Proceedings of the 2025 Winter Simulation Conference
E. Azar, A. Djanatliev, A. Harper, C. Kogler, V. Ramamohan, A. Anagnostou, and S. J. E. Taylor, eds.

MODELING AND SIMULATION OF SURGICAL PROCEDURES WITH AN APPLICATION TO LAPAROSCOPIC CHOLECYSTECTOMY

Yiyu Wang[1], Vincent Augusto[1], Canan Pehlivan[2], Julia L. Fleck[1], and Nesrine Mekhenane[3]

[1]Mines Saint-Etienne, Univ. Clermont Auvergne, INP Clermont Auvergne, CNRS, Saint-Etienne, FRANCE
[2]Dept. of Industrial Engineering, IMT Mines Albi-Carmaux, Albi, FRANCE
[3]Chaire Innovation BOPA, AP-HP, Institut Mines Télécom, Université Paris Saclay, Villejuif, FRANCE

ABSTRACT

Surgeons' actions are key to surgical success. Our objective is to develop a decision-support tool to help prioritize patient safety and reduce risks during surgery. We propose a structured mathematical framework that defines key components of a surgical procedure, making it adaptable to various types of surgeries. Using the CholecT50 dataset, we generate and pre-process event logs to construct a process map that models the surgical workflow through Process Mining techniques. This process map provides insights into procedural patterns and can be visualized at different levels of granularity to align with surgeons' needs. To validate its effectiveness, we simulate synthetic surgeries and assess the process map's performance in replicating real surgical workflows. By demonstrating the generalizability of our approach, this work paves the way for the development of an advanced decision-support tool that can assist surgeons in real-time decision-making and post-operative analysis.

1 INTRODUCTION

Surgical safety remains a priority due to ongoing risks from human error, equipment issues, and patient variability. As modern technologies increase procedural complexity, decision-support tools are essential to help surgeons manage these systems and reduce adverse events.

Several approaches exist to model surgical procedures based on operating rooms workflows (Neumuth 2017). Previous works explored various techniques to understand and predict surgeon activities and surgical workflows (Neumann et al. 2022; Franke et al. 2015; Bieck et al. 2020). In the field of image recognition, recent efforts aim to automatically assess the critical view of safety, a key step for ensuring patient safety in procedures such as laparoscopic cholecystectomy (Mascagni et al. 2022). Existing studies have employed terms such as surgical phase, instrument, anatomic target, surgeon verb, and high-level and low-level surgical tasks. However, the lack of a comprehensive public dataset capturing surgeon activities remains a barrier to developing a generalized framework applicable to various surgical procedures.

Intraoperative process mapping (Chung et al. 2017) is commonly used for postoperative analysis and training. In Central Venous Catheters (CVC) procedures, process mining has been shown to generate effective process maps from event logs (Lira et al. 2018). However, such techniques are largely used retrospectively, and real-time or generalized applications remain limited.

Robot-assisted surgery has driven progress in analyzing surgical gestures using video and kinematic datasets (Anastasiou et al. 2023). Neural network models such as Transformer (Vaswani et al. 2017) have been widely used for skill assessment and workflow prediction. However, most public datasets include fewer than 20 activities—e.g., JIGSAWS has 9 (Gao et al. 2014), SARAS-ESAD has 21 (Bawa et al. 2021)—while real surgeries involve over 100, limiting model applicability to complex procedures.

While prior research has provided valuable observational insights, it has largely stopped short of enabling in-depth analysis or identifying intraoperative errors and adverse events. Most work has focused on recognition and segmentation, without translating findings into actionable feedback. As a result, current studies remain far from delivering real-time decision support tools to assist surgeons. Advancing surgical outcomes requires moving beyond observation toward predictive, interactive systems that offer meaningful intraoperative guidance.

Our study has three main objectives. First, we aim to develop a robust and flexible framework for modeling various surgical procedures as event logs. Second, we seek to establish an efficient methodology for pre-processing these logs to identify the surgical workflow. As a case study, we apply our framework to laparoscopic cholecystectomy, using its data to execute the first two steps and generate a detailed process map of the procedure. Finally, we simulate the surgery to assess the reliability of our process map.

Despite existing studies in surgical modeling and process mapping, many lack precise mathematical definitions to clearly distinguish the various steps of a surgical procedure. In our work, we propose a formal framework of mathematical definitions to address this gap. First, it provides the precision and unambiguity required for informatics systems, enabling machine interpretability and real-time automated analysis—unlike the intuitive, context-dependent descriptions typically used by clinicians. Second, it ensures reproducibility and standardization, which are essential for building robust benchmarks and advancing research in surgical error and adverse event detection. Third, it facilitates consistent data annotation and labeling, allowing the creation of high-quality datasets for training and validating algorithmic models. Finally, the proposed framework is designed to be generalizable across different surgical procedures, laying the groundwork for broad applicability in both research and clinical settings.

This work marks a crucial step towards the development of a decision-support tool that assists surgeons in real time by detecting errors and predicting surgical events, while also enabling comprehensive post-operative analysis.

2 METHODOLOGY

2.1 Framework

In 2022, a public dataset (Nwoye and Padoy 2023) was introduced specifically designed for recognizing surgical triplets. A surgical triplet is defined in the form of $<$instrument, verb, target$>$. Surgical triplets have been widely accepted as an effective way to represent surgeon activities during procedures. Previous works (Sharma et al. 2023) have identified such surgical triplets from video data. In parallel, there has been growing interest in the detection of surgical phases (Lavanchy et al. 2023), which segment a procedure into distinct stages, providing a clearer understanding of the overall workflow.

In this context, we define key concepts relevant to surgical process modeling—namely, surgery, surgical phase, surgical gesture, surgical error, and surgical event—as follows.

