CANADIAN MEDICAL EDUCATION JOURNAL | REVUE CANADIENNE DE L'ÉDUCATION MÉDICALE

2026, In Press: https://doi.org/10.36834/v1anvj69

ORIGINAL RESEARCH

Quantifying delay: modeling the impact of timeliness on narrative feedback for Entrustable Professional Activity assessments in internal medicine training using artificial intelligence

Enoch Yu,[1] Haozhong Tian,[2] Farida Mohamad,[2] Karen Schultz,[3] Laura McEwen,[3] Stephen Gauthier,[1] Heather Braund,[4] Nicholas Cofie,[4] Nancy Dalgarno,[4] Adam Szulewski,[5,6] Farhana Zulkernine,[2] Benjamin YM Kwan[2,7]

*Author information is provided in the back matter of this manuscript

Abstract

Introduction: In competency-based medical education (CBME), specifically entrustable professional activities (EPAs), timely feedback is essential because delays can distort assessor recall and performance documentation. Prior work links longer timelag to reduced feedback quality, but its effects across thresholds and outcomes remain unclear. This study examines these relationships using statistical analyses, incorporating artificial intelligence (AI) tools to standardize extraction of narrative feedback metrics.

Methods: We performed a retrospective cross-sectional study of internal medicine resident EPA assessments, each including two narrative sections ('General Feedback,' 'Next Steps'), entrustment scores (0-4), and completion dates. Pretrained transformer-based models estimated sentiment positivity (0-1) and Quality of Assessment for Learning (QuAL) scores (0-5). We manually refined outputs, subsequently fitting datapoints for ordinal logistic, binary logistic, and linear regression. Exploratory analyses identified timelag thresholds linked to changes in feedback metrics.

Results: 301 assessors evaluated 1,147 EPAs for 111 residents, with a 7-day median timelag (IQR: 21). Longer timelag was associated with lower QuAL (OR=0.997/day; $p < 0.05$), fewer words (-0.07words/day; $p<0.05$), and higher odds of empty comments (OR=1.003/day; $p < 0.01$). Effects emerged sequentially: transient increases in entrustment around seven days, followed by dose-dependent declines in quality from 14 days and reductions in feedback length thereafter. Beyond 100 days, comments became uniformly brief with higher positivity and entrustment, suggesting reduced specificity and recall.

Conclusion: Delayed completion meaningfully affects feedback quality, completeness, and tone. Findings support 7-14 day completion targets and flagging EPA assessments beyond 100 days as unreliable. AI-assisted extraction of narrative metrics may support standardized, scalable EPA assessment analysis in CBME.

Quantification du délai : modélisation de l'impact de la rapidité de la rétroaction narrative dans les évaluations des activités professionnelles confiables au cours de la formation en médecine interne à l'aide de l'intelligence artificielle

Résumé

Introduction : Dans la formation médicale axée sur les compétences, en particulier les activités professionnelles confiables (EPA), une rétroaction rapide est essentielle, car les retards peuvent fausser la mémoire des évaluateurs et la documentation de la performance. Des travaux antérieurs associent un délai plus long à une rétroaction de moindre qualité, mais ses effets selon différents seuils et résultats demeurent incertains. Cette étude examine ces relations au moyen d'analyses statistiques, en intégrant des outils d'intelligence artificielle (IA) pour standardiser l'extraction de mesures issues de la rétroaction narrative.

Méthodes : Nous avons mené une étude rétrospective transversale d'évaluations EPA de résidents en médecine interne, comprenant chacune deux sections narratives (« Commentaires généraux », « Prochaines étapes »), des scores de confiance (0-4) et des dates de réalisation. Des modèles préentraînés fondés sur des transformeurs ont estimé la positivité du sentiment (0-1) et les scores de qualité de l'évaluation pour l'apprentissage (QuAL) (0-5). Nous avons affiné manuellement les résultats, puis ajusté les données au moyen de régressions logistique ordinale, logistique binaire et linéaire. Des analyses exploratoires ont permis d'identifier les seuils de délai associés à des changements dans les mesures de rétroaction.

Résultats : 301 évaluateurs ont évalué 1 147 EPA pour 111 résidents, avec un délai médian de 7 jours (IQR : 21). Un délai plus long était associé à une QuAL plus faible (OR=0,997/jour ; $p<0,05$), à moins de mots (-0,07 mot/jour ; $p<0,05$) et à une probabilité plus élevée de commentaires vides (OR=1,003/jour ; $p<0,01$). Les effets sont apparus de façon séquentielle : augmentation transitoire de la confiance autour de sept jours, suivie d'une baisse dose-dépendante de la qualité à partir de 14 jours, puis d'une réduction de la longueur de la rétroaction. Au-delà de 100 jours, les commentaires devenaient uniformément brefs, avec une positivité et une confiance plus élevées, suggérant une diminution de la spécificité et du rappel.

Conclusion : Les retards de réalisation affectent de façon significative la qualité, l'exhaustivité et le ton de la rétroaction. Les résultats appuient des cibles de réalisation de 7 à 14 jours et le signalement des évaluations EPA dépassant 100 jours comme non fiables. L'extraction assistée par IA de mesures narratives pourrait soutenir une analyse standardisée et évolutive des évaluations EPA dans la formation médicale axée sur les compétences.

Introduction

Competency-based medical education (CBME) is the curriculum model used in resident education in Canada. For specialty programs, as guided by the Royal College of Physicians and Surgeons of Canada (RCPSC), there is a focus on entrustable professional activities (EPA) as the basis to assess competence through frequent, task-specific, and low-stakes assessments.[1–4] EPAs are discrete clinical tasks evaluated by assessors to determine the level of entrustment granted to a resident, first proposed in 2005[5] and adopted in Canada around 2017. EPA assessments represent a key outcomes-based tool to monitor resident progression through training in CBME.

Narrative descriptive feedback plays an essential role within EPA-based assessments to clarify performance, provide guidance for growth, and inform program decisions on resident progress.[4,6] However, while the introduction of EPAs aimed to increase the quality and quantity of narrative feedback available to residents, multiple studies identi--ied improvements are still required in areas related to quality, frequency, and actionability.[1,4,7,8]

Timeliness is a critical determinant of feedback effectiveness.[9] Residents consistently report that timely completion of feedback improves event recall, actionability, and perceived usefulness.[10–12] Yet, in postgraduate medical education (PGME), feedback is often delayed after the clinical event.[9,13] One survey found 58% of residents describing their assessments as rarely or never timely,[10] while an interview study noted 70% of EPAs took over two weeks.[14] Reasons may include limited time[9,15] or competing workloads for assessors.[16]

Untimely feedback affects not only residents' ability to apply it, but also the overall quality and strength of feedback itself. Extended time delays had a significant association with lower feedback quality scores after 14 days[17] or one month,[8] and also with greater frequencies of empty feedback.[17,18] Feedback given after seven days had significantly less actionability in recommendations than before.[4]

When delayed, feedback reliability and accuracy could be implicated, with concerns of memory decay, recall bias, halo effects, and central tendency errors.[19–23] Assessors could default to more positive perceptions with movement towards the mean. Feedback post-14 days contained significantly less item-to-item variation,[18] more global unspecific feedback,[18] and less reliability in evaluated scores.[24] Though, some studies found insignificant trends between timeliness with scoring,[18,25] quality,[1,26] or positivity,[4] thus warranting further investigation.

The increasing awareness of timeliness in PGME feedback has led to frameworks with expiry thresholds ranging from 72 hours[27] to 14 days,[28] representing important steps toward promoting prompt assessment completion. However, existing studies guiding these thresholds examine only a narrow subset of factors per study and often pool diverse assessment types or tasks, risking confounding factors. For instance, complex EPA tasks could warrant both greater feedback quality and urgency for completion. Timeliness has multidimensional impacts beyond feedback quality, influencing evaluation scores, sentiment, and detail, and also varies by EPA assessment initiator and evaluation modality. Robust thresholds should therefore account for these multiple factors simultaneously to remain meaningful and adaptable to program needs.

This study builds on existing medical education work on delayed assessment completion,[4,8,17,18,24–26] extending it to examine the impact of timelag through multiple dimensions together. Using a large single-task EPA cohort to minimize confounding variables, we comprehensively analyzed associations between timeliness and feedback characteristics. To support consistent and scalable measurement of narrative feedback features, we further ap--ly artificial intelligence (AI) methods as a tool to help extract sentiment and quality metrics from written comments, which we then incorporated into statistical analyses. Key objectives for this study included 1) characterizing patterns of delayed narrative feedback, 2) modeling different timeliness thresholds on feedback quality and traits, and 3) drawing on multiple metrics to inform more robust guidelines for timely, meaningful assessment.

Methods

Study design and setting

We performed a single-centre retrospective cross-sectional study analyzing a single Core of Discipline (C1) EPA assessment over 3 years (July 1st 2020 to June 9th 2023) for internal medicine residents at a Canadian university. The C1 EPA assesses the task of assessing, diagnosing, and managing patients with complex or atypical acute medical presentations. We chose to analyze only one EPA type to limit inter-assessment variability. Each EPA involved a clinical encounter assessed by direct observation, case discussion, or both. Assessors included faculty, attending physicians, clinical fellows, or senior residents. All assessors underwent training to ensure consistency, including routine departmental PGME workshops and faculty development sessions. Following completion of the clinical task, the resident, assessor, or program management system (automatically based on scheduled requirements) triggered the EPA assessment form. When triggered, Elentra™, an integrated online clinical education platform,[29] notified and delivered EPA forms to the assessor. The assessor could complete the form online on Elentra™ at any time, with a template viewable in this weblink (https://drive.google.com/file/d/1u1VAMZ24tjOGo60rwDh0BGgesQCuhhyh/view?usp = sharing). Once submitted, Elentra™ would return the filled form to the resident, and the residency program would also regularly review feedback to inform resident progression, remediation, and intervention decisions.

