Journal of Biomedical Informatics 101 (2020) 103312

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Using the PARAFAC2 tensor factorization on EHR audit data to understand PCP desktop work

Ioakeim Perros[a,*,1], Xiaowei Yan[b], J.B. Jones[b], Jimeng Sun[a], Walter F. Stewart[c]

[a] Georgia Institute of Technology, Atlanta, GA, United States
[b] Research Development & Dissemination, Sutter Health, Walnut Creek, CA, United States
[c] HINT Consultants, Orinda, CA, United States

ARTICLE INFO

Keywords:
Workflow analysis
Electronic health records

ABSTRACT

Background: Activity or audit log data are required for EHR privacy and security management but may also be useful for understanding desktop workflow.
Objective: We determined if the EHR audit log file, a rich source of complex time-stamped data on desktop activities, could be processed to derive primary care provider (PCP) level workflow measures.
Methods: We analyzed audit log data on 876 PCPs across 17,455 ambulatory care encounters that generated 578,394 time-stamped records. Each individual record represents a user interaction (e.g., point and click) that reflects all or part of a specific activity (e.g., order entry access). No dictionary exists to define how to combine clusters of sequential audit log records to represent identifiable PCP tasks. We determined if PARAFAC2 tensor factorization could: (1) learn to identify audit log record clusters that specifically represent defined PCP tasks; and (2) identify variation in how tasks are completed without the need for ground-truth labels. To interpret the result, we used the following PARAFAC2 factors: a matrix representing the task definitions and a matrix containing the frequency measure of each task for each encounter.
Results: PARAFAC2 automatically identified 4 clusters of audit log records that represent 4 common clinical encounter tasks: (1) medications' access, (2) notes' access, (3) order entry access, and (4) diagnosis modification. PARAFAC2 also identified the most common variants in how PCPs accomplish these tasks. It discovered variation in how the notes' access task was done, including identification of 9 distinct variants of notes access that explained 77% of the input data variation for notes. The discovered variants mapped to two known workflows for notes' access and to two distinct PCP user groups who accessed notes by either using the Visit Navigator or the Wrap-Up option.
Conclusions: Our results demonstrate that EHR audit log data can be rapidly processed to create higher-level constructed features that represent time-stamped PCP tasks.

1. Introduction

Health care structure (e.g., facilities, staff, financing) defines the context for care processes. In ambulatory care, in particular, an increasing share of provider and staff time are spent at the desktop [1]. Log file data (also known as audit logs) may be used to reveal how desktop work is done and the time required to do this work [2–4]. Log files, however, are voluminous and difficult to decipher. We determined whether tensor factorization methods could be used to cluster log file records into groups that represent distinct PCP tasks. If this were possible, then these records could be used beyond workflow analysis to better understand how work is done, to facilitate efficiency improvement efforts, and to learn best practices for use of the EHR.

EHRs automatically generate time-stamped audit logs that track all interactions that a user has with a patient's record, as well as, other desktop activities (e.g., inbox). While these data are created for privacy and security management they also document how users do their desktop work when using the HER [5]. But, audit log data are inordinately complex and are not organized into self-identifiable clusters that represent defined jobs or tasks. In this context, we use the term "job" as defined by Ulwick [6] to refer to the underlying goal of a sequence of actions being taken or tasks being completed [6]. For example, making a medication order in the EHR may be a task associated with the overall "job" of managing a patient's medication use for a

Abbreviations: EHR, electronic health record; PCP, primary care provider

* Corresponding author.

[1] Part of this work was conducted during an internship at Sutter Health.

https://doi.org/10.1016/j.jbi.2019.103312
Received 13 March 2019; Received in revised form 8 October 2019; Accepted 13 October 2019
Available online 15 October 2019
1532-0464/

specific condition with the intention of optimizing outcomes. The log file documents the electronic action (making the order), whereas the job is inferred from an overall ensemble of related actions (e.g., open the patient's medication list, search the medication database for the appropriate medication, select the dose/quantity/pharmacy) that may be summed and expressed as a series of tasks. If tasks can be readily identified, log files could be used to understand how desktop work is done, including the time required to do specific desktop job (i.e. make diagnosis, medication management, procedure order, progress note).

We determined if machine-learning methods could be used to interpret audit log data and rapidly identify the most common types of clinical desktop tasks and to determine variation in how these tasks are done.

2. Materials and methods

2.1. Overview

In this study, we used Epic audit log records from face-to-face primary care encounters to answer the following questions about what PCPs do during encounters [7–9]: (1) Can raw time-stamped audit log records be organized into clusters of records that represent intuitively meaningful representations of tasks that PCPs do? (2) Can audit log records be automatically organized and interpreted to determine variation in how a specific task is done? For the latter, we specifically focused on progress notes because this work is time-consuming and done for every encounter. Moreover, it is known that there are two dominant ways to use the Epic EHR built-in interface features for this purpose. That is, clinical observation has confirmed that are two distinct methods by which progress notes are accessed and documented. We wanted to determine if machine learning could detect these two distinct methods and the ensemble of related tasks and also cluster PCPs by how progress notes were completed. To this end, the methods section describes the source data, pre-processing of the source data for use in modeling, and then the tensor factorization methods used to summarize audit log records.

2.2. Source of data & data extraction

Audit log data were extracted from Sutter Health's Epic EHR data sources. Sutter Health is a large not-for-profit health system that serves over 100 communities in northern California. Sutter uses a single instance of the Epic Systems Corporation EHR.

Audit log files are very large even for a small number of encounters. We randomly selected two working days in 2016 (Epic 2015) and extracted log records for ambulatory care face-to-face encounters completed by 876 Primary Care Providers (PCPs). PCPs include physician, nurse practitioner, and physician assistants. The two days of log files resulted in 578,394 time-stamped audit log records across 17,455 encounters. Each entry in the audit log is time-stamped at a resolution of one-second, where two or more records occurring less than one second apart will have the same time-stamp, but these records are sequentially ordered as they occur. We restricted analysis to PCP face-to-face encounters that had at least two time-stamped audit log records. This resulted in 16,613 input encounters for PARAFAC2. We describe data pre-processing methods followed by the application of tensor factorization.

2.3. Preprocessing of audit log data

We are interested in using tensor factorization to determine if a schema could be created where clusters of log records could be automatically identified that correspond to discrete tasks. To this end, log data were pre-processed by ordering them within provider, by date, by encounter within date, and last, by the audit log record time stamp and sequence indicator. Encounters that contained only one log record were removed. For those encounters, the most frequently appeared log records are: "In Basket message viewed": 40%, "Printing": 26%, "Patient Appointment Demographics form accessed": 12%, "Patient Station accessed": 8%. Fig. 1 describes an audit log sequence that begins with opening of the encounter visit navigator, where multiple log records with the same timestamp means that they occurred during the same second.