Let:

- S be a finite set indexing all surgical procedures. An element is denoted by $s \in S$.
- For each surgery $s \in S$, let P_s be the finite set indexing all surgical phases of s, with $p \in P_s$.
- For each phase $p \in P_s$, let G_{ps} be the finite set indexing surgical gestures of p, with $g \in G_{ps}$.
- $T = \mathbb{N}$ be the discrete time domain.
- $Pr = \{1.\text{ Primary gesture}, 2.\text{ Auxiliary gesture}\}$ be the set that indexes all gesture properties.
- E and A be finite sets indexing surgical error types and adverse event types independent of s and p, respectively.

Definition 1 A *surgery* denoted as *surge*, is defined as $(s,\ t_b,\ t_e)$ where $s \in S$, and $t_b,\ t_e \in T$ denote the start and end times of the surgery, respectively.

Definition 2 A *surgical phase*, denoted as *pha*, is defined as (s, p, t_b, t_e), where $p \in P_s$ is a phase of surgery s, and $t_b,\ t_e \in T$ denote the start and end times of the surgical phase, respectively. At any single time point, only one phase is possible.

Definition 3 A *surgical gesture*, denoted as *ge*, is defined as (s, p, g, pr, t_b, t_e), where: $g \in G_{ps}$ is a gesture within phase p, and $pr \in Pr$ denotes the gesture type. t_b, $t_e \in T$ denote the start and end times of the surgical gesture, respectively.

At any single time point, many gestures are possible. In our research, a gesture type is a surgical triplet. According to the definitions, surgical phases refer to the higher-level tasks that make up the complete surgical procedure (Garrow et al. 2021), distinguishing them from more granular actions such as gestures.

A primary surgical gesture type generally refers to the essential actions directly involved in performing the main tasks of a procedure, while an auxiliary surgical gesture type would describe supporting actions that assist or complement the primary tasks.

Definition 4 A *surgical error*, denoted as *error*, is defined as (s, p, g, e, t), where $e \in E$ is the error type and $t \in T$ is the time point at which the error is detected. An *adverse event*, denoted as *ae*, is defined as (s, p, g, a, t) where $a \in A$ is the adverse event type and $t \in T$ is the time point at which the adverse event is detected.

In medical terms, a surgical error (Suliburk et al. 2019) refers to either the failure to complete a planned action as intended (i.e., an error of execution) or the implementation of an incorrect plan to achieve a goal (i.e., an error of planning). Surgical errors are preventable mistakes. In contrast, an adverse event (Gawria et al. 2021) is an injury caused by medical care during surgery, rather than the underlying disease. Therefore, in the context of Definition 4, sets A and E represent fundamentally different concepts. For instance, a surgical error may result from a lapse in a surgeon's attention, whereas an adverse event could manifest as unexpected bleeding during the procedure.

As highlighted in these definitions, surgeries, phases, and gestures all have an associated duration, whereas surgical errors and adverse events do not. We discretize the detection of phases, gestures, errors, and adverse events, considering $T = \mathbb{N}$. The discretization interval t_s is defined by the user. At each time step, we update the current gesture, phase, error, and adverse event. If no changes occur in the phase or gesture, their duration is incremented. Otherwise, a new phase or gesture begins, resetting the count.

A surgical phase offers a high-level perspective on the procedure, allowing us to determine the current stage within the entire surgical process. In contrast, analyzing gestures enables the identification of specific errors, which can help predict potential adverse events.

2.2 Surgical Data Modeling

To standardize workflows, detect errors, and optimize performance—ultimately enhancing surgical precision and patient safety—a process map of the surgical procedure is essential. Our approach involves constructing this process map using an existing dataset and the framework presented in Section 2.1.

A process map requires event logs as input. An event log is a structured record of events that occur within a system or process over time. In the context of process mining or workflow analysis, an event log typically includes attributes such as Timestamp, Activity, Case ID, and Resource. Event log data provides a robust foundation for generating process maps, enabling direct observation of procedural flow and dynamics.

As discussed in Section 1, triplets are commonly used to model surgical activities based on formal surgical ontologies (Neumuth et al. 2006). Algorithms have been developed to automatically detect these triplets from surgical videos, supporting the initial level of automation needed for building a decision-support system. In our approach, we incorporate triplets as gestures (as defined in Definition 3) within our event logs.

2.2.1 Data Preparation

We consider that each entry of an event log contains the following elements: the patient identifier (Case ID), the type of gesture (Activity) as described in Definition 3, the beginning and ending timestamps of the gesture (Timestamp), and the type of phase (Resource) as described in Definition 2 that corresponds to the current gesture. Up to three gesture types can be recorded simultaneously within a single entry. As detailed in Section 2.1, the definitions of both a phase and a gesture require the inclusion of both the start time (t_b) and the end time (t_e). An illustrative example is provided in Table 1.

Table 1: Illustrative example of an event log.

ID	begin time	end time	phase	gesture
11	0:25:9	0:25:10	clipping-and-cutting	grasper,retract,gallbladder;bipolar,coagulate,cystic_pedicle
11	0:25:11	0:25:39	clipping-and-cutting	bipolar,coagulate,cystic_pedicle
11	0:25:40	0:25:43	clipping-and-cutting	grasper,retract,gallbladder
11	0:25:44	0:25:48	clipping-and-cutting	grasper,retract,gallbladder;scissors,cut,cystic_duct
11	0:25:49	0:26:19	clipping-and-cutting	scissors,cut,cystic_duct
11	0:26:19	0:26:21	clipping-and-cutting	grasper,retract,gallbladder

Using the framework from Section 2.1, we perform the following procedure to pre-process the raw event log data:

1. Replace “instrument, null_verb, null_target" by earlier or later gestures. When surgeons switch instruments, transitional gestures like “instrument, null_verb, null_target" appear. These transitional gestures actions are replaced by examining the previous and consecutive gestures involving the same instrument.

