Developing Data-driven Clinical Pathways using Electronic Health Records: The Cases of Total Laparoscopic Hysterectomy and Rotator Cuff Tears

Minsu Cho[a]*, Kidong Kim[b]*, Jungeun Lim[a], Hyunyoung Baek[c], Seok Kim[c], Hee Hwang[c], Minseok Song[a], Sooyoung Yoo[c]

[a]Department of Industrial and Management Engineering, Pohang University of Science and Technology, Pohang, South Korea

[b]Department of Obstetrics and Gynecology, Seoul National University Bundang Hospital, Seongnam, South Korea

[c]Office of eHealth Research and Businesses, Seoul National University Bundang Hospital, Seongnam, South Korea

Corresponding and co-corresponding authors:

Minseok Song, mssong@postech.ac.kr

77 Cheongam-Ro. Nam-Gu. Pohang. Gyeongbuk. South Korea 37673

Sooyoung Yoo, yoosoo0@snubh.org

166 Gumi-Ro. Bundang-Gu. Seongnam. Gyeonggi. South Korea 13620

*** These authors contributed equally to this work and are first authors.**

Keywords: clinical pathways, electronic health records(EHR), matching rates, evidence-based approach, total laparoscopic hysterectomy(TLH), rotator cuff tears(RCTs)

Abstract

Objective

A clinical pathway is one of the tools used to support clinical decision making that provides a standardized care process in a specific context. The objective of this research was to develop a method for building data-driven clinical pathways using electronic health record data.

Materials and methods

We proposed a matching rate-based clinical pathway mining algorithm that produces the optimal set of clinical orders for each clinical stage by employing matching rates. To validate the approach, we utilized two different datasets of deidentified inpatient records directly related to *total laparoscopic hysterectomy* (TLH) and *rotator cuff tears* (RCTs) from a hospital in South Korea. The derived data-driven clinical pathways were evaluated with knowledge-based models by health professionals using a delta analysis.

Results

Two different data-driven clinical pathways, i.e., TLH and RCTs, were produced by applying the matching rate-based clinical pathway mining algorithm. We identified that there were significant differences in clinical orders between the data-driven and knowledge-based models. Additionally, the data-driven clinical pathways based on our algorithm outperformed the models by clinical experts, with average matching rates of 82.02% and 79.66%, respectively.

Conclusion

The proposed algorithm will be helpful for supporting clinical decisions and directly applicable in medical practices.

1. Introduction

A clinical pathway is a tool that delivers structured clinical services on the basis of evidence-based healthcare in a specified medical context (e.g., for diseases, diagnostics, and surgeries) [1-8]. It has been developed with the aim of standardizing and optimizing care processes to minimize the undesired practice variability and manage clinical outcomes, e.g., length of stays or rehospitalization [1-8]. As such, the application of clinical pathways has received attention since it shortens the length of hospital stays, lowers costs, reduces complications and lowers mortality [9-12].

Typically, a clinical pathway is organized in a day-by-day format, composed of clinical orders that contain clinical services (e.g., prescriptions and treatments) from medical practitioners [4,13]. Additionally, each clinical pathway has a specified length of hospital stay, and clinical stages are defined such as *regular, preoperation, postoperation,* and *discharge* [4,13]. Therefore, solid theoretical backgrounds are required to develop clinical pathways, and typically, most medical sites have developed clinical pathways based on the knowledge of domain experts, e.g., health professionals [14,15].

The process of developing knowledge-based clinical pathways by discussing with medical professionals is ideal, and clinical pathways have been developed on the basis of this approach for frequently occurring significant diseases [14,15]. However, it was challenging to develop clinical pathways for all clinical contexts using the knowledge-based approach due to the following two reasons: (i) time and efforts are limited since health professionals engaged in the development of clinical pathways are overloaded with medical activities, and (ii) healthcare processes are dynamic and complicated since clinical operations, medicines, and therapies are developed continuously.

As an alternative approach and support for the knowledge-based approach, data-driven methods have been developed using the data from electronic health records (EHR) [4,7,16-23]. Specifically, data-centric techniques, e.g., data mining and process mining, have been utilized to develop more realistic clinical pathways [4,7,16-23].

These approaches have resolved the limitations of knowledge-based clinical pathways addressed above,

however, unfortunately, it still has a couple of challenges to apply them as follows immediately: (i) they only focus on deriving a coarse-grained clinical pathway (i.e., at the activity level) with the traditional process mining approaches, in other words, they do not provide any detailed the steps (i.e., clinical orders) for a specific timeframe [16,20,23], (ii) they focus on finding clinical pathway patterns or a summarized model instead of providing how to develop the standardized clinical order set for a specific surgery or diagnosis [4,7,21,38].

This paper proposes an approach to produce a realistic clinical pathway at the order level, i.e., a matching rate-based clinical pathway mining algorithm. The proposed approach starts with our prior research suggesting a method to compare clinical pathways and the relevant log data with a quantitative approach, i.e., matching rates (see the details in Section 3) [13]. The method aims to derive the optimal set of clinical orders for specific timeframes, i.e., clinical pathways, that maximizes the matching rate using patient data collected in EHR. Specifically, we utilized two different sets of deidentified inpatient records directly related to *total laparoscopic hysterectomy* (TLH) and *rotator cuff tears* (RCTs) collected in a tertiary hospital in South Korea. Based on the proposed approach, we derived two different clinical pathways for each clinical context and compared them with the knowledge-based models created by health professionals to validate our work.

2. Related works

The primary discipline associated with our work, i.e., the development of data-driven clinical pathways, is process mining. It aims at deriving process-related knowledgeable insights from event logs and has been widely applied in a healthcare environment [26,29]. The use of process mining in healthcare has been more focused on process discovery, and there have been approaches for automatically extracting clinical workflow process models in different medical fields [29]. Mans et al. [18] presented the discovered spaghetti-like process model for the gynecological oncology healthcare process, while Rebuge and Ferreira [30] proposed a method to generate emergency process models using process

mining with clustering techniques. Also, there have already been attempted to discover clinical process models for outpatients, inpatients, and surgery [29].

Similar to these works, some researchers have strived to develop data-driven clinical pathways, with the focus on identifying frequent patterns and process models, using process discovery techniques pertaining to process mining [16,22,23]. Lakshmanan et al. [22] proposed a method for mining a clinical pathway mining approach based on clinical outcomes by merging process mining with frequent pattern mining. Also, Xu et al. [16] presented a series of steps to generate topic-based activity clusters using Latent Dirichlet Allocation (LDA) and derive a process model using fuzzy mining. These approaches have contributed to developing a high-level abstraction clinical pathway, i.e., an activity-level process model. However, these approaches do not include the details of how to make clinical orders required by a specific stage or date; thus, a low-level abstraction clinical pathway, i.e., the order-level, is required for practical use.

