Paul A. Bergl*, Neehal Shukla, Jatan Shah, Marium Khan, Jayshil J. Patel and Rahul S. Nanchal

Factors influencing diagnostic accuracy among intensive care unit clinicians – an observational study

https://doi.org/10.1515/dx-2023-0026
Received March 3, 2023; accepted November 2, 2023;
published online November 30, 2023

Abstract

Objectives: Diagnostic errors are a source of morbidity and mortality in intensive care unit (ICU) patients. However, contextual factors influencing clinicians' diagnostic performance have not been studied in authentic ICU settings. We sought to determine the accuracy of ICU clinicians' diagnostic impressions and to characterize how various contextual factors, including self-reported stress levels and perceptions about the patient's prognosis and complexity, impact diagnostic accuracy. We also explored diagnostic calibration, i.e. the balance of accuracy and confidence, among ICU clinicians.

Methods: We conducted an observational cohort study in an academic medical ICU. Between June and August 2019, we interviewed ICU clinicians during routine care about their patients' diagnoses, their confidence, and other contextual factors. Subsequently, using adjudicated final diagnoses as the reference standard, two investigators independently rated clinicians' diagnostic accuracy and on each patient on a given day ("patient-day") using 5-point Likert scales. We conducted analyses using both restrictive and conservative definitions of clinicians' accuracy based on the two reviewers' ratings of accuracy.

Results: We reviewed clinicians' responses for 464 unique patient-days, which included 255 total patients. Attending physicians had the greatest diagnostic accuracy (77–90 %, rated as three or higher on 5-point Likert scale) followed by the team's primary fellow (73–88 %). Attending physician and fellows were also least affected by contextual factors. Diagnostic calibration was greatest among ICU fellows.

Conclusions: Additional studies are needed to better understand how contextual factors influence different clinicians' diagnostic reasoning in the ICU.

Keywords: diagnostic error; prognosis/diagnosis; patient safety; crew resource management; healthcare; clinical reasoning

Introduction

Although progress has been made in intensive care unit (ICU) patient safety [1, 2], diagnostic errors remain a pervasive problem [3–5]. Annually, approximately 400,000 intensive care unit (ICU) patients experience a diagnostic error in the United States alone [4]. Diagnostic errors are discovered in nearly 30 % of autopsied ICU decedents, and an estimated one in 16 ICU deaths are attributable to misdiagnosis [5].

Notwithstanding, few, if any, studies have prospectively how contextual features, such as a clinician's level of stress, confidence, and perceptions of the patient [6], might influence diagnostic accuracy in authentic ICU settings. Existing frameworks for studying diagnostic error among the critically ill require cognitive failures to be judged retrospectively through patient record reviews [3, 7] a process that fails to capture these contextual features. While clinicians report that interpersonal, organizational, and interactional factors may negatively influence diagnostic accuracy [8], their effects on diagnostic accuracy have been incompletely quantified *in situ* in the ICU. Finally, the relationship between ICU clinicians' confidence and diagnostic accuracy, or "diagnostic calibration," [9, 10] has not been previously characterized.

Paul Bergl and Neehal Shukla contributed equally to this work and share first authorship.

This study was completed at Froedtert Hospital, Milwaukee, WI, affiliated to the Medical College of Wisconsin.

***Corresponding author: Paul A. Bergl**, MD, Department of Critical Care, Gundersen Health System, 1900 South Ave, Mail Stop LM3-001, La Crosse 54650, WI, USA; and Department of Medicine, University of Wisconsin-Madison School of Medicine and Public Health, Madison, WI, USA, E-mail: pabergl@gundersenhealth.org
Neehal Shukla, Cleveland Clinic Foundation, Internal Medicine Residency Program, Cleveland, OH, USA
Jatan Shah, University of Pittsburgh Medical Centre Chautauqua, Jamestown, NY, USA
Marium Khan, Medical College of Wisconsin Affiliated Hospitals, Milwaukee, WI, USA
Jayshil J. Patel and Rahul S. Nanchal, Department of Medicine, Division of Pulmonary, Critical Care, and Sleep Medicine, Medical College of Wisconsin, Milwaukee, WI, USA

We aimed to explore how contextual factors potentially influence clinicians' diagnostic accuracy in the ICU. We sought to accomplish this aim by first prospectively collecting clinicians' diagnostic impressions and determining their degree of accuracy through comparisons against final diagnoses adjudicated in retrospective chart review. We then analyzed how contextual factors, such as clinicians' roles on the ICU team, self-reported stress, confidence, and perceptions about the patient's prognosis and complexity impacted their degree of diagnostic accuracy.

Methods

We conducted an observational study in a medical ICU at an academic tertiary care hospital in three phases: (1) prospective collection of participants' responses between June 2019 and August 2019, (2) retrospective, blinded adjudication of final diagnoses through chart review, and (3) retrospective rating of clinicians' diagnostic accuracies, specificities, and team consensus compared to adjudicated diagnoses.

Study subjects

Participants included ICU team members: attending physicians, fellows, resident physicians, advanced practitioners (APs), and nurses.