2. Merge neighboring gestures that have an inclusion relationship. In event logs, variations of gestures with an inclusion relationship, like “gesture A; gesture B; gesture C" and its subsets, often appear continuously. When the same combination is repeated over time, neighboring gestures with this inclusion relationship are merged.

3. Keep primary gestures. As described in the dataset, raw annotations can include up to three simultaneous gestures. Usually, one is the primary gesture, while the others are supporting or auxiliary. To focus on primary gestures in cholecystectomy, we exclude auxiliary gestures like “grasp," “retract," which assist in actions. We also exclude "aspirate" and "irrigate," as they are cleaning actions not directly related to the surgical procedure, even when performed by the main surgeon. However, we retain “grasper, grasp, specimen_bag" as a key surgical activity.

After this pre-processing step, there is typically only one gesture occurring at any given time.

4. Keep the verb and target of the triplets. The verb and target in triplets capture the core actions and objectives of the procedure, crucial for understanding surgical workflows. Since surgeons are familiar with the instruments, focusing on verbs and target provides a clearer, high-level view, simplifying the process map without losing essential details

5. Regroup related gestures. In our research, we regroup repetitive gestures that occur in loops during specific phases of surgery. These systematic actions help capture patterns that are essential to the success of the surgery and prevent self-loops in the process map.

The pre-processing method was developed based on expert input from surgeons. While we hypothesize that it performs well, it has not yet been quantitatively evaluated.

2.2.2 Process Mining

Process mining analyzes event data to improve operations through techniques like Process Discovery, Conformance Checking, and Process Enhancement (van der Aalst 2016).

Process Discovery automatically generates process models from event logs, revealing actual workflows and patterns based on real-time data. In our research, we apply Process Discovery to create a detailed process map of the surgical procedure. We use the software Disco (Günther and Rozinat 2012) for Process Discovery, utilizing the Flexible Heuristics Miner algorithm (Weijters and Ribeiro 2011) to generate a process map from event log data.

Next, we apply Conformance Checking to assess how well the actual procedure aligns with the established process model. Additionally, Conformance Checking supports post-operative analysis by identifying potential errors and adverse events. To evaluate and compare Process Mining Discovery models, we utilize key indicators such as fitness, precision, simplicity, and generalization (Buijs et al. 2012). The mathematical formulations of the four indicators (Rozinat and van der Aalst 2008; van der Aalst et al. 2012; Berti and van der Aalst 2020; Buijs et al. 2012) are not included here due to their complexity; however, a concise description of each metric is provided in this section to support understanding.

Fitness evaluates how well a process model can reproduce the behavior recorded in the event log. In token-based replay (TBR) fitness (Rozinat and van der Aalst 2008), the metric quantifies the proportion of correctly produced and consumed tokens during the simulation of log traces through the model, penalizing missing or remaining tokens. In contrast, alignment-based (Align) fitness (van der Aalst et al. 2012) measures how closely each trace in the event log can be aligned with an execution of the model by computing the cost of deviations; lower alignment costs indicate higher fitness.

Precision measures how well the behavior allowed by a process model matches what actually happened in the event log. In TBR, the model loses precision (Berti and van der Aalst 2020) if it allows steps that are never used when simulating real executions. In alignment-based (Align) precision (van der Aalst et al. 2012), the model is penalized for having transitions (called escaping edges) that are possible in the model but never needed to match the log.

Additionally, the generalization metric evaluates a process model's ability to generalize beyond the observed event log by capturing unseen but plausible behavior. It penalizes models that overfit the log by relying too heavily on infrequent or unique traces, thus assessing the model's robustness to future, similar process executions. Simplicity reflects the structural complexity of the model and is measured by comparing the size of the process tree to the number of activities in the event log. Larger models tend to increase perceived complexity and error likelihood. They are defined both in Buijs et al. (2012).

2.3 Simulation

To assess the performance of the process models generated through Process Mining, we conducted simulations of 1000 complete surgical procedure scenarios for each model using Python. The simulation parameters consist of three main components:

1) Process Model Selection: The process models are derived from the process maps obtained via Process Mining, by selecting specific Path and corresponding Activity percentages. Table 5 presents the detailed configurations of the process models used in the simulations.

2) Activity Duration Modeling: For each activity, all observed durations from the event logs are collected into a list. We then automatically fit the most appropriate probability distribution to these durations using Python. The fitting procedure considers a set of candidate distributions. For each activity within a simulated scenario, the duration is sampled from the fitted distribution corresponding to that activity.

3) The transition probability between activities is calculated as the frequency of each transition divided by the total frequency of all outgoing transitions from a given activity.

The evaluation focused on several key performance metrics: (i) the mean procedure duration, (ii) the number of activities per simulated procedure, (iii) the frequency distribution of each activity across all simulated scenes, and (iv) the statistical similarity between simulated and real-world data, measured using p-values from appropriate statistical tests (e.g., Kolmogorov–Smirnov test).

3 RESULTS AND DISCUSSION

3.1 Dataset and Event Logs Generated by Framework

In this work, we utilize the CholecT50 dataset, which was initially created for the recognition of triplets in endoscopic videos of laparoscopic cholecystectomy surgery. Cholecystectomy is the removal of the gallbladder by dividing the cystic duct and cystic artery, and separating the gallbladder from its bed, without injuring the common bile duct, the liver, or its vascular structures.

CholecT50 (Nwoye and Padoy 2023) comprises 50 endoscopic videos of laparoscopic cholecystectomy surgeries, designed to support research in fine-grained action recognition within laparoscopic procedures.

The videos were recorded in Strasbourg, France, and frames were sampled at a rate of 1 frame per second (fps). Each frame is annotated with triplet information capturing surgical actions in the format: "instrument, verb, target". In addition to triplet annotations, the dataset includes surgical phase labels, as illustrated in Figure 1.