Assessors scored entrustment on a 6-level rating scale (scored 0-4) developed by the Queen's University Internal Medicine program (Table 1), based on the 5-point scale from Chen and colleagues. anchored on the level of supervision required by the assessor.[30] While some institutions use alternative scales, like the 5-point Ottawa scale, which anchors on the level of intervention required instead,[31] the Internal Medicine program developed this scale as it better aligned with the spectrum of supervision provided to residents in training.

This study received approval from the university's Health Sciences and Affiliated Teaching Hospitals Research Ethics Board (File #6029081, 6035804).

Table 1. EPA entrustment scoring scale on resident performance

Score		Resident Performance
0	EPA not achieved	Supervisor actively performs the EPA with resident
0		Supervisor intermittently assists resident to perform the EPA
1		Supervisor outside room, immediately available, double-checks findings
2	EPA achieved	Supervisor outside room, immediately available, checks only key findings
3		Supervisor offsite, available by phone, checks only key findings
4		Distant supervisor, post-hoc debrief available as needed

Data collection

Following form submission, Elentra™ stored assessment data in a secure database. Each form record consisted of the EPA initiator role, date of EPA encounter and feedback submission, entrustment score, and narrative feedback under two comment sections ('General Feedback' and 'Next Steps'). We collected EPA data from Elentra™ as a spreadsheet, de-identified to maintain confidentiality. Timelag represented the number of days between EPA task encounter and EPA assessment completion. We defined empty comments as those that were blank or contained placeholders like "–," "NA," "see below," or "see above."

Feedback analysis

We analyzed four metrics: entrustment score, sentiment positivity, feedback quality, and word count, with entrustment as the main metric of study as it directly reflects supervisor assessment and clinical competency decisions. Assessors scored entrustment on an established zero to four point scale, with ≥2 denoting successfully achieving the EPA (Table 1).[32] We calculated word count separately for each feedback section. We used natural language processing (NLP) AI models on our local system to systematically analyze narrative comments and cap-

ture patterns within language or terminology, predicting measures of positivity and feedback quality.

Sentiment positivity estimation used a pretrained DistilBERT transformer model, fine-tuned for sentiment prediction on the SST-2 Stanford Sentiment Treebank corpus.[33–35] The SST-2 comprises 215,154 distinct phrases derived from 11,855 individual sentences in movie reviews, each annotated by three human adjudicators.[34] The output positivity score ranged from 0 (negative) to 1 (positive), imputed separately for each feedback section.

We measured feedback quality by Quality of Assessment for Learning (QuAL) scores, a validated metric (zero to five) shown to correlate with feedback utility.[36,37] QuAL consisted of three subscores: 1) evidence of resident performance (zero to three), 2) presence of suggestions to improve ('Suggestions'), and 3) presence of connections supporting the suggestion with evidence ('Connections').[36,37] QuAL analysis adopted a pretrained Bio-Clinical BERT transformer model, fine-tuned on two million clinical notes from the MIMIC-III v1.4 database[38,39] and subsequently trained on 2,500 EPAs for emergency medicine residents in Ontario and Saskatchewan by Spadafore and colleagues.[40] The model employed an ensemble of three submodels, each corre-ponding to one QuAL subscore, and processed each feedback section separately. While tested in a similar CBME setting, because of its training on emergency medicine data, we acknowledge it may not have fully captured program-specific feedback trends in internal medicine.

Separately, for each comment, an independent rater (E.Y.) scored sentiment on a 5-point scale (0/0.25/0.5/0.75/1) and QuAL scores by their subdomains. We manually corrected model predictions when the sentiment score differed by >0.25 from human ratings, or if we found the predicted QuAL subdomain to be clearly incorrect upon further review. This integration of manual review helped ensure high data fidelity, providing refined model outputs that we used for subsequent analysis.

Statistical analysis

We summarized feedback characteristics using descriptive statistics, followed by subgroup analysis using established timelag thresholds from prior literature (48hrs, 7d, 14d, 30d).[4,8,17,18] We assessed differences across ordinal timelag groups using trend tests (Cochran-Armitage test, Spearman correlation), and across EPA initiators using group tests (Chi-squared test for independence, Kruskal-Wallis test). Follow-up analyses comparing individual variables used two-sample independent t-tests for means and two-proportion Z-tests for proportions.

Associations between timelag and ordinal outcomes underwent ordinal logistic regression and fit using the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm. Bootstrapping with 300 resampled datasets estimated 95% confidence intervals (CI). We modelled binary outcomes over timelag with binary logistic regression, and continuous outcomes with linear regression using a least-squares approach. A Pearson correlation matrix summarized the overall relationships among feedback variables and timelag.

To identify potential timelag thresholds, exploratory analyses mapped mean differences in feedback metrics before and after successive candidate thresholds. We validated selected thresholds that showed notable shifts using two-sample independent t-tests, with a false discovery rate (FDR) correction via the Benjamini-Hochberg procedure to control for multiple comparisons.

Sensitivity analyses assessed the robustness of results to extreme delays by excluding outlier EPAs using Tukey's method (Q1 - 1.5*IQR; Q3 + 1.5*IQR) and repeating correlation and threshold analyses. We applied a threshold FDR adjustment collectively for all total comparisons, with and without outlier exclusion combined. We evaluated heteroscedasticity using the Breusch-Pagan test to detect changes in feedback variance over timelag.

All analyses used Python 3.11[41,42] and Microsoft Excel, using Matplotlib[43] and Seaborn[44] for plotting, and using SciPy,[45] scikit-learn,[46] and statsmodels[47] for statistical testing.

Results

Baseline characteristics of feedback

Over the three-year period, 301 distinct assessors completed 1,147 EPA-based assessments for the C1 task across 111 distinct residents (Figure 1). Residents initiated most assessments (n = 1079; 94.1%), with 53 (4.6%) initiated by attending physicians and 15 (1.3%) by the program management system. All records contained entrustment and timelag data. However, 243 (21.2%) lacked comments in 'General Feedback' and 543 (47.3%) lacked comments in 'Next Steps.' The median entrustment score was 3 (IQR 1), indicating consistently high confidence in resident performance. Assessors completed assessments with low delay, at a median of 7 days (IQR 21) with positive skewing. Nearly half (n = 548, 47.8%) of submissions occurred within 7 days (Table 2), while one-fifth (n = 229) after 30 days. Program-initiated assessments had a longer median timelag (24 days, IQR 24) than resident-initiated EPAs (7 days, IQR 22). Assessors completed assessor-initiated assessments most promptly (median 0 days, IQR 1).

Among assessments containing comments, the median word count was 33 (IQR 39) in 'General Feedback' and 25 (IQR 33) in 'Next Steps,' with both distributions positively skewed. On refinement of AI model outputs, 80.0 to 96.4% of comments did not require manual correction. 13.9% (n = 126) and 20.0% (n = 121) of comments were manually corrected for sentiment positivity, for 'General Feedback' and 'Next Steps' respectively, often due to conflation of poor patient status with resident performance. This was lower for QuAL subcategories, at 7.9% (n = 71) and 3.6% (n = 22) for level of detail, 6.4% (n = 58) and 17.7% (n = 107) for presence of suggestions, and 4.1% (n = 37) and 11.8% (n = 71) for presence of suggestions connected to feedback.

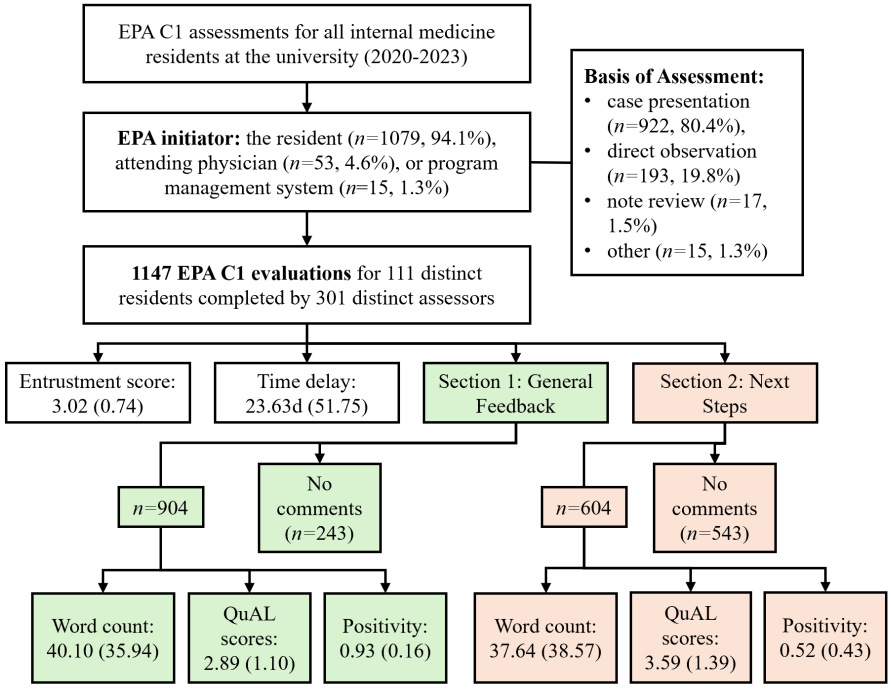


*Figure 1. Flow of EPA C1 assessments and assessor/task details.**n indicates the number of data points, v(v) denotes mean (standard deviation), and d denotes days. Positivity scores range from 0 (negative) to 1 (positive). QuAL scores range from 0 (no feedback) to 5 (feedback with suggestions that link to behavior described).*

Table 2. Descriptive statistics and trend tests of EPA characteristics across groups of different completion time delay#*