Log records were further pre-processed to disassemble the contents of an encounter (sequence of log records) into smaller log record sequences that may reflect common tasks or parts of tasks done by PCPs. We refer to these discrete log record sequences as "potential tasks" that are defined as follows: in the absence of an explicit signal indicating the end of a task, we assume that sequential records that occurred on consecutive seconds (e.g., interval 00:04–00:06 in Fig. 2) and that are separated by a break of no activity (i.e., > 1 s) either before or after the sequence are considered to be part of the same potential task, but not necessarily representing a complete task.

Log records co-occurring with the last timestamp of a potential task indicate actions which occupy the user's attention (e.g., reviewing data retrieved when an EHR features is clicked), or that are consistently followed by a separate user activity (e.g., talking to the patient) that does not involve desktop interactions. Accordingly, we define two variants of audit log records which we call Intermediate (occurring before the last time stamp) and Terminal (co-occurring with the final timestamp before a > 1 s gap between potential tasks) features. We annotate all records occurring during the n-th timestamp as Terminal features and records occurring during the timestamps $1, \cdots, n-1$ as Intermediate features. Also, duplicate features within the same potential task are removed. Such duplicates may arise when we create the multi-hot vector encodings of potential tasks as in Fig. 2, when there may be multiple occurrences of a certain feature (e.g., I Visit Navigator template loaded).

In sum, we transform each k-th ambulatory care encounter to a binary matrix $\mathbf{X}_k$ of size $I_K \times J$, where I_k corresponds to the number of potential tasks for the k-th encounter and J denotes the total number of features derived from raw audit log records (Fig. 3). Importantly, the number of potential tasks per encounter is not known in advance. Finally, in the absence of labeled data, we do not know in advance

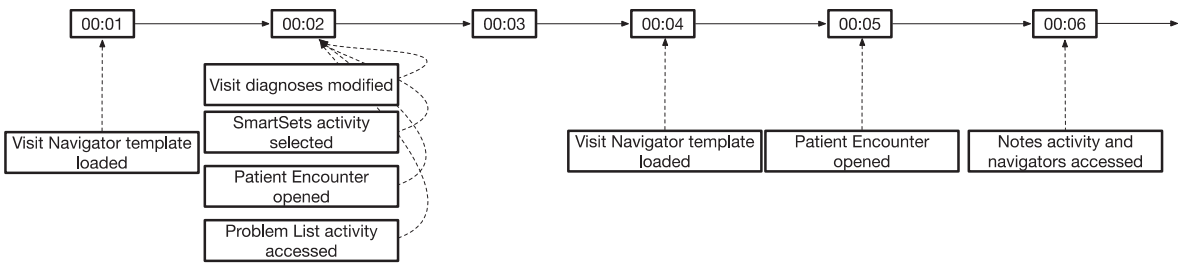

Fig. 1. Example of EHR audit log records corresponding to the opening of the patient record at the start of an encounter. Note that there may be multiple log records with the same timestamp. Timestamps with no records are indicators of no computer related interactions (e.g., doctor is viewing data on screen, talking to patient, etc.).

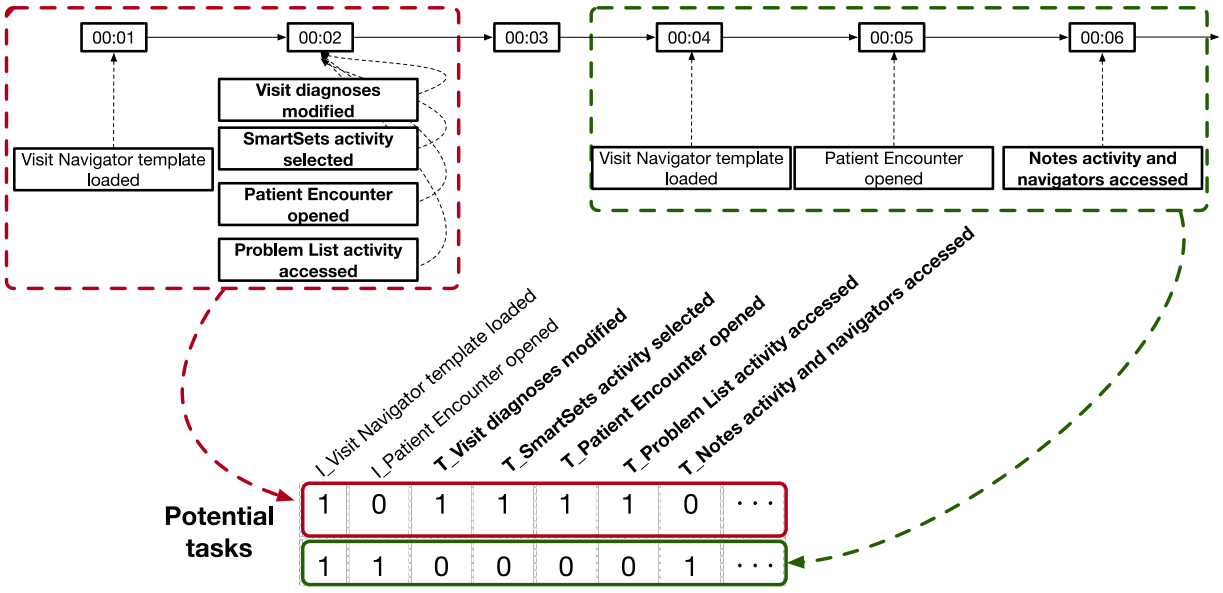

Fig. 2. Transformation of face-to-face encounter audit log records data into a binary matrix. Each row of the matrix is a multi-hot vector encoding of the feature assembly representing a potential PCP task; the row number reflects the chronological task ordering during a single encounter. Each one of the columns corresponds to a feature defined as a variant of a certain audit log record: if an audit log record occurs or co-occurs with the last timestamp of a potential task, then this is marked as **T**erminal. Otherwise, it is marked as **I**ntermediate. Terminal features are in bold font.

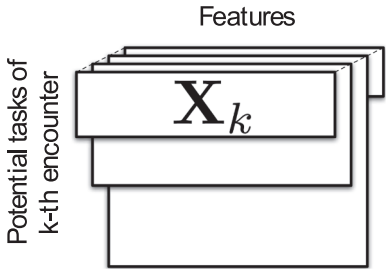

Fig. 3. A visual representation of the collection of matrices created through from pre-processing to organize raw log records into potential PCP tasks. Each matrix corresponds to a single encounter. The rows correspond to potential PCP tasks recorded (which can vary across encounters) and the columns to the features constructed from log records.

whether a group of log records represents a systematic way of accomplishing a PCP task. PARAFAC2 tensor factorization [7] was used to address this challenge.