Human subject safety and consent

The Medical College of Wisconsin/Froedtert Hospital Institutional Review Board (IRB) reviewed this project (PRO00034633) and deemed it exempt from full IRB oversight. Written informed consent with participants was waived, and we notified participants by informational letters. The IRB did not consider patients as subjects, thereby eliminating the need for consent. We received a waiver from HIPAA authorization given the limited patient health information required.

Inclusions and exclusions

We included all admissions to the MICU service between June 16, 2019 and August 1, 2019. We excluded admissions in which only one team member rendered a diagnosis because this study was part of a larger observational study designed to also examine how team members' diagnoses aligned. There were no other inclusion or exclusion criteria applied; participants continued to give their response even for patients in a whom a clear primary diagnosis had emerged.

Definitions

Commonly accepted definitions of diagnostic error include failure to establish a timely diagnosis and/or effectively convey the diagnosis to the patient or surrogate [3, 7]. Because we did not seek to determine timeliness of the diagnosis and did not directly involve patients, we defined our variables as follows:

(1) *Diagnostic accuracy* – How well working hypotheses matched the ultimate adjudicated diagnosis
(2) *Diagnostic specificity* – How well working hypotheses suggested a clear, limited diagnosis and/or conveyed an unambiguous idea with a narrow array of interpretations

Primary data collection and variables

We interviewed participants about their patients' primary diagnoses and contextual factors. We collected diagnoses as open-ended responses and all other variables on ordinal 5-point Likert scales. Interview questions and their relationship to contextual factors are shown in Table 1.

To gather these data, an investigator (NS) conducted brief interviews on or after rounds with each team member to complete the questionnaire. Nurses, residents, and APs were approached immediately after the team rounded on each patient. Attendings and fellows were interviewed at the conclusion of rounds, and the night fellow completed the questionnaire for new overnight admissions at their shift's end. We attempted to query all subjects each day, but participants were not interviewed if competing clinical demands rendered them unavailable. Fellows were explicitly asked if they were present on rounds as they may have departed rounds to address other patient care situations.

Table 1: A full list of variables and sample questions that subjects were surveyed on daily for the first 72 h of patient admissions.

All subjects
Accuracy "What is the primary diagnosis for which this patient is being treated?"
Confidence "How confident are you that this diagnosis is correct?"
Stress "How would you rate your level of stress right now?"
Prognosis "How likely is this patient to survive to hospital discharge?"[a] Or "How likely is this patient to be alive and without significant cognitive or physical limitations in 1 month?"[a] [a]Our designated institutional official in graduate medical education requested that trainees (residents and fellows) not be asked about survival per se.
Nurses, residents, and APs
Complexity "How complex would you consider this patient?"
Nurses only
Teamwork "To what extent has the team incorporated you into today's care plan for this patient?"
Residents and APs only
Open-mindedness "To what extent is the team pursuing alternate diagnoses for this patient?"
Attendings, fellows, and night fellows
Presence on rounds "Were you present on rounds when this patient was discussed?"

We did not gather any identifying information on participants, such as demographic data or years of clinical experience, beyond their professional role (nurse, AP, etc.). At the time of this study, all APs had fewer than 4 years of ICU experience as an AP.

Interviews were only conducted on the first three days of patients' ICU admissions as we posited, from experience, that most patients would have established diagnoses by this time. For patients with ICU length of stay shorter than three days, we only collected data for the duration of the ICU stay. The unit of time for this study was a "patient-day" defined as all of the data gathered from available participants (i.e. nurses, residents, etc.) on a given patient on a given day. Thus, a patient with an ICU length of stay of less than 24 h would only have one patient-day in our data set while patients with ICU length of stay greater than three days would have the maximum of three patient-days in our data set.

To qualify as a patient-day, at least two members of the care team (i.e. nurses, residents, etc.) needed to partake in interviews. We did not track whether the same participants responded on subsequent days for the same patient, and, again, we could not assure that all members of the care team would be available for interviews on each patient-day.

Primary and secondary outcomes

Our primary outcomes were individual clinicians' diagnostic accuracy and diagnostic specificity. Secondary outcomes included the interactions among clinicians' diagnostic accuracy, specificity, and the contextual variables in Table 1.

Of note, in final analyses, resident and AP responses were combined into "primary clinician" responses for two reasons. First, patients were assigned to a resident or AP but never both. Secondly, we did not identify any statistically significant differences between these groups for any variable of interest during preliminary analyses.

Adjudication of diagnoses

We adjudicated the ultimate primary diagnosis for each patient's ICU stay based on retrospective chart review using all available clinical data in the medical record (e.g. subsequent clinician documentation, discharge diagnoses, pathology reports, culture data, imaging results, etc.). Adjudication occurred at least one month, but up to 18 months, after patients were discharged from the ICU and was performed by an intensivist (PB, 160 patients' records reviewed), critical care fellow (JS, 47 records reviewed), and/or internist (MK, 50 records reviewed). For cases with an unclear diagnosis (2 patients), at least two team members independently reviewed the chart and, through discussion, agreed upon a final adjudicated diagnosis. While performing chart review, adjudicators were blinded to diagnoses provided by participants on rounds. Investigators did not make any systematic attempt to link participants' responses on questionnaires to the unblinded data collected through chart review.

