Classifying Clinical Work Settings Using EHR Audit Logs: A Machine Learning Approach

Seunghwan Kim, MS; Sunny S. Lou, MD, PhD; Laura R. Baratta, BS; and Thomas Kannampallil, PhD

ABSTRACT

OBJECTIVES: We used electronic health record (EHR)–based raw audit logs to classify the work settings of anesthesiology physicians providing care in both surgical intensive care units (ICUs) and operating rooms.

STUDY DESIGN: Observational study.

METHODS: Attending anesthesiologists who worked at least 1 shift in 1 of 4 surgical ICUs in calendar year 2019 were included. Time-stamped EHR-based audit log events for each week were used to create event frequencies and represented as a term frequency–inverse document frequency matrix. Primary classification outcome of interest was a physician's clinical work setting. Performance of multiple supervised machine learning classifiers were evaluated.

RESULTS: A total of 24 attending physicians were included; physicians performed a median (IQR) of 2545 (906-5071) EHR-based actions per week and worked a median (IQR) of 5 (3-7) weeks in a surgical ICU. A random forest classifier yielded the best discriminative performance (mean [SD] area under receiver operating characteristic curve, 0.88 [0.05]; mean [SD] area under precision-recall curve, 0.72 [0.13]). Model explanations illustrated that clinical activities related to signing of clinical notes, printing handoff data, and updating diagnosis information were associated with the positive prediction of working in a surgical ICU setting.

CONCLUSIONS: A random forest classifier using a frequency-based feature engineering approach successfully predicted work settings of physicians with multiple clinical responsibilities with high accuracy. These findings highlight opportunities for using audit logs for automated assessment of clinician activities and their work settings, thereby affording the ability to accurately assess context-specific work characteristics (eg, workload).

Am J Manag Care. 2023;29(1):e24-e30. doi:10.37765/ajmc.2023.89310

Modern clinical work is documented using electronic health records (EHRs).[1] Given the complexities of clinical practice, physicians often work in different roles across multiple settings (eg, a trainee working in inpatient, emergency care, and outpatient settings over the course of their rotations). The tasks of tracking the sequence of clinical workflows and measuring clinician workload associated with different clinical work settings and responsibilities are challenging. Although observational methods have been used for evaluating workflows in the past,[2,3] these approaches are time- and effort-intensive and prone to observational biases.[4]

Audit log files—trails of clinician interactions with an EHR—have provided considerable opportunities for unobtrusively tracking clinician actions, associated workloads, and downstream outcomes.[5-9] Although much of the prior research on audit logs has focused on descriptive characterization of clinician workload and its impact,[10,11] recent research has used machine learning approaches to place clinician work activities within specific contextual situations. For example, in a recent study, Mai et al[12] developed a predictive model to classify patient encounters associated with a cohort of pediatric residents. This model relied on audit log data to accurately classify physician-patient encounters in inpatient, primary care, and emergency department settings for trainees who worked across these 3 settings. Moreover, the algorithm was used to develop and track clinical work metrics (eg, number of patients seen) during various time periods.[13] Several similar studies utilized audit logs to ascertain clinical work practices and behaviors, including determining actual resident duty hours in comparison with self-reported duty hours,[14] assessing time spent on primary care patient exams,[15] and investigating clinical note writing practices and their impact on note writing efficiency.[16]

A similar, but unexplored and important, class of workflow identification problem with audit logs is ascertaining the clinical work settings and associated work responsibilities of clinicians who practice in multiple roles. For example, determining the clinical service and practice setting of a clinician with multiple practices and expertise (eg, an anesthesiologist seeing patients

in the surgical intensive care unit [ICU] and in an operating room) often requires complex proxy metrics and heuristics.

The differences in clinical work settings introduces diverging patterns of work responsibilities and actions. These differences in work patterns can potentially be ascertained through the analysis of actions that are performed in the EHR. Such an approach has relevance in the modern clinical work environment, where EHR-based actions are often tracked to assess clinician performance metrics to manage both their workload and their wellness (eg, Epic's Signal platform providing time spent on the EHR). Appropriately classifying the clinical service setting during clinical care is key to assessing clinical practice patterns (eg, documentation), behavioral patterns, workload (eg, time spent on the EHR), and other downstream effects (eg, errors) associated with clinical service (eg, for a dually certified physician in the emergency and ICU settings).

In this study, we developed a data pipeline and an analytical approach to track the work activities of physicians to classify their associated clinical work settings using raw audit log files. Our primary hypothesis was that EHR-based raw clinical activity logs—a proxy for the clinical activities associated with setting-specific work responsibilities—could be used for discerning clinical work settings. Specifically, we investigated the use of supervised machine learning classifiers to ascertain the clinical work settings based on a sequence of EHR-based work activities of anesthesiology physicians who provided care in both surgical ICUs and operating rooms.