In the raw video data and corresponding annotations, up to three triplets may co-occur simultaneously within a single frame, reflecting the complexity and multi-instrument nature of real-world surgical procedures.

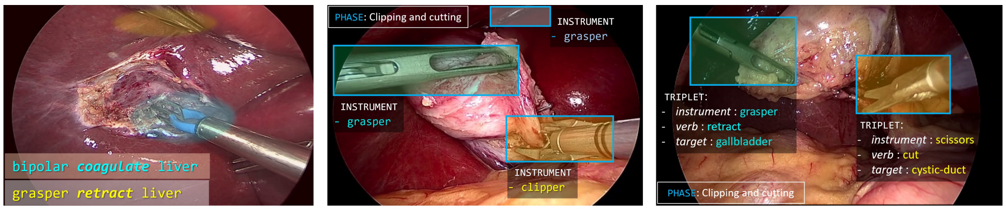

Figure 1: Surgical triplets and phase annotation from the CholecT50 dataset.

We begin by transforming the annotations of the 50 videos from the CholecT50 dataset into event logs that incorporate the defined terms for phases (Definition 2) and gestures (Definition 3) following the framework outlined in Section 2.1. This process results in raw event logs with fine granularity, capturing the intricate details of the cholecystectomy procedures, as illustrated in Table 1.

Table 2: Illustrative example of a pre-processed event log.

ID	begin time	end time	phase	gesture
11	0:25:9	0:25:43	clipping-and-cutting	coagulate,cystic_pedicle
11	0:25:44	0:26:21	clipping-and-cutting	cut,cystic_duct

Next, we pre-process the raw event logs according to the steps outlined in Section 2.2.1. Table 2 presents the pre-processed version of the data shown in Table 1. Compared to Table 1, the gestures in Table 2 have been simplified and cleaned to ensure greater consistency and clarity.

3.2 Process Mining Discovery and Conformance Checking

3.2.1 Process Map

The process maps, generated at varying levels of granularity using Disco, are then analyzed based on the pre-processed event logs.

In Disco, the level of detail in a process map can be adjusted by modifying the parameters for Activity and Path. An Activity represents a single step within the process, while a Path refers to the sequence of activities followed by a specific case throughout the process. For a process map, we can individually adjust the percentage of Activity and Path from 0% to 100% depending on the user's needs to increase the model's flexibility and allow different levels of granular analysis. This ensures the map accurately represents the process without any data leakage. A higher Activity percentage will display more activities, while a higher Path percentage will reveal additional possible pathways in the procedure.

In our case, Activities refer to gestures performed during laparoscopic cholecystectomy, extracted from the pre-processed event logs of the CholecT50 dataset. Each gesture consists of a verb and a target, without instrument details, and is labeled as ge1, ge2, etc.

In our laparoscopic cholecystectomy process map, 0% Paths retains only the strongest (most common) connections, while 100% Paths includes all transitions, even rare ones, resulting in a highly complex, spaghetti-like model. Given the inherent complexity of surgical workflows, even at 0% Paths, some low-frequency paths (e.g., occurring once) can still appear. To maintain clarity and avoid excessive complexity, we recommend keeping the Paths percentage minimal, ideally below 10%, or even 5% for intricate models.

After pre-processing, 50 distinct gestures were identified across 50 laparoscopic cholecystectomies. In Disco, the Activity slider filters gestures by frequency. At 10%, only the 5 most frequent gestures are shown; at 15%, 8 gestures appear, though some phases remain unrepresented. At 25%, 11 gestures are displayed, covering all surgical phases. Higher thresholds add more detail but increase map complexity.

Based on the feedback from surgeons at l'Hôpital Paul-Brousse AP-HP, we selected 25% Activity and 0% Path to ensure the map is representative. These parameter values represent a trade-off between capturing essential variations while keeping the process map clear and manageable. Figure 2 and Figure 3 present the result. Figure 2 displays the total frequency of each gesture and its transitions in 50 cholecystectomy procedures. Figure 3 displays the average duration for each gesture and transition. The time next to each path represents the duration of the transition, which in our case is 1 second due to the frequency of our dataset.

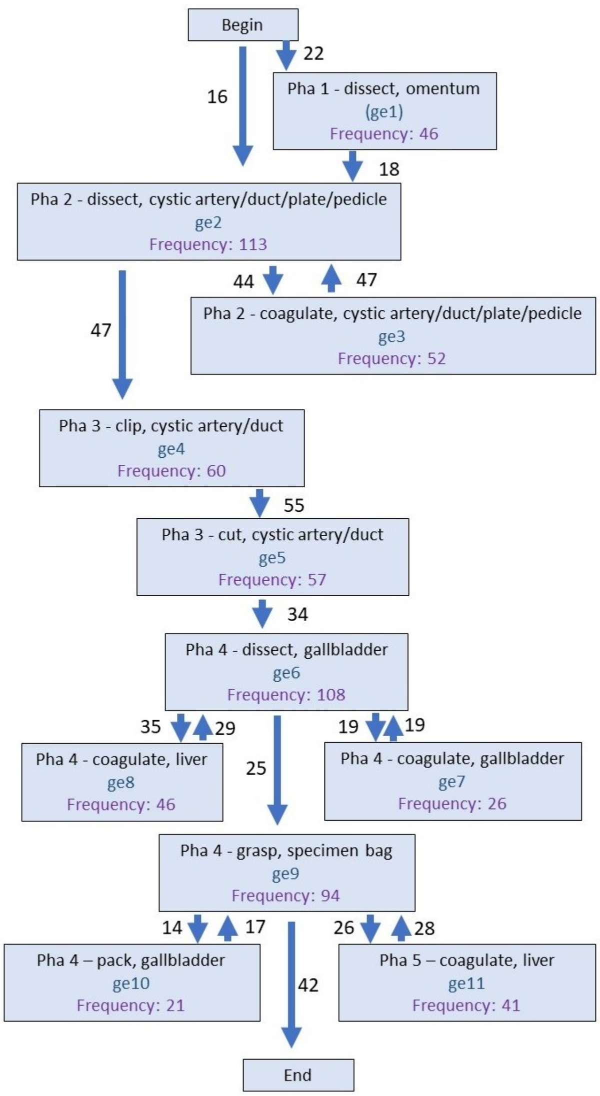

Figure 2: Process map of a laparoscopic cholecystectomy representing the frequency of gestures (25% Activity and 0% Path).