	Full Cohort (n = 1147)	<48hrs (n = 260)	2-7d (n = 288)	7-14d (n = 177)	14-30d (n = 193)	>30d (n = 229)	Trend Testing
Time Delay (days)	23.63 (51.75)	0.37 (0.48)	3.88 (1.47)	9.41 (2.13)	20.41 (4.73)	88.58 (89.03)	
Entrustment Score (0-4)	3.02 (0.74)	2.88 (0.80)	3.03 (0.69)	3.06 (0.68)	3.15 (0.69)	3.03 (0.78)	ρ=0.09, ***p* = 0.001**
General Feedback (n = 904)		225	229	141	143	166	
Empty Feedback	243 (21.2%)	35 (13.5%)	59 (20.5%)	36 (20.3%)	50 (25.9%)	63 (27.5%)	χ^2=533.00, ***p* = 0.000**
QuAL Score (0-5)	2.89 (1.10)	2.76 (1.16)	3.01 (1.13)	2.94 (1.05)	3.08 (1.04)	2.70 (1.05)	ρ=0.01, p = 0.759
with suggestions	136 (15.0%)	32 (14.2%)	45 (19.7%)	19 (13.5%)	24 (16.8%)	16 (9.6%)	χ^2=219.00, p = 0.144
with connections	124 (13.7%)	29 (12.9%)	39 (17.0%)	19 (13.5%)	23 (16.1%)	14 (8.4%)	χ^2=202.00, p = 0.221
Sentiment Positivity (0-1)	0.93 (0.16)	0.94 (0.16)	0.93 (0.16)	0.94 (0.15)	0.92 (0.18)	0.93 (0.17)	ρ=-0.06, p = 0.072
Word Count	40.10 (35.94)	35.66 (34.87)	42.66 (36.28)	40.74 (32.15)	45.78 (44.74)	37.16 (30.46)	ρ=0.06, p = 0.086
Next Steps (n = 604)		170	147	87	87	113	
Empty Feedback	543 (47.3%)	90 (34.6%)	141 (49.0%)	90 (50.8%)	106 (54.9%)	116 (50.7%)	χ^2=1103.00, ***p* < 0.001**
QuAL Score (0-5)	3.59 (1.39)	3.49 (1.42)	3.55 (1.46)	3.94 (1.27)	3.57 (1.33)	3.52 (1.40)	ρ=0.02, p = 0.553
with suggestions	441 (73.0%)	123 (72.4%)	110 (74.8%)	69 (79.3%)	63 (72.4%)	76 (67.3%)	χ^2=741.00, p = 0.386
with connections	312 (51.7%)	79 (46.5%)	76 (51.7%)	56 (64.4%)	44 (50.6%)	57 (50.4%)	χ^2=548.00, p = 0.444
Sentiment Positivity (0-1)	0.52 (0.43)	0.58 (0.43)	0.48 (0.43)	0.48 (0.43)	0.53 (0.42)	0.50 (0.43)	ρ=-0.07, p = 0.092
Word Count	37.64 (38.57)	38.51 (42.83)	33.56 (31.41)	54.61 (51.22)	36.08 (34.82)	29.77 (26.41)	ρ=0.01, p = 0.818

**mean (SD) for continuous variables, n (%) for binary variables. Timeliness intervals were left-inclusive and right-exclusive. We conducted tests for trend using the Cochran–Armitage test (binary variables) and Spearman correlation (continuous variables) across ordinal timeliness classes. #QuAL score sub-components included 'with suggestions' (suggestions for improvement) and 'with connections' (connections from suggestions to evidence). We defined sentiment positivity as 0 (negative) to 1 (positive).*

Sentiment positivity was higher in 'General Feedback' at a median of 0.99 (IQR 0.02) than 'Next Steps' with a median of 0.39 (IQR 0.95), suggesting a more critical tone (Appendix A). Positivity showed a bimodal pattern in 'Next Steps,' clustering toward both extremes. QuAL scores demonstrated a similar bimodal distribution, with a median score of 3 (IQR 0) in 'General Feedback,' where 15.0% (n = 136) included suggestions and 91.2% (n = 124) of those linked the suggestions to evidence. 'Next Steps' had a higher median score of 4 and greater variability in QuAL scores (IQR 3), with 73.0% (n = 441) of assessments having suggestions and 70.7% (n = 312) of those suggestions connected to provided evidence.

Subgroup analyses of feedback by time delay and initiator

Subgroup analysis by timelag groups revealed that entrustment increased modestly with longer delays, with a mean entrustment of 3.15 (14-30 days) compared to 2.88 (<48 hours) (Table 2). This trend was significant (p = 0.001). The proportion of empty comments also rose in a dose-dependent manner, increasing from 13.5% (n = 35/260) to 27.5% (n = 63/229) in 'General Feedback' and from 34.6% (n = 90/260) to 50.7% (n = 116/229) in 'Next Steps,' between <48 hours and >30 days (both p<0.001 for trend). Although QuAL scores and the presence of suggestions or connections showed no consistent

temporal trend, they tended to peak around 2-7 days post-task. We observed no strong associations for sentiment or word count, though word count dropped markedly after 30 days in both sections.

In EPA assessments of extended timelag, such as after 100 days, 'General Feedback' sentiment became almost uniformly positive with limited negative comments, while 'Next Steps' entries remained more balanced (Appendix A). At these extreme delays, word counts also shifted toward brevity. All comments given post-100 days were ≤100 words and ≤50 words in 'General Feedback' and 'Next Steps,' respectively.

Further analysis confirmed that empty narrative feedback had an association with longer completion delays. Assessments lacking 'General Feedback' had a mean timelag 10.3 days longer than those with comments (31.8 vs 21.4 days; 95%CI: 1.68,18.98; $p = 0.019$), which was significant. The difference was also significant in 'Next Steps' at 6.43 days (27.0 vs 20.6 days; 95%CI: 0.37,12.48; $p = 0.038$).

Regression modeling of timelag

To evaluate timeliness more granularly and over extended durations, we conducted regression analyses using timelag as a continuous predictor across feedback metrics. For feedback quality, ordinal regression showed significantly lower quality scores with increasing delay in both sections ($p<0.01$ for 'General Feedback'; $p<0.05$ for 'Next Steps') (Figure 2). The model estimated an odds ratio (OR) of 0.997 per day, indicating every 10 additional days of delay corresponded to 3% lower odds of receiving higher-quality feedback. Increases in empty feedback (score 0) primarily drove this decline, accompanied by a reduction in detailed comments without suggestions (score 3) in 'General Feedback' or a reduction in detailed comments with suggestions and connections (score 5) in 'Next Steps.'

Examining QuAL subsections, binary logistic regression confirmed that longer timelag predicted a higher likelihood of empty 'General Feedback' (OR 1.003 per day, $p = 0.009$), equivalent to 3% greater odds of missing comments for every 10 days of delay. We observed a similar significant trend for 'Next Steps' (OR 1.003 per day; $p = 0.041$). Conversely, increased timelag associated with fewer suggestions in 'Next Steps' (OR 0.996 per day, $p = 0.017$) and 'General Feedback' (OR 0.994 per day, $p = 0.054$), equivalent to 4% or 6% decreased odds of a suggestion per 10 days delayed. The presence of evidence-connected suggestions likewise dropped when the presence of suggestions dropped, with no significant trends. For entrustment scores, ordinal regression revealed no significant association overall, though with an upward shift in the highest entrustment level (score 4) at extreme delays (Figure 2). Follow-up analysis identified that 37.7% ($n = 20/53$) of assessments completed at or after 100 days received a score of 4, compared with 23.6% ($n = 258/1094$) before 100 days, suggesting overrepresentation of high-end scores at extended timelag (14.1% mean difference; two-proportion z-test: $Z=2.35$, $p = 0.019$).

Linear regression models examining continuous outcomes (Figure 3) showed that longer timelag predicted shorter comments in 'Next Steps' with approximately seven fewer words per 100 days (95%CI; 13.6,0.5; $p < 0.05$). In contrast, the trend for 'General Feedback' was balanced and insignificant; however, with inclusion of invalid feedback, we identified a significant negative trend in both sections. Sentiment analysis showed a weak, non-significant trend toward increased positivity in 'General Feedback' at higher timelag. Analyses identified no associations between sentiment and timelag in 'Next Steps.'

Across all metrics, Pearson correlations between timelag and feedback features were weak (all $|r| <= 0.10$; Appendix B). The only significant correlations, of lower word count being associated with longer timelag in both sections, became non-significant after outlier exclusion of $n = 123$ EPAs (upper-bound: 54.5 days). This suggested a small subset of extreme delays likely led to the observed effect. When excluding outliers, lower sentiment had a significant correlation to longer timelag; the initial analysis including outliers did not observe this