TAKEAWAY POINTS

Physicians often perform different roles in multiple settings; automatically classifying physician activities from raw electronic health record (EHR)–based audit logs can help accurately assess their work-related behaviors associated with each setting (eg, an intensive care unit vs an operating room). We developed a data pipeline and associated machine learning algorithms for automatic classification of clinical work settings.

- A random forest classifier had a high discriminative performance (accuracy = 0.92), and model explanations showed discriminant validity.
- Automatically classifying work settings helps accurately assess setting-specific work patterns; this has implications for accurate workload measurements, targeted physician support, and EHR design aligned with work-related contexts.

METHODS

Participants and Study Design

This study was conducted at Barnes-Jewish Hospital in St Louis, Missouri, a tertiary care hospital that is part of the academic medical center associated with the Washington University School of Medicine. The study population included attending anesthesiology physicians who worked at least a single shift in 1 of 4 surgical ICUs between January 1 and December 31, 2019.

Clinical responsibilities for critical care–trained anesthesiologists are typically split between work weeks serving as the supervising attending physician in a surgical ICU and weeks supervising the provision of anesthesia in operating rooms. Based on our informal discussions with anesthesiologists (and coauthor S.S.L., a practicing anesthesiologist), an attending physician's job responsibilities generally differ in these 2 roles. For example, in the surgical ICU, attending physicians care for up to 35 critically ill patients at once; typical EHR work activities include reviewing patient information and cosigning or creating addenda to notes written by resident physicians or advanced practice providers. In the operating room, attending physicians supervise up to 4 operating rooms simultaneously, each staffed by a nurse anesthetist or a resident physician; typical EHR activities include reviewing patient information, writing preoperative and postoperative assessments, documenting intraoperative events, and placing medication and laboratory orders. These are examples of expected clinical work activities and are not a comprehensive list of all possible activities in either of these settings.

The data for this study were part of a larger study evaluating the work practices of anesthesia clinicians in perioperative settings.[6] This study was approved by the institutional review board of Washington University (IRB #202009032) with a waiver of informed consent.

Data Collection

As mandated by the Health Insurance Portability and Accountability Act, all EHR-based activities are recorded to monitor access to protected patient health information. These data—commonly referred to as audit logs—are trails of clinician activity and are stored in structured databases. Each user action on the EHR to export, modify, or view creates corresponding audit log events that include data on the user performing the action, a time stamp of the event, accessed EHR component (eg, reviewing patient chart, ordering medication), and associated patient identifiers. We extracted EHR-based raw audit logs for each attending physician in the study from Epic's Clarity database (Epic Systems) for calendar year 2019. We did not utilize vendor-derived audit log aggregated metrics (eg, Epic Signal) because these measures are subject to change and do not have an associated anesthesia data model, making them unsuitable for week-level classification.

Primary Outcome

The primary outcome of interest was the identification of whether an attending physician was working in a surgical ICU on a given week. A master schedule from the institutional billing office was used as the "ground truth" to identify weeks during which a physician was working in this setting. Based on the billing data, a physician's work week in the ICU was assigned a positive binary label and work weeks in the non-ICU setting (ie, operating room) were assigned a negative binary label.

Feature Engineering

Audit log augmentation. Raw audit logs lack details regarding EHR components related to each access event. Therefore, we retrieved additional metadata from Epic's Clarity database for actions related to notes and reports, to populate a "report name" field with granular information on the type of report or note that the physician accessed (eg, patient chart advisories report, progress notes).[17] The "metric name" field extracted with the raw audit log files, which represents an action performed in the EHR, was combined with the report name field to represent distinct EHR actions. We refer to these as EHR action pairs.

Frequency encoding. The raw audit log data set was segmented into unique clinician work weeks (in the surgical ICU vs operating room, where the same anesthesiologist provided care), matching our primary outcome definition of a week in a clinical setting. We created representations of work weeks by computing the relative frequency of each EHR action pair (ie, metric name and report name combination) for each week. This approach was analogous to a bag-of-words approach used in natural language processing (NLP). The goal of this approach was to encode the appearance of tokens (ie, words for NLP, EHR action pairs for raw audit logs) by assigning numerical values that measured the frequency of appearance within a sequence of tokens (ie, documents for NLP, EHR action pairs recorded over each work week for raw audit logs).