Determination of diagnostic accuracy and specificity

Two investigators (PB and NS) independently scored clinicians' diagnostic accuracy and specificity for each patient-day on 5-point Likert scales (Supplemental Table 1). The investigators were blinded to each other's ratings and did not access the patient's record during these ratings. Raters simply determined accuracy by comparing each clinician's diagnostic impression against the ultimate adjudicated diagnosis. To ensure consistency, the two investigators performed multiple rating calibrations with each other as outlined in the Supplemental Appendix.

Prior to additional data analysis, investigators' scores were also dichotomized with values of 1–2 being reassigned a value of 0 (i.e. inaccurate or nonspecific) and scale values of 3–5 reassigned a value of 1 (i.e. accurate or specific).

Statistical analyses

Statistical analyses were performed on both dichotomized and ordinal data. We calculated agreement between investigators' ratings of accuracy and specificity using Cohen's kappa statistics [11, 12] on dichotomized data and unweighted Cohen's kappa, linearly weighted Cohen's kappa, and intra-class correlation coefficients (ICC) on ordinal data.

For diagnostic accuracy and specificity, we report two proportions based on whether both investigators found the diagnosis accurate, specific, and/or agreed upon by the team (a "restrictive" definition of accuracy, specificity, or agreement) or only one investigator found the diagnosis accurate, specific, and/or agreed upon by the team ("conservative" definition).

We made pairwise comparisons to examine relationships between diagnostic accuracy, specificity, and contextual factors using chi-square tests on dichotomized data and Spearman's rho using a mean accuracy and specificity score from both raters against the ordinal data. Similarly, we made pairwise comparisons for diagnostic accuracy between clinician groups using the Kruskal–Wallis test with subsequent two-tailed Mann–Whitney tests for direct comparisons. As mentioned previously and as elaborated in the results, we combined resident and AP responses into a single "resident/AP" clinician type for most analyses after demonstrating that these clinicians had no significant differences in accuracy, specificity, or confidence.

In this study, we calculated diagnostic calibration scores to determine the relationships between accuracy and confidence and between specificity and confidence as detailed in the Supplemental Appendix. We compared clinicians' calibration scores using Kruskal–Wallis tests.

Finally, for variables significantly associated with diagnostic accuracy, we computed diagnostic test performance characteristics.

Results

We analyzed ICU clinicians' responses for 464 unique patient-days that included 255 patients. Diagnoses from primary daytime fellows were available on 456 patient-days (98.3 %), residents or APs on 448 patient-days (96.6 %), attending physicians on 383 patient-days (82.5 %), nurses on 307 patient-days (66.1 %), and night fellows on 94 patient-days (20.2 %). The Supplemental Appendix contains a comprehensive list of final adjudicated diagnoses.

Individual clinicians' accuracy

Using dichotomized data, the two raters' agreement on diagnostic accuracy scores was moderate with a simple agreement of 85.5 % and Cohen's kappa of 0.577 for 1,677

Table 2: Individual clinicians' accuracy and specificity. Accuracy and specificity are first listed as a distribution of ratings and then as a percentage using both restrictive (lower bound) and conservative (upper bound) definitions (see text for definitions).

Clinician	Diagnostic accuracy, distribution of ratings	Diagnostic accuracy, %	Diagnostic specificity, distribution of ratings	Diagnostic specificity, %	Sample size in patient-days
Attending	**1:** 54 (6.4 %)	77.3–89.6	**1:** 27 (3.3 %)	75.7–92.0	383
	2: 73 (8.7 %)		**2:** 91 (11.0 %)		
	3: 223 (26.6 %)		**3:** 215 (26.1 %)		
	4: 270 (32.2 %)		**4:** 269 (32.6 %)		
	5: 146 (17.4 %)		**5:** 157 (19.0 %)		
Fellow	**1:** 73 (7.8 %)	73.2–87.5	**1:** 73 (8.1 %)	65.1–85.9	456
	2: 103 (11.0 %)		**2:** 139 (15.5 %)		
	3: 310 (33.0 %)		**3:** 284 (31.6 %)		
	4: 260 (27.7 %)		**4:** 240 (26.7 %)		
	5: 148 (15.8 %)		**5:** 151 (16.8 %)		
Night fellow	**1:** 15 (8.1 %)	74.5–84.0	**1:** 5 (2.7 %)	67.7–92.4	94
	2: 22 (11.8 %)		**2:** 26 (13.8 %)		
	3: 48 (25.8 %)		**3:** 40 (21.3 %)		
	4: 59 (31.7 %)		**4:** 63 (33.5 %)		
	5: 32 (17.2 %)		**5:** 40 (21.3 %)		
Resident/AP (combined)	**1:** 70 (7.7 %)	69.0–85.9	**1:** 92 (10.6 %)	62.5–83.3	449
	2: 125 (13.8 %)		**2:** 138 (15.9 %)		
	3: 281 (31.0 %)		**3:** 274 (31.5 %)		
	4: 242 (26.7 %)		**4:** 232 (26.7 %)		
	5: 156 (17.2 %)		**5:** 136 (15.6 %)		
Resident	**1:** 47 (6.9 %)	70.6–86.2	**1:** 66 (10.3 %)	62.1–82.8	340
	2: 94 (13.9 %)		**2:** 111 (17.3 %)		
	3: 212 (31.3 %)		**3:** 204 (31.8 %)		
	4: 200 (29.5 %)		**4:** 190 (29.6 %)		
	5: 109 (16.1 %)		**5:** 91 (14.2 %)		
AP	**1:** 23 (10.1 %)	63.3–85.3	**1:** 26 (11.4 %)	63.8–84.8	109
	2: 31 (13.6 %)		**2:** 27 (11.8 %)		
	3: 69 (30.3 %)		**3:** 70 (30.7 %)		
	4: 42 (18.4 %)		**4:** 42 (18.4 %)		
	5: 47 (20.6 %)		**5:** 45 (19.7 %)		
Nurse	**1:** 93 (16.0 %)	58.3–75.9	**1:** 120 (23.1 %)	44.0–71.9	307
	2: 110 (19.0 %)		**2:** 132 (25.4 %)		
	3: 209 (36.0 %)		**3:** 187 (36.0 %)		
	4: 120 (20.7 %)		**4:** 114 (22.0 %)		
	5: 79 (13.6 %)		**5:** 49 (9.4 %)		