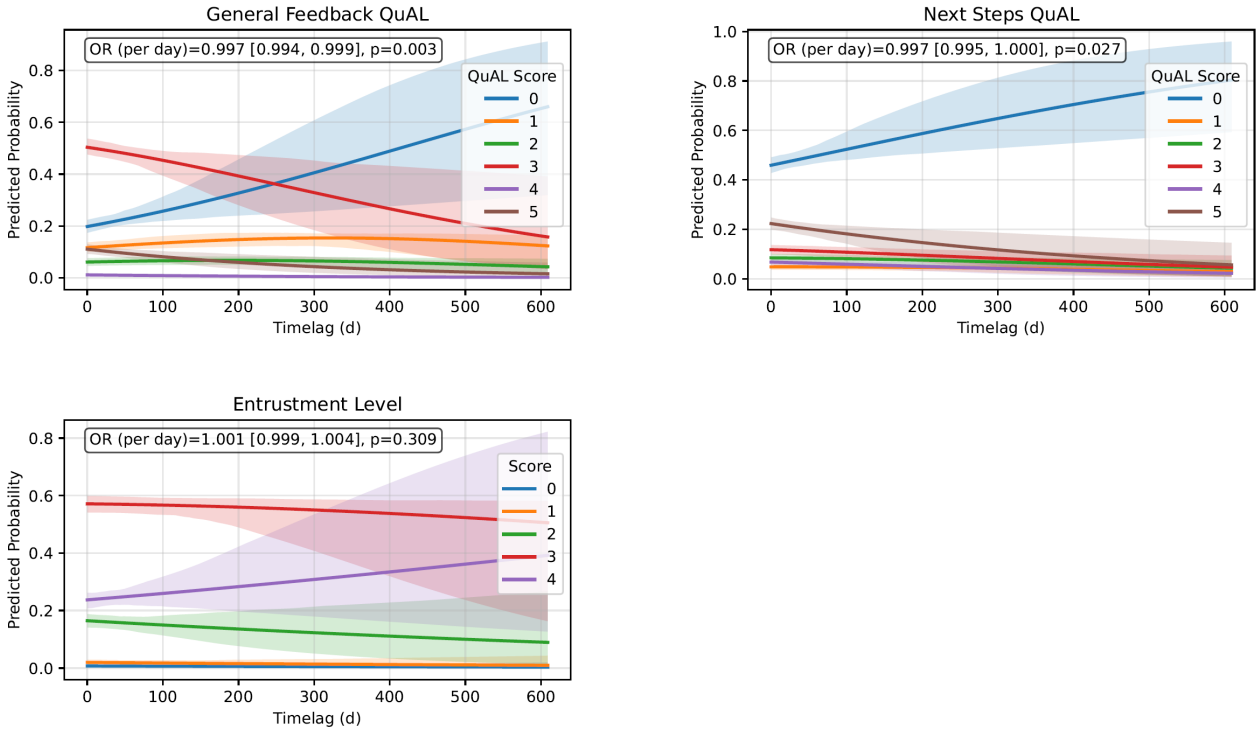


Figure 2. Predicted probabilities of QuAL scores and entrustment over timelag from ordinal regression models. Shaded regions represent 95%CI imputed from bootstrapping (300 rounds). Odds ratios (OR) are interpreted as the outcome change per one-day increase in timelag.

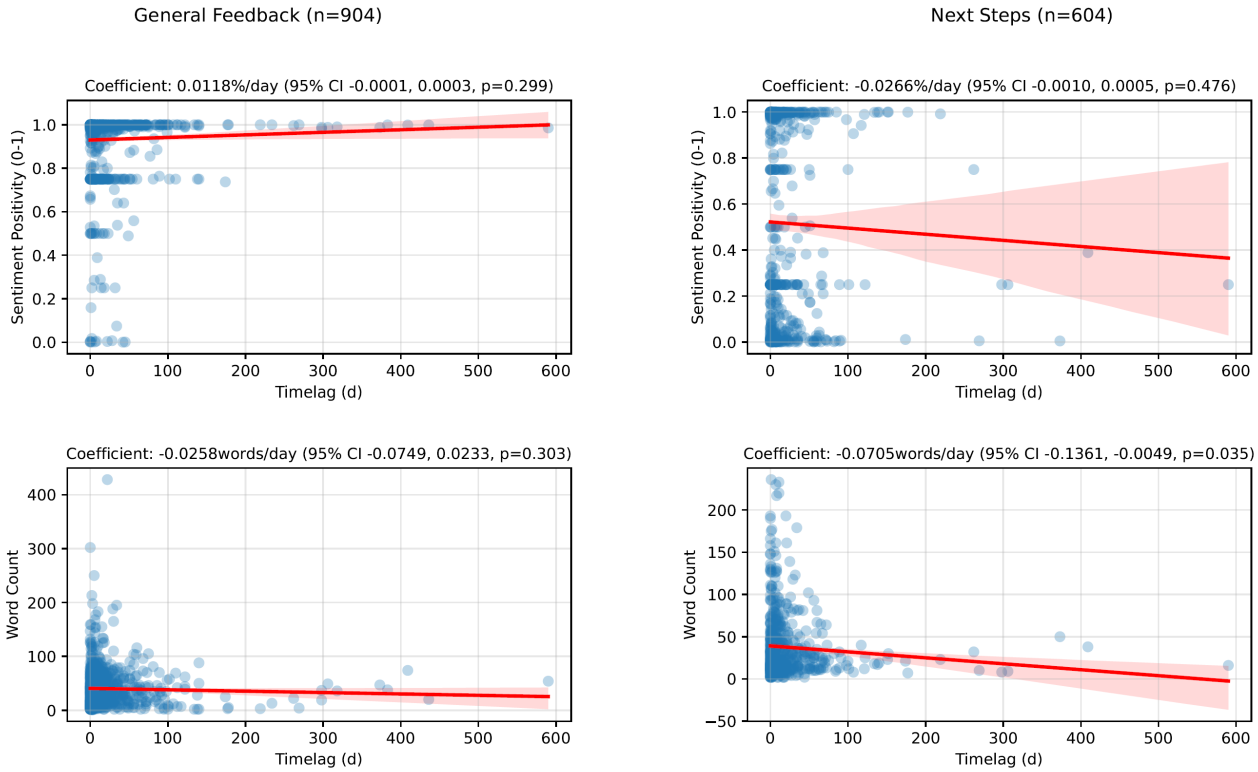


Figure 3. Linear regression model of time delay with sentiment positivity and word count. *Shaded regions represent 95%CI (t-distribution) of the model. Coefficient represents outcome change per one-day increase in completion timelag.*

Exploratory threshold testing and comparisons

To identify potential thresholds where timelag may correspond to meaningful changes in outcomes, we conducted exploratory analyses comparing mean differences in each metric before and after successive timepoints. This revealed several transition zones, or intervals where thresholding led to large differences in outcome. For entrustment, assessments completed after three to seven days had higher scores than those completed sooner, with a secondary zone around 100 days. For feedback quality, assessments completed after 14-150 days showed lower QuAL scores in 'General Feedback' than those before, and 70-90 days for 'Next Steps' (Appendix C). In sentiment, transition zones for higher positivity appeared after around 50 and 150 days in 'General Feedback,' and between 80-100 days in 'Next Steps' (Appendix D). Feedback length had transition zones of lower word count within the 40-100 day range in both sections. After outlier exclusion, early transition zones in entrustment and QuAL scores ('General Feedback') persisted. No zones for sentiment positivity, word count, or QuAL ('Next Steps') remained, aside from a mild 35-45 day transition for lower word count in 'General Feedback.'

Based on these exploratory trends, we selected three thresholds: seven days (entrustment), 14 days (QuAL scores, empty feedback), and 45 days (QuAL scores, word count). Higher thresholds such as 150 days may capture more pronounced changes in word count or sentiment, but we considered them too delayed to be practical. To validate these thresholds, we performed t-tests with FDR multiple comparisons correction (Table 3). The seven-day threshold showed significantly higher entrustment in EPAs completed after the delay; this effect persisted after outlier exclusion but lost significance. QuAL scores showed a dose-response trend in 'General Feedback' and had greater reductions in quality at higher thresholds, with or without outlier exclusion. EPAs completed after 14- and 45-day thresholds had significantly lower quality in 'General Feedback.' 'Next Steps' showed a similar directional trend without reaching significance. Word counts also exhibited a dose-response pattern, with progressively shorter feedback at longer delays. We observed significant reductions at 14- and 45-day thresholds in 'Next Steps,' and only 45-day in 'General Feedback.' These effects remained consistent but insignificant post-outlier exclusion, except for the 14-day threshold in 'Next Steps.' Sentiment positivity did not significantly change across tested thresholds.

Tests for heteroscedasticity revealed a mild but significant variance decrease in positivity and word count in 'Next Steps' with increasing timelag, but this effect disappeared after outlier exclusion (Appendix E).

In summary, delays beyond roughly two weeks had an association with a noticeable drop in feedback quality and length, with extreme delays beyond three months resulting in uniformly brief feedback with higher entrustment levels and, for 'General Feedback,' a greater positive skew.

*Table 3. Mean differences in feedback outcomes after versus before timelag thresholds**