We considered each week of clinical work as a bag-of-words, and the frequency of each unique EHR action pair (ie, a word) in that week was computed. Next, we computed the term frequency–inverse document frequency (TF-IDF) statistic, commonly used in information retrieval and NLP, for all weeks such that rare EHR action pairs were given greater weight and thus assigned a higher relative frequency value. The resulting matrix was then used as a frequency-based feature matrix, with each row representing a unique clinician's work week and each column a unique EHR action pair.

Machine Learning Model Development and Training

Logistic regression, multinomial naïve Bayes, and the random forest classifiers were used for supervised classification. For all supervised classification models, we used a nested, stratified cross-validation approach for training, hyperparameter tuning, model selection, and model evaluation. The stratification of cross-validation splits ensured that class imbalance was preserved. Within the nested cross-validation structure, the outer cross-validation consisted of 10 folds. Each of the 10 folds was used as a held-back test set while all other folds collectively were used for training for a total of 10 outer iterations. During each outer cross-validation iteration, for each training data set (ie, all folds except for a single fold used as the held-back test data set), a 5-fold inner cross-validation was performed to select the best performing model hyperparameters identified from grid search. This approach therefore resulted in 10 best models (ie, 1 for each of the 10-fold cross validation iterations).

As a baseline, a logistic regression model was used. For this model, the following hyperparameters were tuned: regularization method, stopping criteria, regularization strength, and maximum number of iterations until convergence. Then, a multinomial naïve Bayes classifier with a tuned additive smoothing hyperparameter and learned class prior probabilities was also trained and tested, as it is known to be suitable for text classification with discrete features and works for the fractional counts produced by TF-IDF. Finally, a random forest classifier, an advanced ensemble of decision tree learners, was used to fit a group of decision trees on the data set to control overfitting and improve classification accuracy. The following hyperparameter sets were tuned for the random forest model: number of decision trees and maximum number of features considered for each split. Ranges of hyperparameter spaces searched are listed in the **eAppendix Table** (available at **ajmc.com**).

Supervised machine learning classifiers often perform poorly when the number of features exceeds the number of training examples.[18] Therefore, principal component analysis (PCA) was used for dimensionality reduction. PCA is an orthogonal linear transformation of the original data onto a lower-dimension linear space that maximizes variance of projected data and minimizes information loss.[19,20] We applied PCA to our feature matrix by including the principal components that have a cumulative explained variance above the 80% threshold. All supervised learning models were applied to both the original feature matrix and the PCA-reduced feature matrix, except for the multinomial naïve Bayes, which cannot handle negative feature values produced by the PCA.

All analyses were performed using custom Python code with Python 3.8.13 and sci-kit learn 1.0.2 versions.

Model Evaluation

Each model was evaluated using performance metrics—accuracy score, area under the receiver operating characteristic curve (AUROC), area under precision-recall curve (AUPRC), precision or positive predictive value, recall (sensitivity), and F1 score—and averaged across 10-fold cross validation loops. For the best performing model, we used the Shapley additive explanations (SHAP)[21] to identify the top 20 EHR action pairs that contributed to the overall classification outcome. SHAP is a model-agnostic explanation technique that helps in determining the contribution of each feature to the model's prediction using a game-theoretic approach.

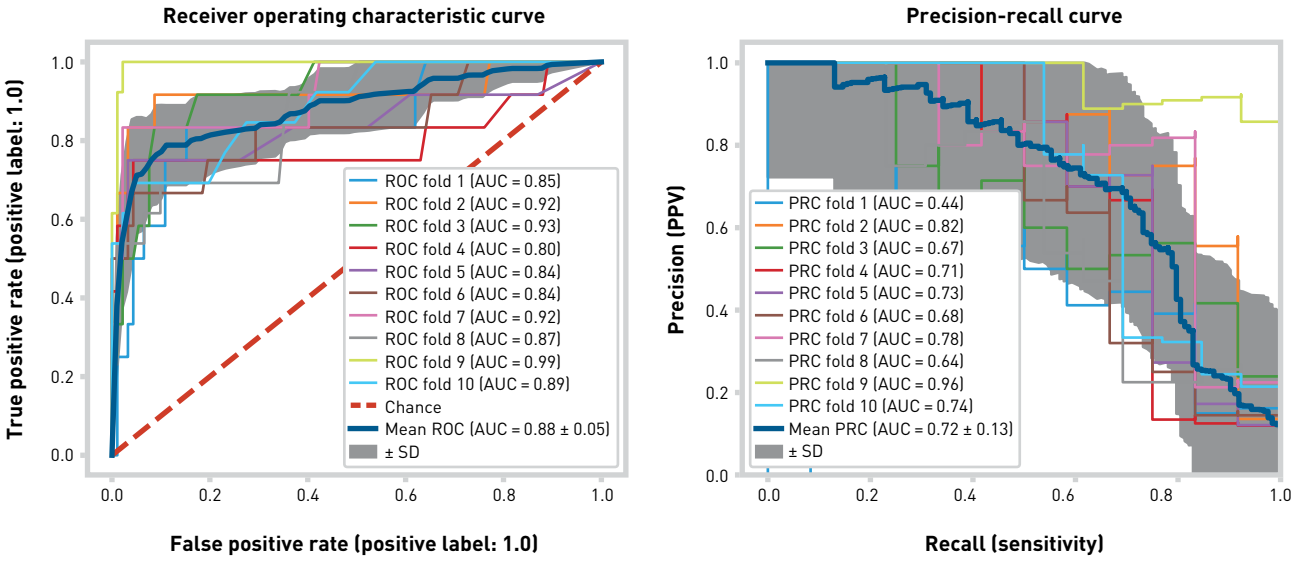
In addition, we identified the top 5 EHR action categories associated with each clinical work setting. The rank of contribution was determined using the magnitude of the mean SHAP value, and the clinical action categories were manually produced by grouping actions that were associated with a single clinical task (eg, cosigning clinical note, cosigning clinical note with attestation, and signing a clinical note were grouped into a category called "signing clinical notes").