AP, advanced practitioner.

independent ratings. Using the ordinal data on a 5-point scale, agreement was fair to moderate with unweighted Cohen's kappa of 0.329, linearly weighted Cohen's kappa of 0.534, and intra-class correlation coefficient of 0.706.

When combining all clinician types, individual diagnostic accuracy ranged from 70.8–85.3 % (restrictive and conservative definitions). Table 2 demonstrates diagnostic accuracy by clinician, and Table 3 includes results of Kruskal–Wallis tests comparing clinicians' accuracy. We observed the following differences in accuracy rates:

- **Attending physicians** were the most accurate using either a restrictive or conservative definition. Their mean accuracy ratings were 3.5.
- Both daytime and night float **fellows** were slightly less accurate compared to attending physicians with mean accuracy ratings of 3.4, but this difference was not statistically significant in Mann–Whitney tests. Fellows were slightly more accurate when present on rounds for the patient's presentation (87.0 % accurate vs. 77.2 %, significant only using conservative definition, $p<0.05$, chi-square test).
- **Residents and APs** did not demonstrate any significant difference in accuracy using conservative or restrictive definitions or the 5-point rating scales (Mann–Whitney tests). Residents and APs were significantly less accurate compared to attending physicians in the Kruskal–Wallis

Table 3: This table depicts p-values for results of the Kruskall–Wallis test comparing clinicians' accuracies, specificities, and confidence against each other.

Results of Kruskall–Wallis tests for accuracy, specificity and confidence amongst clinicians					
	Accuracy (restrictive definition)				
	Nurse	**Resident/APP**	**Fellow**	**Night fellow**	**Attending**
Nurse	1	0.100	**<0.0001**	**0.038**	**<0.0001**
Resident/AP	0.100	1	0.238	0.620	**0.009**
Fellow	**<0.0001**	0.238	1	0.999	0.662
Night Fellow	**0.038**	0.620	0.999	1	0.978
Attending	**<0.0001**	**0.009**	0.662	0.978	1
	Accuracy (mean rating for two reviewers, 5-point Likert)				
	Nurse	**Resident/APP**	**Fellow**	**Night fellow**	**Attending**
Nurse	1	**<0.0001**	**<0.0001**	**0.005**	**<0.0001**
Resident/AP	**<0.0001**	1	1.000	0.940	0.184
Fellow	**<0.0001**	1.000	1	0.935	0.986
Night Fellow	**0.005**	0.940	0.935	1	0.970
Attending	**<0.0001**	0.184	0.986	0.970	1
	Specificity (restrictive definition)				
	Nurse	**Resident/APP**	**Fellow**	**Night fellow**	**Attending**
Nurse	1	**<0.0001**	**<0.0001**	**0.001**	**<0.0001**
Resident/AP	**<0.0001**	1	0.450	0.595	**<0.0001**
Fellow	**<0.0001**	0.450	1	0.988	**0.007**
Night Fellow	**0.001**	0.595	0.988	1	0.514
Attending	**<0.0001**	**<0.0001**	**0.007**	0.514	1
	Specificity (mean rating for two reviewers, 5-point Likert)				
	Nurse	**Resident/APP**	**Fellow**	**Night fellow**	**Attending**
Nurse	1	**<0.0001**	**<0.0001**	**<0.0001**	**<0.0001**
Resident/AP	<0.0001	1	0.792	**0.039**	**<0.0001**
Fellow	**<0.0001**	0.792	1	0.174	**0.001**
Night Fellow	**<0.0001**	**0.039**	0.174	1	1.000
Attending	**<0.0001**	**<0.0001**	**0.001**	1.000	1
	Confidence (5-point Likert scale)				
	Nurse	**Resident/APP**	**Fellow**	**Night fellow**	**Attending**
Nurse	1	**<0.0001**	**<0.0001**	0.785	0.993
Resident/AP	**<0.0001**	1	1.000	**<0.0001**	**<0.0001**
Fellow	**<0.0001**	1.000	1	**<0.0001**	**<0.0001**
Night Fellow	0.785	**<0.0001**	**<0.0001**	1	0.901
Attending	0.993	**<0.0001**	**<0.0001**	0.901	1

AP, advanced practitioner. At an alpha < 0.05, significant differences are shown with bolded values.

tests, but this difference was only significant when the restrictive definition of diagnostic accuracy was applied in Mann–Whitney tests ($p<0.05$). Significant differences were not detected in residents' or APs' accuracy when compared to fellows' accuracy (Mann–Whitney and Kruskal–Wallis tests). Again, because we found no differences in residents' and APs' accuracy at this stage in our analysis, these clinicians are combined as "residents/APs" henceforth unless specifically denoted.

