Journal of the American Medical Informatics Association, 28(6), 2021, 1168–1177
doi: 10.1093/jamia/ocaa338
Advance Access Publication Date: 12 February 2021
Research and Applications

Research and Applications

Mining tasks and task characteristics from electronic health record audit logs with unsupervised machine learning

Bob Chen[1,2], Wael Alrifai[3,4], Cheng Gao[3], Barrett Jones[3], Laurie Novak[3], Nancy Lorenzi[3], Daniel France[5], Bradley Malin[3,6,7], and You Chen[3,7]

[1]Epithelial Biology Center, Vanderbilt University Medical Center, Nashville, Tennessee, USA, [2]Program in Chemical and Physical Biology, Vanderbilt University, Nashville, Tennessee, USA, [3]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, Tennessee, USA, [4]Department of Pediatrics, Vanderbilt University Medical Center, Nashville, Tennessee, USA, [5]Department of Anesthesiology, Center for Research and Innovation in Systems Safety, Vanderbilt University Medical Center, Nashville, Tennessee, USA, [6]Department of Biostatistics, School of Medicine, Vanderbilt University Medical Center, Nashville, Tennessee, USA, and [7]Department of Electrical Engineering and Computer Science, School of Engineering, Vanderbilt University, Nashville, Tennessee, USA

Corresponding Author: You Chen, PhD, Department of Biomedical Informatics, Vanderbilt University Medical Center, 2525 West End Ave, Suite 1475, Nashville, TN 37203, USA; you.chen@vanderbilt.edu

Received 26 June 2020; Editorial Decision 13 December 2020; Accepted 17 December 2020

ABSTRACT

Objective: The characteristics of clinician activities while interacting with electronic health record (EHR) systems can influence the time spent in EHRs and workload. This study aims to characterize EHR activities as tasks and define novel, data-driven metrics.

Materials and Methods: We leveraged unsupervised learning approaches to learn tasks from sequences of events in EHR audit logs. We developed metrics characterizing the prevalence of unique events and event repetition and applied them to categorize tasks into 4 complexity profiles. Between these profiles, Mann-Whitney *U* tests were applied to measure the differences in performance time, event type, and clinician prevalence, or the number of unique clinicians who were observed performing these tasks. In addition, we apply process mining frameworks paired with clinical annotations to support the validity of a sample of our identified tasks. We apply our approaches to learn tasks performed by nurses in the Vanderbilt University Medical Center neonatal intensive care unit.

Results: We examined EHR audit logs generated by 33 neonatal intensive care unit nurses resulting in 57 234 sessions and 81 tasks. Our results indicated significant differences in performance time for each observed task complexity profile. There were no significant differences in clinician prevalence or in the frequency of viewing and modifying event types between tasks of different complexities. We presented a sample of expert-reviewed, annotated task workflows supporting the interpretation of their clinical meaningfulness.

Conclusions: The use of the audit log provides an opportunity to assist hospitals in further investigating clinician activities to optimize EHR workflows.

Key words: Unsupervised learning, electronic health records, metrics, tasks, audit logs, human-computer interaction, clinician activities

INTRODUCTION

Clinician activities involving electronic health record (EHR) systems can influence the time spent in EHRs and affect their workload, which can induce stress and burnout.[1–6] Clinicians use EHRs for various functions, including chart review, documentation, messaging, orders, patient discovery, medication reconciliation, etc.[7] Healthcare organizations (HCOs) and EHR vendors have previously investigated such usages to understand clinician EHR activities and efficiency.[8,9] These investigations measure the time spent on each EHR function to build provider efficiency profiles.[8–10]

In recent years, EHR audit logs have become valuable resources for the investigation of clinician efficiency in EHRs.[8,11,12] When a clinician accesses or moves between modules in the EHR interface, such as moving from Progress Notes to Order Entry screens, a timestamped record of that action is documented, along with clinician and patient identifiers.[13–16] One example study leveraged audit logs to identify 15 tasks, including clerical (eg, assigning Current Procedural Terminology and International Classification of Diseases–Tenth Revision codes), medical care (eg, reviewing an encounter note), and inbox tasks (eg, developing a letter to a patient) completed by family medicine physicians.[11] Sinsky et al.[12] created a set of metrics to quantify the time spent by a physician in an EHR by using audit logs.

Our approach applies unsupervised learning methods to audit log event sequences to identify and characterize putative EHR tasks performed by clinicians. The goal of our work is to provide an informatics framework for mining audit log data to discover EHR event and session patterns, which we call tasks. We developed hierarchical metrics to describe these tasks and leveraged them to investigate task complexity and efficiency. As a pilot study, this work focuses on task complexity, task efficiency, and task prevalence among clinicians. We stratify tasks by complexity and investigate differences in task efficiency and clinician prevalence between each complexity profile. Our methods can potentially guide HCOs to optimize EHR activities or clinical workflows, by highlighting specific inefficient tasks. To test our methods, we applied our approach to identify and characterize EHR tasks for nurses involved in the care of surgical cases in the neonatal intensive care unit (NICU).

MATERIALS AND METHODS

Here, we present our approach and the metrics we developed to describe sessions and the tasks they comprise; we also characterize task complexity and efficiency profiles. We use a formal hypothesis testing framework to assess the relationship between task complexity and efficiency. Finally, we present evidence supporting the clinical meaningfulness of our discovered tasks by applying process mining algorithms and by examining the clinical workflow annotations of a subset of the involved EHRs.

Characterizing events in EHR audit logs

There are multiple types of events in EHR audit logs, such as placing medication orders, creating progress notes, assigning diagnosis, and so on.[13–18] Each event belongs to 1 of the 4 access types, a categorization inherent to Epic audit log systems (Epic Systems, Verona, WI): export (eg, operating report printed), modify (eg, flowsheets data saved), system (eg, barcode scanned), or view (eg, form viewed). Intrinsically, EHR audit log data describe who did what, when it was performed, and to which patient records. An illustration of these data is provided in Table 1. The definitions of event terminology are provided in Supplementary Table A1. Because the event terminology in this study is specific to the Epic system, we provide examples of generic representations of such event terms in Supplementary Table A2, as interpreted by our NICU experts. Further clinical definitions of the Epic terms included in this study can be found at Epic's EHR UserWeb (userweb.epic.com).