RESULTS

General Characteristics

A total of 24 attending anesthesiology physicians who worked at least 1 shift in the surgical ICU in 2019 were included. Each physician

FIGURE 1. Predictive Performance of the Best Performing Random Forest Model on Original TF-IDF Feature Matrix[a]



AUC, area under the curve; PPV, positive predictive value; PRC, precision-recall curve; ROC, receiver operating characteristic curve; TF-IDF, term frequency–inverse document frequency.

[a]We plotted (A) ROC and (B) PRC. Figures show curves averaged across 10-fold cross validation (blue), SD of the mean (shaded grey), and curves for each cross-validation fold. AUC and SD of each ROC and PRC are included in the legends.

TABLE 1. Predictive Performances for Supervised Classification Models[a]

Model	Features	Accuracy	AUROC	AUPRC	Precision (PPV)	Recall (sensitivity)	F1
Logistic regression	TF-IDF	0.90 (0.02)	0.85 (0.06)	0.55 (0.11)	0.69 (0.22)	0.31 (0.12)	0.41 (0.13)
Logistic regression	PCA-reduced	0.87 (0.04)	0.63 (0.07)	0.16 (0.04)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
Naïve Bayes	TF-IDF	0.88 (0.00)	0.78 (0.05)	0.33 (0.10)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
Random forest	**TF-IDF**	**0.92 (0.01)**	**0.88 (0.05)**	**0.72 (0.13)**	**0.90 (0.11)**	**0.35 (0.08)**	**0.50 (0.09)**
Random forest	PCA-reduced	0.87 (0.02)	0.70 (0.07)	0.27 (0.08)	0.34 (0.33)	0.11 (0.12)	0.16 (0.16)

AUPRC, area under the precision-recall curve; AUROC, area under the receiver operating characteristic curve; PCA, principal component analysis; PPV, positive predictive value; TF-IDF, term frequency–inverse document frequency.

[a]Three supervised machine learning classifiers were evaluated: logistic regression, multinomial naïve Bayes classifier, and the random forest classifier. All supervised models were fit to the original TF-IDF feature matrix and PCA-reduced matrix except for the multinomial naïve Bayes classifier, which cannot handle negative feature values produced by the PCA. Mean (SD) testing performance values across nested 10-fold cross-validation are shown for all metrics. The best performing random forest classifier is in bold.

