Validation of knowledge acquisition for surgical process models.

Thomas Neumuth, Pierre Jannin, Gero Strauss, Juergen Meixensberger, Oliver Burgert

To cite this version:

Thomas Neumuth, Pierre Jannin, Gero Strauss, Juergen Meixensberger, Oliver Burgert. Validation of knowledge acquisition for surgical process models.. Journal of the American Medical Informatics Association, 2008, 16 (1), pp.72-80. 10.1197/jamia.M2748 . inserm-00344260

HAL Id: inserm-00344260
https://inserm.hal.science/inserm-00344260

Submitted on 4 Dec 2008

Manuscript type

Research paper

Title

Validation of Knowledge Acquisition for Surgical Process Models

Authors

Thomas Neumuth[1], MEng
Pierre Jannin[2,3,4], PhD
Gero Strauss[1,5], MD, PhD
Juergen Meixensberger[1,6], MD, PhD
Oliver Burgert[1], PhD

Departments

[1] University of Leipzig, Innovation Center Computer Assisted Surgery (ICCAS), Leipzig, Germany
[2] INSERM, U746, Faculty of Medicine, Rennes, France
[3] INRIA, VisAGeS Unit/Project, Rennes, France
[4] University of Rennes I, CNRS, UMR 6074, IRISA, Rennes, France
[5] University Hospital Leipzig, Department of Otorhinolaryngology, Leipzig, Germany
[6] University Hospital Leipzig, Department of Neurosurgery, Leipzig, Germany

Contact

Thomas Neumuth
University of Leipzig
Innovation Center Computer Assisted Surgery (ICCAS)
Semmelweisstr. 14
D-04103 Leipzig
Germany
thomas.neumuth@iccas.de
Phone: +49-341-97-12010
Fax: +49-341-97-12009

ABSTRACT

Objective: Surgical Process Models (SPMs) are models of surgical interventions. The objectives of this study are to validate acquisition methods for Surgical Process Models and to assess the performance of different observer populations.

Design: The study examined 180 SPM of simulated Functional Endoscopic Sinus Surgeries (FESS), recorded with observation software. About 150,000 single measurements in total were analyzed.

Measurements: Validation metrics were used for assessing the granularity, content accuracy, and temporal accuracy of structures of SPMs.

Results: Differences between live observations and video observations are not statistically significant. Observations performed by subjects with medical background gave better results than observations performed by subjects with technical background. Granularity was reconstructed correctly by 90%, content by 91%, and the mean temporal accuracy was 1.8 *s*.

Conclusion: The study shows the validity of video as well as live observations for modeling Surgical Process Models. For routine use, we recommend live observations due to their flexibility and effectiveness. If high precision is needed or the SPM parameters are altered during the study, video observations are the preferable approach.

Keywords

Validation [V03.950],

Validation Studies as Topic [E05.337.925],

Observation [H01.770.644.241.800],

Surgical Procedures, Operative [E04]

I. INTRODUCTION

Surgery is a clinical specialty with a long history, but surgical techniques are learned in an apprentice-master model that leads to several surgical schools treating the same disease in different ways.

There is no explicit methodology available, which prevents an objective comparison of surgical strategies at a fine-grained level. Using process models with fine-grained descriptions of surgical interventions as the processes, surgeons get a powerful tool for the discussion of different surgical approaches and scientifically sound process models of their surgical work steps.

A detailed Surgical Process Model (SPM) may help in understanding a procedure, especially in difficult cases. Such a detailed model must be available for a broad variety of similar interventions to cover all clinically relevant deviations from the standard procedure.

Furthermore, a collection of verified and valid SPMs of surgical processes, especially for rare cases, could help in the implementation of new surgical techniques (e.g., minimally invasive surgery or computer assisted interventions) that require a detailed understanding of the intervention course in order to optimally assist the surgeon.