Learning sessions from events through sessionization

The notion of a session is based on the definition used in web analytics as described by Arlitt.[19] A session represents a group of individual user interactions performed within a certain time frame to accomplish some given clinical function using the EHR. In our case, this group of user interactions may be EHR workflow activities (eg, administering medication orders, chart reviewing notes, inputting data into flowsheets). We utilize these sessions, each of which consists of a series of consecutive audit logged EHR events, to abstract clinically meaningful functions. We employ time-oriented sessionization to segregate sequences of events. This framework assumes that multiple consecutive events can be aggregated into a single session, given some time threshold. This assumption is consistent with our observation that clinicians perform EHR tasks in discrete sessions of consecutive actions. Further, we extend our time-oriented sessionization procedure by initializing new sessions each time a clinician accesses a different patient's records.

Still, a challenge in time-oriented sessionization is the determination of a time interval threshold. We used a data-driven heuristic, described by Satopää et al.[20] as "knee point" finding, to estimate

Table 1. Example sessions identified from EHR audit logs

Clinician ID	Patient ID	Timestamp	Session Action/Event	Access Type
Session 1				
A	1	2:10 PM	IN BASKET MESSAGE OF ANY TYPE DISPLAYED IN HYPERSPACE	VIEW
A	1	2:11 PM	VISIT NAVIGATOR TEMPLATE LOADED	VIEW
A	1	2:11 PM	THE PROBLEM LIST IS VIEWED	VIEW
A	1	2:11 PM	PEND A NOTE	MODIFY
A	1	2:12 PM	IN BASKET MESSAGE OF ANY TYPE DISPLAYED IN HYPERSPACE	VIEW
A	1	2:12 PM	AN EXISTING PATIENT IS SELECTED FROM PATIENT LOOKUP	VIEW
A	1	2:14 PM	PATIENT SIDEBAR REPORT ACCESSED	VIEW
Session 2				
A	2	2:18 PM	IN BASKET MESSAGE OF ANY TYPE DISPLAYED IN HYPERSPACE	VIEW
A	2	2:19 PM	PATIENT SIDEBAR REPORT ACCESSED	VIEW
A	2	2:19 PM	A FORM IS VIEWED	VIEW
A	2	2:21 PM	PATIENT DEMOGRAPHICS FORM ACCESSED	VIEW

EHR: electronic health record.

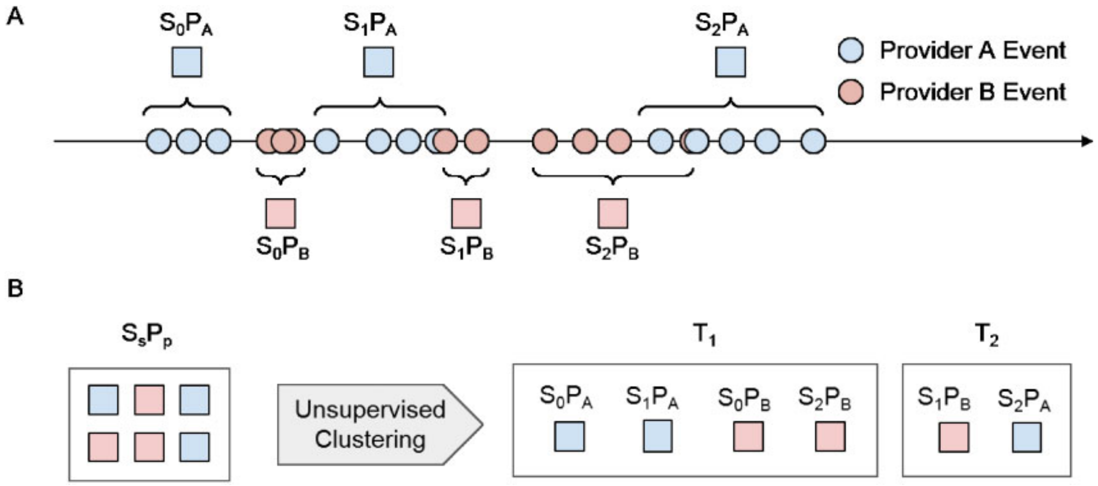

Figure 1. (A) An illustration of using events to learn sessions. A session corresponds to a sequence of consecutive events committed by a healthcare professional to the electronic health records of a patient. Each circle is an event and each square is a session. There are 4 sessions in this figure. Sessions S_0P_A and S_1P_A were performed by healthcare professional P_A, and S_0P_B and S_1P_B were performed by P_B. Events sharing the same timestamp are parsed into different sessions per healthcare professional and new sessions are established under 2 conditions: if the sessionization time threshold is passed or the healthcare professional accesses a different patient's electronic health record. (B) The relationship between events, sessions, and tasks (eg, T_1 and T_2) is enabled through unsupervised clustering of sessions on the basis event count similarities after sessionization.

this threshold and ensure generalizability with other systems or organizations, where sessions may have different lengths. This heuristic is often used to find operating points in complex systems and this "knee" is akin to geometric curvature, formally described by Satopää et al for any continuous function *f* as:

$$K_f(x) = \frac{f''(x)}{\left(1 + f'(x)^2\right)^{1.5}}$$

Where $K_f(x)$ represents a standard closed-form defining the curvature of *f* at any point as a function of its first and second derivative. We find some time threshold, *x*, for our sessionization function which maximizes curvature (of the resulting normalized session number curve) through the Kneedle algorithm.[20] Still, each chosen threshold needs to be validated for clinical meaningfulness in its respective setting, which we do for a subset of selected tasks and sessions through process mining and expert review.

Figure 1 illustrates our process for the sessionization of events and their relationship to tasks. Clinicians P_A and P_B committed 2 sessions represented by squares (S_0P_A and S_1P_A for $\mathbf{P_A}$, and S_0P_B and S_1P_B for $\mathbf{P_B}$) to the EHR of a patient. In Figure 1A, the interval between sessions S_0P_A and S_1P_A is greater than a threshold. Examples of sessions identified from EHR audit logs are provided in Table 1. Figure 1B depicts the aggregation of functionally similar sessions through unsupervised clustering, on the basis of their event count similarities.

Learning tasks from sessions

Though sessionization helps to abstract clinical meaning from individual events, information from multiple sessions can be further integrated to define action patterns, which we call tasks. Each session can be thought of as an observation with N features represented by the number of times an event was performed within that session. Table 2 shows an example of a session feature matrix.