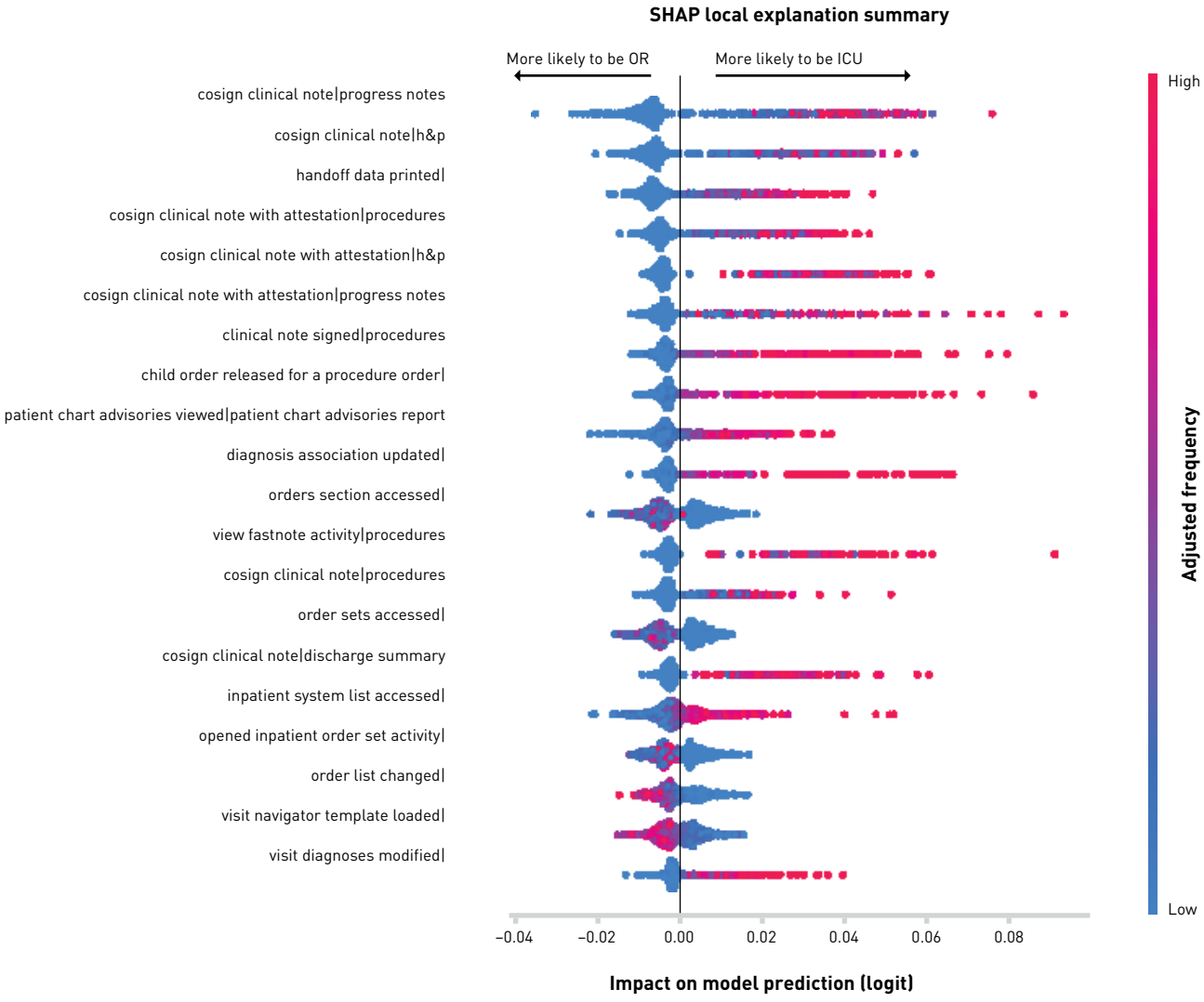
worked a median (IQR) of 47 (38-50) weeks per year, with a median (IQR) of 5 (3-7) weeks in the ICU. A total of 1042 work weeks were observed, with each week containing a median (IQR) of 2545 (906-5071) audit log activities. A total of 1177 unique EHR action pairs were identified in the audit log data.

Model Performance

The random forest classifier on the original TF-IDF matrix yielded the highest discriminative performance, with a mean (SD) AUROC of 0.88 (0.05) and a mean (SD) AUPRC of 0.72 (0.13), and was selected as the best performing model (**Figure 1**). Baseline logistic regression model also yielded a high predictive performance, with a mean (SD) AUROC of 0.85 (0.06) and a mean (SD) AUPRC of 0.55 (0.11). Multinomial naïve Bayes classifier performance on the original feature matrix yielded a mean (SD) AUROC of 0.78 (0.05) and a mean (SD) AUPRC of 0.33 (0.10). Multinomial naïve Bayes classifier was not applied to the PCA-reduced matrix due to its inability to process negative feature values created from PCA. All evaluation metrics were averaged across 10 cross-validation loops, for which each cross-validation iteration had a best model selected through a nested 5-fold cross-validation of the training folds. Performance of all supervised classifiers are summarized in **Table 1**.