In this regard, to overcome this limitation, several approaches have devoted to creating the order-level clinical pathway using process mining [4,7,21,38]. Huang et al. [4] suggested a mining approach to derive summarized clinical pathways given minimum support threshold and event logs, while Iwata et al. [21] suggested the similarity-based visualization approach that provides a compressed model based on the probabilistic threshold from users. Huang et al. [7] focused on finding clinical pathway patterns, not the homogenized model. Also, Huang et al. [38] proposed a method to detect anomalies in clinical pathway patterns. To recap, those methods were effective in analyzing clinical pathways such as identifying patterns and deriving summarized information, but they had a limitation that is not applicable to find a more standardized clinical pathway presented in this study.

The other related discipline to our research is healthcare data mining, which has more focused on discovering clinical order sets necessitated for clinical decision support tools [31-35]. More in detail, they developed several approaches to provide the patient-personalized clinical order sets using Hidden Markov Model [31], K-means clustering [32,33], recommendation systems [34], and frequent itemset mining and association rule mining [35]. These approaches are quite similar to the data-driven clinical

pathway mining in that both approaches produce a series of order sets as output. However, they devote to identifying patient-personalized clinical order sets with the aim of reducing cognitive click costs of experts and providing personalized healthcare [34,35], while clinical pathway mining focuses on deriving comprehensive order sets for a specific clinical context in an unfavorable economic scenario with the aims of minimizing variations of the clinical results. Thus, it is required to suggest a different method for clinical pathway development.

3. Materials and Methods

3.1. Clinical pathways

This section introduces the features and overall structure of clinical pathways developed in this research. To determine the format of clinical pathways, we have employed the existing works of literature and discussion with health professionals. In this regard, we aimed at developing the most standardized clinical order set for a specific surgery or diagnosis; in other words, our models did not cover the diversity, i.e., conditional branching in models, caused from the comorbidities, e.g., characteristics or history of patients. To this end, an alternative method was considered as building branch CPs by distinguishing from the parent CP, if needed to create the model for a specific comorbidity. In this regard, we utilized a simple two-step algorithm [37]; (i) the statistical analysis is performed to identify the relationship between patient characteristics and clinical outcomes (e.g., length of stays, readmission rates, and matching rates) and (ii) if exists, the proposed approach in this study is applied to two patient cohorts based on patient characteristics (i.e., holding and non-holding), and then a branch CP is defined in the case that the derived two order sets have a clear difference.

Regarding the structure of the clinical pathways, it has the specified length, i.e., duration, for a particular surgery or diagnosis, and a set of clinical orders is constructed for each day, i.e., a day-by-day format. Also, relative dates based on surgery dates (e.g., *OP day, 1 day before,* and *1 day after*) were utilized instead of the actual dates on admission (e.g., *day 1* and *day 2*) since the difference in medical orders is

clear based on the time of the surgery. Furthermore, clinical pathways hold clinical stages such as *regular, pre-operation, post-operation,* and *discharge* for each day, and only unique orders can be included for a specific stage and date. Thus, it does not take into account that a single order appears multiple times in a particular stage and date. This is because historical data may enclose biased and suboptimal behaviors including overutilization of orders [36]. Lastly, with the same reason, only medication and test orders are included in clinical pathways based on the opinion of health professionals that they do not cause this issue.

3.2. Data

Two different sets of deidentified inpatient records related to total laparoscopic hysterectomy and rotator cuff tears were collected from the Seoul National University Bundang Hospital, a tertiary hospital in South Korea. TLH is one of the surgeries performed in obstetrics and gynecology [24], and we collected records from 520 inpatients who received the surgery from January 2012 to May 2014. RCT is a common disease managed in orthopedics [25], and data from 360 inpatients between June 2014 and 2015 were extracted to develop a data-driven clinical pathway.

These records included patient information, hospitalization, operation, diagnosis, and orders. Specifically, concerning the clinical orders, i.e., basic units of clinical pathways, 18115 and 14862 events were extracted, respectively. Additionally, each event included order attributes such as stages (e.g., preoperation, operation, and postoperation) and types (e.g., medicines and tests). Table 1 provides detailed information about the collected data.

The present study was approved (IRB No. B-1609/361-105) by the Institutional Review Board of the Seoul National University Bundang Hospital, which waived patients' informed consent. All deidentified EHR data were then provided to the researchers for this study.

Table 1. Types and attributes of the collected data.

Type	Attribute
Patient Information	Patient ID, Age, Sex, Drinking, Smoking, Allergy, Disease history, Operation history, Drug history, Family disease history, Diabetes, Hypertension, Hyperlipidaemia, Cardiovascular, Cerebrovascular
Hospitalization	Hospitalization ID, Patient ID, Admission date, Discharge date, Admission type, Discharge schedule type, Assigned physician ID, Department code, Department name
Operation	Operation ID, Patient ID, Hospitalization ID, Operation date, Operation code, Operation name
Diagnosis	Diagnosis ID, Patient ID, Hospitalization ID, Diagnosis code, Diagnosis code classification, Diagnosis name, Physician ID, Department code, Department name, Diagnosis date
Order	Order ID, Patient ID, Hospitalization ID, Order code, Order name, Order date, PRN status, Order stage, Order type, Order interruption classification code, Order interruption date

Hospitalization ID: a unique ID for identification of inpatients.

3.3. Data re-engineering

In a clinical pathway, the same orders can be repeatedly involved in different dates or stages; thus, a single order can be published multiple times according to its usage. Thus, to distinguish the same clinical order in building a clinical pathway, we performed a data re-engineering approach. More in detail, we determined the clinical orders as the combination of order names, relative dates based on surgery dates, and order types (i.e., *'order_date_type'*). Here, regarding the relative dates based on surgery dates, the day of the operation was calculated as 0, the next day as 1, and the one day before surgery as -1, while the order types were defined as regular, pre-operation, post-operation, operation, and discharge. Thus, if order A was used as a normal order on the day of surgery, the code would be expressed as *'A_0_regular'*.