We apply term frequency–inverse document frequency[21] and principal component analysis[22] to normalize the data and to reduce noise and dimensionality. The principal components learned from principal component analysis are used to initialize a t-distributed stochastic neighbor embedding (t-SNE), which is a machine learning algorithm for visualizing high-dimensional datasets.[23] For unsupervised clustering, we used the Leiden community detection algorithm[24] to learn session clusters from the low-dimensional representation of the data generated by t-SNE; each detected cluster of sessions is considered a task in this study.

Visualizing task event flows using process mining and sampled session validation

Because it is difficult to infer clinical meaningfulness from task-aggregated sessions, which consist of long sequences of events, we use process mining to orient each task into a process tree for workflow review. We apply an inductive miner algorithm implemented in ProM, which is an open-source framework for process mining and process tree generation.[25] A process tree is a directed hierarchical graph in which each node is an event with child events representing potential subsequent events within the same session, thus showing the ordered relations among events. Because a task can include hundreds to thousands of sessions, we downsample tasks of interest before process tree generation, by randomly selecting 1000 sessions without replacement. Fundamentally, a task is oriented as a process using ProM, in which batches of sessions are grouped to form a single process tree. In our NICU case study, we leverage the expertise of a NICU nurse and a neonatologist with information derived from EHRs to validate a subset of sessions and tasks for clinical meaningfulness. This focused annotation was accomplished by reviewing the events in a session and analyzing the activities of clinicians that occur during the session through reviews of EHRs (eg, clinical notes, orders, measurements, diagnoses, procedures). For each session, the NICU experts provide a summary statement. We further incorporate 4 event access types (export, modify, system, and view) to subclassify the events and aid in the interpretability of clinical function per session.

Developing hierarchical metrics to describe sessions and tasks

Our metrics' definitions depend on structured layers of abstraction, starting from raw, unabstracted audit log entries to abstracted sessions and tasks. This study examined metrics on 3 levels of abstraction: system level, session level, and task level.

System-level metrics are at the lowest level of abstraction and simply describe events' characteristics on a system-wide level. Specifically, these system-level metrics describe audit log entries given the clinician and patient involved, the type of event recorded, and its timestamp.

Session-level metrics operate on groups of observed events. For example, we define a session's duration as the sum of its constitutive event durations. These session-level abstractions arise from time-oriented sessionization performed on unabstracted audit log entries. Outside of time metrics, we use the number of unique events to quantify the event diversity of a session. Table 3 shows the formal definition of this session diversity coefficient (*session_diversity_coef*). Similarly, specific event types can appear in a single session repeatedly. We developed a metric to describe event repetitiveness within a session, which is formally defined in Table 3 (*session_repetition_coef*).

Task-level metrics are derived from the aggregation of sessions and their respective metrics. Definitions and examples of these metrics are shown in Table 3. We also describe the prevalence of a task by calculating the number of unique clinicians whose performed sessions are clustered within that task. A task affiliated with many unique clinicians demonstrates that it is highly prevalent in the workflows of a given clinical setting. This metric is designed to help HCOs identify tasks that impact many clinicians. The investigation and optimization of such tasks may be used to maximize clinician benefit compared with less prevalent tasks. A complete list of the metrics we developed at different levels can be found in our online repository (https://github.com/bobchen1701/Single_Session_Analysis).

Table 2. An example of a session feature matrix, with corresponding event metadata

	Session 0	Session 1	Session . . .	Session S
Event 1	0	0	30	12
Event 2	0	0	5	1
Event . . .	3	2	1	0
Event E	0	10	0	4

	Provider	Patient	Timestamp
Session S	1	B	01:07:00
Event E	1	B	01:09:00
	1	B	01:11:00
	1	B	01:44:00

Stratifying tasks into their respective complexity profiles

Intuitively, we provide a 2-dimensional definition of task complexity, with task diversity and task repetitiveness. We define highly complex tasks as diverse, involving many unique event types, and repetitive, involving sequences of repeated actions. Our task-level metrics can then be assembled to further stratify learned tasks with respect to complexity, formally defined by the relative rank of the **task_session_repetition_coef** and **task_session_diversity_coef** metrics. Based on this rank-based framework, we divided the learned tasks into quadrants, representing 4 complexity profiles: low diversity/low repetition (LDLR), low diversity/high repetition (LDHR), high diversity/high repetition (HDHR), and high diversity/low repetition (HDLR).

Examining differences in duration, clinician prevalence, and event access type between tasks of different complexity profiles

This project aims to identify which types of tasks are time-consuming and impact large portions of clinicians. We hypothesized that the diversity and repetition coefficients of a task may impact the time spent on it as well as its prevalence among clinicians. Additionally, we measure the differences in event access types between tasks of different complexity profiles, which demonstrate variations in clinical workflows in utilizing EHRs to complete tasks. We conducted pairwise tests to compare the differences in the duration, the event access type representation, and the number of clinicians affiliated with tasks belonging to the 4 defined complexity profiles. Our null hypotheses are that between any pair of task complexity profiles (1) there is no difference in the task durations, (2) there is no difference in the proportion of event access types, and (3) there is no difference between the number of affiliated clinicians. Because the distributions of durations, the proportion of event access types, and

Table 3. Metrics characterizing the time duration, diversity, and repetitiveness of sessions and tasks

Metric	Abbreviation	Definition
Session level		
session_time	s_time	Time to complete a session (seconds). Ex. A NICU nurse completed a session related to the in-basket-message, and the session took 56 seconds.
session_repetition_coef	s_rep_co	Total events per session/total unique events per that same session. Ex. A NICU nurse completed a session consisting of 100 events, and the unique number of events is 10, then the repetition coefficient is 100/10 = 10.
session_diversity_coef	s_dst_co	The diversity of events within a session, relative to the diversity of events in the system. The number of unique events in a session/total number of unique events in the system. Ex. A NICU nurse completed a session with 100 unique events. The total number of unique events in the system is 300. Then the diversity coefficient is 100/300.
Task level		
task_session_time	t_s_time	Mean value of "session_time"
task_session_repetition_coef	t_s_rep_co	Mean value of "session_repetition_coef"
task_session_diversity_coef	t_s_dst_co	Mean value of "session_diversity_coef"
task_provider_diversity	t_p_dst	For a given task, the number of unique healthcare professionals who performed it. Ex. if 30 NICU nurses perform a task, then the provider diversity coefficient is 30 for the task.

NICU: neonatal intensive care unit.