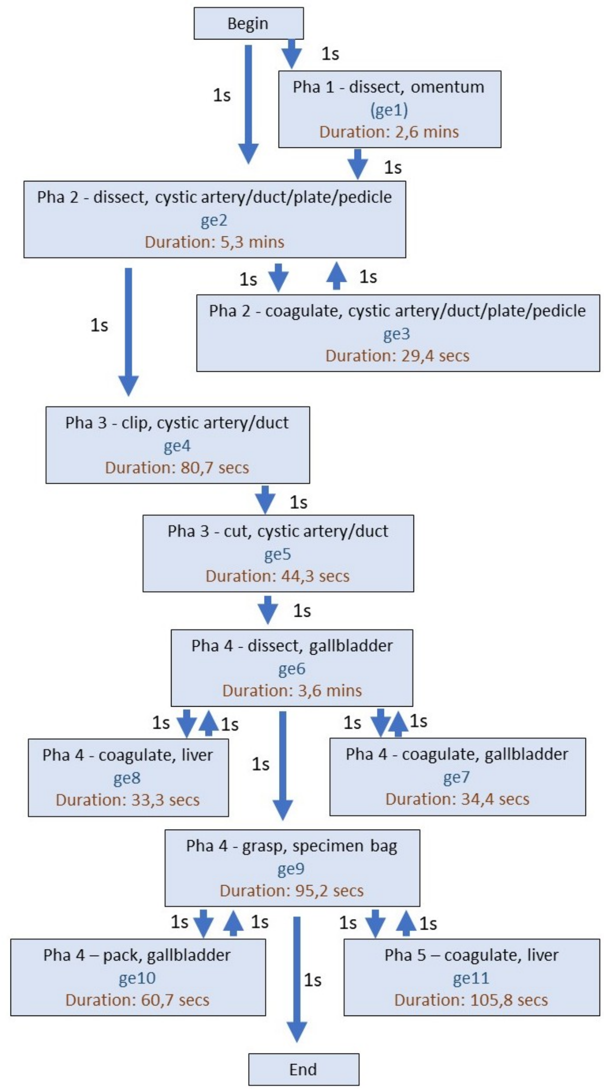

Figure 3: Process map of a laparoscopic cholecystectomy representing the duration of gestures (25% Activity and 0% Path).

Although this process map includes only 25% of the gestures, it effectively represents the key steps in a laparoscopic cholecystectomy surgery. As shown in Table 3, the workflow recommended by expert surgeons is shown in the column "Suggested main gestures" and the workflow generated by process mining is given in the column "Process map main gestures". The map perfectly aligns with the workflow recommended by expert surgeons, making it a realistic and reliable model. Table 4 presents the names of the phases and gestures from Table 3.

Table 3: Process map conformance checking.

Phase (pha)	Suggested main gestures (ge)	Process map main gestures (ge)
pha1	ge1	ge1
pha2	ge2, ge3	ge2, ge3
pha3	ge4, ge5	ge4, ge5
pha4	ge6, ge7, ge9, ge10	ge6, ge7, ge8, ge9, ge10
pha5	ge11	ge11

Table 4: Phase and gesture names.

phase	name	gesture	name
pha1	preparation	ge1	pha1 - dissect, omentum
pha2	calot-triangle-dissection	ge2	pha2 - dissect, cystic artery/duct/plate/pedicle
		ge3	pha2 - coagulate, cystic artery/duct/plate/pedicle
pha3	clipping-and-cutting	ge4	pha3 - clip, cystic artery/duct
		ge5	pha3 - cut, cystic artery/duct
pha4	gallbladder-packaging	ge6	pha4 - dissect, gallbladder
		ge7	pha4 - coagulate, gallbladder
		ge8	pha4 - coagulate, liver
		ge9	pha4 - grasp, specimen bag
		ge10	pha4 - pack, gallbladder
pha5	cleaning-and-coagulation	ge11	pha5 - coagulate, liver

3.2.2 Conformance Checking and Analysis

After identifying the most representative model for a standardized surgical workflow, we modify the parameter configurations (Activity and Path) to conduct Conformance Checking. The performance of each model is evaluated using quantitative metrics, including number of gestures, number of paths, fitness, precision, simplicity, and generalization. The results, presented in Table 5, were generated with the PM4Py (Berti et al. 2019) library in Python.

Table 5: Metrics to evaluate model performance.

Model	Activity (%)	Path (%)	nb gestures	nb paths	fitness		precision		simplicity	generalization
					TBR	align	TBR	align		
Model 1	25	0	11	18	0.86	0.75	0.57	1	0.76	0.83
Model 2	40	0	17	28	0.82	0.78	0.53	1	0.72	0.75
Model 3	50	10	25	47	0.84	0.81	0.53	1	0.65	0.57
Model 4	15	0	143	227	0.41	0.75	0.57	1	0.76	0.83

In Table 5, Models 1 to 3 are based on pre-processed data, with the indicators of fitness, precision, simplicity, and generalization derived from the pre-processed event logs. In contrast, Model 4 is generated using the raw data, and the corresponding indicators are calculated from the original event log.