3.4. Order-level matching rate

Prior to explain our algorithm for deriving clinical pathways, we first explain our prior research that compares clinical pathways and the relevant log data with a quantitative approach [13]. More specifically, it generates a numerical value of comparison between clinical orders from clinical pathways and logs. The order-level matching rate has employed conformance checking in process mining discipline, and it signifies the extent to which the log is related to the possible behaviors in the process model. Here, we have defined the matching rate by replacing with the clinical pathway. Formula (1) provides the proposed matching rate.

$$Matching\ rate = \frac{1}{2}\left(1 - \frac{M_{CP}}{N_{CP}}\right) + \frac{1}{2}\left(1 - \frac{R_{Log}}{N_{Log}}\right) \quad (1)\ [13]$$

- M_{CP}: The number of orders included in the clinical pathway but not shown in the log data
- N_{CP}: The number of orders included in the clinical pathway
- R_{Log}: The number of orders included in the log data but not shown in the clinical pathway
- N_{Log}: The number of orders included in the log data

As shown in this formula, the matching rate is composed of the application rate of orders in the clinical pathway (i.e., $\frac{1}{2}\left(1 - \frac{M_{CP}}{N_{CP}}\right)$) and the matched ratio of orders in the log data (i.e., $\frac{1}{2}\left(1 - \frac{R_{Log}}{N_{Log}}\right)$). Thus, it covers both how the orders included in the clinical pathway are applied to the patients and how the orders used by the patients are different from the clinical pathway.

Table 2 provides an example of how to measure the matching rate with the clinical pathway and log data. In this example, the clinical pathway is composed of 12 clinical orders (i.e., T1, T2, T3, Te1, Te2, Te3, M1, M2, M3, I1, I2, and I3), while the different order codes for four patients are contained. As provided in the Table, P1's orders are the same as those of the clinical pathway. Thus, M_{CP} and R_{Log} are 0, and the matching rate becomes 1.00. In the case of P2, compared to those in the clinical pathway,

some orders, e.g., Te3, I2, and I3, are missing (i.e., expressed as "–"). As such, M_{CP} and N_{Log} are 3 and 9, respectively, and the matching rate becomes 0.88. Compared to P2, the orders for P3 include some orders not defined in the clinical pathway, e.g., T4, I4, and I5. Thus, N_{Log} and R_{Log} are 15 and 3, respectively, and these values decrease the matching rate to 0.90. Lastly, P4 lacks some of the required orders and the addition of orders not defined in the clinical pathway. As such, the matching rate becomes the lowest value among the four patients.

Table 2. An example of how to measure the matching rate.

CP	*T1*	*T2*	*T3*		*Te1*	*Te2*	*Te3*	*M1*	*M2*	*M3*	*I1*	*I2*	*I3*			N_{CP}	M_{CP}	N_{Log}	R_{Log}	*Matching rate*
P1	T1	T2	T3		Te1	Te2	Te3	M1	M2	M3	I1	I2	I3			12	0	12	0	1.00
P2	T1	T2	T3		Te1	Te2	–	M1	M2	M3	I1	–	–			12	3	9	0	0.88
P3	T1	T2	T3	T4	Te1	Te2	Te3	M1	M2	M3	I1	I2	I3	I4	I5	12	0	15	3	0.90
P4	T1	–	T3	T4	Te1	Te2	–	M1	–	M3	I1	I2	–	I4		12	4	10	2	0.73

3.5. Matching rate-based clinical pathway mining algorithm

The matching rate-based clinical pathway mining algorithm is outlined in Figure 1. Our algorithm takes log data (i.e., *L*) as an input and produces the clinical pathway (i.e., *CP*) and its matching rate (i.e., *m*). In the proposed algorithm, the first step is to define and initialize the required variables (e.g., *m*) and sets (e.g., *CP*, O_L, and AR_O). Then, as shown in lines 3-7, a unique clinical order set (i.e., O_L) is prepared and included in the log data. Next, we calculate the application rates for clinical orders involved in O_L. In the following steps (i.e., line 9-12), the application rate is defined as the number of applied clinical orders (i.e., $\sum_{0<k\leq p}\sum_{0<l\leq n}\begin{cases}1 & if\ o_{k,l}=o_j\\ 0 & otherwise\end{cases}$) divided by the number of patients (i.e., *p*). As a result of this step, the list of application rates for each clinical order (i.e., AR_O) is prepared. Next, all clinical orders are sorted in descending order based on their application rates, and O_L' is generated (line 13). Lines 14-22 provide an iterative approach to find the optimal clinical pathway. In the sorted clinical order set O_L', we first select a clinical order that has the highest application rate and is included in the clinical pathway set *CP*. If the application rates of different orders are the same, then the order is

randomly selected. Then, we measure the average value of matching rates (i.e., m_t) using $\frac{1}{2}\left(1-\frac{\sum_{0<k\leq p} M_{k,CP}}{\sum_{0<k\leq p} N_{k,CP}}\right)+\frac{1}{2}\left(1-\frac{\sum_{0<k\leq p} R_{k,Log}}{\sum_{0<k\leq p} N_{k,Log}}\right)$. Here, let $N_{k,CP}$ denote the number of clinical orders in the clinical pathway for patient k. $M_{k,CP}$, $N_{k,Log}$, and $R_{k,Log}$ are defined in a similar fashion, e.g., $R_{k,Log}$ is the number of orders included in the data of patient k but is not shown in the clinical pathway. After that, the measured m_t is compared with the current maximum value m. If m_t is larger than m, the selected order is unchanged in CP. In the opposite case, however, it is removed from the clinical pathway. By conducting this step iteratively, the proper set of clinical orders is determined and becomes the optimal clinical pathway.

Algorithm 1 *Matching rate-based clinical pathway mining*

Input

A log data L

Output

A clinical pathway (i.e., a list of clinical orders) CP

A matching rate of the clinical pathway m

1. $CP, O_L, AR_O \leftarrow \{\}$
2. $m \leftarrow 0$
3. ***for all*** o_i *in* L ***do***
4. ***if*** o_i *does not exist in* O_L ***then***
5. $O_L \leftarrow O_L \cup o_i$
6. ***end if***
7. ***end for***
8. *Let n be the number of unique clinical orders in* O_L *and p be the total number of patients in L*
9. ***for all*** o_j *in* O_L ***do***
10. *Calculate the application rate* $ar_j = \frac{\sum_{0<k\leq p}\sum_{0<l\leq n}\begin{cases}1 & if\ o_{k,l}=o_j\\ 0 & otherwise\end{cases}}{p}$
11. $AR_O \leftarrow AR_O \cup ar_j$
12. ***end for***
13. *Sort* O_L *by* AR_O *in the descending order to get the* $O_L' = \{o_1', o_2', \ldots, o_n'\}$
14. ***for all*** o_j' *in* O_L' ***do***
15. $CP \leftarrow CP \cup o_j'$
16. *Calculate the average of matching rates* $m_t = \frac{1}{2}\left(1-\frac{\sum_{0<k\leq p} M_{k,CP}}{\sum_{0<k\leq p} N_{k,CP}}\right)+\frac{1}{2}\left(1-\frac{\sum_{0<k\leq p} R_{k,Log}}{\sum_{0<k\leq p} N_{k,Log}}\right)$
17. ***if*** $m_t > m$ ***then***
18. $m \leftarrow m_t$
19. ***else***
20. $CP \leftarrow CP \otimes o_j'$
21. ***end if***
22. ***end for***