	All Data			Outliers Excluded		
Threshold	7d	14d	45d	7d	14d	45d
En-trustment (0-4)	0.12 (95% CI: 0.04, 0.21; p = 0.005, **FDR=0.030**)	0.10 (95% CI: 0.02, 0.19; p = 0.021, FDR=0.078)	0.06 (95% CI: -0.07, 0.20; p = 0.344, FDR=0.482)	0.11 (95% CI: 0.02, 0.20; p = 0.013, FDR=0.059)	0.09 (95% CI: -0.01, 0.19; p = 0.063, FDR=0.168)	-0.08 (95% CI: -0.39, 0.24; p = 0.633, FDR=0.732)
General Feedback						
QuAL Score	-0.22 (95% CI: -0.39, -0.04; p = 0.017, FDR=0.075)	-0.27 (95% CI: -0.46, -0.09; p = 0.004, **FDR=0.025**)	-0.48 (95% CI: -0.73, -0.23; p = 0.000, **FDR=0.002**)	-0.14 (95% CI: -0.33, 0.05; p = 0.137, FDR=0.271)	-0.19 (95% CI: -0.40, 0.02; p = 0.082, FDR=0.189)	-0.63 (95% CI: -1.23, -0.02; p = 0.042, FDR=0.116)
Sentiment Positivity	-0.01 (95% CI: -0.03, 0.01; p = 0.434, FDR=0.545)	-0.01 (95% CI: -0.04, 0.01; p = 0.204, FDR=0.334)	0.02 (95% CI: -0.01, 0.05; p = 0.200, FDR=0.334)	-0.02 (95% CI: -0.04, 0.01; p = 0.133, FDR=0.271)	-0.03 (95% CI: -0.06, -0.00; p = 0.025, FDR=0.084)	-0.06 (95% CI: -0.18, 0.05; p = 0.266, FDR=0.402)
Word Count	-1.64 (95% CI: -5.80, 2.52; p = 0.439, FDR=0.545)	-2.33 (95% CI: -6.71, 2.05; p = 0.296, FDR=0.427)	-7.74 (95% CI: -12.71, -2.77; p = 0.002, **FDR=0.019**)	-0.35 (95% CI: -4.87, 4.17; p = 0.879, FDR=0.899)	-0.56 (95% CI: -5.80, 4.67; p = 0.832, FDR=0.872)	-11.29 (95% CI: -20.99, -1.59; p = 0.024, FDR=0.083)
Next Steps						
QuAL Score (0-5)	-0.28 (95% CI: -0.52, -0.04; p = 0.021, FDR=0.078)	-0.33 (95% CI: -0.58, -0.09; p = 0.008, **FDR=0.041**)	-0.26 (95% CI: -0.60, 0.08; p = 0.132, FDR=0.271)	-0.23 (95% CI: -0.49, 0.02; p = 0.071, FDR=0.182)	-0.29 (95% CI: -0.56, -0.02; p = 0.039, FDR=0.110)	0.07 (95% CI: -0.66, 0.81; p = 0.837, FDR=0.872)
Sentiment Positivity	-0.03 (95% CI: -0.10, 0.04; p = 0.382, FDR=0.515)	-0.01 (95% CI: -0.08, 0.07; p = 0.838, FDR=0.872)	-0.03 (95% CI: -0.13, 0.08; p = 0.637, FDR=0.732)	-0.02 (95% CI: -0.10, 0.05; p = 0.530, FDR=0.648)	0.01 (95% CI: -0.07, 0.09; p = 0.835, FDR=0.872)	0.05 (95% CI: -0.15, 0.25; p = 0.604, FDR=0.727)
Word Count	-2.16 (95% CI: -6.07, 1.75; p = 0.279, FDR=0.413)	-6.99 (95% CI: -10.69, -3.29; p = 0.000, **FDR=0.002**)	-7.88 (95% CI: -11.62, -4.13; p = 0.000, **FDR=0.001**)	-0.45 (95% CI: -4.78, 3.87; p = 0.837, FDR=0.872)	-5.67 (95% CI: -9.93, -1.40; p = 0.009, **FDR=0.045**)	-4.73 (95% CI: -12.93, 3.47; p = 0.249, FDR=0.387)

**Mean difference (95%CI; raw p-value, FDR-adjusted p-value). Positive differences indicate higher values in EPAs completed at or after the delay threshold, than before the threshold. We calculated the False Discovery Rate-adjusted p-value using the Benjamini–Hochberg procedure across all comparisons.*

Discussion

This study examined how timeliness influences the quality, content, and tone of narrative feedback in EPA-based assessments for a large internal medicine cohort. While meaningful assessment takes time, limited assessor availability for completion remains a persistent barrier in CBME,[48,49] and delays in assessment can further diminish written feedback quality. We show that timelag meaningfully influenced multiple aspects of narrative assessment, though the strength and timing of effects varied.

Narrative quality was generally high, comparable to other studies using QuAL scoring.[49] QuAL scores decreased in a dose-dependent manner with increasing timelag, particularly in 'General Feedback'. Declines emerged at 14 days and continuing to 150 days, consistent with 14 to 30-day thresholds in prior studies[8,18] but later than a 7-day actionability threshold in one study.[4] In contrast, 'Next Steps' was more resilient, showing decreases after 60-80 days, possibly due to the section's robustness in containing more suggestions. Regression modeling estimated roughly a 3% reduction in the odds of higher-quality feedback every 10 days in both sections, significant in both sections.

Empty comments largely drove the QuAL decline, with more empty comments as early as after 48 hours and again at 14 days. Assessors completed EPA assessments lacking 'General Feedback' 10.3 days later on average, supporting prior findings that

delayed entries yield more omissions.[17,18] Logistic models showed a 2-3% increase in odds of empty feedback per 10 days. The decline in QuAL scores was also attributable to shifts from detailed to low-detail comments in 'General Feedback'.

Entrustment showed the earliest sensitivity to delay, with scores higher in EPA assessments completed after seven days than before. This effect re-emerged at extreme timelags (≥100 days), where the proportion of top-level scores was 14.1% higher than pre-100 days. Such elevation likely reflected recall bias and reduced reliability,[18] consistent with rater leniency patterns where assessors may default to more positive or central-tendency judgments[50,51] when details fade over time. The early increase could be from this or alternative patterns where assessors considered poor performances more urgent and completed earlier. Although some studies found no significant association between entrustment and timelag,[18,25] these often examined shorter durations or time windows, which may have obscured late effects.

Sentiment and word count also followed temporal patterns, with most shifts emerging at extended timelags. Sentiment positivity rose subtly over time, with sentiment in 'General Feedback' completed after 100 days being mostly uniformly positive. While one study found no significant association between sentiment and timelag,[4] this may reflect limited observation windows for delays. The word count of 'Next Steps' declined steadily, with only brief feedback (≤50 words) beyond 100 days. Word count and positivity also saw less variance at higher timelags, consistent with anticipated declines in variability due to loss of recall.[17] Regression modeling estimated 'Next Step' comments having seven fewer words per 100 days, with a dose-dependent drop in length at higher thresholds seen as early as 14 days. However, significance was lost upon outlier exclusion, suggesting the effect was attributable to outliers of greater delay. This aligns with prior studies where higher word counts have been indirectly associated with higher-quality feedback,[52,53] which in turn has been linked to lower timelag.[4,8,17]

When considered collectively, the duration of delay differentially affected entrustment early with a transient rise near seven days, followed by a dose-dependent decline in QuAL beginning around 14 days ('General Feedback') and 21 days ('Next Steps'). There was a reduction in word count early after 14 days, with more pronounced feedback brevity and sentiment positivity beyond 100 days. While timelag showed limited predictive strength overall, we saw its strongest influences in the rise of empty feedback and decline in QuAL scores, with less impact on word count and entrustment, then sentiment.

These findings raise important questions about the quality of feedback associated with delayed EPA-based assessment. Incorporating our results with prior work, we support the utility of a 14-day threshold to mitigate empty feedback, QuAL declines, and shortened word counts. Feedback beyond 100 days showed consistent bias toward positivity, brevity, and high entrustment. Given the growing backlog of pending assessments in many programs, a 100-day upper limit may serve as a practical starting point beyond which EPA assessments can be deleted or deprioritized, with a gradual approach toward a 14-day completion target.

Our results offer strong utility for optimizing timely, high-quality feedback practices. We identified several timelag thresholds, each associated with distinct changes in EPA feedback metrics. These thresholds can be embedded within electronic portfolios and assessment guidelines to prevent feedback degradation and guide timely completion. These findings also underscore the importance of complementary strategies, such as targeted education initiatives and performance reviews, to promote timely documentation.[28,54]

The study further demonstrates how artificial intelligence (AI) can be a tool to enhance assessment analysis in CBME, through both feedback review and trend identification. We applied pretrained NLP models to evaluate narrative comments, generating sentiment and QuAL scores that we subsequently verified manually. Prior studies have validated these models, reporting high accuracies of 91.3%[35] for sen-

timent classification and 87% (within one point) for QuAL scores.[40] Comparably, for our dataset, 80.0% to 96.4% of feedback did not require manual correction, which varied depending on the predicted metric. We also integrate these outputs into machine learning regression models to predict trends and anticipate patterns of feedback degradation. Combined, this approach provides an efficient and scalable framework for automated EPA assessment analysis, which can inform ongoing quality improvement processes and prospective feedback monitoring systems to better identify delayed or low-quality EPA feedbacks.

This study had several limitations. The proportion of narrative feedback that was empty may have introduced response bias in entrustment and sentiment outcome measures. Although focusing on a single EPA task minimized inter-assessment variability and reduced confounding from task heterogeneity, it also limited generalizability to other specialties. It is important to note the current study fo--uses on quality of documented feedback associated with assessments, and does not account for verbal feedback provided at the time of the activity. It also remains unknown whether the quality of scores used for entrustment decisions (assessment of learning) declines with time in the same way as feedback (assessment for learning). Multiple other factors further contribute to observations being non-independent, such as within-resident or within-assessor correlations and comparisons across postgraduate years. Ideally, approaches such as sandwich-estimators, specifically Huber-White standard errors, or random-effects models would be used to evaluate and adjust for clustering. However, because the research team anonymized all data prior to analysis for confidentiality, we could not conduct such subgroup analyses.

Furthermore, we conducted sentiment analysis using a model fine-tuned on the SST-2 dataset, which has been validated across settings, including clinical entrustment and feedback,[55,56] but its testing in this resident PGME context remains limited. The SST-2 dataset also has limited representation of neutral sentiment, contributing to the bimodal sentiment pattern observed. Similarly, prior researchers fine-tuned the QuAL scoring model on emergency medicine residency data, which may not capture program-specific trends and fully generalize to internal medicine. However, the QuAL scoring criteria itself is context-independent and applied across other specialties with success, like anesthesiology[37] or ophthalmology.[52] While the vast majority of model predictions required no correction following manual rating and review, accuracy was variable across metrics and up to 20% comments required refinement. This highlighted the importance of cautious interpretation of outputs and future independent validation studies to explore prediction patterns, validity, and acceptable accuracy thresholds for deployment.