– **Nurses' reported diagnoses** were significantly less accurate than attending physicians and fellows. Their accuracy was slightly worse than resident/APs' accuracy with a significant difference detected when the mean ratings on 5-point scales were used (Kruskal–Wallis and Mann–Whitney tests, $p<0.0001$ for both tests) or when a restrictive definition of accuracy was used in Mann–Whitney tests ($p<0.05$) (see Tables 2 and 3).

Table 4: Correlation coefficients (Spearman's rho) for association between individual clinician diagnostic performance and contextual variables that we posited might influence accuracy.

Diagnostic performance			Contextual factors					
Clinician accuracy	**Specificity**	**Confidence**	**Perception of teamwork**	**Perceived complexity**	**Perceived prognosis**	**Alternate diagnoses being considered**	**Self-reported stress**	**ICU day**
Nurse	r=0.592[a]	r=0.202[a]	r=0.070	r=−0.176[b]	r=0.219[a]		r=−0.068	r=−0.067
Resident/AP	r=0.623[a]	r=0.190[a]		r=−0.193[a]	r=0.213[a]	r=−0.236[a]	r=−0.098	r=−0.085
Fellow	r=0.573[a]	r=0.337[a]			r=0.180[a]		r=−0.021	r=−0.098
Night fellow	r=0.669[a]	r=0.261			r=0.090		r=−0.026	
Attending	r=0.556[a]	r=0.177[a]			r=0.152[c]		r=−0.008	r=−0.026

Significant associations are bolded ([a]p-value<0.0001, [b]p-value<0.001, [c]p-value<0.01).

Individual clinicians' specificity

Agreement on dichotomized specificity ratings between the two investigators was moderate with simple agreement of 79.4 % and Cohen's kappa of 0.472 for 1,657 independent ratings (dichotomous data). Using the ordinal data, we observed only slight agreement (Cohen's κ=0.179) on specificity ratings, which improved to moderate agreement with linearly weighted Cohen's kappa of 0.436 and ICC of 0.648.

Attending physicians were most specific (Table 2 and 3) and significantly more so than any other clinicians besides the night fellows (Kruskal–Wallis test). Both attending physicians and night fellows had mean specificity ratings of 3.6 and had specificity exceeding 90 % using the conservative definition (Table 2). Nurses were the least specific with mean rating of 2.7 (Kruskal–Wallis test). Residents, APs, and fellows' diagnoses exhibited similar specificities (Mann–Whitney and Kruskal–Wallis tests). Tables 2 and 3 summarize these data.

In chi-square analyses on dichotomized data and correlation analyses calculated by Spearman's rho, we found a statistically significant association between accuracy and specificity for all clinicians (Table 4 and Table S2) with Spearman's rho ranging from 0.556 to 0.623 (p<0.0001).

Table 5: Depicts p-values for results of the Kruskal–Wallis test comparing clinicians' accuracies, specificities, and confidence against each other.

Results of Kruskal–Wallis tests for accuracy, specificity, and confidence amongst clinicians					
	Nurse	**Resident/ APP**	**Fellow**	**Night fellow**	**Attending**
Accuracy (restrictive definition)					
Nurse	1	0.100	0.000[a]	0.038[a]	<0.0001[a]
Resident/AP	0.100	1	0.238	0.620	0.009[a]
Fellow	0.000[a]	0.238	1	0.999	0.662
Night fellow	0.038[a]	0.620	0.999	1	0.978
Attending	<0.0001[a]	0.009[a]	0.662	0.978	1
Specificity (restrictive definition)					
Nurse	1	0.000[a]	<0.0001[a]	0.001[a]	<0.0001[a]
Resident/AP	0.000[a]	1	0.450	0.595	<0.0001[a]
Fellow	<0.0001[a]	0.450	1	0.988	0.007[a]
Night fellow	0.001[a]	0.595	0.988	1	0.514
Attending	<0.0001[a]	<0.0001[a]	0.007[a]	0.514	1
Confidence (5-point Likert scale)					
Nurse	1	<0.0001[a]	<0.0001[a]	0.785	0.993
Resident/AP	<0.0001[a]	1	1.000	0.000[a]	<0.0001[a]
Fellow	<0.0001[a]	1.000	1	0.000[a]	<0.0001[a]
Night fellow	0.785	0.000[a]	0.000[a]	1	0.901
Attending	0.993	<0.0001[a]	<0.0001[a]	0.901	1

At an alpha <0.05, significant differences are shown with superscript a.