For the random forest model applied to the original TF-IDF feature matrix, 20 features with the greatest impact on the model's prediction were identified (**Figure 2**). In any given week, higher adjusted frequencies of activities related to signing clinical notes increased the model's likelihood of predicting a week in the ICU

FIGURE 2. SHAP Local Explanation Summary Plot of Overall Top 20 Contributive Features (ascending order) to the Classification Outcome Across 10-Fold Cross-Validation[a]



Child order, related order; EHR, electronic health record; h&p, history and physical; ICU, intensive care unit; OR, operating room; SHAP, Shapley additive explanations; TF-IDF, term frequency–inverse document frequency.

[a]Each feature is an EHR action pair, which is the metric name of the action combined with a report name (whenever it is available), separated by a pipe symbol "|" after the metric name. Each dot represents a clinician's work week. Vertical location shows what feature it is depicting, color shows whether that feature had high or low adjusted frequency (based on TF-IDF statistic) for that work week, and horizontal location shows whether the effect of that value caused a prediction of surgical ICU or OR setting. For example, higher adjusted frequency of an EHR action pair named "cosign clinical note|progress notes" increased the model's likelihood of predicting an ICU work week; higher frequency of an "order list changed|" increased the model's likelihood of predicting a work week in the OR.

(ie, positive label). Similarly, higher frequencies of activities related to printing handoff data and updating diagnosis associations (see Figure 2 for a detailed list) also increased the model's likelihood of predicting a work week in the ICU. In contrast, higher frequencies of activities related to accessing order sets, modifying orders, and loading the navigator template had the most impact on predicting a work week in the operating room). The top 5 action categories contributing to predicting each setting are provided in **Table 2**.

DISCUSSION

In this study, we developed an automated process for classifying the clinical work settings of anesthesiologists who had dual responsibilities of working in both surgical ICUs and operating rooms. Our primary hypothesis was that EHR-based raw clinical activity logs—a proxy for the clinical activities associated with setting-specific clinical work responsibilities—could be used for discerning clinical work settings. Toward this end, we developed

an automated prediction pipeline using raw audit logs and applied various supervised machine learning algorithms. We found that a random forest classifier could identify physicians' work setting (ie, surgical ICU) with 92% accuracy on average. This highlights the ability of models that rely on unobtrusively collected sequences of clinical work activities to automatically classify activities of the same physician across different clinical work settings (ie, surgical ICU, operating room).

For the best-performing random forest models across nested cross-validation, explanations derived using SHAP values showed discriminant validity: differentiating between tasks performed in the surgical ICU and the operating room; for example, physicians often signed residents' clinical notes in the ICU. Other features of surgical ICU work were also prominent and included printing handoff data, viewing patient chart advisories report, viewing note activity related to procedures, and updating diagnosis information.

There are a limited number of studies applying machine learning techniques to audit logs; a majority of these studies have relied on unsupervised clustering techniques to generate exploratory groupings of audit log actions.[12-16] Moreover, to the best of our knowledge, there are no known studies of automated work setting classification using audit log events. These findings have significant implications for burgeoning research on the use of raw audit logs for unobtrusively ascertaining clinical work activities and settings. Several groups of physicians work across multiple settings and modalities; for example, physicians work in both outpatient and inpatient settings, often on the same day, and trainees often transition among different settings based on their clinical rotations. Identifying the appropriate settings and context of work is important for accurate measure computation, given the considerable attention on standardized metrics related to EHR-based work,[22] such as time spent on the EHR, documentation time, and inbox time.

From a methodological standpoint, our approach relied on "engineered" features from raw sequences of audit log events. This feature engineering approach was adopted from the word tokenization and frequency computation methods of text analysis techniques; such methods have been used to represent the frequency of EHR-based events within short time frames. The use of word tokenization and frequency measurements was particularly suitable for raw audit log data because just like words in a sentence, each EHR action pair was recorded sequentially to form a longitudinal set of action events.

This study highlights the potential of audit logs to ascertain clinical actions and settings associated with those actions, which is often difficult to accomplish using clinical data alone. For instance, within a work week, a physician in the operating room can access multiple patients, and some of these may be patients they cared for during previous encounters (eg, holdover from the previous week). EHR-based data are also hampered by incomplete data regarding physician schedules or changes in their schedules, which are often managed external to an EHR. As such, approaches relying on alternate data sources, such as EHR-based audit logs, provide potential directions for future research.

TABLE 2. Top 5 Contributive Features for Each Clinical Work Setting[a]

Work setting	Rank of contribution	Clinical action category
ICU	1	Signing clinical notes
	2	Printing handoff data
	3	Order release
	4	Chart access
	5	Diagnosis modification
OR	1	Order access
	2	Order modification
	3	Loading navigator template
	4	Chart review
	5	Clinical note review

ICU, intensive care unit; OR, operating room; SHAP, Shapley additive explanations.