Surgical Process Models may be used to facilitate the development of technical components for surgical assist systems (SAS) (1; 2) and to support standardization efforts for desired functionalities of SAS, such as future extensions of Digital Imaging and Communications in Medicine (DICOM) for surgery (3; 4).

The ultimate purpose of SPMs is the generation of these descriptions for technical requirement analysis, evaluation, and systems comparison.

For the modeling, data must be at an adequate level of granularity. The modeling must address behavioral, anatomical, and pathological aspects and surgical instruments (5).

Accuracy is crucial. This is why the modeling must be rigorously validated. The objective of our study was the validation of data acquisition for SPMs. The research question was, "*How accurate are observations of Surgical Processes by human observers?*" We designed a rigorous validation strategy that assessed the accuracy of SPMs that were acquired from simulated interventions in a controlled environment. We studied several validation criteria: granularity, content accuracy, and temporal accuracy, using video and live observation as data acquisition strategies and using medical and technical students as the observer populations. For assessing the validation criteria, metrics have been defined and applied to the SPMs. Secondary questions of interest included time to complete observations, the subjective workload estimation of observers, and the level of surgical knowledge required by observers.

II. BACKGROUND

The amount of information available from surgical processes is large and complex, although the knowledge of the surgeon is mostly implicit and hidden from formal assessment. Data may be acquired by using two main strategies: sensor systems or, in a more classical way, human observation.

Only a few sensor technologies are available for application in the sensitive operating room (OR) environment. These technologies are not suitable for uniform acquisition of data such as work-step information, inter-device communication, human-device behavior or inter-human behavior for modeling due to missing information models, network communication, and interfaces. It is necessary to use human recognition and perception capabilities for parts of the data acquisition, which is a common strategy in biomedicine (6) and empirical social sciences (7).

Only a few approaches for modeling surgical processes are described in the literature. MacKenzie et al. (8) performed iteratively top-down and bottom-up analyses for assessing laparoscopic Nissen fundoplications for training residents. The data acquisition was performed based on video observations. Münchenberg et al. (9) modeled surgical procedures of Frontal Orbital Advancements to treat craniosynostosis for the purpose of planning and technical intra-operative support for the surgeon; the data acquisition methodology was not mentioned. Jannin et al. (5) modeled surgical procedures in the context of multimodal image-guided neurosurgery. Data were acquired pre- and post-operatively via questionnaires. None of the previous work validated the data acquisition process. Validation of data acquisition in the clinical domain has been performed by Vawdrey et al. (10), who assessed the data quality of ventilators operated by respiratory therapists. Data were acquired by electronic medical records. Rosenbloom et al. (11; 12) evaluated the interface terminologies of clinical interfaces. These studies were adequate for medical patient records, but they did not provide an overall measure of the accuracy.

Working definitions used in this article are strongly related to Business Process Modeling and Workflow Management Systems (13). By analogy, we define a *Surgical Process (SP)* as *a set of one or more linked procedures or activities that collectively realize a surgical objective within the context of an organizational structure defining functional roles and relationships*. The surgical objective is the correction of an undesirable state of the patient's body, which is performed in the organizational structure of a hospital. The responsible surgeon coordinates the performance of the surgical procedure. We define a *Surgical Process Model (SPM)* as *a simplified pattern of a Surgical Process that reflects a predefined subset of interest of the SP in a formal or semi-formal representation* (14). The working definitions are also provided to clarify the relationship to the frequently used term Surgical Workflow, which relates to the performance of a Surgical Process with support of a Workflow Management System (15).