Conclusion

The findings of this study highlight timelag as a key determinant of feedback quality, length, scoring, and sentiment within EPA-based assessments. We saw increased delay associated with progressive declines in narrative quality, shorter and more frequently empty comments, and a shift toward uniformly positive sentiment, indicating potential loss of specificity and reliability over time. While trends were not strong or always consistent, their potential impact underscores the importance of monitoring timeliness as part of assessment quality assurance. Establishing evidence-based timelag thresholds, like 7 and 14 days, may prevent feedback degradation and promote timely, actionable guidance for learners. Faculty should aim to complete assessments within two weeks to preserve feedback quality, and institutions should consider capping or reconsidering extremely late assessments. Future research should refine optimal timelag ranges, evaluate interventional strategies, examine effects on resident development and feedback engagement, and comprehensively assess NLP-based methods on their validity and effectiveness. Integrating timelag monitoring into electronic portfolios may ultimately en--ance the efficiency and educational impact of as--essments in CBME.

Author information:

1- Department of Medicine, Queen's University, Ontario, Canada

2- School of Computing, Queen's University, Ontario, Canada

3- Postgraduate Medical Education (PGME), Queen's University, Ontario, Canada

4- Office of Professional Development & Educational Scholarship (OPDES), Queen's University, Ontario, Canada

5- Department of Emergency Medicine, Queen's University, Ontario, Canada

6- Department of Psychology, Queen's University, Ontario, Canada

7- Department of Diagnostic Radiology, Queen's University, Ontario, Canada

Correspondence to:

Enoch Yu

email: 20ey3@queensu.ca

Published ahead of issue:

Jun 30, 2026

Conflict of Interest:

Adam Szulewski is a Principal Strategist, Continuing Professional Development, as well as producer and host of the medical education podcast KeyLIME+; both for the Royal College of Physicians and Surgeons of Canada. Heather Braud is a section editor for the CMEJ. She has adhered to the CMEJ policy for editors as authors. The remaining authors have no conflicts of interest to declare.

Funding:

None received.

References

1. Leclair R, Ho JSS, Braund H, et al. Exploring the quality of narrative feedback provided to residents during ambulatory patient care in medicine and surgery. *J Med Educ Curric Dev.* 2023 May 16;10:23821205231175734. https://doi.org/10.1177/23821205231175734.
2. Brown A, Currie D, Mercia M, et al. Does the implementation of competency-based medical education impact the quality of narrative feedback? a retrospective analysis of assessment data in a Canadian internal medicine residency program. *Can J Gen Intern Med.* 2022 Nov;17(4):67–85. https://doi.org/10.22374/cjgim.v17i4.640.
3. Liu L, Jiang Z, Qi X, et al. An update on current EPAs in graduate medical education: a scoping review. *Med Educ Online.* 26(1):1981198. https://doi.org/10.1080/10872981.2021.1981198.
4. Madrazo L, DCruz J, Correa N, Puka K, Kane SL. Evaluating the quality of written feedback within entrustable professional activities in an internal medicine cohort. *J Grad Med Educ.* 2023 Feb;15(1):74–80. https://doi.org/10.4300/JGME-D-22-00222.1.
5. ten Cate O. Entrustability of professional activities and competency-based training. *Med Educ.* 2005 Dec;39(12):1176–7. https://doi.org/10.1111/j.1365-2929.2005.02341.x.

6. Klein R, Snyder ED, Koch J, et al. Analysis of narrative assessments of internal medicine resident performance: are there differences associated with gender or race and ethnicity? *BMC Med Educ.* 2024 Jan 17;24(1):72. https://doi.org/10.1186/s12909-023-04970-2.
7. Branfield Day L, Miles A, Ginsburg S, Melvin L. Resident perceptions of assessment and feedback in competency-based medical education: a focus group study of one internal medicine residency program. *Acad Med.* 2020 Nov;95(11):1712–7. https://doi.org/10.1097/ACM.0000000000003315
8. Fernandes RD, de Vries I, McEwen L, et al. Evaluating the quality of narrative feedback for entrustable professional activities in a surgery residency program. *Ann Surg.* 2024 Apr 25. https://doi.org/10.1097/SLA.0000000000006308.
9. Marcotte L, Egan R, Soleas E, et al. Assessing the quality of feedback to general internal medicine residents in a competency-based environment. *Can Med Educ J.* 2019 Nov 28;10(4):e32–47. https://doi.org/10.36834/cmej.57323
10. Gundle KR, Mickelson DT, Hanel DP. Reflections in a time of transition: orthopaedic faculty and resident understanding of accreditation schemes and opinions on surgical skills feedback. *Med Educ Online.* 2016 Jan 1. https://doi.org/10.3402/meo.v21.30584.
11. Gordon M, Daniel M, Ajiboye A, et al. A scoping review of artificial intelligence in medical education: BEME guide no. 84. *Med Teach.* 2024 Apr 2;46(4):446–70. https://doi.org/10.1080/0142159X.2024.2314198.
12. Bello RJ, Sarmiento S, Rosson GD, et al. Understanding the role for operative performance rating tools in meeting surgical trainee feedback needs: a qualitative study. *Plast Reconstr Surg Glob Open.* 2016 Jun 29;4(6):e780. https://doi.org/10.1097/GOX.0000000000000777.
13. Tuma F, Nassar AK. Feedback in medical education. In: *StatPearls.* Treasure Island (FL): StatPearls Publishing; 2025.
14. Ahn E, LaDonna KA, Landreville JM, Mcheimech R, Cheung WJ. Only as strong as the weakest link: resident perspectives on entrustable professional activities and their impact on learning. *J Grad Med Educ.* 2023 Dec;15(6):676–84. https://doi.org/10.4300/JGME-D-23-00204.1.
15. Nathwani JN, Glarner CE, Law KE, et al. Integrating post-operative feedback into workflow: perceived practices and barriers. *J Surg Educ.* 2017;74(3):406–14. https://doi.org/10.1016/j.jsurg.2016.11.001.
16. Abeid M, Mbekenga C, Jusabani AM, et al. "It's the time they take to give feedback; it has become a problem" exploring barriers and facilitators for research supervision in postgraduate medical education in Tanzania: a qualitative study. *Adv Med Educ Pract.* 2025 Apr 28;16:685–98. https://doi.org/10.2147/AMEP.S512498.
17. Hanmore T, Moon C, Curtis R, Hopman W, Baxter S. Is time really of the essence? timeliness of narrative feedback in ophthalmology CBME assessments. *Med Teach.* 2024;46(5):705–10. https://doi.org/10.1080/0142159X.2023.2274286.
18. Williams RG, Chen X (Phoenix), Sanfey H, et al. The measured effect of delay in completing operative performance ratings on clarity and detail of ratings assigned. *J Surg Educ.* 2014 Nov 1;71(6):e132–8. https://doi.org/10.1016/j.jsurg.2014.06.015.
19. Chen JX, Kozin E, Bohnen J, et al. Tracking operative autonomy and performance in otolaryngology training using smartphone technology: a single institution pilot study. *Laryngoscope Investig Otolaryngol.* 2019;4(6):578–86. https://doi.org/10.1002/lio2.323.
20. Williams RG, George BC, Bohnen JD, et al. A proposed blueprint for operative performance training, assessment, and certification. *Ann Surg.* 2021 Apr;273(4):701. https://doi.org/10.1097/SLA.0000000000004467.
21. Koduri S, Altshuler DB, Khalsa SSS, et al. Using a mobile application for evaluation of procedural learning in neurosurgery. *World Neurosurg.* 2020 Jun 1;138:e124–50. https://doi.org/10.1016/j.wneu.2020.02.049.
22. Abbott KL, Krumm AE, Kelley JK, et al. Surgical trainee performance and alignment