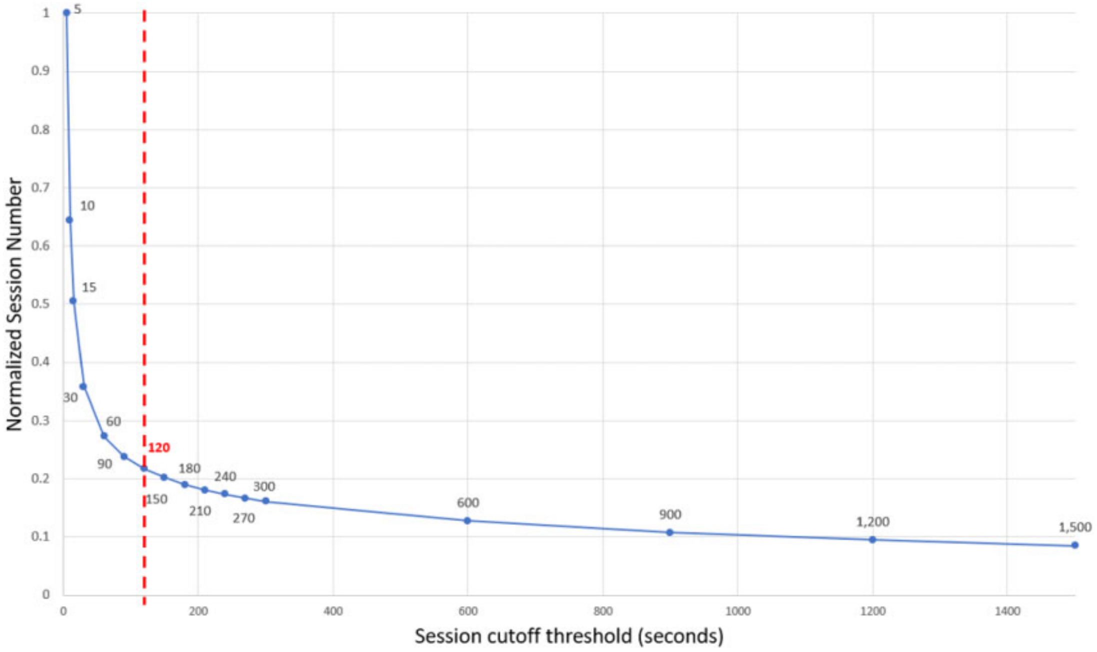

Figure 2. The data-driven detection of a stable sessionization threshold, by finding the point of maximum curvature in the normalized session number curve. This curve was derived through testing 17 thresholds, which yielded a diminishing number of sessions, each represented by a labeled point ranging from 5 to 1500 seconds. Here, the selected sessionization time threshold is highlighted in red, being 120 seconds, and was the point of maximum curvature estimated through the Kneedle algorithm.

the number of clinicians per task are non-Gaussian, we use the Mann-Whitney *U* test, which is a nonparametric test for non-Gaussian distributions, and Bonferroni corrections at a significance level of .05.[26]

Our analyses were performed using Python 3.7. Libraries used include scikit-learn, scipy, pandas, numpy, anndata, scanpy, kneedle, and pegasuspy.[20,27] Further documentation of our methods can be found o (https://github.com/bobchen1701/Single_Session_Analysis).

RESULTS

NICU case study

We introduce a case study on the management of patients, by nurses, who received surgeries and had stayed in the NICU. We studied 3 months of EHR audit logs generated by 33 NICU nurses. The total number of events, sessions, and tasks are 1 130 589, 57 234, and 81, respectively. We used a time interval threshold of 120 seconds for sessionization; this was the point of maximum curvature, or "knee point," determined through the Kneedle algorithm for the normalized curve generated from 17 tested thresholds, ranging from 5 to 1500 seconds (Figure 2; Supplementary Table A3). In the audit logs, we detected 326 uniquely performed events. The resulting 57 234 sessions were treated as observations defined by these 326 unique events.

We calculated a Pearson's correlation coefficient between the session-ending event distribution (frequency of an event serving as endings of sessions) and global event distribution (frequency of an event in all sessions). We observed an r value of 0.96 with a *P* value <.0001, which indicated that the session-ending event distribution matched global event distributions with no skew introduced through sessionization.

We selected 50 principal components, which covered over 90% of the data's variation. These components were used to initialize a t-SNE embedding with the following parameters: perplexity = square root of the number of total observations, early exaggeration = 12, and learning_rate = 1000. We used the Leiden community detection algorithm with a resolution of 0.1 and the number of nearest neighbors set to the square root of the number of total observations.

Session- and task-level metrics describe NICU inpatient EHR activities

Figure 3 shows the 81 tasks that we learned for NICU nurses (Figure 3A) and the distribution of multiple, session-level metrics: time to perform a session (session_time, Figure 3B), repetitiveness in events (session_repetition_coef, Figure 3C), and diversity in events (session_diversity_coef, Figure 3D) listed in Table 3. As shown in Figures 3B and 3C, sessions on the right-hand side of the embedding exhibit higher repetition coefficients (Figure 3C) and also take more time to complete (Figure 3B). As shown in Figure 3D, sessions with high diversity in events are distributed across the whole embedding.

Figures 3E to 3H show the distributions of abstracted task-level metrics (described in Table 3): time to perform a task (task_session_time) (Figure 3F), repetitiveness in events (task_session_repetition_coef) (Figure 3G), and diversity in events (task_session_diversity_coef) (Figure 3H). Figure 3E shows the number of unique clinicians associated with each task, most of which have a large number of clinicians (25-30 NICU nurses) involved. There are some tasks at the bottom of the embedding that have only a small number of nurses involved, which are introduced in the following section.