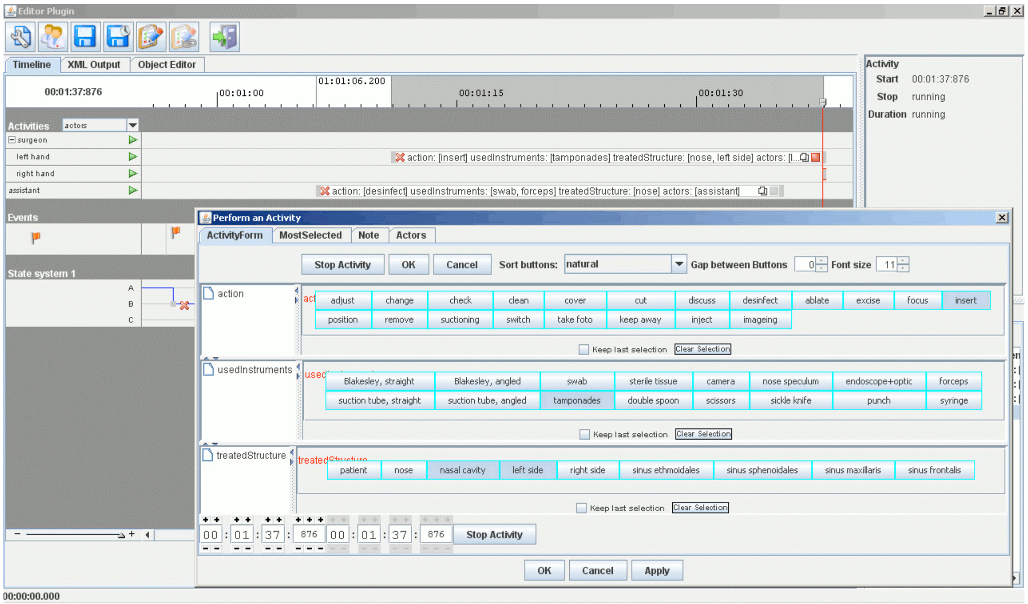
The objective of this work was to perform a validation study for assessing data acquisition results of SPMs by human observers with specialized software. The SPs consisted of simulations of Functional Endoscopic Sinus Surgeries (FESS).

III. METHODS

First, the data acquisition software and its underlying ontological concepts are introduced. Then, the experimental setup and post-processing are described. The notion of variables that might influence a validation study for SPMs is discussed in a separate section. These variables were divided into three groups: extraneous variables that need to be held constant, independent variables that were manipulated according to the experimental design, and dependent variables that were affected by the manipulation of the independent variables. Finally, the validation metrics quantified the manipulation effects.

A. Data Acquisition Software and Fundamental Concepts

The data were acquired with a JAVA software application, the Surgical Workflow Editor (16; 17). The objective of the software is to devise ontological concepts used for describing the SP to the observer and to ask him or her for the instantiation of these concepts to create an observation protocol. A screenshot of the Surgical Workflow Editor is shown in Figure 1.



Activity
Start 00:01:37:876
Stop running
Duration running

action: [insert] usedInstruments: [tamponades] treatedStructure: [nose, left side] actors: [l...

action: [desinfect] usedInstruments: [swab, forceps] treatedStructure: [nose] actors: [assistant]

Sort buttons: natural
Gap between Buttons 0
Font size 11

00 : 01 : 37 : 876 00 : 01 : 37 : 876

Figure 1: A screenshot of the Surgical Workflow Editor.

The data acquisition process begins with the definition of the structure of the SPM. The structure is described by the *structural ontology* and specifies how information of the SP is represented in the SPM. During actual data acquisition, specific concepts of the observed SP, described by the *content ontology* (e.g., surgical actions, participants, or instruments) are instantiated by the observer.

Our structural ontology contains three types of flow objects (18): *activities*, *state transitions*, and *events*. Each SPM consists of these flow objects.

Activities represent manual work steps performed during the interventions. To structure their content, we used the factual perspectives for workflow schema proposed in (19), modified them, and added the spatial perspective. An activity consists of five perspectives, which decompose the observer's view into various viewpoints:

- the functional perspective describing *what* is done in a surgical work step;

- the organizational perspective describing *who* is performing a work step;
- the operational perspective to describe *instruments* used in performing a work step;
- the spatial perspective describing *where* a work step is performed;
- and the behavioral perspective describing *when* a work step is performed.