23. ***return*** *CP, m*

Figure 1. Matching rate-based clinical pathway mining algorithm

We provide the procedure of the suggested matching rate-based clinical pathway mining algorithm with the running example in Table 2. In the table, the log data and the clinical orders of four patients are included. Based on the log, we first identify the set of unique clinical orders, i.e., O_L = {T1, T2, T3, T4, Te1, Te2, Te3, M1, M2, M3, I1, I2, I3, I4, I5}. Then, the application rates for clinical orders are calculated, e.g., the values of T1 and T2 are 1.0 and 0.75, respectively. Among them, a clinical order with the highest application rate, i.e., T1, is added into the clinical pathway *CP*, and the average matching rate is calculated as 0.55. In the second iteration, T3 is added in the clinical pathway, i.e., *CP* = {T1, T3}. Then, the average matching rate becomes higher than the value from the first iteration. By conducting this process continuously, we can obtain the highest matching rate as 0.88 and the relevant clinical pathway, i.e., *CP* = {T1, T3, Te1, Te2, M1, M3, I1, T2, Te3, M2, I2, T4}.

3.6. Evaluation

In this section, we explain how to evaluate the derived clinical pathways by comparing with knowledge-based models from health professionals. In this research, knowledge-based clinical pathways have been developed through the CP task force team (TFT) committee consisting of clinical departments, nursing departments, pharmacy departments, insurance review teams, medical information teams, and management innovation teams. If a target is selected according to the consensus of the clinical department, the initial CP is manually designed based on existing order sets and then assumed to be subject to CP TFT. After identifying the appropriate medication, antibiotics appropriateness, and insurance cutbacks by the CP TFT committee, CP is developed in the EHR system with final approval of the committee. In such a process, the committee meets regularly once a month periodically monitors the results of utilizing the developed CP, receives feedback, and updates the order sets.

Figure 2 provides a schema for deriving clinical pathways using the matching rate-based clinical

pathway mining algorithm for TLH and evaluating the derived model. We used the delta analysis that identifies the difference with a qualitative approach by domain experts [26]. Thus, we thoroughly compared the difference between the knowledge-based models from domain experts and the data-driven models.

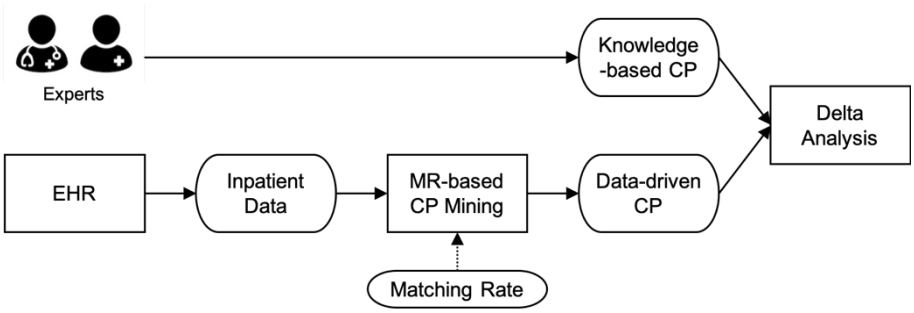
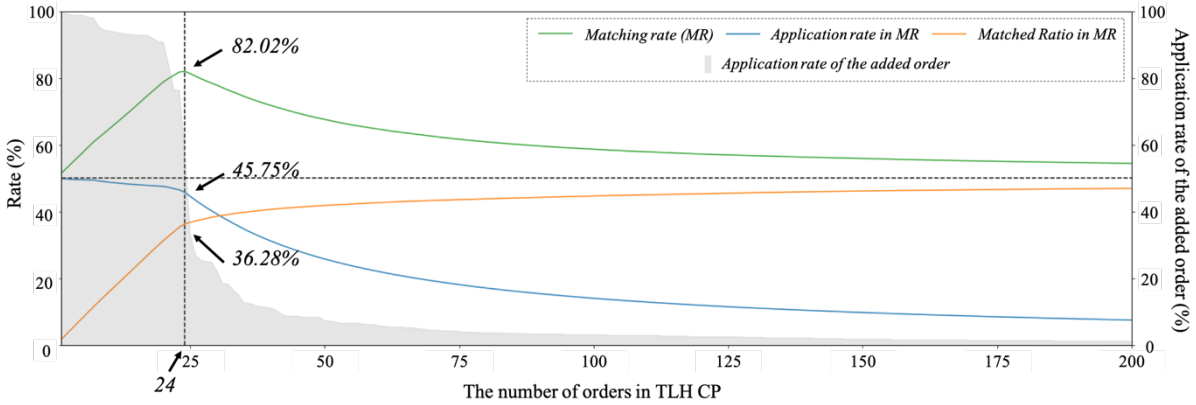


Figure 2. A schema for deriving clinical pathways and evaluation of them.

