other conflicts of interest to disclose. There are no conflicts of interest to disclose for any of the other authors.*

REFERENCES

1. **Burke HB**, **Hoang A**, **Becher D**, et al. QNOTE: an instrument for measuring the quality of EHR clinical notes. J Am Med Inform Assoc. 2014;21(5):910–6.
2. **Burke HB**, **Sessums LL**, **Hoang A**, et al. Electronic health records improve clinical note quality. J Am Med Inform Assoc. 2014;22(1):199–205.
3. **Englander R**, **Flynn T**, **Call S**. Toward Defining the Foundation of the MD Degree: Core Entrustable Professional Activities for Entering Residency. Acad Med. 2016;91(10):1352–8.
4. The Internal Medicine Milestone Project. 2015. Available at: http://www.acgme.org/Specialties/Milestones/pfcatid/2/Internal%20Medicine. Accessed March 26, 2018.
5. **Kuhn T**, **Basch P**, **Barr M**, **Yackel T**. Clinical documentation in the 21st century: executive summary of a policy position paper from the American College of Physicians. Ann Intern Med. 2015;162(4):301–3.
6. **Stetson PD**, **Morrison FP**, **Bakken S**, Johnson SB; eNote Research Team. Preliminary development of the physician documentation quality instrument. J Am Med Inform Assoc. 2008;15(4):534–541.
7. **Dean SM**, **Eickhoff JC**, **Bakel LA**. The effectiveness of a bundled intervention to improve resident progress notes in an electronic health record (EHR). J Hosp Med. 2015;10(2):104–7.
8. **Stetson PD**, **Bakken S**, **Wrenn JO**, **Siegler EL**. Assessing electronic note quality using the physician documentation quality instrument (PDQI-9). Appl Clin Inf. 2012;3(2):164–174.
9. **Embi PJ**, **Yackel TR**, **Logan JR**, **Bowen JL**, **Cooney TG**, **Gorman PN**. Impacts of computerized physician documentation in a teaching hospital: perceptions of faculty and resident physicians. J Am Med Inform Assoc. 2004;11(4):300–9.
10. **O'Donnell HC**, **Kaushal R**, **Barron Y**, **Callahan MA**, **Adelman RD**, **Siegler EL**. Physicians' attitudes towards copy and pasting in electronic note writing. J Gen Intern Med. 2009;24(1):63–8.
11. **Aylor M**, **Campbell EM**, **Winter C**, **Phillipi CA**. Resident Notes in an Electronic Health Record: A Mixed-Methods Study Using a Standardized Intervention With Qualitative Analysis. Clin Pediatr. 2017;56(3):257–262.
12. **Hommos MS**, **Kuperman EF**, **Kamath A**, **Kreiter CD**. The Development and Evaluation of a Novel Instrument Assessing Residents' Discharge Summaries. Acad Med. 2017;92(4):550–5.
13. **Bierman JA**, **Hufmeyer KK**, **Liss DT**, **Weaver AC**, **Heiman HL**. Promoting Responsible Electronic Documentation: Validity Evidence for a Checklist to Assess Progress Notes in the Electronic Health Record. Teach Learn Med. 2017;29(4):420–432.
14. **Cook DA**, **Beckman TJ**. Current Concepts in Validity and Reliability for Psychometric Instruments: Theory and Application. Am J of Med. 2006;199(2):166.e7–177.e16.
15. **Innes JE**, **Booher DE**. Consensus Building and Complex Adaptive Systems. J Am Plan Assoc. 1999;65(4):412–423.
16. **Norcini J**, **Anderson B**, **Bollela V**, et al. Criteria for good assessment: Consensus statement and recommendations from the Ottawa 2010 Conference. Med Teach. 2011;33(3):206–214.
17. Medicare Learning Network. Evaluation and Management Services. Available at: https://www.cms.gov/Outreach-and-Education/Medicare-Learning-Network-MLN/MLNProducts/Downloads/eval-mgmt-serv-guide-ICN006764.pdf. Accessed June 19, 2019.
18. **de Vet HCW**, **Terwee CB**, **Knol DL**, **Bouter LM**. When to use agreement versus reliability measures. J Clin Epidemiol. 2006;59(10):1033–9.
19. R Core Team. R: A language and environment for statistical computing. 2012. Available at: http://www.R-project.org/. Accessed June 21, 2019.
20. **Lockyer J**, **Carraccio C**, **Chen MK**, et al. Core principles of assessment in competency-based medical education. Med Teach. 2017;39(6):609–16.
21. **Holmboe ES**, **Sherbino J**, **Englander R**, **Snell L**, **Frank JR**, ICBME Collaborators. A call to action: The controversy of and rationale for competency-based medical education. Med Teach. 2017;39(6):574–81.
22. **Warm EJ**, **Mathis BR**, **Held JD**, et al. Entrustment and Mapping of Observable Practice Activities for Resident Assessment. J Gen Intern Med. 2014;29(8):1177–82.
23. **Warm EJ**, **Held JD**, **Hellmann M**, et al. Entrusting Observable Practice Activities and Milestones Over the 36 Months of an Internal Medicine Residency. Acad Med. 2016;91(10):1398–1405.
24. **Kinnear B**, **Bensman R**, **Held J**, **O'Toole J**, **Schauer D**, **Warm E**. Critical Deficiency Ratings in Milestone Assessment: A Review and Case Study. Acad Med. 2017;92(6):820–6.