Perspectives are extended by perspective attributes. They decompose perspectives further (e.g., indicating that a surgeon is performing a work step with his right hand, where both perspective attributes belong to the organizational perspective). More examples may be found in Table 2.

For recording work steps with no measurable time extension, we defined the concepts of *state transitions* and *events*. State transitions are changing variables between predefined values, e.g. observable on monitors in the operating room or the phases of an intervention. Events might describe the content of messages, such as the surgeon's instruction to administer a drug. State transitions and events each include the functional and behavioral perspective.

The purpose of the content ontology is to determine the correct intervention-specific relations between perspective contents, e.g., for *suctioning* (functional perspective) only a *suction tube* (operational perspective) may be used. The development of the content ontology is based on expert knowledge.

B. Experimental Setup

The validation procedure consisted of recording the simulated SP and comparing the resulting SPM to a reference afterward. The main steps for the experiments are shown in Table 1.

The validation was performed on simulated Functional Endoscopic Sinus Surgeries (FESS) as SPs. A FESS intervention has the objective of removing polyps from nasal cavities. During the core part of the intervention, the surgeon holds an endoscope with one hand while his or her other hand performs the actual work steps.

The processes to be simulated were built based on real FESS intervention recordings. For the study, the FESS-specific content ontology contained concepts of two participants, two used body parts, twelve actions, 13 surgical instruments, three instrument attributes, and seven treated structures. The concepts were chosen based on routine daily clinical terminology.

In preparation for defining the Gold Standards for the study, flow object patterns of the structural ontology and work step information of the content ontology were used to construct FESS-specific terminology. This was composed of flow object patterns for 41 different activities, which represented surgical work steps, three state transitions, and three events. Pattern examples are shown in Table 2. The patterns of the Gold Standard terminology were used to design three different Gold Standards as simulation scripts that served as references for assessing the accuracy of the simulations. First, the prototype Gold Standard SPM was generated. It contained a typical sequence of work steps with predefined timestamps. From this, two more simulation scripts, the second and the third Gold Standard, were derived by adding noise. The noise additions included modifying the treatment order of nasal cavities, increasing work speed, and switching the surgeon and assistant roles temporarily. The created simulation scripts were checked by two ENT-surgeons for clinical realism. Each of the simulations was 21 minutes in length and was limited to 60 to 90 activities.

Table 1: Step descriptions for the Experimental Design.

Experiment Preparation	1 Select structural ontology of the SPM 2 Define concepts of content ontology 3 Design terminology patterns for Gold Standards 4 Design Gold Standards 5 Speak and record audio representation of Gold Standards 6 Perform simulations without observers and video record them for later use in the video observations and to serve as the Bronze Standards for video observations 7 Code these videos as XML-protocols
Data Acquisition Sessions	8 Perform simulations with live observations, record these simulations on video as the Bronze Standards for the live observations 9 Perform observations of video simulations (recorded in Step 6)
Post Processing	10 Register Bronze Standard protocols to respective Gold Standard protocols 11 Register observation protocols to Bronze Standard protocols 12 Extract Bronze Standard terminology pattern and observation terminology pattern 13 Register Bronze Standard terminology pattern to Gold Standard terminology pattern 15 Register observation terminology pattern to Bronze Standard terminology pattern
Observation Validation	16 Calculate vm_i for observation protocols by comparing observation protocols to corresponding Bronze Standards as references 17 Perform statistical analysis

Simulation Validation	18 Calculate vm_i for Bronze Standards by comparing them to corresponding Gold Standards as references 19 Perform statistical analysis

The three different Gold Standards were spoken and recorded as audio files, containing detailed instructions for the work steps to be performed by the actors. One simulation for each Gold Standard was performed without observers, recorded with multiple video cameras, synchronized, and cut as a video representation of the simulation for later use in video observations. After these protocols of the Bronze Standards were coded in XML-format, they served as reference SPMs for validation of the simulations against the Gold Standard simulation scripts and for validation of the video observations by medical and technical observers.