4. Results

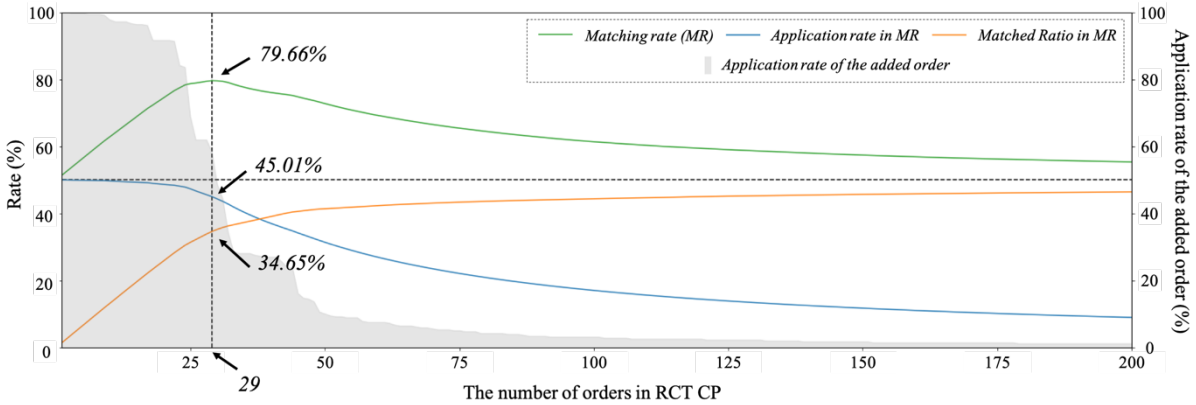
4.1. Matching rate-based CP mining algorithm results

This subsection explains the results for deriving data-driven clinical pathways using the proposed matching rate-based clinical pathway mining algorithm. As a result of applying our algorithm, two different sets of clinical orders that maximize the matching rate were prepared for the TLH and RCT clinical pathways, respectively. Figure 3 provides a graphical representation of the process for deriving clinical pathways.



283

284 **(a) The result for deriving the TLH clinical pathway**



285

286 **(b) The result for deriving the RCT clinical pathway**

287 **Figure 3. The results for deriving the clinical pathways**

288

289 As provided in Figure 3a, in the process of deriving the TLH clinical pathway, a clinical order with the
290 highest application rate, i.e., *Acetylcysteine*, was first included in the clinical pathway (application rate:
291 99.62%). Since its application rate was almost 1.0, the application rate of orders in the clinical pathway
292 $(\frac{1}{2}\left(1-\frac{M_{CP}}{N_{CP}}\right))$, i.e., the left side of the matching rate, was close to 50.0% (more precisely, 49.81%).
293 However, the other side of the matching rate, i.e., the matched ratio of orders in the log data
294 $(\frac{1}{2}\left(1-\frac{R_{Log}}{N_{Log}}\right))$, was almost 0 (more precisely, 1.64%) since the one-order clinical pathway is not
295 sufficient to reflect the whole log data. As a result, the matching rate became 51.45% by adding the two
296 values. Then, we included an additional order with the second highest application rate of 99.42% in the

clinical pathway. This decreased the average application rate (49.76%) and increased the matched ratio of orders (3.28%). As such, the matching rate was also increased by 53.04%. Through the repetition of this process, we identified that 24 clinical orders maximized the matching rate as 82.02% and became the optimal set for the clinical pathway; the average application rate and the matched ratio of orders were 45.75% and 36.28%, respectively.

Adding one more clinical order with the 25th highest application rate in the clinical pathway, the matching rate was decreased by 81.37% because the average application rate had a greater decrease than the increase in the matched ratio. If we include all 1259 orders in the clinical pathway, the average matched ratio of clinical orders becomes the perfect score of 50%. Even if it has the maximum value, however, the average matching rate becomes at most 51.31% since the application rate has the lowest value of 1.31%.

Similar to the procedure of deriving the TLH clinical pathway, we discovered the RCT clinical pathway. Figure 3b provides the results for deriving the RCT model. As a result of applying the proposed algorithm, it was determined that a clinical pathway with the top 29 clinical orders within the application rate had the highest matching rate of 79.66%; the average application rate and the matched ratio of orders were 45.01% and 34.65%, respectively.

4.2. Comparison with knowledge-based clinical pathways

4.2.1. The statistical comparative analysis results

From the derived data-driven clinical pathway, we performed a comparative analysis with the knowledge-based model to show the outperformance of the proposed algorithm. Table 3 provides the statistical comparative analysis results for the TLH and RCT clinical pathways. Concerning the TLH clinical pathway, the model from our algorithm was relatively more straightforward than the existing model. In detail, one order was newly added in the data-driven clinical pathway, while four existing orders were deleted. Both the average and median matching rates of the derived model were increased

by approximately 5% compared to those of the existing model. Additionally, as a result of conducting the statistical hypothesis testing, e.g., a *t-test* [27], it was determined that the matching rate of the data-driven clinical pathway was significantly higher than that of the knowledge-based model (*p-value* < 0.001). Regarding the RCT clinical pathway, there was a dramatic reduction in the number of clinical orders in the derived model. In the newly developed clinical pathway, 33 existing orders were removed, while 5 orders were newly included. Regarding the matching rate, the average and median from the model applied in the proposed algorithm were increased by approximately 23% and 24%, respectively, compared to those from the knowledge-based model. Furthermore, similar to the TLH case, the statistical result showed that the data-driven clinical pathway significantly outperforms the model from domain experts (*p-value* < 0.001).

Table 3. The statistical comparative results for the TLH and RCT clinical pathways

	TLH		RCT	
	Knowledge-based	Data-Driven	Knowledge-based	Data-Driven
Number of orders	27	24	57	29
Added	-	1	-	5
Removed	-	4	-	33
Average of matching rates (%)	76.97	82.02	56.64	79.66
95% CI of matching rates (%)	(75.89-78.05)	(80.89-83.16)	(55.82-57.47)	(78.52-80.79)
T-test (*p-value*)	-	<0.001	-	<0.001
Median of matching rates (%)	80.03	85.12	57.12	81.17
SD of matching rates (%)	12.51	13.21	7.93	10.90

4.2.2. The delta analysis results

We performed the order-level delta analysis to identify the difference between clinical pathways within an order level, and Figure 4 provides the results of the analysis. Regarding the TLH clinical pathway, two models from the domain experts and log data were composed of 27 and 24 clinical orders over four days, respectively. Specifically, *photography* was newly added to the data-driven model, while four medicine orders (e.g., *ketorolac, aceclofenac, and multienzymes)* were deleted on the operation day and one day after. Regarding the RCT clinical pathway, there was a significant difference between the newly

developed and the existing model. In contrast to the existing clinical pathway being a five-day schedule, the data-driven clinical pathway was developed over six days. To this end, *site marking* was moved from one day before to two days before the operation day. Additionally, *oxycodone-naloxone* and three orders for *shoulder Rt* were added one day before and three days after the operation, respectively. Furthermore, 33 clinical orders, including *ondansetron, tramadol, fentanyl, and morphine,* were removed from the clinical pathway.