NICU nurse EHR tasks can be divided into 4 complexity profiles

Figures 4 and 5 show the distributions of tasks across 4 complexity profiles. We discover 28, 27, 14, and 12 HDHR, LDLR, LDHR, and HDLR tasks, respectively. The number of HDHR and LDLR tasks is much larger than the LDHR and HDLR tasks. HDHR tasks have 34 403 sessions in total (see Supplementary Table A4), consisting of

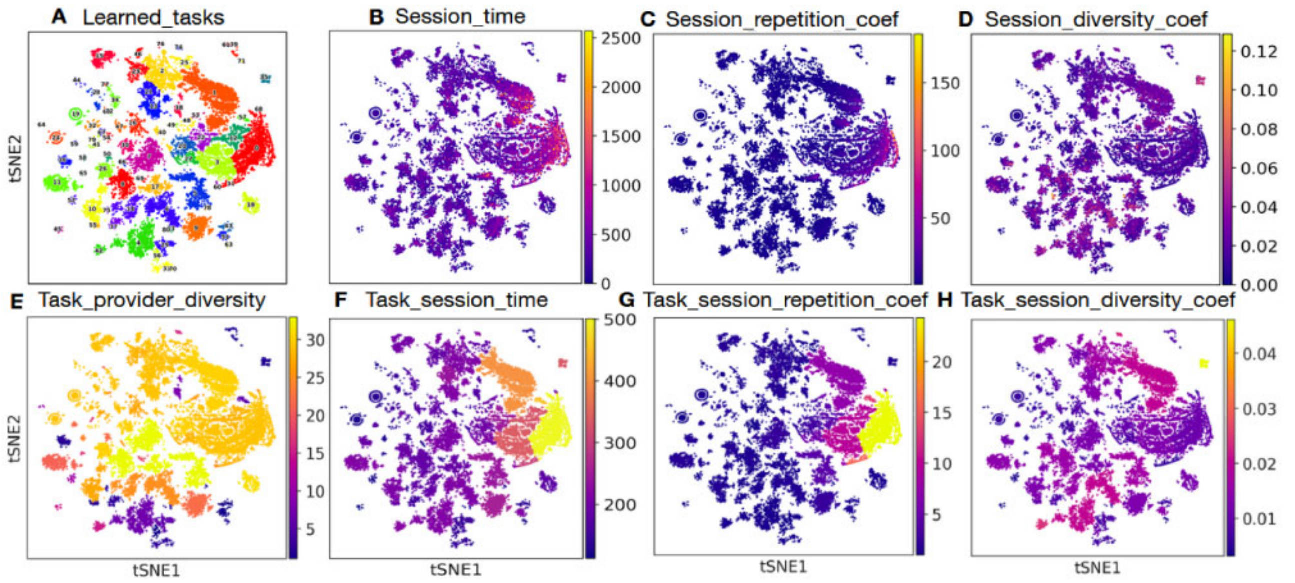

Figure 3. (A) The 81 learned tasks for neonatal intensive care unit nurses are visualized using t-distributed stochastic neighbor embedding (t-SNE). Tasks are separated by colors and labeled by task IDs. Each point is a session. (B-D) Distributions of the 3 session-level metrics (duration, repetitiveness, and diversity) across individual sessions. (B) The ***session_time*** is measured as seconds. (C) The ***session_repetition_coef*** is calculated using total events per session/total unique events per that same session. (D) The ***session_diversity_coef*** is measured as the number of unique events in a session/total number of events in the system. (E-H) Distributions of the 4 task-level metrics (number of unique clinicians involved, duration, repetition, and diversities in events). (F) The ***task_session_time*** is measured as the mean of ***session_time*** of all sessions within a task. (G) The ***task_session_repetition_coef*** is measured as the mean of ***session_repetition_coef*** of all sessions with a task. (H) The ***task_session_diversity_coef*** is calculated as the mean of ***session_diversity_coef*** of all sessions in a task. (E) The ***task_provider_diversity*** is measured as the number of clinicians observed performing sessions associated with a particular task.

60.1% of the total sessions. This suggests that most of the activities performed by NICU nurses are highly complex in the context of event diversity and repetition. The number of NICU nurses involved in HDHR and LDLR are similar (27 vs 26, as shown in Supplementary Table A5). Many more NICU nurses are involved in HDHR and LDLR tasks than those involved in HDLR and LDHR. The mean session times are 292, 144, 185, and 190 seconds in the HDHR, LDLR, LDHR, and HDLR profiles, respectively. The descriptive statistics of each complexity profile, in terms of session times and the clinician prevalence, are shown in Supplementary Tables A4 and A5.

Pairwise statistical testing suggests significant differences in task duration but not in clinician prevalence per task complexity profile

The non-Gaussian distributions of session times and unique clinician numbers per task are depicted in Supplementary Figure A1. We performed pairwise Mann-Whitney *U* tests to examine the statistical differences between our 4 task complexity profiles. With the exception of the LDHR and HDLR profile pair, all 5 of the other profile pairs had significant differences in the duration at the corrected significance level of 0.008333, with a *P* value <.0001. The results are shown in Table 4. We confirmed that there is no significant difference (with a *P* value >.008333) in clinician prevalence between task groups, with the exception of the HDHR and LDLR profile pair. The test results are shown in Table 4.

Pairwise statistical testing suggests significant differences in "exporting" and "system" event access types but not in "view" and "modify" per task complexity profile

The non-Gaussian distributions of proportions of event access types per task are depicted in Supplementary Figure A2. Generally, "view" and "modify" were the 2 most common event access types (Supplementary Table A6), which encompassed common activities involving flowsheet interactions (eg, flowsheets activity opened, flowsheets data saved). With the exception of the HDHR and LDHR profile pair, none of the other 5 profile pairs exhibited statistically significant differences in the "view" events. From Table 4, it can be seen that no significant differences were identified for any of the 6 profile pairs in the "modifying" events. Further disparities involved significant differences in "exporting" and "system" event access types, primarily between task profiles with different diversity rankings (HD vs LD in contrast with HDHR vs HDLR or LDHR vs LDLR); though these event access types were far less common than "view" and "modify," these results suggest that our diversity stratification of task profiles may be useful in differentiating tasks with respect to their functional niches in a clinical workflow.