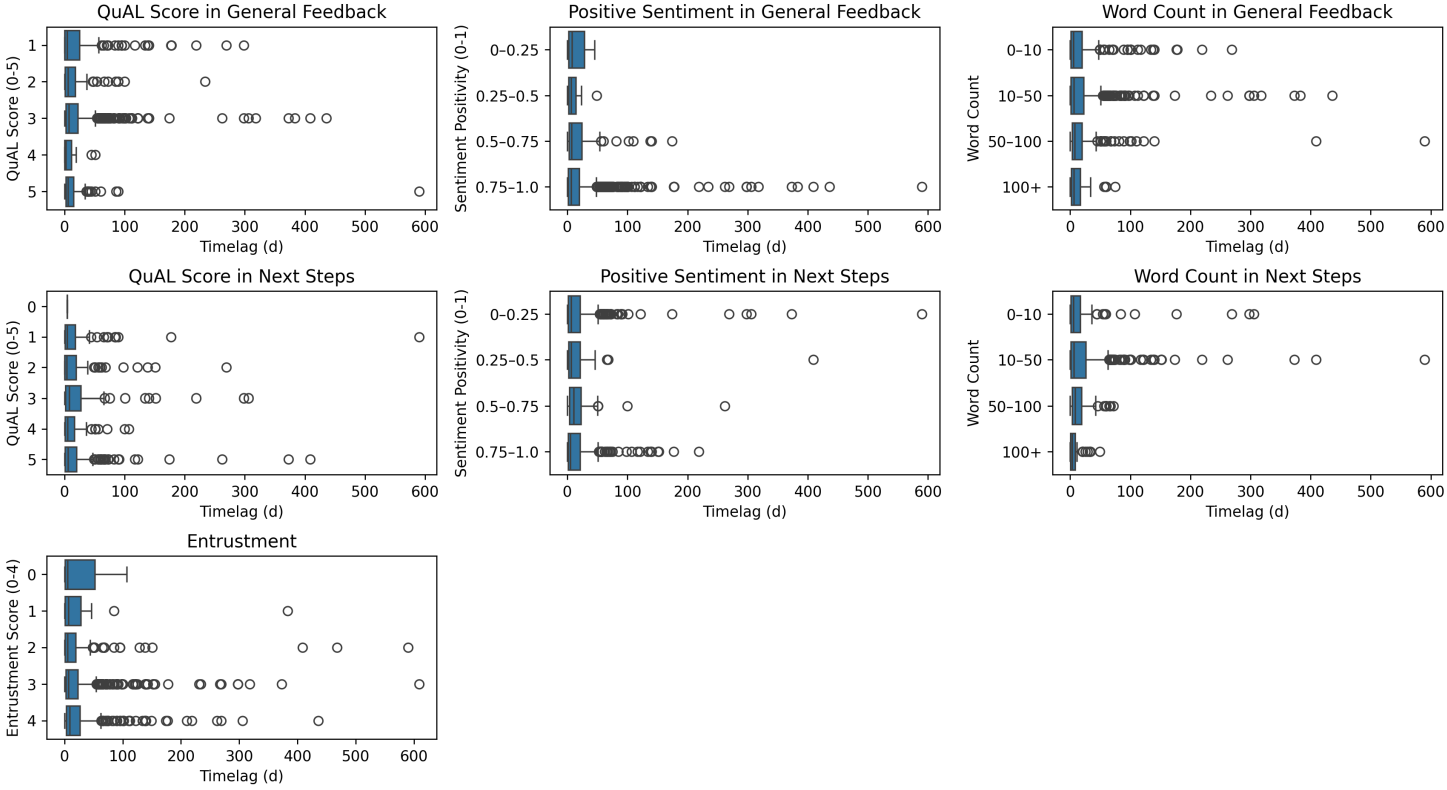
with surgical program director expectations. *Ann Surg.* 2022 Dec;276(6):e1095. https://doi.org/10.1097/SLA.0000000000004990.
23. Abdelsattar JM, AlJamal YN, Ruparel RK, et al. Correlation of objective assessment data with general surgery resident in-training evaluation reports and operative volumes. *J Surg Educ.* 2018 Nov 1;75(6):1430–6. https://doi.org/10.1016/j.jsurg.2018.04.016.
24. Bello RJ, Meyer ML, Cooney DS, et al. The reliability of operative rating tool evaluations: how late is too late to provide operative performance feedback? *Am J Surg.* 2018 Dec 1;216(6):1052–5. https://doi.org/10.1016/j.amjsurg.2018.04.005.
25. Hiller KM, Waterbrook A, Waters K. Timing of emergency medicine student evaluation does not affect scoring. *J Emerg Med.* 2016 Feb 1;50(2):302–7. https://doi.org/10.1016/j.jemermed.2015.09.010.
26. Traser CJ. "Do I really have to complete another evaluation?" exploring relationships among physicians' evaluative load, evaluative strain, and the quality of clinical clerkship evaluations. Indiana University–Purdue University Indianapolis; 2017. http://dx.doi.org/10.7912/C2/2110
27. Abdou H, Kidd-Romero S, Brown RF, Kavic SM, Kubicki NS. Keep it SIMPL: improved feedback after implementation of an app-based feedback tool. *Am Surg.* 2022 Jul 1;88(7):1475–8. https://doi.org/10.1177/00031348221082279.
28. Faiella W, Navjot S, Ramer S. Competency-based cardiology training: a simple approach to improve supervisor completion of entrustable professional activities. *CJC Open.* 2024 Oct 1;6(10):1248–53. https://doi.org/10.1016/j.cjco.2024.07.007.
29. Elentra https://elentra.com
30. Chen HC, van den Broek WES, ten Cate O. The case for use of entrustable professional activities in undergraduate medical education. *Acad Med.* 2015 Apr;90(4):431–6. https://doi.org/10.1097/ACM.0000000000000586
31. Ryan MS, Khan AR, Park YS, et al. Workplace-based entrustment scales for the core EPAs: a multisite comparison of validity evidence for two proposed instruments using structured vignettes and trained raters. *Acad Med.* 2022 Apr 1;97(4):544–51. https://doi.org/10.1097/ACM.0000000000004222
32. ten Cate O. A primer on entrustable professional activities. *Korean J Med Educ.* 2018 Mar;30(1):1–10. https://doi.org/10.3946/kjme.2018.76.
33. Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv; 2020. https://doi.org/10.48550/arXiv.1910.01108.
34. Pang B, Lee L. Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales. *ACL.* 2005:115–24. https://doi.org/10.3115/1219840.1219855.
35. DistilBERT fine-tuned SST2. Hugging Face; 2024.
36. Chan TM, Sebok-Syer SS, Sampson C, Monteiro S. The quality of assessment of learning (Qual) score: validity evidence for a scoring system aimed at rating short, workplace-based comments on trainee performance. *Teach Learn Med.* 2020;32(3). https://doi.org/10.1080/10401334.2019.1708365.
37. Choo EK, Woods R, Walker ME, O'Brien JM, Chan TM. The quality of assessment for learning score for evaluating written feedback in anesthesiology postgraduate medical education: a generalizability and decision study. *Can Med Educ J.* 2023;14(6):78–85. https://doi.org/10.36834/cmej.75876.
38. Johnson AEW, Pollard T, Shen L et al. MIMIC-III, a freely accessible critical care database. *Sci Data.* 2016;3:160035. https://doi.org/10.1038/sdata.2016.35.
39. Alsentzer E, Murphy JR, Boag W, et al. Publicly available clinical BERT embeddings. arXiv; 2019. https://doi.org/10.48550/arXiv.1904.03323.
40. Spadafore M, Yilmaz Y, Rally V, et al. Using natural language processing to evaluate the quality of supervisor narrative comments in competency-based medical education. *Acad Med.* 2024;99(5):534. https://doi.org/10.1097/ACM.0000000000005634
41. Harris CR, Millman KJ, van der Walt SJ, et al. Array programming with NumPy. *Nature.* 2020;585:357–62. https://doi.org/10.1038/s41586-020-2649-2.

42. Pandas Development Team. Pandas. *Zenodo*; 2025. https://doi.org/10.5281/zenodo.17229934.
43. Hunter JD. Matplotlib: a 2D graphics environment. *Comput Sci Eng*. 2007;9(3):90–5.
44. Waskom ML. Seaborn: statistical data visualization. *J Open Source Softw*. 2021;6(60):3021. https://doi.org/10.21105/joss.03021.
45. Virtanen P, Gommers R, Oliphant TE et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods*. 2020;17:261–72. https://doi.org/10.1038/s41592-019-0686-2.
46. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res*. 2011;12:2825–30.
47. Seabold S, Perktold J. *Statsmodels: econometric and statistical modeling with Python*. 2010.
48. Massie J, Ali JM. Workplace-based assessment: a review of user perceptions and strategies to address the identified shortcomings. *Adv Health Sci Educ Theory Pract*. 2016 May;21(2). https://doi.org/10.1007/s10459-015-9614-0.
49. Poeppelman RS, Cho J, Nachbor K, et al. "But why?": explanatory feedback is a reliable marker of high-quality narrative assessment of surgical performance. *Acad Med*. 2025 May;100(5):614–20. https://doi.org/10.1097/ACM.0000000000005985
50. Dudek NL, Marks MB, Regehr G. Failure to fail: the perspectives of clinical supervisors. *Acad Med*. 2005 Oct;80(10 Suppl):S84–7. https://doi.org/10.1097/00001888-200510001-00023.
51. Dexter F, Vasilopoulos T, Fahy BG. Adjusting for resident rater leniency or severity improves the reliability of routine resident evaluations of faculty anesthesiologists. *Cureus*. 2025 Jun 19;17(6). https://doi.org/10.7759/cureus.86366.
52. Curtis R, Moon CC, Hanmore T, Hopman WM, Baxter S. Use the right words: evaluating the effect of word choice and word count on quality of narrative feedback in ophthalmology competency-based medical education assessments. *Can Med Educ J*. 2024 Apr 29. https://doi.org/10.36834/cmej.76671.
53. Page M, Gardner J, Booth J. Validating written feedback in clinical formative assessment. *Assess Eval High Educ*. 2020 Jul 3;45(5):697–713. https://doi.org/10.1080/02602938.2019.1691974.
54. Marty AP, Linsenmeyer M, George B, et al. Mobile technologies to support workplace-based assessment for entrustment decisions: guidelines for programs and educators: AMEE guide no. 154. *Med Teach*. 2023 Nov 2;45(11):1203–13. https://doi.org/10.1080/0142159X.2023.2168527.
55. Kit Y, Mokji MM. Sentiment analysis using pre-trained language model with no fine-tuning and less resource. *IEEE Access*. 2022;10:107056–65. https://doi.org/10.1109/ACCESS.2022.3212367.
56. Gin BC, ten Cate O, O'Sullivan PS, Boscardin C. Assessing supervisor versus trainee viewpoints of entrustment through cognitive and affective lenses: an artificial intelligence investigation of bias in feedback. *Adv Health Sci Educ*. 2024;29(5):1571–92. https://doi.org/10.1007/s10459-024-10311-9

Appendices.

Appendix A. Boxplot descriptive statistics for timelag at different levels of feedback metrics.

Sentiment and word count displayed only EPAs with non-empty comments. We defined metric intervals as left-inclusive and right-exclusive.