[a]Based on SHAP values and the direction of influence on predicted outcome, top 5 contributive features for each clinical work setting (ie, surgical ICU, OR) were extracted.

Limitations

This study has several limitations. This was a single-center study with a small number of physicians working in a specific set of clinical settings (ie, surgical ICU and the operating room), and as such the findings may not be generalizable to other settings or physician groups. Temporality of audit log sequences was not considered; incorporation of such information may better represent contextual information and improve predictive performance. Furthermore, there is a need to validate the top contributive EHR-based activities from the SHAP analysis to better interpret their meaning in the clinical workflow context. Finally, there is a large class imbalance between the class labels; oversampling techniques may reduce the class imbalance problem and improve predictive performance.

The use of the TF-IDF statistic in audit log–based feature creation may introduce an unintended bias by upweighting rare actions that are not representative of work responsibilities (eg, holdover activities from the previous week). However, often there are common work activities regardless of work settings, and the markers of setting-specific unique activities could also be rare. Thus, we avoided masking the effect of discriminative ability of those setting-specific unique activities. As a result, we did not incorporate any additional weighting schemes to the actions (eg, weighting based on the action content).

CONCLUSIONS

For clinicians who have various duties across multiple clinical settings, determining clinical work activities and roles is important to study clinical practice patterns and behaviors, workload, and other downstream effects such as errors pertaining to each setting. Toward this end, our focus was on developing an automated process to classify clinical work settings based on raw audit log data. We found that supervised machine learning techniques along with a frequency-based feature engineering approach can successfully

predict clinical work settings with reasonably high accuracy and face validity. These findings provide an opportunity for an automated approach to characterize clinician activities and work settings in other domains. ■

Acknowledgments

The authors would like to thank Derek Harford for his assistance in obtaining the data used in this study.

Author Affiliations: Institute for Informatics (SK, SSL, LRB, TK), Division of Biology and Biomedical Sciences (SK, LRB, TK), and Department of Anesthesiology (SSL, TK), School of Medicine, Washington University in St. Louis, St Louis, MO; Department of Computer Science and Engineering, McKelvey School of Engineering (TK), Washington University in St. Louis, St Louis, MO.

Source of Funding: None.

Author Disclosures: The authors report no relationship or financial interest with any entity that would pose a conflict of interest with the subject matter of this article.

Authorship Information: Concept and design (SK, SSL, TK); acquisition of data (SSL, TK); analysis and interpretation of data (SK, SSL, TK); drafting of the manuscript (SK, LRB, TK); critical revision of the manuscript for important intellectual content (SK, SSL, LRB, TK); statistical analysis (SK); obtaining funding (SSL); administrative, technical, or logistic support (SK, LRB, TK); and supervision (SK, SSL, TK).

Address Correspondence to: Thomas Kannampallil, PhD, Washington University in St. Louis, 660 S Euclid Ave, Campus Box 8054, St Louis, MO 63110. Email: thomas.k@wustl.edu.