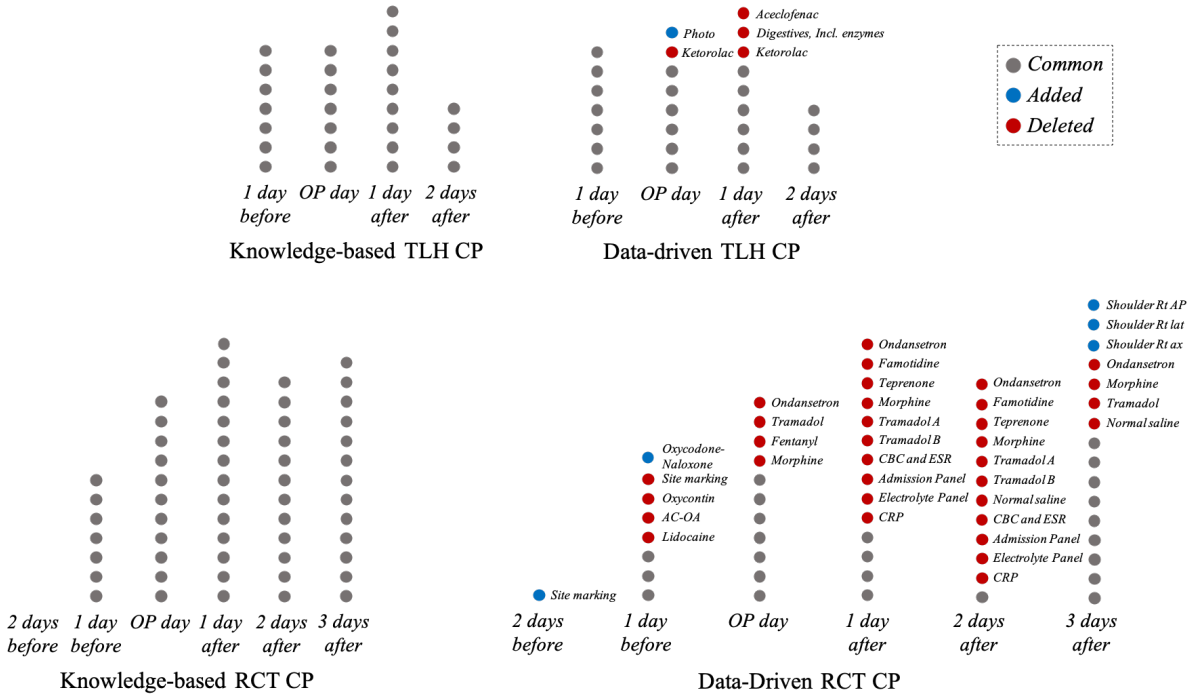


Figure 4. The order-level comparative analysis

We performed further comparative analysis on the basis of dates (e.g., 1 day before, OP day, and 1 day after), order stages (e.g., regular, pre-operation, post-operation, and discharge) and order types (e.g., medications and tests) to identify the details of the difference between two models.

The results of the detailed delta analysis for the TLH clinical pathway is provided in Table 4. Comparing two different TLH models, it was identified that most of the clinical orders are included in the medication type and that there is no significant difference, i.e., commonly utilized orders in both models.

Regarding the deleted orders in the data-driven model, it was recommended to remove *Ketorolac* included in the post-operation of the operation day and the regular stage of the 1 day after. Also, the model suggested deleting two medication orders in the discharge stage of the 1 day after since there were not enough patients who leave out the hospitals on the next day after the surgery. As far as the test type orders were concerned, both models included two orders, e.g., *CBC* and *Urine analysis*, in the regular stage of the 1 day after, and the supplement of a single order, e.g., *Photography*, in the OP day was recommended in the data-driven model.

Table 4. The detailed delta analysis results for the TLH clinical pathway

Day	Stage	Medication			Test		
		Common	Added	Deleted	Common	Added	Deleted
1 day before	Regular	Electrolytes, Thioglycolic, Bisacodyl, Sodium Phosphate, Multienzymes, Magnesium Citrate, Cefotetan					
OP Day	Regular					Photo	
	Pre-OP	Electrolytes, Cafotetan, Tissue exam.					
	Post-OP	Acetylcystein, Carbohydrates, Famotidine		Ketorolac			
1 day after	Regular	Acetylcystein, Famotidine, Carbohydrates, Cefotetan		Ketorolac	CBC, Urine Analysis		
	Discharge			Aceclofenac, Multienzymes			
2 days after	Regular	Aceclofenac, Multienzymes					
	Discharge	Aceclofenac, Multienzymes					

As a result of the detailed delta analysis for the RCT clinical pathway (presented in Table 5), medication type orders occupied the majority of the models, and all stages hold them. Comparing two different RCT models, there were three significant differences; (i) deleting a series of medication orders (e.g., *Ondansetron, Tramadol, Morphine, Fentanyl,* and *Famotidine*) engaged in the post-operation and regular stages from the OP day to 3 days after, (ii) removing some medication (e.g., *Tramadol &*

Paracetamol and *Teprenone*) and test (e.g., *Admission Panel, CBC & ESR, CRP,* and *Electrolyte Panel)* orders connected to the discharge stage at 1 day and 2 days after the surgery, and (iii) adding three test orders (e.g., *Shoulder Rt AP, Shoulder Rt lat,* and *Shoulder Rt ax*) in the discharge stage at the last day.

Table 5. The detailed delta analysis results for the RCT clinical pathway

Day	Stage	Medication			Test		
		Common	Added	Deleted	Common	Added	Deleted
2 days before	Regular					Site Marking	
1 day before	Regular	Acetaminophen, Pregabalin, Celecoxib	O-N	Lidocaine, Oxycontin			AC-OA, Site Marking
OP Day	Pre-OP	Cefazolin, Electrolytes					
	Post-OP	Palonosetron, Sodium Chloride, EwC, Famotidine, Cefazolin		Ondansetron, Tramadol, Morphine, Fentanyl			
1 day after	Regular	Palonosetron, Sodium Chloride, Cefazolin, T&P		Famotidine, Ondansetron, Tramadol, Morphine			
	Discharge			T&P, Teprenone			AP, CBC & ESR, CRP, EP
2 days after	Regular	T&P		Famotidine, Ondansetron, Tramadol, Morphine			
	Discharge			T&P, Teprenone			AP, CBC & ESR, CRP, EP
3 days after	Regular	Teprenone, Afloqualone, T&P		Ondansetron, Tramadol, Morphine, Sodium Chloride			
	Discharge	Teprenone, T&P			AP, CBC & ESR, CRP, EP	Shoulder Rt AP, Shoulder Rt lat, Shoulder Rt ax	

(O-N: Oxycodone-Naloxone, EwC: Electrolytes with Carbohydrates, T&P: Tramadol & Paracetamol, AP: Admission Panel, EP: Electrolyte Panel)

4.3. Organizational relevance

The data-driven CPs were reviewed and commented by domain experts including the obstetrics and orthopedic clinicians. As far as the TLH clinical pathway was concerned, the clinical order that needed to be added was revealed after visual inspection of the resected pathologic tissue after the operation.