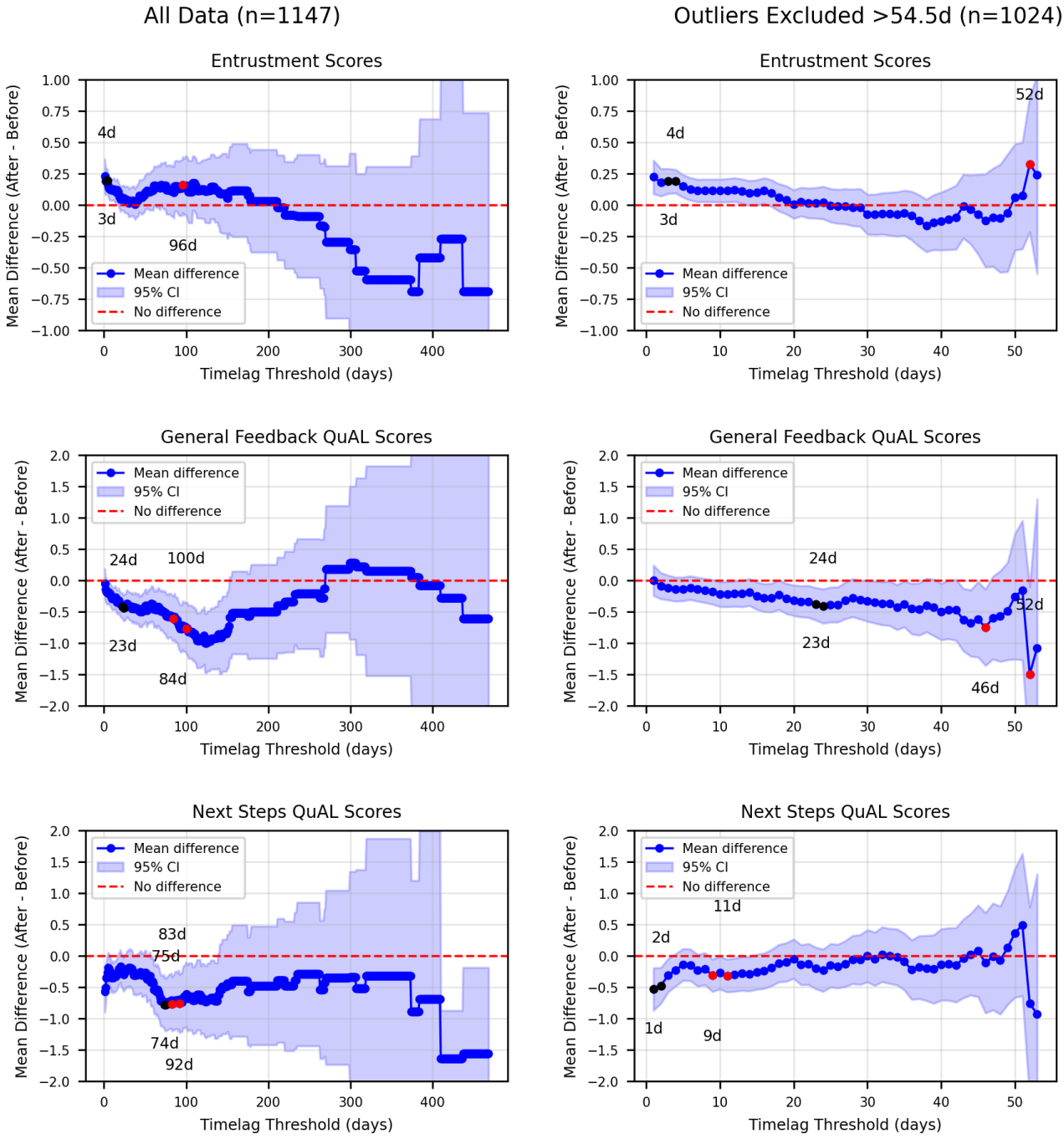
Appendix B. Pearson correlation matrix of timelag to feedback characteristics, with sensitivity analysis of outlier removal*

Correlation with Timelag	*n* (all data)	coefficient	p-val	*n* (outliers removed)	coefficient	p-val
Entrustment	1147	r = 0.02	0.492	1024	r = 0.04	0.156
General Feedback						
QuAL Score	904	r = -0.05	0.132	813	r = 0.03	0.386
Sentiment Positivity	904	r = 0.03	0.299	813	r = -0.10	**0.005**
Word Count	1147	r = -0.06	**0.035**	1024	r = -0.04	0.257
Next Steps						
QuAL Score	604	r = -0.05	0.249	550	r = 0.02	0.651
Sentiment Positivity	604	r = -0.03	0.476	550	r = 0.01	0.854
Word Count	1147	r = -0.08	**0.006**	1024	r = -0.05	0.082

*Outliers were computed for timelag based on Tukey's method with an upper bound of 54.5 days.

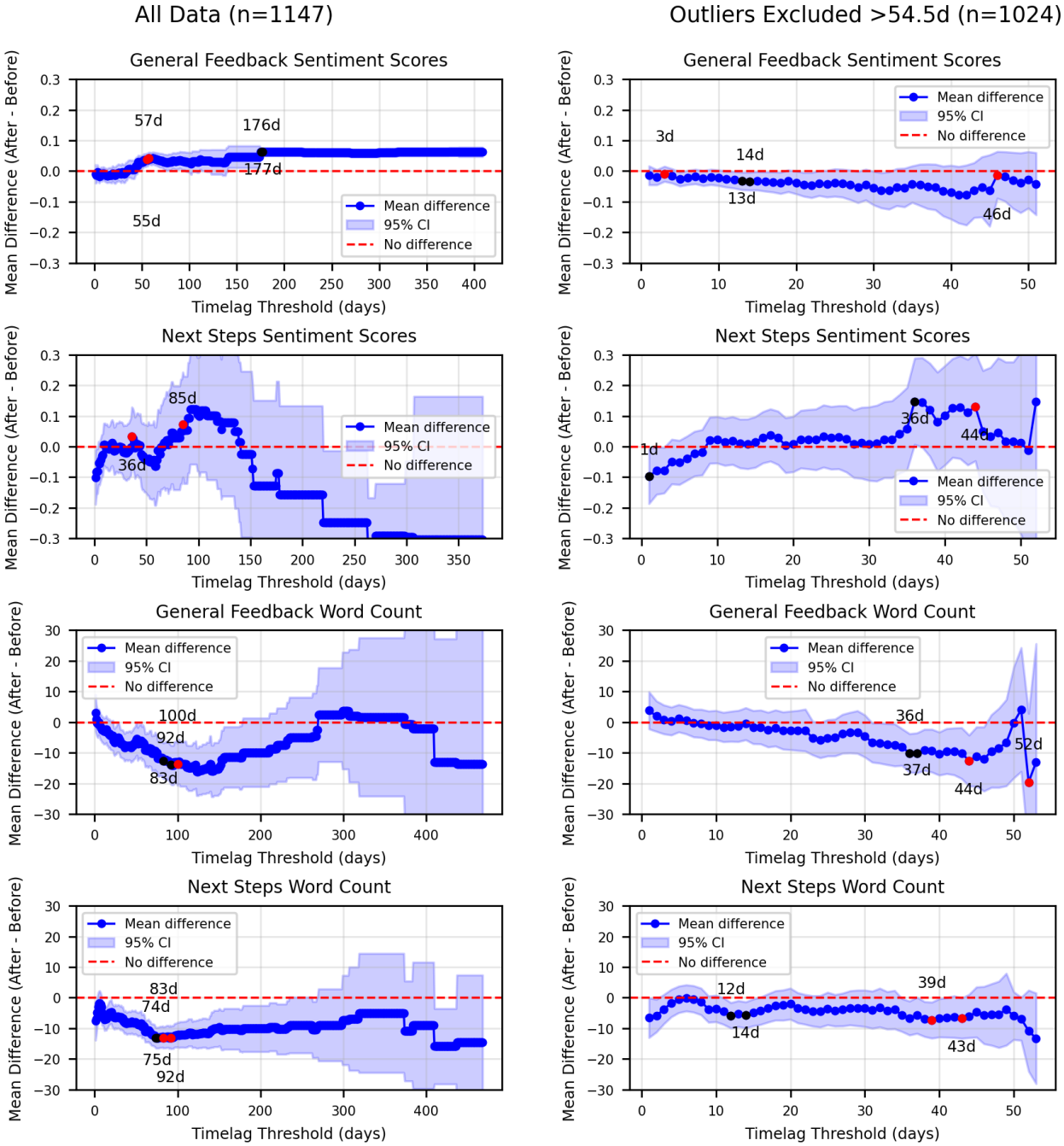
Appendix C. Exploratory threshold analysis for mean differences of QuAL and word count before and after time delay thresholds.

Shaded areas represent 95%CIs (t-distribution) for the mean difference between EPAs after and before or at the threshold. Right plots represent comparisons across all data, while left plots represent comparisons with outlier exclusion (Tukey's method; 54.5d upper bound). Listed dates represent days of time delays with greatest significance (black dot) or mean difference (red dot).



Appendix D. Exploratory threshold analysis for mean differences of sentiment positivity and word count before and after time delay thresholds.

Shaded areas represent 95%CIs (t-distribution) for the mean difference between EPAs after and before or at the threshold. Right plots represent comparisons across all data, while left plots represent comparisons with outlier exclusion (Tukey's method; 54.5d upper bound). Listed dates represent days of time delays with greatest significance (black dot) or mean difference (red dot).



Appendix E. Breusch–Pagan heteroscedasticity tests for variance over time-lag

Correlation with Timelag	*n* (all data)	LM χ^2	p-val	*n* (outliers removed)	LM χ^2	p-val
Entrustment	1147	1.36	0.244	1024	0.07	0.795
General Feedback						
QuAL Score	904	1.42	0.233	2.62	2.62	0.105
Sentiment Positivity	904	0.81	0.369	3.42	3.42	0.0643
Word Count	1147	0.54	0.464	0.04	0.04	0.843
Next Steps						
QuAL Score	604	2.84	0.092	0.41	0.41	0.521
Sentiment Positivity	604	4.36	**0.037** ↓	1.73	1.73	0.189
Word Count	1147	4.19	**0.041** ↓	2.01	2.01	0.156

**We computed outliers for timelag based on Tukey's method with an upper bound of 54.5 days. Arrows signify direction of variance change in higher time delays, for significant results.*