The metrics from Model 4 indicate that the raw data without pre-processing is highly complex. Even with parameter values set to 15% for Activity and 0% for Path, the various combinations of gestures result in numerous activities and paths. As a result, replay fitness, precision, simplicity, and generalization are significantly lower compared to models generated using pre-processed data. This demonstrates that our pre-processing method is effective in simplifying the data and improving model performance.

As shown in Table 5, reducing the percentages for Activity and Paths leads to improved simplicity and generalization. Models 1, 2, and 3 all exhibit strong fitness, indicating that they accurately capture the actual workflow. Given that these models are based on pre-processed event logs, their alignment precision reaches 1 in all cases. Although TBR precision is lower in certain instances due to the inclusion of infrequent activities and a wide variety of paths, the models still perform well in terms of generalization. It is clear that, to better observe the randomness in a surgical procedure, we can increase the percentage of Activities and Paths, which highlights variations and unpredictability. Notably, Model 1, which was selected by expert surgeons, demonstrates the best overall performance across key metrics.

3.3 Simulation Performance Analysis

Using the generated models (Model 1, Model 2, and Model 3), we simulated 1,000 synthetic surgical procedures for each model. Although the process maps are not strictly acyclic, the simulation regulates repetition by leveraging the transition probabilities from Figure 3. Additionally, to prevent excessive loops, we impose a constraint ensuring that no repetition exceeds five occurrences.

Table 6 presents a comparison of the procedure between the synthetic data and the real raw data after pre-processing. Models 1-3 represent the same models as in Table 5, and Model 5 represents the real raw data after pre-processing. The “number of gestures" refers to the count of distinct, individual gestures, while the “mean number of gestures per scene" represents the total count of gestures within a scene, including any repeated actions.

The mean duration of procedures exhibits minimal differences across different models. To assess the statistical significance of these variations, we conducted a hypothesis test and evaluated the p-values. Since all p-values exceed 0.05, we fail to reject the null hypothesis, indicating that the simulated data closely align with the real observations.These findings indicate that our simulation approach is reliable, particularly for Model 1, which demonstrates the strongest alignment with real-world data.

These results indicate that Models 1-3 exhibit a relatively high standard deviation in procedure duration, with Model 1 and Model 3 also showing shorter average durations compared to the real data. This variability can be attributed to two main factors. First, the process discovery algorithm introduces inconsistencies in loop probabilities and non-representative paths, increasing procedural variability and contributing to the observed standard deviation. Second, the models contain fewer gestures than the actual procedures, which inherently reduces total duration. This simplification prevents the models from fully capturing the procedural complexity, leading to a less accurate representation of real-world surgical timelines. Furthermore, Model 3 includes a higher number of non-representative paths, further affecting its alignment with actual surgical workflows.

However, Models 1-3 have successfully captured the most frequent pathways and activities, suggesting that even the simpler models are capable of representing the core structural elements of the surgical process.

Table 6: Simulation result to evaluate model performance.

	Model 1	Model 2	Model 3	Model 5
nb gestures	11	17	25	40
mean duration (min)	33.0	34.6	32.2	33.7
median duration (min)	27.6	30.4	27.8	32.9
standard deviation duration (min)	28.1	20.6	26.2	10.3
mean nb of gestures/scene	14	16	14	15.9
p-value	0.684	0.590	0.365	

Table 7 presents the frequency of gestures per scene across the three models. As model complexity increases, the number of gestures also rises. For clarity and consistency in comparison, we focus on the 11 key gestures that are present in all three models. This selection ensures that the analysis remains meaningful and allows for a direct comparison of gesture distribution across different model configurations.

Overall, the results consistently show that Model 1 has the smallest deviation from Model 5, further confirming that Model 1 is the most representative in reflecting real surgical workflows.

In particular, gestures ge6 and ge8 appear more frequently in the simulated models than in the real data. This discrepancy arises because the simulated models contain fewer activities overall, making ge6 a necessary step in all process maps. Additionally, the ge6 $\rightarrow$ ge8 loop is the most frequently observed transition. According to surgeons, ge4 and ge5 are the most stable gestures, occurring in every cholecystectomy. This is reflected in Models 1, 2, and 5, where the frequency of ge4 and ge5 exceeds 1, aligning with real-world expectations. However, in Model 3, the frequency falls below 1, which is not realistic. This suggests that Model 3 is overly complex or contains excessive "spaghetti" behavior, making it less representative of the actual procedure.

Table 7: Frequency of gestures per scene.

gesture (ge)	1	2	3	4	5	6	7	8	9	10	11
model 1	0.6	2.0	1.0	1.0	1.0	3.1	0.8	1.3	2.0	0.4	0.6
model 2	1.6	1.9	0.9	1.0	1.0	3.2	0.8	1.4	1.9	0.3	0.6
model 3	1.5	2.0	0.9	0.9	0.9	2.3	0.5	1.2	1.9	0.3	0.6
model 5	0.9	2.3	1.0	1.2	1.1	2.1	0.5	0.9	1.9	0.4	0.8

3.4 Errors and adverse events analysis based on the framework

Process maps and simulations are done base on our framework. When the level of detail in the process map is increased—with more granular activities and paths—it becomes possible to identify potential surgical anomalies, particularly through collaboration with surgeons. For instance, an error appears on the process map as an additional isolated phase and as an extra gesture compared to the global process map.

Some gestures containing terms like “coagulation” might indicate a response to unexpected bleeding. If such actions appear infrequently or outside typical phases, they can signal the occurrence of an adverse event. Likewise, the emergence of additional gestures not present in a reference process map may imply a deviation caused by complications, where these extra actions serve to manage or mitigate the issue.