2.4. PARAFAC2 factorization: variation in how PCPs complete tasks

We determined whether the PARAFAC2 factorization [7] could identify tasks and variation in how tasks are done. We selected the PARAFAC2 framework (i.e., model & factorization algorithm) because it can effectively account for the folded structure of our intermediate data. That is, our raw input data is from a set of PCP encounters, where the audit data from each represents a sequence of records corresponding to underlying potential tasks, each of which can ultimately be represented by an ensemble of features (i.e., audit data entries). Such a folded structure can be concisely modeled by a multi-dimensional data array, containing a matrix of sequenced potential tasks X features for each one of the PCP encounters; such an array fits the expected input for the PARAFAC2 factorization. PARAFAC2 also offers an easily-interpretable approach to dimensionality reduction, as noted below. Notably, in contrast to other simpler multi-dimensional dimensionality reduction approaches (e.g., the CP/PARAFAC decomposition [10]), PARAFAC2 does not require any ad-hoc alignment of tasks among encounters. This is crucial because it is unrealistic to assume that all encounters will share the same number of potential tasks. PARAFAC2 also preserves model uniqueness under mild assumptions, boosting reliability of model interpretation (we elaborate on that below) and also it natively handles input data sparsity, which is a by-product of the fact that a few features will be observed for every potential task. Finally, specialized algorithms have recently been developed [9] scaling up PARAFAC2 for sparse input data; the use of scalable frameworks is crucial in order to cope with the voluminous nature of log files.

The input data for this decomposition (Fig. 4) are organized as a

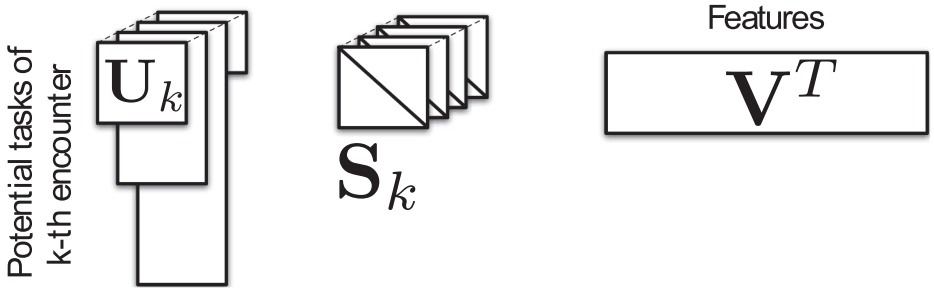

Fig. 4. PARAFAC2 factorization given an input collection of matrices (as in Fig. 3). The chosen model can handle a varying number of potential tasks across different encounters. The matrix contains the most prevalent variation patterns of the most common PCP tasks.

binary multi-way array (tensor) of size $I_k \times J \times K$, where K is the number of encounters. PARAFAC2 approximates the input $\{\mathbf{X}_k \in R^{I_k \times J}\}$ as:

$$\mathbf{X}_k \approx \mathbf{U}_k \mathbf{S}_k \mathbf{V}^T$$

where $k = 1, ...,K$, $\mathbf{U}_k \in R^{I_k \times R}$, $\mathbf{S}_k \in R^{R \times R}$ is diagonal and $\mathbf{V} \in R^{J \times R}$, $\mathbf{V}^T$ is the transpose of $\mathbf{V}$. R is a user-defined parameter which denotes the number of tasks extracted from the model. We exploit the following factors of the PARAFAC2 model: *a)* a matrix revealing the most representative variations of PCP task definitions across all encounters ($\mathbf{V}$ matrix in the model), and *b)* a matrix containing the frequency of each task extracted for each k-th encounter ($\mathbf{S_k}$ matrix in the model). The model can be extended so that non-negative constraints are imposed on $\{\mathbf{S}_k\}$, $\mathbf{V}$ factors [11], a constraint we employ to enhance interpretability.

2.5. Model and algorithm

PARAFAC2 preserves uniqueness of the model output by imposing the constraint that the cross product $\mathbf{U}_k^T\mathbf{U}_k$ is constant regardless of which k is involved [7]. This implies that the correlations between the tasks extracted are constant for all the encounters [8]. Preserving uniqueness means that the model is pursuing the *actual latent factors,* rather than an arbitrary rotation of them, which boosts the reliability of the model interpretation [12].

The state-of-the-art framework for fitting the PARAFAC2 model uses Alternating Least Squares (ALS) [8]. This corresponds to optimizing a least-squares criterion, in an alternating fashion, where we solve for a subset of factors by fixing all others. Our input data are inherently sparse, meaning that there are a few non-zero elements recorded as compared to the total size of the input matrices. Thus, we use the algorithm proposed by Perros et al., which also follows the ALS framework, but is optimized for sparse datasets [9,13]. See Appendix A for the details of the algorithm.

2.6. Automatic determination of the number of task definitions extracted

In practice, the user-defined parameter R, which denotes the number of tasks extracted, is unknown. In this Section, we describe how to extend [9] to automatically determine R [14]. In particular, we summarize the two main steps of the PARAFAC2-ALS algorithm [8] and illustrate how an intermediate result of the second step is exploited to assess the validity of the R-dimensional model solution.

As noted above, $\{\mathbf{X}_k \in R^{I_k \times J}\}$ is the input collection of matrices and $\mathbf{U}_k \in R^{I_k \times R}$, $\{\mathbf{S}_k \in R^{R \times R}\}$ (restricted to be diagonal) and $\mathbf{V} \in R^{J \times R}$ are the output model factors. To preserve uniqueness, Harshman imposed the constraint that $\mathbf{U}_k^T\mathbf{U}_k$ is constant regardless of which k is involved [7]. For this constraint to hold, each $\mathbf{U}_k$ factor is decomposed as: $\mathbf{U}_k = \mathbf{Q}_k\mathbf{H}$, where $\mathbf{Q}_k$ is of size $I_k \times R$ and has orthonormal columns and $\mathbf{H}$ is an $R \times R$ matrix, which is constant across k. Then, the targeted invariance constraint is implicitly enforced since: $\mathbf{U}_k^T\mathbf{U}_k = \mathbf{H}^T\ \mathbf{Q}_k^T\ \mathbf{Q}_k\mathbf{H} = \mathbf{H}^T\mathbf{H} = \Phi$. Since we follow an Alternating Least Squares (ALS) protocol, we first solve for a certain subset of factors by fixing all others. The first step solves for $\{\mathbf{Q}_k\}$, by fixing $\mathbf{H}$, $\{\mathbf{S}_k\}$ and $\mathbf{V}$, where $k \in \{1, \cdots, K\}$; it can be shown that each one of k subproblems is an Orthogonal Procrustes problem which has a closed-form solution [8]. The second step solves for $\mathbf{H}$, $\{\mathbf{S}_k\}$ and $\boldsymbol{V}$, by fixing $\{\mathbf{Q}_k\}$**.** This step is equivalent to:

$$\min_{\boldsymbol{H},\{\boldsymbol{S_k}\},\boldsymbol{V}} \sum_{k=1}^{K} ||\boldsymbol{Q_k^T X_k} - \boldsymbol{H S_k}\ \boldsymbol{V}^T||_F^2 \quad (1)$$

Eq. (1) CP decomposition as the second step of PARAFAC2-ALS.