Since it should be decided whether or not the order was issued according to the surgical result, expert commented to exclude it from the CP. Additionally, The four medication orders recommended for removing (eg, ketorolac, aceclofenac, and digestives, incl. Enzymes) were analgesic and digestive system orders to prescribe to patients with pain and dyspepsia, which can occur frequently after surgery. As a result, it was concluded that we need to improve the system to make it possible to provide those orders only to the necessary patients rather than issuing them regularly.

As far as the RCT clinical pathway was concerned, the US Extremity site marking test was added two days before the operation, which provoked the change in the total schedule of the clinical pathway. Additionally, we recognized that three orders for X-ray photography were regularly implemented before the patients were discharged. In addition, it was discovered that 33 orders recommended to be removed were used to avoid bleeding, prevent infection, relieve pain, and improve digestion. As with the TLH clinical pathway, we concluded that those orders should not be commonly applied to all patients.

Overall, both clinical departments agreed that the results of this study could be well accepted and reflected in the clinical setting. The orders derived from the data-driven CPs provide reliable results, but the final decision should be made in a semi-automatic manner after expert reviews according to the various context of the patient or medical practice patterns of the healthcare organization. Also, considering these results, the hospital prepared the revised clinical pathway by modifying clinical orders and changing the schedule, and we finally arrived at a conclusion that it is necessary to develop a new system that can recommend the orders appropriately according to the patients' symptoms or test results.

5. Discussion

An existing study presented four criteria to define a new clinical pathway: (i) a structured multidisciplinary plan of care, (ii) translating guidelines into local structures, (iii) detailed steps in a course of care in a plan, and (iv) aiming to standardized care for a specific population [28]. Based on these criteria, we evaluated whether our algorithm is suitable for deriving a clinical pathway. First, the output of the algorithms, i.e., a set of clinical orders, are the clinical activities that members in multiple disciplines are intimately involved in. In particular, this study utilized data from inpatients who have

410 undergone TLH and RCT surgery performed with multidisciplinary care; the same is true for the derived
411 models from data. Also, our algorithm takes EHR data, i.e., evidence, and provides a structured and
412 detailed order plan for a specific clinical context (e.g., TLH or RCT). Therefore, we argue that clinical
413 pathways from our algorithm address the full four criteria.

414 This research makes a significant contribution of automatically developing clinical pathways based on
415 the collected data in electronic health records. Hospitals generally cannot build clinical pathways due
416 to an insufficient workforce, time, and costs. The proposed algorithm will enable the preparation of
417 more accurate clinical pathways without any human intervention. In other words, this paper is of value
418 because it is useful for supporting decision making with an evidence-based approach.

419 The proposed algorithm produces a fine-grained model (i.e., at the order-level), not a coarse-grained
420 model (i.e., at the activity-level). Thus, it is directly applicable to medical practices. Additionally, it is
421 extensible where various clinical pathways can be developed. For example, we can develop a clinical
422 pathway that considers operations and patient characteristics together, e.g., a TLH-diabetic-female
423 clinical pathway.

424 Despite these contributions, this study has several limitations. First, this research highly depends on
425 data variability. Specifically, it would be difficult to identify clinical pathways based on data from
426 internal medicine departments that have a high variability of patient behaviors. To overcome this
427 challenge, future studies should develop a method to determine in advance whether a clinical pathway
428 needs to be created and an approach for building a clinical pathway considering both data and domain
429 knowledge. Second, this research does not cover distinguishing of infrequent events from random noise
430 since our primary research goal is to derive a standardized order set. However, it may require including
431 a couple of orders for specific patients who have the same comorbidity with the first-diagnosis. To deal
432 with this, future works should present a method for analyzing the need for clinical pathway
433 segmentation by analyzing the relationship between patient characteristics and their clinical orders and
434 developing the relevant branch CPs. Additionally, the analysis results presented in this paper were only
435 from a single hospital. Thus, this study may lack generalizability since clinical pathways and their data

can differ among hospitals. Future studies should perform more case studies using data from multiple hospitals.

6. Conclusion

This paper suggested a matching rate-based clinical pathway mining algorithm to automatically develop clinical pathways based on collected data in electronic health records. The practical applications at a real site resulted in a rise in matching rates of two different clinical pathways, i.e., the TLH and RCT clinical pathways. This research will be helpful in supporting clinical decision making and can be directly applied in medical practices.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science, and Technology (No.NRF-2016R1C1B2016346) and Korea Institute for Advancement of Technology(KIAT) grant funded by the Korea Government(MOTIE) (N0008691, The Competency Development Program for Industry Specialist).

Statement on conflicts of interest

None.

Summary table

What was already known on the topic

- Clinical pathways deliver structured clinical services by evidence-based healthcare in a specified medical context.
- Knowledge-based models by health professionals are prominent; however, nowadays, data-

driven clinical pathways have been implemented with the data from electronic health records.

What this study added to our knowledge

- Our algorithm, i.e., matching rate-based clinical pathway mining algorithm, can automatically develop the fine-grained clinical pathway at the clinical order level.
- In the real-life cases of TLH and RCTs, we identified that data-driven clinical pathways using our algorithm could complement the deficiencies of the knowledge-based models. Therefore, this research will help support clinical decision making and can be directly applied in medical practices.

REFERENCES

[1] Waters RC, Toy JM, Drechsler A. Clinical pathways for primary care: current use, interest and perceived usability. *Journal of the American Medical Informatics Association*. 2018;25(7):901-906.

[2] Panella M, Marchisio S, Di Stanislao F. Reducing clinical variations with clinical pathways: Do pathways work? *International Journal for Quality in Health Care*. 2003;15(6):509-521.

[3] Romeyke T, Stummer H. Clinical pathways as instruments for risk and cost management in hospitals - a discussion paper. *Global journal of health science*. 2012;4(2):50-59.

[4] Huang Z, Lu X, Duan H, Fan W. Summarizing clinical pathways from event logs. *Journal of Biomedical Informatics*. 2013;46(1):111–27.