The simulation component also provides valuable metrics, including average durations and standard deviations for each gesture. Abnormal phase durations—whether unusually short or unusually long—as well as additional pathways or gesture sequences, may indicate errors, whether cognitive or technical. Instrument failure can also be reflected through these parameters.

By integrating structural analysis, gesture frequency, and temporal metrics, our approach facilitates the precise identification of deviations within surgical workflows. This level of insight enhances the detection of critical issues and deepens our understanding of how complications impact procedural dynamics, ultimately supporting safer and more efficient surgical practices. However, this analysis has not yet been formally validated. Further discussions with surgeons are necessary to develop quantitative methods for a more robust evaluation.

4 CONCLUSION AND PERSPECTIVES

Our work highlights three key contributions: 1) a formal mathematical framework for generating event logs that comprehensively describe the entire surgical procedure, applicable for modeling a variety of surgical types, 2) an event log pre-processing method designed to streamline and simplify the surgical process, and 3) the application of this approach to real-world data from laparoscopic cholecystectomy surgeries using process mining techniques, and the simulation of synthetic surgical procedures. These contributions provide a structured foundation for better understanding surgical workflows, improving decision-making, and enhancing the analysis of procedures.

A limitation is that, at this stage, our framework has been validated using only a single surgical procedure—cholecystectomy—due to the limited availability of annotated datasets for other types of surgeries. Future work will focus on validating the framework across a broader range of surgical procedures and integrating surgeon expertise into the design and evaluation process to enhance the system's robustness and generalizability. Another limitation is the lack of quantitative evaluation of the event log pre-processing method. The current assessment relies primarily on informal discussions with surgeons, which introduces potential subjective bias due to the absence of a standardized validation protocol. In the future, we aim to develop a standardized evaluation framework with objective metrics and structured validation —such as completeness, accuracy, and consistency—by multiple surgeons. This will reduce subjective bias and improve the reliability of the pre-processing method.

In our future research, we anticipate focusing on three key objectives. Firstly, by generating an event log from real-time detection of triplets within surgical videos, we will pre-process the event data using our proposed method. Next, we aim to classify these detailed triplets into the broader activities depicted in our process map. This approach will enable us to emphasize the surgeon's primary actions at any given moment, providing clearer insights into the procedural flow. Secondly, we will perform a more detailed analysis of gesture triplets to quantitatively identify errors and adverse events, and further to predict the occurrence of these adverse events, and forecast upcoming activities. By leveraging our process map, simulations, and machine learning techniques, we aim to develop a real-time risk warning system for surgeons, enhancing patient safety during the procedure. Finally, we will incorporate patient-specific information to further refine the methods outlined earlier. By including factors such as the patient's medical background, health status, and other relevant data, we can enhance the system's ability to recognize potential complications and anticipate next steps more accurately. Additionally, analyzing how procedural variations across different patient profiles impact the framework's performance will provide valuable insights into its adaptability. This personalized integration will enable more precise guidance during surgery, supporting better decision-making and minimizing risks for each patient.

REFERENCES

Anastasiou, D., Y. Jin, D. Stoyanov, and E. Mazomenos. 2023. "Keep Your Eye on the Best: Contrastive Regression Transformer for Skill Assessment in Robotic Surgery". *IEEE Robotics and Automation Letters* 8(3):1755–1762.

Bawa, V. S., G. Singh, F. KapingA, I. Skarga-Bandurova, E. Oleari, and A. L. et al. 2021. "The SARAS Endoscopic Surgeon Action Detection (ESAD) Dataset: Challenges and Methods". *arXiv preprint arXiv: 2104.03178*.

Berti, A., and W. van der Aalst. 2020. "A Novel Token-Based Replay Technique to Speed Up Conformance Checking and Process Enhancement". *arXiv preprint arXiv: 2007.14237*.

Berti, A., S. J. van Zelst, and W. van der Aalst. 2019. "Process Mining for Python (PM4Py): Bridging the Gap Between Process- and Data Science". *arXiv preprint arXiv: 1905.06169*.

Bieck, R., K. Heuermann, M. Pirlich, J. Neumann, and T. Neumuth. 2020. "Language-Based Translation and Prediction of Surgical Navigation Steps for Endoscopic Wayfinding Assistance in Minimally Invasive Surgery". *International Journal of Computer Assisted Radiology and Surgery* 15(12):2089–2100.

Buijs, J. C. A. M., B. F. van Dongen, and W. M. P. van der Aalst. 2012. "On the Role of Fitness, Precision, Generalization and Simplicity in Process Discovery". In *Confederated International Conferences: CoopIS, DOA-SVI, and ODBASE 2012*. September 10^{th}-14^{th}, Rome, Italy, 305–322.

Chung, R., D. Hunter-Smith, R. Spychal, V. Ramakrishnan, and W. Rozen. 2017. "A Systematic Review of Intraoperative Process Mapping in Surgery". *Gland Surgery* 6(6):715–725.

Franke, S., J. Meixensberger, and T. Neumuth. 2015. "Multi-Perspective Workflow Modeling for Online Surgical Situation Models". *Journal of Biomedical Informatics* 54:158–166.

Gao, Y., S. S. w. Vedula, C. E. Reiley, N. Ahmidi, B. Varadarajan, and H. C. L. et al. 2014. "The JHU-ISI Gesture and Skill Assessment Working Set (JIGSAWS): A Surgical Activity Dataset for Human Motion Modeling". In *Modeling and Monitoring of Computer Assisted Interventions (M2CAI) – 2014 MICCAI Workshop*. September 14^{th}-18^{th}, Boston, US.

Garrow, C. R., K. Kowalewski, L. Li, M. Wagner, M. W. Schmidt, and S. E. et al. 2021. "Machine Learning for Surgical Phase Recognition: A Systematic Review". *Annals of Surgery* 273(4):684–693.