The solution minimizing the above objective is given by a CANDECOMP/PARAFAC (CP) decomposition [10,15,16] on a tensor $\boldsymbol{Y} \in R^{R \times J \times K}$ with frontal slices $\mathbf{Y}_k = \mathbf{Q}_k\mathbf{X}_k$.[8] CP decomposes an input tensor as a sum of R rank-one vector outer products, i.e., $\boldsymbol{Y} \approx \sum_{r=1}^{R} \mathbf{h}_r\mathbf{v}_r\mathbf{w}_r$, where $\mathbf{h}_r \in R^R$, $\mathbf{v}_r \in R^J$, $\mathbf{w}_r \in R^K$ are column vectors and $^\circ$ is the outer product operator. The assembly of vectors $\mathbf{h_r}$, $\mathbf{v}_r$, $\mathbf{w}_r$ as: $\mathbf{H} = [\mathbf{h}_1,\mathbf{h}_2,\cdots,\mathbf{h_R}] \in R^{R \times R}$, $\mathbf{V} = [\mathbf{v}_1,\mathbf{v}_2,\cdots,\mathbf{v}_R] \in R^{J \times R}$, $\mathbf{W} = [\mathbf{w}_1,\mathbf{w}_2,\cdots,\mathbf{w}_R] \in R^{K \times R}$, gives the set of $\mathbf{H}$, $\mathbf{V}$, $\mathbf{W}$ factor matrices of the CP model. Note that Eq. (1) follows the "slice-wise" CP formulation, where each matrix $\mathbf{S}_k$ contains the row vector $\mathbf{W}(k, :)$ along its main diagonal [8].

We demonstrated how the PARAFAC2-ALS algorithm leads to a CP decomposition as the second of the alternating steps. A method to assess the suitability of the user-defined parameter R of the CP model is described in [11,17]; we exploit this method to assess the validity of the intermediate CP decomposition, thus indirectly assessing the validity of our PARAFAC2 model [14]. We provide a description of this method below and showcase all pseudocode details in Appendix A.

Before describing how the validity of the intermediate CP model is assessed, we need to introduce the Tucker model [18], which is a generalization of CP. The Tucker decomposition approximates a tensor as a sum of rank-one factors, weighted by the elements of a core tensor:

$$\boldsymbol{X} \approx \sum_{p=1}^{R}\sum_{q=1}^{R}\sum_{r=1}^{R} \boldsymbol{G}(p, q, r)\mathbf{h}_p \cdot \mathbf{v}_q \cdot \mathbf{w}_r$$

In the case of the CP model, the core tensor $\boldsymbol{G}$ is super-diagonal [19] (i.e., non-zero values are only in entries $\boldsymbol{G}(p, q, r)$where $p = q = r$). Note that in general the Tucker model permits the factor matrices to have varying number of columns; in this specific case, each of the $\mathbf{H}$, $\mathbf{V}$, $\mathbf{W}$ matrices have R columns.

Considering that $\mathbf{H}$, $\mathbf{V}$, $\mathbf{W}$ are computed via the intermediate CP model with components on a tensor $\boldsymbol{Y}$ of size $R \times J \times K$, the core consistency diagnostic (CORCONDIA) [11,17] computes the corresponding core tensor$\boldsymbol{G}$, assumed to be super-diagonal. If the computed core tensor$\boldsymbol{G}$ contains significant deviations from a super-diagonal core tensor, then this would imply that the CP decomposition is not optimal because the *a priori* number of components selected is not appropriate or because the CP model cannot accurately summarize the data. To compute the core tensor$\boldsymbol{G}$, the following least-squares problem needs to be solved:

$$\min_{\boldsymbol{G}} ||vec(\boldsymbol{Y}) - (\mathbf{H} \otimes\ \mathbf{V} \otimes\ \mathbf{W})vec(\boldsymbol{G})||$$

with solution: $vec(\boldsymbol{G}) = (\mathbf{H} \otimes\ \mathbf{V} \otimes\ \mathbf{W})^+ vec(\boldsymbol{Y})$, where $\otimes$ stands for the Kronecker product and + stands for the matrix pseudoinverse [20]. Then, if $\boldsymbol{T}$is defined as a super-diagonal tensor of ones, the percentage of core consistency is defined as:

$$1 - \frac{\sum_{p=1}^{R}\sum_{q=1}^{R}\sum_{r=1}^{R} (\boldsymbol{G}(p, q, r) - \boldsymbol{T}(p, q, r))^2}{R}$$

To increase our chances of avoiding local minima in our experiments, we run the PARAFAC2 fitting algorithm 10 times for each target rank R (i.e., number of task definitions pursued) using random initialization of factors and with the Singular Value Decomposition (SVD)-based initialization suggested in [21]. Then, we chose the solution with the minimum cost function error, among the ones achieving a diagnostic score of at least 90%. If the diagnostic score was lower than 90% for all 10 runs, then we simply picked the result with the best diagnostic score.

The validity of our model is measured through the CORCONDIA score on the intermediate CP as described above. The percentage of model fit is measured as the proportion of the total sum of squares that is explained by the model: [8]

$$Fit(R) = 1 - \frac{\sum_{k=1}^{K} ||\mathbf{X}_k - \mathbf{U}_k\mathbf{S}_k\mathbf{V}^T||_F^2}{\sum_{k=1}^{K} ||\mathbf{X}_k||_F^2}$$

A reasonable expectation is that as long as we increase the number of representative task definitions pursued, the fit percentage will increase as we are providing the model with more factors to describe data variation.

2.7. Proposed model interpretation

Below, we summarize the proposed approach to interpret model parameters. Each one of the R columns of the **V** matrix of the model provides a certain task definition – to extract this definition, we order the weights of each column in decreasing order and list the associated features. Due to the non-negativity constraints imposed on **V**, features associated with higher weights are expected to be more representative of the corresponding task definition.

The diagonal $\mathbf{S_k}$ matrix in the model is representative of the frequency of each task extracted for each k-th encounter. To estimate the task utilization for each user (i.e. PCP), we propose to consider the mean of all the R-dimensional encounters for each user. The resulting R-dimensional vector for each user can be used to identify variations among the users, with respect to task utilization. To visually identify clusters of users with similar utilization profiles, we use the tSNE [22] software, which can reduce the R-dimensional vectors to the 2-dimensional space.