[5] Wakamiya S, Yamauchi K. What are the standard functions of electronic clinical pathways? *International Journal of Medical Informatics*. 2009;78(8):543–50.

[6] Lenz R, Blaser R, Beyer M, Heger O, Biber C, Bäumlein M, et al. IT support for clinical pathways—Lessons learned. *International Journal of Medical Informatics*. 2007;76:S397–402.

[7] Huang Z, Lu X, Duan H. On mining clinical pathway patterns from medical behaviors.

Artificial Intelligence in Medicine. 2012;56(1):35–50.

[8] Cooch P, Roudsari A. Computerization of workflows, guidelines, and care pathways: a review of implementation challenges for process-oriented health information systems. *Journal of the American Medical Informatics Association*. 2011;18(6):738-748.

[9] Lin F, Chou S, Pan S, Chen Y. Mining time dependency patterns in clinical pathways. *International Journal of Medical Informatics*. 2001;62(1):11–25.

[10] Shi J, Su Q, Zhao Z. Critical Factors for the Effectiveness of Clinical Pathway in Improving Care Outcomes. *2008 International Conference on Service Systems and Service Management*. 2008;(1).

[11] Caminiti C, Scoditti U, Diodati F, Passalacqua R. How to promote, improve and test adherence to scientific evidence in clinical practice. *BMC Health Services Research*. 2005;5(1):62.

[12] Macario A, Horne M, Goodman S, Vitez T, Dexter F, Heinen R, et al. The effect of a perioperative clinical pathway for knee replacement surgery on hospital costs. *Anesthesia and analgesia*. 1998;86(5):978–84.

[13] Yoo S, Cho M, Kim S, et al. Conformance analysis of clinical pathway using electronic health record data. *Healthcare Informatics Research*. 2015;21(3):161-166.

[14] Yang W, Su Q. Process Mining for Clinical Pathway Literature Review and Future Directions. *Service Systems and Service Management (ICSSSM), 2014 11th International Conference*. 2014;(2010):1–5.

[15] Hua-qiong W, Tian-shu Z, Li-li T, Yang-ming Q, and Jing-song L. Creating hospital-specific customized clinical pathways by applying semantic reasoning to clinical data. *Journal of Biomedical Informatics*. 2014;52:354-363..

[16] Xu X, Jin T, Wei Z, Wang J. Incorporating Topic Assignment Constraint and Topic Correlation Limitation into Clinical Goal Discovering for Clinical Pathway Mining. *Journal of*

Healthcare Engineering. 2017.

[17] Lang M, Bürkle T, Laumann S, Prokosch H-UU. Process mining for clinical workflows: challenges and current limitations. *Studies in health technology and informatics*. 2008;136:229–34.

[18] Mans R, Schonenberg H. Process mining techniques: an application to stroke care. *Studies in health technology and informatics*. 2008;136:573–8.

[19] Mans R, Schonenberg MH, Song M, Aalst W van der, Bakker PJM. Application of Process Mining in Healthcare – A Case Study in a Dutch Hospital. *Communications in Computer and Information Science*. 2011;127:224–37.

[20] Perer A, Wang F, Hu J. Mining and exploring care pathways from electronic medical records with visual analytics. *Journal of Biomedical Informatics*. 2015;56:369–78.

[21] Iwata H, Hirano S, Tsumoto S. Construction of clinical pathway based on similarity-based mining in hospital information system. *Procedia Computer Science*. 2014;31:1107–15.

[22] Lakshmanan GT, Rozsnyai S, Wang F. Investigating clinical care pathways correlated with outcomes. *2013 International Conference on Business Process Management*. 2013;323–38.

[23] Xu X, Jin T, Wei Z, et al. TCPM: Topic-Based Clinical Pathway Mining. *2016 IEEE 1st International Connected Health: Applications, Systems, and Engineering Technologies, CHASE 2016*. 2016:292-301.

[24] Einarsson JI, Suzuki Y. Total laparoscopic hysterectomy: 10 steps toward a successful procedure. *Rev Obstet Gynecol*. 2009;2(1):57-64.

[25] Ellman H. Diagnosis and treatment of incomplete rotator cuff tears. *Clinical orthopaedics and related research*. 1990;254:64-74.

[26] van der Aalst WMP, Weijters AJMM. Process mining: a research agenda. *Computers in industry.* 2004;53(3):231-244.

[27] Kaur H, Wasan SK. Empirical study on applications of data mining techniques in healthcare. *Journal of Computer science.* 2006;2(2):194-200.

[28] Adegboyega KL, Thomas R, Leigh K, et al. What is a clinical pathway? Refinement of an operational definition to identify clinical pathway studies for a Cochrane systematic review. *BMC Medicine*. 2016;14(35).

[29] Rojas E, Munoz-Gama J, Sepulveda M, et al. Process mining in healthcare: A literature review. *Journal of biomedical informatics.* 2016;61: 224-236.

[30] Rebuge A and Ferreira DR, Business process analysis in healthcare environments: A methodology based on process mining. *Information systems*. 2012;37(2):99-116.

[31] Lin FR, Chou SC, Pan SM, et al. Mining time dependency patterns in clinical pathways. *International journal of medical informatics.* 2001;62(1):11-25.

[32] Zhang Y and Padman R. Data driven order set development using metaheuristic optimization. *Conference on Artificial Intelligence in Medicine in Europe.* 2015:47-56.

[33] Zhang Y, Padman R, and Levin JE. Paving the COWpath: data-driven design of pediatric order sets. *Journal of the American Medical Informatics Association*. 2014;21(e2):e304-e311.

[34] Chen JH, Podchiyska T, and Altman RB. OrderRex: clinical order decision support and outcome predictions by data-mining electronic medical records. *Journal of the American Medical Informatics Association.* 2015;23(2):339-348.

[35] Wright A and Sittig DF. Automated development of order sets and corollary orders by data mining in an ambulatory computerized physician order entry system. *AMIA annual symposium proceedings.* 2006;2006:819.

[36] Zhang Y, Trepp R, Wang W, et al. Developing and maintaining clinical decision support using clinical knowledge and machine learning: the case of order sets. *Journal of the American Medical Informatics Association*. 2018;25(11):1547-1551.

[37] Lim J, Kim K, Cho M, et al. Deriving a sophisticated clinical pathway based on patient conditions from electronic health record data. (*Working Paper).*

[38] Huang Z, Dong W, Ji L, et al. On local anomaly detection and analysis for clinical pathways. *Artificial Intelligence in Medicine*. 2015;65(3):167-177.