Individual clinicians' confidence

Nurses, night fellows, and attending physicians had nearly identical high levels of confidence with mean values of 4.6, 4.7, and 4.6 respectively (no significant difference in Kruskal–Wallis or Mann–Whitney tests, distributions in Supplemental Table S3). Residents, APs, and daytime fellows were significantly less confident than other clinicians with mean confidences of 4.4, 4.4, and 4.3 respectively (p<0.001 in Kruskal–Wallis test) but were not significantly different from each other (p>0.05, Mann–Whitney tests). We found a weak but significant correlation between confidence and accuracy, with Spearman's rho ranging from 0.177 to 0.337 across all clinicians (Table 4).

Fellows were significantly more likely to be accurate when confident about the diagnosis; in addition, the correlation between accuracy and confidence was highest for fellows (Spearman's rho=0.337, p<0.0001). Fellows' confidence had a positive LR of 2.36 (95 % CI 1.33–4.20) and negative LR of 0.36 (95 % CI 0.26–0.49) for an accurate diagnosis (Table 5).

Residents/APs were significantly more likely to be accurate when confident (70.4 % accurate vs. 40.9 %, restrictive definition of accuracy, p<0.01, chi-square test). Residents'/APs' lack of confidence was highly sensitive (97.1 %, 95 % CI 94.5–98.7 %) for diagnostic inaccuracy with negative LR of 0.31 (95 % CI 0.14–0.71) for an accurate diagnosis.

Contextual factors associated with individual clinicians' accuracies

Among all clinicians, self-reported stress and the patient's length of stay in the ICU had no correlation with diagnostic accuracy (Table 4). Most associations between contextual factors and accuracy were weakly correlated at best (Table 4). The effects of contextual variables on diagnostic accuracy varied by clinician:

- **Attending physicians'** accuracies were largely uninfluenced by contextual factors. While statistically significant, associations were weak between their accuracy and confidence (r=0.177, p<0.001) and between their accuracy and perceptions about the patient's prognosis (r=0.152, p<0.01).
- **Residents/APs** were significantly more likely to be when they reported that the team was *not* considering other diagnoses (78.3 vs. 58.9 %, restrictive definition of accuracy, p<0.0001, chi-square test). They were also significantly less accurate when they perceived the patient to be complex (64.4 vs. 78.0 %, restrictive definition of accuracy, p<0.01, chi-square test) or that the patient's prognosis was poor (54.3 vs. 73.3 %, restrictive definition, p<0.0001, chi-square test). All these associations were weak correlations (Spearman's rho values in Table 4).
- **Nurses** were significantly less likely to report an accurate diagnosis when they perceived that the patient was complex or had a poor prognosis, both in chi-square tests on dichotomized data and tests of correlation with Spearman's rho.

Table 5 includes the full results for test performance characteristics for select variables affecting individual clinician accuracy. Supplemental Table 2 demonstrates the results of the chi-square analyses for all potential factors that we posited would affect diagnostic accuracy. A comprehensive reporting of all participants' subjective ratings of contextual factors is found in Supplemental Table S3.

Individual clinicians' diagnostic calibration

We found that all clinicians tended toward overconfidence relative to their accuracy as depicted in Supplemental Figure S1. In Kruskal–Wallis tests, APs/residents and primary daytime fellows had significantly better diagnostic accuracy calibration than other clinicians. Except for nurses, all other clinicians exhibited similar specificity calibration. Additional results of calibration tests are included in Supplemental Figure S2 and Table 2.

Discussion

In our study of clinicians' diagnostic impressions on ICU patients, attending physicians were the most accurate and specific and appeared to be the least influenced by contextual variables, including confidence, perceived patient prognosis, or stress level. Fellows had similar accuracy to attending physicians, but their accuracy was significantly higher when confident. The association between accuracy and specificity, while strong for all clinicians, may have simply resulted from their inherent interconnectedness. Importantly, this study does not seek to answer whether high diagnostic accuracy, specificity, and confidence are inherently desirable but rather describes patterns of accuracy and related variables.

Diagnostic calibration has emerged as a description of the balance between clinicians' accuracy and confidence [9, 10]. While previous studies on calibration rely exclusively on clinicians solving cases in simulated settings [10, 13, 14], our study includes an exploratory analysis of calibration in actual work environments. In this ICU, we found that residents/APs and fellows exhibited the best calibration although all clinicians tended toward overconfidence. Because achieving calibration likely results from consistent feedback on individual diagnostic performance over time [9, 15], trainees may have more opportunities to improve calibration. Attending physicians, while diagnostically accurate and specific, lack these feedback systems. Moreover, attending physicians might also express greater confidence to fulfill professional expectations about their competence, whereas trainees may be more open to admitting a lack of confidence. Whether fellows represent ideally calibrated diagnosticians on ICU teams – or whether their role simply positions them to recognize when the team's diagnostic hypotheses are accompanied by appropriate uncertainty – merits further study.

We are not aware of any studies exploring how clinicians' perceptions about ICU patients' prognoses affect diagnostic accuracy. In our study, fellows and attending physicians' accuracy was not influenced by perceived prognosis; thus, amassing clinical experience may allow clinicians to compartmentalize judgments about diagnosis and prognosis. Alternatively, clinicians with lower

diagnostic acumen may reflexively predict a poorer prognosis among critically ill patients whom they cannot accurately diagnose.