3. Results

For our experiments, we used the Matlab implementation of the PARAFAC2 algorithm designed for sparse input, provided in [9]. Among a total of 132 features, derived as Intermediate or Terminal variants of audit log records (refer to the Methods Section for details), we excluded 84 that accounted for less than 1% of all feature occurrences across all the encounters. This resulted in J = 48 unique features as input to the model. No significant change in fit (e.g., 53.98% before vs 54.53% after excluding rare features the chosen target rank of R = 5) was observed before and after excluding the uncommon features.

Fig. 5 summarizes the percentages of fit and consistency diagnostic for various target ranks (i.e., number of tasks). As suggested in [14], the final solution comprising five tasks has the highest fit, as well as, a high consistency diagnostic (above 90%). Thus, for our experiment, the solution with 5 tasks is considered the most suitable.

Table 1 presents the most common tasks discovered using PARAFAC2 in our first experiment. The name assigned to each task was derived from the log record with highest weight in each cluster. We have used the term "access" in Table 1 as a general term for the cluster of EHR features accessed by the user; this does not imply that the task is a read-only view of the EHR; the audit log source file does not record more granular activities, such as keyboard typing, data entry, copy/paste activities, etc.

Table 1
Task definitions discovered via PARAFAC2. The first part of each entry denotes whether the feature corresponds to an audit log record that is an intermediate (I) or a terminal one (T). The terminal features are in bold font.

Task	Features[1]
1. Medication access	**T_Medications activity accessed**
	T_Patient Encounter opened
	I_Visit Navigator template loaded
	T_Chief Complaint navigator section accessed
	T_PCP History accessed
	I_Data from related encounters accessed through VB
2. Note access	**T_Notes activity and navigators accessed**
	T_Patient Encounter opened
	T_Problem List activity accessed
3. Order access	**T_Order Entry activity accessed**
	T_Diagnosis association updated
4. Diagnosis modification	**T_Visit diagnoses modified**
	T_Patient Encounter opened
	I_Visit Navigator template loaded

[1] Features are listed in order of frequency of occurrence as indicated by the PARAFAC2 model weights. Features associated with weights > 0.2 are listed.

3.1. Discovery of most representative PCP tasks

Without prior information, PARAFAC2 identified five groups of features accounting for 55% of the variation in audit log data (Fig. 5). One of those 5 groups contained only intermediate features. The latter might represent searching for information that is not linked to a specific or known task. We excluded this task. For each of the remaining 4 discovered task definitions, we list the identified features, in order of frequency as indicated by the PARAFAC2 model weights, and also as indicated by whether they are intermediate (I) or terminal (T). We annotate each one of the feature groups according to their most frequent terminal feature (e.g., we consider the first feature group in Table 1 to reveal the most common variation pattern of completing the Medication access task).

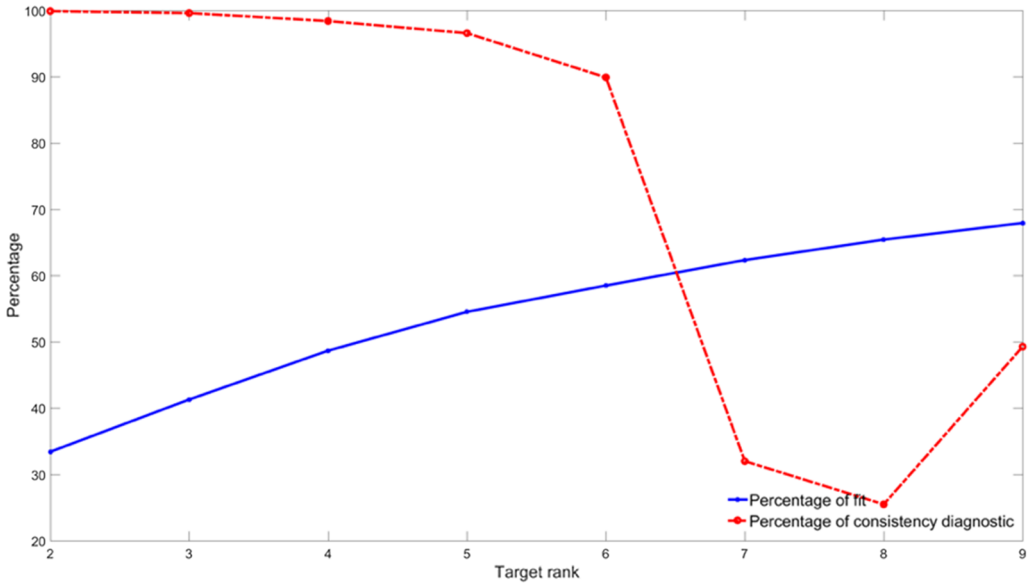

Fig. 5. Percentage of consistency diagnostic and fit. We chose a solution with five tasks that had the highest fit (i.e., lowest error) and a well-specified model reflected by the consistency diagnostic (over 90 percent).

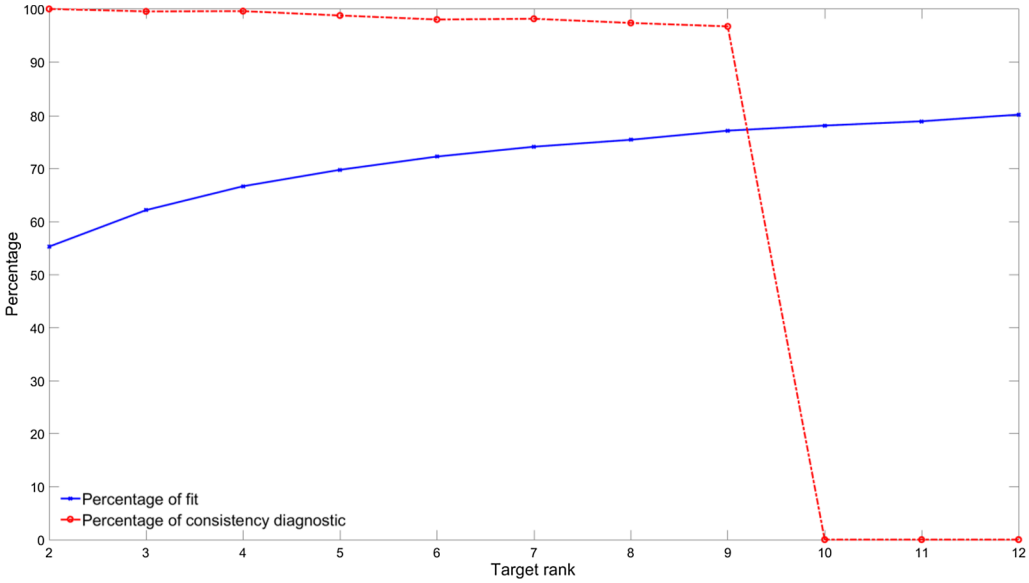

Fig. 6. Plot of the percentage of consistency diagnostic and fit for the **Notes activity access**. We chose a solution with 9 notes access patterns as it provided the highest fit (i.e., lowest error) and most well-specified model reflected by the consistency diagnostic (over 90 percent). For illustration purposes, we have set the values of consistency diagnostic less than zero to zero.