Process mining and reviewing clinical workflow annotations suggest clinically meaningful task learning

By illustrating these 4 task complexity profiles, we further investigate tasks within each of these categories. We randomly selected 1 task from each of the 4 profiles to visualize as a process tree (see Supplementary Figures A3-A6). To investigate the clinical workflows involved, we randomly selected a session from each of these tasks for the expert review of their clinical workflow annotations (listed in Supplementary Tables A8-A11). We include definitions of terms used in these reviews in Supplementary Tables A1 and A2. These workflow annotations, derived from task sessions, suggest that clinically meaningful processes can be learned through our methods. The annotations of 4 task sessions are described in Supplementary Table A7. Notably, a comparison of the annotations between an HDHR session (Supplementary Table A8) and an LDLR session (Supplementary Table A11) suggests that the patient associ-

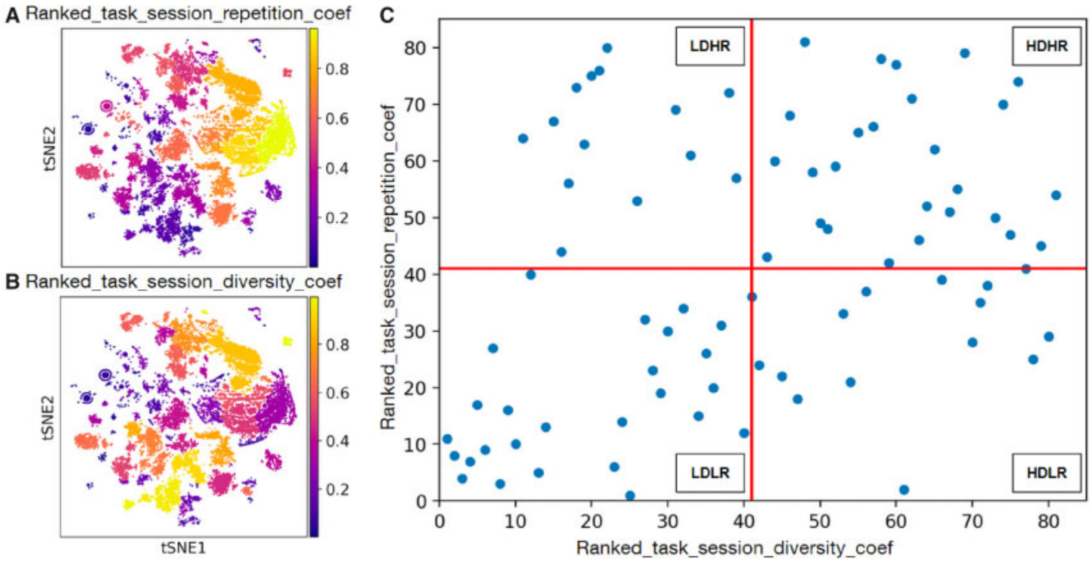

Figure 4. Task complexity profiles. Tasks are categorized into 4 complexity profiles based on their diversity and repetition coefficients. The 4 complexity profiles are high diversity/high repetition (HDHR), low diversity/low repetition (LDLR), high diversity/low repetition (HDLR), and low diversity/high repetition (LDHR). t-SNE: t-distributed stochastic neighbor embedding.

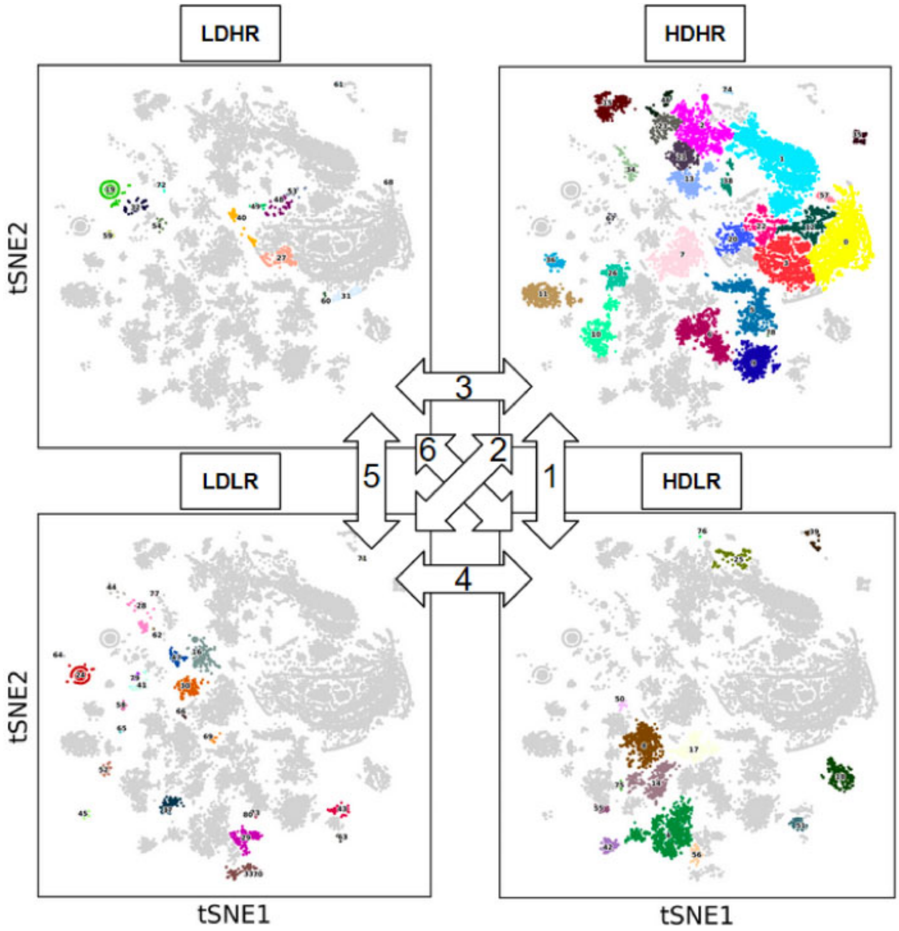

Figure 5. Tasks across the 4 complexity profiles. Each embedding highlights the tasks associated with a particular task complexity profile. Also, the 6 pairs between the 4 complexity profiles are depicted. HDHR: high diversity/high repetition; HDLR: high diversity/low repetition; LDHR: low diversity/high repetition; LDLR: low diversity/low repetition.

Table 4. The statistical significance of differences in the duration, the proportion of event access types, and the number of clinicians between pairs of tasks with different complexity levels

Pairwise Test of Differences	Individual Session Time		Clinician Prevalence per Task		Proportion of View Events per Task		Proportion of Modify Events per Task		Proportion of Export Events per Task		Proportion of System Events per Task	
(α = 0.0083)	MWU	*P* Value	MWU	*P* Value	MWU	*P* Value	MWU	*P* Value	MWU	*P* Value	MWU	*P* Value
1. HDHR vs HDLR	1.031×10^8	<.0001*	137.5	.785	182	.4291	172	.3250	153	.1632	166	.2680
2. HDHR vs LDLR	4.232×10^7	<.0001*	211	.006*	292	.1489	324	.3186	197	.0019*	124	2.377×10^{-5}*
3. HDHR vs LDHR	4.854×10^7	<.0001*	167.5	.278	87	.0026*	108	.0134	80	.0009*	40	2.132×10^{-5}*
4. HDLR vs LDLR	2.29×10^7	<.0001*	173.5	.40	159	.2617	175	.4269	79.5	.0007*	79	.0014*
5. LDHR vs LDLR	1.073×10^7	<.0001*	106.5	.016	119	.0381	130	.0719	168.5	.3139	168	.3394
6. LDHR vs HDLR	2.338×10^7	.0143	7.5	.105	47	.0101	55	.0254	34.5	.0009*	27.5	.0005*

HDHR: high diversity/high repetition; HDLR: high diversity/low repetition; LDHR: low diversity/high repetition; LDLR: low diversity/low repetition; MWU: Mann-Whitney *U*.