Stress and time pressures are often identified as barriers to diagnosis [3, 8, 16–18], but this finding is not consistent in all studies that assess clinical skills or reasoning [19, 20]. We found that self-reported stress was largely unassociated with clinician accuracy though we quantified stress with a potentially imprecise single-item survey question. Perhaps the stress of the ICU promotes a vigilance that counterbalances its tendency to reduce cognitive performance. Whether stress might hinder ICU clinicians' diagnostic accuracy under some conditions, for example in complex patients with uncertain diagnoses, cannot be answered by our methods.

Nurses had the lowest overall diagnostic performance as determined by our methods. However, nurses' reported diagnoses might be better viewed as a surrogate of team communication because our physician-centric definition for diagnostic accuracy would inherently disadvantage nurses. Nursing diagnosis tends toward symptom-based labels rooted in patient behaviors, and nurses view diagnoses as patients' responses to their medical conditions [21]. Medical diagnosis, conversely, approaches a patient's symptoms through anatomy and physiology to identify disease states [21]. With these considerations, our results suggest that, if nurses' accuracy is a rough marker of team communication, nurses are not fully incorporated into the idealized multidisciplinary diagnostic process [3]. Future studies should explore these disparities in nurses' abilities to accurately report their patients' primary diagnoses.

Our study has several strengths. First, we approached diagnostic accuracy with a greater tolerance to uncertainty than other validated measures of diagnostic errors such as the Safer Dx or SPADE frameworks [7, 22]. Further, prospectively collecting clinicians' diagnostic impressions "off the record" allowed us quantify clinicians' confidence and calibration, which are incompletely reflected in clinical documentation. This study is also the first to compare diagnostic performance for multiple ICU team members and to describe their diagnostic calibration. Additionally, simultaneously studying the contextual factors that might influence diagnostic accuracy, such as stress or perceived prognosis, provides insights into the complex diagnostic process in high-acuity care settings.

There are also obvious limitations. First, as participants were surveyed on a volunteer basis for each patient-day, data were not always complete. Second, we compared clinicians' diagnostic impressions against a "gold standard" of adjudicated primary diagnoses based on chart review. These adjudicated diagnoses may have contained their own inaccuracies, owing to our reliance on retrospective review of records, timing of the reviews, and the variable clinical experience of adjudicators. Third, we cannot determine whether clinicians' diagnoses reflected their own independent judgments or simply reflected the team's diagnostic impressions. Fourth, our measures of contextual variables (stress, perceived complexity, etc.) were not collected using psychometrically validated instruments and may have been influenced by social desirability and the Hawthorne effect. Fifth, we did not quantify the number of patient handovers or identify when patients were recently handed over in this study. Fellows' superior diagnostic performance may reflect increased familiarity with patients as these clinicians had the greatest continuity in our medical ICU. Finally, we did not seek to measure rates of diagnostic error *a priori* using commonly accepted definitions [3, 7]; accordingly, our definition of "diagnostic accuracy" may not fully align with contemporary understanding of diagnostic error.

Nonetheless, this study might inform immediate improvements to the diagnostic process in academic ICUs. While attending physicians are the most likely to have the correct diagnosis, they should recognize that uncertainty among the team, as evidenced by residents/APs reporting that multiple diagnoses are being considered or trainees reporting low confidence in the diagnosis, should prompt more rigorous consideration of alternate diagnoses. In addition, our finding of lower diagnostic calibration among attending physicians argues for enhanced methods to promote feedback to the most senior clinicians in the ICU [23, 24], particularly since such efforts have shown promise among trainees [25, 26].

Acknowledgments: This work was supported by the Medical College of Wisconsin Department of Medicine.

Research ethics: Informed consent was obtained from all subjects as outlined in the manuscript's methods. Research involving human subjects complied with all relevant national regulations, institutional policies and is in accordance with the tenets of the Helsinki Declaration (as revised in 2013). This study was reviewed by the Medical College of Wisconsin/Froedtert Hospital Institutional Review Board (IRB) and was deemed exempt from full IRB review (internal project #34633).

Informed consent: Informed consent was obtained from all individuals included in this study as outlined in the Methods.

Author contributions: Neehal Shukla, Paul Bergl, Jayshil Patel, and Rahul Nanchal made substantial contributions to the study design and data analysis and interpretation. Jatan Shah and Marium Khan contributed substantially to the acquisition, analysis, and interpretation of the data. All authors contributed

to the drafting and revising of the manuscript for intellectual content and approved this version. All authors have accepted responsibility for the entire content of this manuscript and approved its submission.

Competing interests: The authors state no conflict of interest. The authors have no financial disclosures relevant to this study to report.

Research funding: Author NS was provided a stipend of $3600.00 in July/August 2019 for her efforts via the Medical Student Summer Research Project Program. Funding came from the Department of Medicine, Medical College of Wisconsin. There is no other research funding to declare.

Data availability: The raw data can be obtained upon request to the corresponding author.