3.2. Variation in notes' access task

We further determined whether PARAFAC2 could be used to extract variation in audit log records that correspond to *Note Access* (Table 1) with the intention of detecting differences among PCPs in how notes are accessed and done.

Beginning with the 17,455 unique input encounters, we only retained tasks that contained the "T_Notes activity and navigators accessed" feature, indicating the inclusion of Notes' access as a terminal log record. We then follow the same pre-processing strategy as above, by excluding very infrequent features (< 1% of the total feature frequency) and excluding encounters with less than two records. This resulted in 14,659 encounters each described by 51 features.

Fig. 6 summarizes the percentages of fit and consistency diagnostic for various target ranks (i.e., number of pursued notes' access variation patterns). We extracted 9 distinct patterns associated with the notes' access task (see Appendix B), that explained approximately 77% of the input data variation.

We examined whether the 9 notes access pattern variants were specific to PCP groups using the pattern utilization indicator model factor ($\mathbf{S}_k$ in Fig. 4). This factor describes the pattern utilization of each k-th encounter for all R variation patterns identified above. To estimate the pattern utilization for each user (i.e. PCP), we simply estimate the mean of all the R-dimensional encounters for each user. We used the tSNE [22] software to visualize user representations, by reducing the 9-dimensional vectors to the 2-dimensional space.

Fig. 7 shows a clear separation of PCPs into two distinct groups (denoted as Groups A and B on the left and bottom right part of Fig. 7 respectively). We separately examined the log records of groups A and B to understand variation in accessing notes (see Appendix B) and task patterns followed by each one of the groups (Appendix C).

PCP Groups A and B were distinct in that the 6th variation pattern (i.e. see Appendix B, T_Communication Management navigator section accessed) is rarely used by group A (4%) but is often used by group B (29%). The 6th task variation pattern corresponds to a "Wrap-Up" tag in the EHR that is available to directly access notes before closing a patient encounter in the EHR.

Group B PCPs rarely used the patterns 2, 5, and 7 (Fig. 7). A common log record among those patterns is the loading of the "Visit Navigator", which is used by 30% of group A PCPs, and only 0.5% of group B PCPs. The latter corresponds to use of the "Visit Navigator" tag to access progress note in the EHR.

Finally, we independently worked with an EPIC expert and asked for a demonstration of ways to access clinic notes. For each approach that was demonstrated we recorded the underlying log record sequence of EPIC interface tags. Two dominant approaches to access clinic notes were identified, one via the "Visit navigator" tag and one via "Wrap-Up" tag.

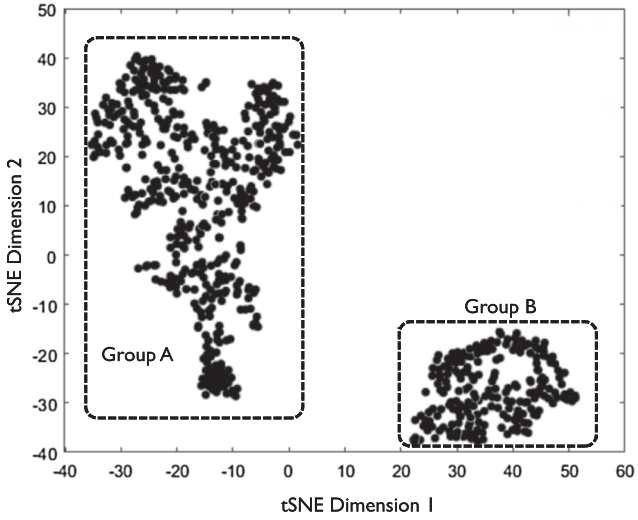

Fig. 7. tSNE visualization of the PCP patterns for Notes' access. Each point in the 2-D space corresponds to a PCP's relative position with respect to her pattern of utilization. The key conclusion is that there do exist 2 distinct clouds of points corresponding to 2 groups of primary care providers, based on the 9 variation patterns of notes' access. Individual axes can be safely ignored, as tSNE only preserves neighborhood information between data points.

4. Discussion

Electronic health records (EHR) data are increasingly used to understand the relation between the care that was delivered and the change in patient outcomes. But, workflow improvement efforts have a different focus, attempting to increase efficiency and effectiveness of work processes within and between encounters and to monitor progress towards efficiency [23]. Clinical data, alone, are usually insufficient to understand whether workflow efforts like "Lean process improvement" is working and, importantly, to understand how it is working. In fact, a common criticism of healthcare workflow improvement work is that it

lacks scientific measurement principles supported by readily accessible continual data. Audit log data may be useful in providing a scientific foundation for workflow management. Clinic practice workflow improvements (e.g., Lean principles) have not consistently translated into sustainable efficiency improvements [24]. In part, this may be because workflow efforts in health care lack a systematic and scalable approach to quantifying processes before and after a workflow improvement are implemented [1,25].

How work gets done has been described by Ulwick [6] and Christensen [26] as fundamentally important to understanding how to innovate in ways that have a high probability of success [26,27]. While these methods have been used in a diversity of business sectors, there are few applications to healthcare data. The volume and availability of audit log files offers a means to understand how healthcare is delivered by PCPs and others and to understand how work actually gets done. Access to this high-resolution data may offer a means to accelerate innovations to making healthcare work easier and more effective.