*Indicates the difference is significant.

ated with the HDHR session had significantly more complex clinical interactions than the patient associated with the LDLR session (Supplementary Table A7).

DISCUSSION

EHR utilization is one of the contributory factors for clinician workload, stress, and burnout.[1,28] Studies to date have created metrics to measure the amount of time spent by clinicians in EHRs and the tasks that clinicians perform while interacting with EHR systems, primarily in outpatient and ambulatory settings. Yet few studies have focused on inpatient settings or have developed metrics to model task complexity in EHR utilization.[29–31] This study created an unsupervised learning framework to identify clinician tasks performed in EHRs and developed hierarchical metrics to describe EHR task complexity, which is lacking in the existing literature. We tested the effectiveness of our metrics and approach in learning EHR tasks and task complexities for nurses in the NICU. Our hierarchical metrics capture contextual information of a task beyond its explicit, session-level content. For instance, we confirmed that clinicians require much more time to perform HDHR tasks. In our case study, over 60% of sessions exhibited this complexity profile. We further demonstrated that these tasks were associated with complex patient health conditions (patient 1 in Supplementary Table A7). Ultimately, by applying the approaches described in this study, investigators can better understand clinician activities in EHRs with reduced manual effort with the utilization of machine learning.

The scope of this study and its limitations

Our project is a pilot study, and there are limitations that we want to acknowledge as guidelines for further studies on the EHR audit logs.

First, an EHR task performed by a clinician is determined by the setting, the clinician's role, and the patients managed. Thus, the learning of tasks should be fixed to a specific setting and specific healthcare professional. Our case study was limited to NICU nurses and their management of patients undergoing surgery; however, our approach is generalizable to other settings and clinician roles if adjustments are made to the sessionization and event annotation procedures. Similarly, task validation strategies will differ depending on the target clinician role, clinical setting, and EHR system used. This case study focuses on task complexity metrics, rather than on the content of tasks, and further validation, such as a formalized qualitative study, examining each of the 81 identified tasks is necessary to fully realize task function in a clinical workflow.

Second, some Epic terms do not have one-to-one mappings to generic terms that are utilized by non-Epic systems. The mapping of system-specific terms is an open question in EHR interoperability, but we note that specifications detailed in ASTM E2147-18,[32] which was the basis of this case study's generic term mapping, may be helpful in this type of mapping. Further limitations specific to our case study's Epic EHR originate from the limited granularity of certain logged actions. For instance, the event "A PRINT GROUP BASED REPORT (LRP) IS VIEWED" is common in our case study's example sessions and reflects users viewing consolidated patient information reports in a number of potential locations, such as summary reports, patient lists reports, or snapshots in the EHR. This event lacks the specificity needed to describe precise clinician actions, but highlights critical areas that audit logging systems can be improved.

Third, we note that flowsheet data in our case study are limited to manually entered vital signs and custom structured data, owing to the EHR system used. Our results show that EHR tasks involve nursing activities related to flowsheet data entry, saving, and viewing (Supplementary Table A13), but these activities do not use streams of mapped data from monitors, infusion pumps, ventilators, or other medical devices. Our analysis would benefit from the integration of more granular and diverse sources of human-computer interaction data.

Fourth, it should be recognized that clinicians may operate in care teams, performing EHR tasks simultaneously. In this study, we only focused on the learning of nonsimultaneous, individual clinician tasks. This aspect of care team–oriented tasks and the development of corresponding metrics is a natural next step for this line of research.

Last, we acknowledge that there may exist clinical events that are not captured by the EHR audit log, occurring within intervals between our defined sessions or wholly undocumented in our EHR system. To address this, we show that the majority of intersession intervals are short, which reduces the windows of time that potentially contain unrecorded clinician activities (Supplementary Figure A11 and Supplementary Table A12). Still, our event duration calculation assumes there are no time gaps between sequential actions performed within the same session, which may not be true in clinical practice. We refer to a study conducted at Vanderbilt University Medical Center[33] investigating the capability of audit logs in capturing clinician activities. This study shows that the log-generated breadcrumbs encounter summary can capture all interactions documented in clinical notes, with the exception of physical exams. Based on their observations, we assume that audit logs can provide functionally accurate representations of clinician activities in EHR systems.

CONCLUSION

EHR audit logs are a rich resource that can be leveraged to understand clinician EHR activities as well as their efficient performance. Such a resource could potentially be used to model clinician workload, stress, and burnout. Without efficient informatics tools, it is difficult to explore audit logs and mine EHR usage patterns. We developed unsupervised learning approaches and applied them to EHR audit logs to learn tasks and develop metrics describing their complexities. We designed multiple hypothesis tests based on these novel metrics to evaluate task efficiency, and we tested the effectiveness of our framework for specific healthcare professionals—NICU nurses—in the management of patients receiving surgeries, staying in the NICU. The results of the case study show that our approaches and metrics can identify complex and time-consuming tasks for future optimizations.

FUNDING

This research was supported, in part, by the National Library of Medicine of the National Institutes of Health under Award Number R01LM012854 and T32LM012412.

AUTHOR CONTRIBUTIONS

BC and YC conceived the presented idea and performed the data collection and analysis, method and metric design and development, experiment design, evaluation and interpretation of the experiments, and writing of the manuscript. WA, CG, BJ, LN, NL, DF, and BM performed evaluation and interpretation of the experiments, and editing of the manuscript.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

DATA AVAILABILITY

The datasets generated and/or analyzed during the current study are not publicly available due to patient and provider private information investigated but are available from the corresponding author on reasonable request.