Gawria, L., R. Rosenthal, H. van Goor, and S. Dell-Kuster. 2021. "Classification of Intraoperative Adverse Events in Visceral Surgery". *Surgery* 171(6):1570–1279.

Günther, C. W., and A. Rozinat. 2012. "Disco: Discover Your Processes". In *Proceedings of the Demonstration Track of the 10th International Conference on Business Process Management (BPM 2012)*. September 4th, Tallinn, Estonia, 40-44.

Lavanchy, J. L., S. Ramesh, D. Dall'Alba, C. Gonzalez, P. Fiorini, and B. M.-S. et al. 2023. "Challenges in Multi-Centric Generalization: Phase and Step Recognition in Roux-en-Y Gastric Bypass Surgery". *arXiv preprint arXiv: 2312.11250.*

Lira, R., J. Salas-Morales, R. D. L. Fuente, R. Fuentes, M. Sepúlveda, and M. A. et al. 2018. "Tailored Process Feedback Through Process Mining for Surgical Procedures in Medical Training: The Central Venous Catheter Case". In *2018 Business Process Management Workshops*. September 9th-14th, Sydney, Australia, 163–174.

Mascagni, P., A. Vardazaryan, D. Alapatt, T. Urade, T. Emre, and C. F. et al. 2022. "Artificial Intelligence for Surgical Safety: Automatic Assessment of the Critical View of Safety in Laparoscopic Cholecystectomy Using Deep Learning". *Annals of Surgery* 275(5):955–961.

Neumann, J., A. Uciteli, T. Meschke, R. Bieck, S. Franke, and H. H. et al. 2022. "Ontology-Based Surgical Workflow Recognition and Prediction". *Journal of Biomedical Informatics* 136:104240.

Neumuth, T. 2017. "Surgical Process Modeling". *Innov Surg Sci* 2(3):123–137.

Neumuth, T., N. Durstewitza, M. Fischera, G. Strauß, A. Dietza, and J. M. et al. 2006. "Structured Recording of Intraoperative Surgical Workflows". *Medical Imaging 2006: PACS and Imaging Informatics, SPIE* 6145:54–65.

Nwoye, C., and N. Padoy. 2023. "Data Splits and Metrics for Method Benchmarking on Surgical Action Triplet Datasets". *arXiv preprint arXiv: 2204.05235*.

Rozinat, A., and W. van der Aalst. 2008. "Conformance Checking of Processes Based on Monitoring Real Behavior". *Information Systems* 33(1):64–95.

Sharma, S., C. I. Nwoye, D. Mutter, and N. Padoy. 2023. "Surgical Action Triplet Detection by Mixed Supervised Learning of Instrument-Tissue Interactions". *arXiv preprint arXiv: 2307.09548*.

Suliburk, J. W., Q. M. Buck, C. J. Pirko, N. N. Massarweh, N. R. Barshes, and H. S. et al. 2019. "Analysis of Human Performance Deficiencies Associated with Surgical Adverse Events". *JAMA Network Open* 2(7):e198067.

van der Aalst, W. 2016. *Process Mining - Data Science in Action*. 2nd ed. Heidelberg: Springer.

van der Aalst, W., A. Adriansyah, and B. van Dongen. 2012. "Replaying History on Process Models for Conformance Checking and Performance Analysis". *WIREs Data Mining and Knowledge Discovery* 2(2):182–192.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L.Jones, and A. N. G. et al. 2017. "Attention is All You Need". *arXiv preprint arXiv: 1706.03762*.

Weijters, A., and J. Ribeiro. 2011. "Flexible Heuristics Miner (FHM)". In *2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*. April 11th-15th, Paris, France.

AUTHOR BIOGRAPHIES

YIYU WANG is currently a Ph.D student at Mines Saint-Étienne, working on a project related to the development of a digital twin for surgical procedures. Her research is part of a collaboration between Mines Saint-Étienne, the BOPA Chair, and AP-HP. Her email address is yiyu.wang@emse.fr and his website is https://www.linkedin.com/in/yiyu-wang-157500a4/.

VINCENT AUGUSTO Vincent Augusto received his Ph.D. from École des Mines de Saint-Étienne in 2008. He is currently a professor of industrial engineering at EMSE, affiliated with the Center for Health Engineering and the CNRS UMR 6158 LIMOS. His research focuses on modeling, simulation, and optimization of healthcare systems and their supply chains. His e-mail is augusto@emse.fr and his web addresses is https://www.mines-stetienne.fr/author/augusto/

CANAN PEHLIVAN a faculty member in Industrial Engineering at Yeditepe University, Istanbul, and previously held positions at IMT Atlantique and IMT Mines Albi in France. Her research focuses on modeling and optimizing complex systems under uncertainty, particularly in healthcare logistics and data science. She holds a Ph.D. from École des Mines de Saint-Étienne. Her email address is canan.pehlivan@yeditepe.edu.tr and her website is https://yeditepe.edu.tr/en/academic-staff/3812.

JULIA L. FLECK is a permanent faculty member at Mines Saint-Étienne, Center for Biomedical and Healthcare Engineering. Her research focuses on healthcare data analytics, dynamic system modeling, and simulation-based optimization. She holds a Ph.D. in Systems Engineering from Boston University and has received awards from CISE, NCI, AACR, and the Alexander von Humboldt Foundation. Her e-mail address is julia.fleck@emse.fr and her website is https://www.mines-stetienne.fr/author/julia-fleck/.

NESRINE MEKHENANE is a pediatric surgeon and PhD student at Université Paris Cité's Learning Planet Institute. Her research at the BOPA Chair focuses on intraoperative adverse events and their postoperative impact, aiming to improve surgical safety and performance. Her e-mail address is mekhenane.nm@gmail.com and her website is https://www.linkedin.com/in/nesrine-mekhenane-4196b3131/?originalSubdomain=fr.