Audit log file data have not been widely used to understand care processes. Adler-Milstein et al. examined the impact of EHR platform usage on PCP productivity [28]. Tai-Seale et al. analyzed the time allocation patterns of PCPs between seeing patients and using the EHR platform [29]. Hirsch et al. used audit log file data to better understand clinical workflow practices and the corresponding patient experience [5]. Hribar et al. exploited the audit log file to calculate timings of clinical workflows [30]. Wetterneck et al. present a PCP task list derived from human observations of encounters; such an approach though has limited scalability due to the laborious nature of manual encounter observations [31]. Using Epic audit log data, Wang et al. estimated the time that PCPs spent on the most common types of activities (i.e., denoted “action categories”) that binned audit data into actions and into several behaviors (e.g.,) defined by internal medicine physicians. However, to complete a clinical task, a clinician may need to engage in multiple actions and related tasks [32]. None of these previous studies used audit log data to reveal PCP tasks and to understand how tasks actually get done (e.g., workflow variation patterns), nor to enable classification of EHR users based on their EHR access variation patterns to complete clinical tasks. While user behavior analysis methods based on clickstream data could also be applied (e.g., [33]), those methods impose much stricter order constraints among the log records of a sequence as compared to our work. Those constraints make it increasingly harder to capture the exact same order and time gaps among clickstreams from different users. Also, our approach which is based on dimensionality reduction permits a direct interpretation of the compressed audit log records enabling the extraction of meaningful representation of physician tasks, which aligns with one of the main objectives of this work. Overall, our work is a first step towards exploiting audit logs to understand PCP work processes. To this end, we applied advanced machine learning methods to identify the most common tasks that PCPs do during office encounter, as well as, common variants of those tasks. Our work opens the possibility of developing scalable quantitative measurement tools to both understand clinical tasks and the time required for these tasks.

One logical next step to this work is to link the task variation patterns with PCPs to clinical efficiency measures (e.g. minimal time spending in specific EHR area): given identification of user EHR utilization patterns to complete tasks, we could then estimate which group is more efficient in term of time spent in the tasks and timely completeness of encounter documentation. This type of work could offer rapid means of improving how EHRs could be more effectively used by PCPs, as well as, identifying EHR tasks that are a challenge to complete or too time-consuming, as priority candidates for vendor improvements.

Audit log data metrics could also be linked with clinical outcomes for defined patient segments that vary by disease burden to study variation in clinical processes and demands. In so doing, it may be possible to identify optimal approaches to use of the EHR for patient subgroups. For example, patients with severe or complicated conditions require much more time and effort by the PCPs during the encounter than less severe patients. Examination of audit logs for encounters of patients who are very demanding may reveal specialized workflows that facilitate the PCP’s use of the EHR and an increase in the volume of work done with less effort.

Understanding audit log data is in its infancy. Our work and that of others is only beginning to reveal the potential power of these data. Our preliminary use of PARAFAC2 revealed only five tasks. These were the most common. Future work will likely involve an iterative process where the most common tasks are identified and documented. Once these tasks are identified, the audit log records for these known, common tasks can be eliminated so that less common tasks can be revealed. To this end, we recently eliminated the audit log records associated with these first five tasks and then applied PARAFAC2 to the remaining log records. Not surprising, a new set of tasks were identified. While we partially demonstrated this methodology by focusing on the Notes’ access variation patterns, a systematic application of this idea is an important future direction to pursue in ultimately creating an ontology that could be used to represent the content of audit log data.

4.1. Limitations

There are several limitations to the current work. We did not consider the frequency of features within a task in each of our experiments. Duplicate features may represent ~5% of consecutive activities. Retaining repeated features may increase the variation of patterns, but it does not impact the identification of features that were used to represent a pattern variation. The current task variation patterns captured ~55% of EHR audit log record variation; We have not investigated the remaining 45% of variation, other than to note that iterative application of PARAFAC2 reveals new tasks. As we have noted, removal of the audit log records associated with these common tasks with re-application of PARAFAC2 to the remaining log records identifies less common tasks. In addition to the four well-defined clinical tasks we have identified, in-basket work (i.e. managing patient phone call, order results, order authorization, emails with patients or other PCPs, referral, etc.) accounts for 10–15% of PCP time. But, in-basket work is often done in fragments where individual tasks are interspersed among others. This makes it difficult to dissect these tasks from log records. Future work on iterative extraction of tasks may reveal the details of how in-basket work is done as a step towards reducing the growing demand that this work imposes on PCPs.

In the absence of a signal or dictionary indicating the end of a task, we adopted the simplifying assumption that all log records that are separated by less than a second are part of the same potential task. While we acknowledge this simplification, we argue that our approach fully utilizes log records at the available time-stamped resolution (i.e. one-second) and is robust to the two scenarios that may occur from this assumption. First, records grouped together may not belong to the same potential task. For example, in Fig. 2, it is possible that “Visit Navigator template loaded” is an irrelevant log record to the second potential task (timestamps 00:04–00:06). Still, PARAFAC2 would still capture the (I_Patient Encounter Opened, T_Notes activity and navigators accessed) task if this was prevalent in other encounters, whether or not it had the same timestamp as other features in the set. Second, groups of records that should be collectively considered to indicate a single task may be assigned to two different tasks by the model. For example, in Fig. 2, it is possible that both of the potential tasks we extracted should actually be grouped together to form a single task. Still, PARAFAC2 would capture more granular tasks if log records composed to the tasks are commonly clustered together in other encounters.

Finally, clinical validation is needed to ensure that the labels we have applied to each identified task indeed map to what a PCP is trying to accomplish following the corresponding log record sequence. However, we believe that assigning the label based on the log record

with the highest weight (thus, indicating the most frequent record), as we did, is the most reasonable data-driven way to label each task.

5. Conclusions

Our application of PARAFAC2 to audit log data offers an analytic approach that can facilitate the creation of ontologies or libraries that describe how healthcare work is done. We believe this is a critical step towards creating innovations that successfully equip PCPs and others to get more done with less effort and with a better experience.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the National Science Foundation, award IIS-1418511, IIS-1838042 and CCF-1533768, the National Institute of Health award 1R01MD011682-01, R01-HL116832 and R56HL138415.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jbi.2019.103312.

References

[1] P. Mazzocato, R.J. Holden, M. Brommels, H. Aronsson, U. Bäckman, M. Elg, J. Thor, How does lean work in emergency care? A case study of a lean-inspired intervention at the Astrid Lindgren Children's hospital, Stockholm, Sweden, Feb 1, BMC Health Services Research 12 (28) (2012) 28, https://doi.org/10.1186/1472-6963-12-28.

[2] B.G. Arndt, J.W. Beasley, M.D. Watkinson, J.L. Temte, W.-J. Tuan, C.A. Sinsky, V.J. Gilchrist, Tethered to the EHR: primary care physician workload assessment using EHR event log data and time-motion observations, Ann. Fam. Med. 15 (5) (2017) 419–426 PMID:28893811.

[3] G. Hripcsak, D.K. Vawdrey, M.R. Fred, S.B. Bostwick, Use of electronic clinical documentation: time spent and team interactions, J. Am. Med. Inform. Assoc. 18 (2) (2011 Apr) 112–117 PMID:21292706.

[4] S. Read-Brown, M.R. Hribar, L.G. Reznick, L.H. Lombardi, M. Parikh, W.D. Chamberlain, S.T. Bailey, J.B. Wallace, T.R. Yackel, M.F. Chiang, Time requirements for electronic health record use in an academic ophthalmology center, JAMA Ophthalmol. 135 (11) (2017) 1250–1257 Nov 1, PMID:29049512.