References

1. Rothschild JM, Landrigan CP, Cronin JW, Kaushal R, Lockley SW, Burdick E, et al. The Critical Care Safety Study: the incidence and nature of adverse events and serious medical errors in intensive care. Crit Care Med 2005;33:1694–700.
2. Bergl PA, Nanchal RS, Singh H. Diagnostic error in the critically III: defining the problem and exploring next steps to advance intensive care unit safety. Ann Am Thorac Soc 2018;15:903–7.
3. National Academies of Sciences, Engineering, and Medicine. Improving diagnosis in health care. Washington, DC: The National Academies Press; 2015.
4. Bergl PA, Taneja A, El-Kareh R, Singh H, Nanchal RS. Frequency, risk factors, causes, and consequences of diagnostic errors in critically ill medical patients: a retrospective cohort study. Crit Care Med 2019;47: e902–10.
5. Winters B, Custer J, Galvagno SM Jr., Colantuoni E, Kapoor SG, Lee H, et al. Diagnostic errors in the intensive care unit: a systematic review of autopsy studies. BMJ Qual Saf 2012;21:894–902.
6. Merkebu J, Battistone M, McMains K, McOwen K, Witkop C, Konopasky A, et al. Situativity: a family of social cognitive theories for understanding clinical reasoning and diagnostic error. Diagnosis 2020;7:169–76.
7. Singh H, Khanna A, Spitzmueller C, Meyer AND. Recommendations for using the Revised Safer Dx Instrument to help measure and improve diagnostic safety. Diagnosis 2019;6:315–23.
8. Barwise A, Leppin A, Dong Y, Huang C, Pinevich Y, Herasevich S, et al. What contributes to diagnostic error or delay? A qualitative exploration across diverse acute care settings in the United States. J Patient Saf 2021:239–48. https://doi.org/10.1097/pts.0000000000000817.
9. Meyer AND, Singh H. The path to diagnostic excellence includes feedback to calibrate how clinicians think. JAMA 2019;321:737–8.
10. Meyer AND, Payne VL, Meeks DW, Rao R, Singh H. Physicians' diagnostic accuracy, confidence, and resource requests: a Vignette study. JAMA Intern Med 2013;173:1952–8.
11. Cohen J. A coefficient of agreement for nominal scales. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960;20:37–46.
12. Cohen J. Weighed kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull 1968;70:213–20.
13. Kuhn J, van den Berg P, Mamede S, Zwaan L, Bindels P, van Gog T. Improving medical residents' self-assessment of their diagnostic accuracy: does feedback help? Adv Health Sci Educ Theory Pract 2022; 27:189–200.
14. Hautz WE, Schubert S, Schauber SK, Kunina-Habenicht O, Hautz SC, Kämmer JE, et al. Accuracy of self-monitoring: does experience, ability or case difficulty matter? Med Educ 2019;53:735–44.
15. Singh H, Connor DM, Dhaliwal G. Five strategies for clinicians to advance diagnostic excellence. BMJ 2022;376:e068044.
16. Croskerry P. Diagnostic failure: a cognitive and affective approach. In: Henriksen K, Battles JB, Marks ES, Lewin DI, editors. Advances in patient safety: from research to implementation (volume 2: concepts and methodology). Rockville: Agency for Healthcare Research and Quality (US); 2005.
17. ALQahtani DA, Rotgans JI, Mamede S, Mahzari MM, Al-Ghamdi GA, Schmidt HG. Factors underlying suboptimal diagnostic performance in physicians under time pressure. Med Educ 2018;52: 1288–98.
18. Blascovich J, Tomaka J. The biopsychosocial model of arousal regulation. Adv Exp Soc Psychol 1996;28:1–51.
19. Pottier P, Dejoie T, Hardouin JB, Le Loupp AG, Planchon B, Bonnaud A, et al. Effect of stress on clinical reasoning during simulated ambulatory consultations. Med Teach 2013;35:472–80.
20. Pottier P, Hardouin JB, Dejoie T, Castillo JM, Le Loupp AG, Planchon B, et al. Effect of extrinsic and intrinsic stressors on clinical skills performance in third-year medical students. J Gen Intern Med 2015;30: 1259–69.
21. Chiffi D, Zanotti R. Medical and nursing diagnoses. J Eval Clin Pract 2015; 21:1–6.
22. Liberman AL, Newman-Toker DE. Symptom-Disease Pair Analysis of Diagnostic Error (SPADE): a conceptual framework and methodological approach for unearthing misdiagnosis-related harms using big data. BMJ Qual Saf 2018;27:557–66.
23. Dhaliwal G. Web exclusives. Annals for hospitalists inpatient notes – diagnostic excellence starts with an incessant watch. Ann Intern Med 2017;167:HO2–3.
24. Bowen JL, O'Brien BC, Ilgen JS, Irby DM, ten Cate O. Chart stalking, list making, and physicians' efforts to track patients' outcomes after transitioning responsibility. Med Educ 2018;52:404–13.
25. Shenvi EC, Feupe SF, Yang H, El-Kareh R. Closing the loop": a mixed-methods study about resident learning from outcome feedback after patient handoffs. Diagnosis 2018;5:235–42.
26. Brisson GE, Barnard C, Tyler PD, Liebovitz DM, Neely KJ. A framework for tracking former patients in the electronic health record using an educational registry. J Gen Intern Med 2018;33:563–6.

Supplementary Material: This article contains supplementary material (https://doi.org/10.1515/dx-2023-0026).