ACKNOWLEDGMENTS

We thank Carrie Reale, a board-certified nurse informaticist, for providing supervision in evaluating and annotating sessions and tasks learned from the NICU setting. We also thank all researchers from the EHR audit log National Research Network led by Dr Julia Adler-Milstein and Task Type Workgroup led by Dr You Chen for providing constructive directions on the development of the unsupervised learning framework. Finally, we thank Kirill Mangutov for helpful discussions regarding human-computer interactions.

CONFLICT OF INTEREST STATEMENT

The authors have no competing interests to declare.

REFERENCES

1. Kroth PJ, Morioka-Douglas N, Veres S, *et al.* Association of electronic health record design and use factors with clinician stress and burnout. *JAMA Netw Open* 2019; 2 (8): e199609.
2. Adler-Milstein J, Huckman RS. The impact of electronic health record use on physician productivity. *Am J Manag Care* 2013; 19(10 Spec No): SP345–52.
3. Chen L, Guo U, Illipparambil LC, Netherton MD, *et al.* Racing against the clock: internal medicine residents' time spent on electronic health records. *J Grad Med Educ* 2016; 8 (1): 39–44.
4. Babbott S, Manwell LB, Brown R, *et al.* Electronic medical records and physician stress in primary care: results from the MEMO Study. *J Am Med Inform Assoc* 2014; 21 (e1): e100–6.
5. Card AJ. Physician burnout: resilience training is only part of the solution. *Ann Fam Med* 2018; 16 (3): 267–70.
6. Robertson SL, Robinson MD, Reid A. Electronic health record effects on work-life balance and burnout within the I3 population collaborative. *J Grad Med Educ* 2017; 9 (4): 479–84.
7. Overhage JM, McCallie D. Jr., Physician time spent using the electronic health record during outpatient encounters: a descriptive study. *Ann Intern Med* 2020; 172 (3): 169–74.
8. 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): 480–90.
9. Cohen GR, Friedman CP, Ryan AM, *et al.* Variation in physicians' electronic health record documentation and potential patient harm from that variation. *J Gen Intern Med* 2019; 34 (11): 2355–67.
10. Monsen CB, Singh A, Fellner J, Jackson SL. Measuring provider efficiency in Epic: a preliminary mixed-methods exploration of the provider efficiency profile. *J Gen Intern Med* 2018; 33: S253.
11. Arndt BG, Beasley JW, Watkinson MD, *et al.* Tethered to the EHR: primary care physician workload assessment using EHR event log data and time-motion observations. *Ann Fam Med* 2017; 15 (5): 419–26.
12. 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): 639–43.
13. Adler-Milstein J, Adelman JS, Tai-Seale M, *et al.* EHR audit logs: A new goldmine for health services research? *J Biomed Inform* 2020; 101: 103343.
14. Chen Y, Xie W, Gunter CA, *et al.* Inferring clinical workflow efficiency via electronic medical record utilization. *AMIA Annu Symp Proc* 2015; 2015: 416–25.
15. Chen Y, Lorenzi N, Nyemba S, *et al.* We work with them? healthcare workers interpretation of organizational relations mined from electronic health records. *Int J Med Inform* 2014; 83 (7): 495–506.
16. Chen Y, Patel MB, McNaughton CD, *et al.* Interaction patterns of trauma providers are associated with length of stay. *J Am Med Inform Assoc* 2018; 25 (7): 790–9.
17. Chen Y, Yan C, Patel MB. Network analysis subtleties in ICU structures and outcomes. *Am J Respir Crit Care Med* 2020; 202 (11): 1606–7.
18. Chen Y, Lehmann CU, Hatch LD, *et al.* Modeling care team structures in the neonatal intensive care unit through network analysis of EHR audit logs. *Methods Inf Med* 2019; 58 (4/5): 109–23.
19. Arlitt M. Characterizing web user sessions. *SIGMETRICS Perform Eval Rev* 2000; 28 (2): 50–63.
20. Satopää V, Albrecht J, Irwin D, Raghavan B. Finding a "kneedle" in a haystack: Detecting knee points in system behavior. In: *Proceedings of the 31st International Conference on Distributed Computing Systems Workshops*; 2011: 166–71.
21. Ramos J. Using TF-IDF to determine word relevance in document queries. In: *Proceedings of the 20th International Conference on Machine Learning*; 2003: 133–42.
22. Kim KI, Jung K, Kim HJ. Face recognition using kernel principal component analysis. *IEEE Signal Process Lett* 2002; 9 (2): 40–2.
23. Belkina AC, Ciccolella CO, Anno R, *et al.* Automated optimized parameters for T-distributed stochastic neighbor embedding improve visualization and analysis of large datasets. *Nat Commun* 2019; 10 (1): 5415.
24. Traag VA, Waltman L, van Eck NJ. From Louvain to Leiden: guaranteeing well-connected communities. *Sci Rep* 2019; 9 (1): 5233.
25. Rozinat A, van der Aalst WM. Decision mining in ProM. In: *Proceedings of the International Conference on Business Management*; 2006; 4102: 420–5.
26. Ruxton GD. The unequal variance t-test is an underused alternative to Student's t-test and the Mann–Whitney U test. *Behav Ecol* 2006; 17 (4): 688–90.
27. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol* 2018; 19 (1): 15.
28. Tran B, Lenhart A, Ross R, Dorr DA. Burnout and EHR use among academic primary care physicians with varied clinical workloads. *AMIA Jt Summits Transl Sci Proc* 2019; 2019: 136–44.
29. Varela LO, Wiebe N, Niven DJ, *et al.* Evaluation of interventions to improve electronic health record documentation within the inpatient setting: a protocol for a systematic review. *Syst. Rev* 2019; 8 (1): 54.
30. Ozkaynak M, Reeder B, Hoffecker L, *et al.* Use of electronic health records by nurses for symptom management in inpatient settings: a systematic review. *Comput Inform Nurs* 2017; 35 (9): 465–72.
31. Cruz-Correia R, Boldt I, Lapão L, *et al.* Analysis of the quality of hospital information systems audit trails. *BMC Med Inform Decis Mak* 2013; 13 (1): 84.
32. ASTM E2147-18. Standard Specification for Audit and Disclosure Logs for Use in Health Information Systems. West Conshohocken, PA: ASTM International; 2018.
33. Tang LA, Johnson KB, Kumah-Crystal YA. Breadcrumbs: assessing the feasibility of automating provider documentation using electronic health record activity. *AMIA Annu Symp Proc* 2018; 2018: 1008–17.