[5] A.G. Hirsch, J.B. Jones, V.R. Lerch, X. Tang, A. Berger, D.N. Clark, W.F. Stewart, The electronic health record audit file: the patient is waiting, J. Am. Med. Inf. Assoc.: JAMIA 24 (e1) (2017) e28–e34.

[6] A.W. Ulwick, Jobs to be Done: Theory to Practice, Idea Bite Press, 2016.

[7] R.A. Harshman, PARAFAC2: Mathematical and technical notes. UCLA working papers in phonetics 1972;22(3044):122215.

[8] H.A.L. Kiers, J.M.F. Ten Berge, R. Bro, PARAFAC2-Part I. A direct fitting algorithm for the PARAFAC2 model, J. Chemom. 13 (3–4) (1999) 275–294.

[9] I. Perros, E.E. Papalexakis, F. Wang, R. Vuduc, E. Searles, M. Thompson, J. Sun, SPARTan: Scalable PARAFAC2 for large & sparse data, Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining [Internet] Halifax, NS, Canada, ACM, 2017, pp. 375–384 [cited 2017 Sep 18], [doi: 10.1145/3097983.3098014].

[10] R. Harshman, {Foundations of the PARAFAC procedure: Models and conditions for an" explanatory" multi-modal factor analysis}. UCLA Working Papers in Phonetics 1970;16(1):84.

[11] R. Bro, Multi-way analysis in the food industry - models, algorithms, and applications. MRI, EPG and EMA, Proc ICSLP 2000, (1998).

[12] E.E. Papalexakis, C. Faloutsos, N.D. Sidiropoulos, Tensors for data mining and data fusion: models, applications, and scalable algorithms, ACM Trans. Intell. Syst. Technol. 8 (2) (2016) 16:1–16:44, https://doi.org/10.1145/2915921.

[13] I. Perros, E.E. Papalexakis, R. Vuduc, E. Searles, J. Sun, Temporal phenotyping of medically complex children via PARAFAC2 tensor factorization, Feb 8, 103125, J. Biomed. Inform. (2019), https://doi.org/10.1016/j.jbi.2019.103125.

[14] M.H. Kamstrup-Nielsen, L.G. Johnsen, R. Bro, Core consistency diagnostic in PARAFAC2, May 1, J Chemometrics 27 (5) (2013) 99–105, https://doi.org/10.1002/cem.2497.

[15] F.L. Hitchcock, The expression of a tensor or a polyadic as a sum of products, Apr 1, J. Math. Phys. 6 (1–4) (1927) 164–189, https://doi.org/10.1002/sapm192761164.

[16] J.D. Carroll, J.-J. Chang, Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart-Young" decomposition, Sep 1, Psychometrika 35 (3) (1970) 283–319, https://doi.org/10.1007/BF02310791.

[17] R. Bro, H.A.L. Kiers, A new efficient method for determining the number of components in PARAFAC models, J. Chemom. 17 (5) (2003) 274–286.

[18] L.R. Tucker, Some mathematical notes on three-mode factor analysis, Sep 1, Psychometrika 31 (3) (1966) 279–311, https://doi.org/10.1007/BF02289464.

[19] T. Kolda, B. Bader, Tensor decompositions and applications, Aug 5, SIAM Rev 51 (3) (2009) 455–500, https://doi.org/10.1137/07070111X.

[20] E.E. Papalexakis, C. Faloutsos, Fast efficient and scalable core consistency diagnostic for the parafac decomposition for big sparse tensors, 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015, pp. 5441–5445, , https://doi.org/10.1109/ICASSP.2015.7179011.

[21] R. Bro, C.A. Andersson, H.A.L. Kiers, PARAFAC2—Part II. Modeling chromatographic data with retention time shifts, J. Chemometrics 13 (3–4) (1999) 295–309 May 1, [doi: 10.1002/(SICI)1099-128X(199905/08)13:3/4 < 295::AID-CEM547 > 3.0.CO;2-Y].

[22] L. van der Maaten, G. Hinton, Visualizing data using t-SNE, 9(Nov), J. Machine Learning Res. (2008) 2579–2605.

[23] J.J. Cimino, Improving the electronic health record—are clinicians getting what they wished For? Mar 13, JAMA 309 (10) (2013) 991–992, https://doi.org/10.1001/jama.2013.890.

[24] T. Melton, The benefits of lean manufacturing: what lean thinking has to offer the process industries, Chem. Eng. Res. Des. 83 (6) (2005) 662–673.

[25] J. Moraros, M. Lemstra, C. Nwankwo, Lean interventions in healthcare: do they actually work? A systematic literature review, Apr 1, Int. J. Qual. Health Care 28 (2) (2016) 150–165, https://doi.org/10.1093/intqhc/mzv123.

[26] C.M. Christensen, K. Dillon, T. Hall, D.S. Duncan, Competing against luck: The story of innovation and customer choice, Harper Business (2016).

[27] A.W. Ulwick, What Customers Want Using Outcome-driven Innovation to Create Breakthrough Products and Services [Internet], McGraw-Hill, New York, 2005 [cited 2018 Nov 10]. Available from: http://www.ebrary.comISBN:978-0-07-140867-7.

[28] J. Adler-Milstein, R.S. Huckman, The impact of electronic health record use on physician productivity, Am. J. Manag. Care (2013) Nov;19(10 Spec No):SP345-352. PMID: 24511889.

[29] M. Tai-Seale, C.W. Olson, J. Li, A.S. Chan, C. Morikawa, M. Durbin, W. Wang, H.S. Luft, Electronic health record logs indicate that physicians split time evenly between seeing patients and desktop medicine, Health Aff. 36 (4) (2017) 655–662.

[30] M.R. Hribar, S. Read-Brown, I.H. Goldstein, L.G. Reznick, L. Lombardi, M. Parikh, W. Chamberlain, M.F. Chiang, Secondary use of electronic health record data for clinical workflow analysis, J. Am. Med. Inform. Assoc. (2017) Sep 26; PMID:29036581.

[31] Development of a primary care physician task list to evaluate clinic visit workflow | BMJ Quality & Safety [Internet]. [cited 2019 Mar 1]. Available from: https://qualitysafety.bmj.com/content/21/1/47.long.

[32] J.K. Wang, D. Ouyang, J. Hom, J. Chi, J.H. Chen, Characterizing electronic health record usage patterns of inpatient medicine residents using event log data, e0205379, PLoS ONE 14 (2) (2019 Feb 6), https://doi.org/10.1371/journal.pone.0205379.

[33] G. Wang, X. Zhang, S. Tang, H. Zheng, B.Y. Zhao, Unsupervised clickstream clustering for user behavior analysis, Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems New York, NY, USA, ACM, 2016, pp. 225–236.