REFERENCES

1. Jha AK. Meaningful use of electronic health records: the road ahead. *JAMA*. 2010;304(15):1709-1710. doi:10.1001/jama.2010.1497.
2. Abraham J, Reddy MC. Challenges to inter-departmental coordination of patient transfers: a workflow perspective. *Int J Med Inform*. 2010;79(2):112-122. doi:10.1016/j.ijmedinf.2009.11.001
3. Abraham J, Reddy MC. Re-coordinating activities: an investigation of articulation work in patient transfers. In: *CSCW '13: Proceedings of the 2013 Conference on Computer Supported Cooperative Work*. Association for Computing Machinery; 2013:67-78. doi:10.1145/2441776.2441787
4. Sedgwick P, Greenwood N. Understanding the Hawthorne effect. *BMJ*. 2015;351:h4672. doi:10.1136/bmj.h4672
5. Adler-Milstein J, Adelman JS, Tai-Seale M, Patel VL, Dymek C. EHR audit logs: a new goldmine for health services research? *J Biomed Inform*. 2020;101:103343. doi:10.1016/j.jbi.2019.103343
6. Lou SS, Kim S, Harford D, et al. Effect of clinician attention switching on workload and wrong-patient errors. *Br J Anaesth*. 2022;129(1):e22-e24. doi:10.1016/j.bja.2022.04.012
7. Lou SS, Lew D, Harford DR, et al. Temporal associations between EHR-derived workload, burnout, and errors: a prospective cohort study. *J Gen Intern Med*. 2022;37(9):2165-2172. doi:10.1007/s11606-022-07620-3
8. Lou SS, Liu H, Warner BC, Harford D, Lu C, Kannampallil T. Predicting physician burnout using clinical activity logs: model performance and lessons learned. *J Biomed Inform*. 2022;127:104015. doi:10.1016/j.jbi.2022.104015
9. Kannampallil T, Abraham J, Lou SS, Payne PR. Conceptual considerations for using EHR-based activity logs to measure clinician burnout and its effects. *J Am Med Inform Assoc*. 2021;28(5):1032-1037. doi:10.1093/jamia/ocaa305
10. Rule A, Chiang MF, Hribar MR. Using electronic health record audit logs to study clinical activity: a systematic review of aims, measures, and methods. *J Am Med Inform Assoc*. 2020;27(3):480-490. doi:10.1093/jamia/ocz196
11. Rule A, Melnick ER, Apathy NC. Using event logs to observe interactions with electronic health records: an updated scoping review shows increasing use of vendor-derived measures. *J Am Med Inform Assoc*. Published online September 29, 2022. doi:10.1093/jamia/ocac177
12. Mai MV, Orenstein EW, Manning JD, Luberti AA, Dziorny AC. Attributing patients to pediatric residents using electronic health record features augmented with audit logs. *Appl Clin Inform*. 2020;11(3):442-451. doi:10.1055/s-0040-1713133
13. Mai MV, Muthu N, Carroll B, Costello A, West DC, Dziorny AC. Measuring training disruptions using an informatics based tool. *Acad Pediatr*. Published online March 16, 2022. doi:10.1016/j.acap.2022.03.006
14. Lin JA, Pierce L, Murray SG, et al. Estimation of surgical resident duty hours and workload in real time using electronic health record data. *J Surg Educ*. 2021;78(6):e232-e238. doi:10.1016/j.jsurg.2021.08.011
15. Neprash HT, Everhart A, McAlpine D, Smith LB, Sheridan B, Cross DA. Measuring primary care exam length using electronic health record data. *Med Care*. 2021;59(1):62-66. doi:10.1097/MLR.0000000000001450
16. Gong JJ, Soleimani H, Murray SG, Adler-Milstein J. Characterizing styles of clinical note production and relationship to clinical work hours among first-year residents. *J Am Med Inform Assoc*. 2022;29(1):120-127. doi:10.1093/jamia/ocab253
17. Lou SS, Liu H, Harford D, Lu C, Kannampallil T. Characterizing the macrostructure of electronic health record work using raw audit logs: an unsupervised action embeddings approach. *J Am Med Inform Assoc*. Published online December 8, 2022. doi:10.1093/jamia/ocac239
18. Hughes G. On the mean accuracy of statistical pattern recognition. *IEEE Trans Inform Theory*. 1968;14(1):55-63. doi:10.1109/TIT.1968.1054102
19. Hotelling H. Analysis of a complex of statistical variables into principal components. *J Educ Psychol*. 1933;24(6):417-441. doi:10.1037/h0071325
20. Jolliffe IT. Principal component analysis for special types of data. In: Jolliffe IT. *Principal Component Analysis*. Springer; 2002:338-372.
21. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: *Advances in Neural Information Processing Systems*. Neural Information Processing Systems; 2017.
22. Sinsky CA, Rule A, Cohen G, et al. Metrics for assessing physician activity using electronic health record log data. *J Am Med Inform Assoc*. 2020;27(4):639-643. doi:10.1093/jamia/ocz223

Visit **ajmc.com/link/89310** to download PDF and eAppendix

Supplemental Materials

eAppendix Table. Range of hyperparameters during the supervised model selection process. For all supervised classification models, we used a nested, stratified cross validation approach for training, hyperparameter tuning, model selection and model evaluation. Within the nested cross validation structure, the outer cross validation consisted of 10 folds; each of the 10 folds was used as a held back test set while all other folds collectively were used for training. During each outer cross validation iteration, for each training data set (i.e., all folds except for a single fold used as the held back test data set), a 5-fold inner cross validation was performed to select the best performing model hyperparameter set identified from grid search.

Model	**Hyperparameter**	**Range of Values**
Logistic regression	Regularization penalty; *penalty*	{l2, none}
	Tolerance for stopping criteria; *tol*	$\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$
	Inverse of regularization strength; *C*	$\{10^{-3}, 10^{-2}, 10^{-1}, 1\}$
	Maximum number of iterations until convergence; *max_iter*	{10, 100, 500}
Multinomial naïve Bayes classifier	N/A	N/A
Random forest classifier	Number of decision trees; *n_estimators*	{10, 100, 500}
	Maximum number of features considered for each best split; *max_features*	{sqrt, log2}