Table 2: Example flow object pattern used for Gold Standard terminology.

perspective	**perspective attribute**	**example activity**	**example activity**	**example activity**	**example event**	**example state transition**
functional	action	disinfect	dissect	insert	event 1	A -> B
organizational	participant	assistant	surgeon	surgeon	-	-
	used bodypart	-	right hand	left hand	-	-
operational	main instrument	swab	Blakesley	nose speculum	-	-
	supporting instrument	forceps	-	-	-	-
	property of main instrument	-	straight	-	-	-
spatial	anatomical structure patient	patient	patient	patient	-	-
	anatomical structure nose	nose	nasal cavity	nasal cavity	-	-
	anatomical structure side	-	right side	right side	-	-
	anatomical structure nasal cavity	-	c. ethmoidales	-	-	-
behavioral	starttime	00:00:00	00:00:00	00:00:00	00:00:00	00:00:00
	stoptime	00:00:20	00:00:20	00:00:20	-	-

The data acquisition sessions were performed in the ICCAS-demonstration OR in Leipzig and consisted of one educational session day for the uniform training of the observers and three data acquisition sessions days for each observer group. The educational session introduced the purpose of data acquisition for SPM, the Surgical Workflow Editor software, the surgical objectives of FESS procedures, the typical intervention course, and the content ontology to the observers to establish a common context of use. The objective of this session was to simulate the situation for observing real Surgical Processes, where the observer needs to understand the procedure in depth before he or she begins to record data.

Ten observers performed nine observations for each data acquisition strategy with tablet-PCs. Video observations were conducted based on the performance of the simulation scripts without observers during the experimental preparation. The live observations were based on live simulations by the actors. The work steps of the live simulations were recorded by endoscope and two video cameras. After the recorded videos were synchronized, they served as the Bronze Standards for the live observations.

After each observation, the observers performed a workload assessment, the Task Load Index (TLX) test (20) of the National Aeronautics and Space Administration (NASA), for describing their subjective workload feeling, and they continued with acquiring data by the respective other data acquisition strategy of video and live observation. Additionally, the observers were required to pass a knowledge test twice per data acquisition day.

C. Post-Processing

Before analysis, post-processing was required to link each SPM to its reference. Post-processing started with the manual association of each flow object of an observer protocol to its corresponding reference flow object in a Bronze Standard protocol. By performing this association between flow objects, registration matrices of the protocols were created.

Subsequently, the protocols and registration matrices were transferred to a database, where the terminology patterns of the Bronze Standard terminologies and the terminologies of the observations were extracted from the respective protocols and automatically compared to each other.

Finally, the validation metrics were calculated, and the statistical analysis was performed. The statistical analysis was done using multivariate Generalized Linear Models (GLM) for the data acquisition strategy and the observer population. The simulated Gold Standard and the repetition of the measurements were considered as covariates. All statistical tests were performed with a significance level of 0.05 and computed with the SPSS software (SPSS Inc., Chicago, IL).

D. Analysis

Preliminary identification of factors that may influence such a validation is required. Inspired by Shah and Darzi (21), we classified the influence factors for SPs by distinguishing *surgeon-specific* factors S, *technology-specific* factors T, and *patient-specific* factors P (see Table 3 for an overview of used symbols). Generally, we consider a surgical treatment to be a Surgical Process SP, which is a function of the outlined factors.

Technically, a Surgical Process SP is recorded by a measurement system, influenced by *measurement system* factors M. The measurement system factors M therefore influence the representation of a Surgical Process SP by a Surgical Process Model SPM: $M : SP \to SPM$.

Table 3: Symbol Overview.

Symbol	Meaning
$s_i \in S$	Surgeon-specific factors
$p_i \in P$	Patient-specific factors
$t_i \in T$	Technology-specific factors
$m_i \in M$	Measurement systems factors
$sp_i \in SP$	Surgical Process
$spm_i \in SPM$	Surgical Process Model
$vm_i \in VM$	Validation metrics

Additionally, we arranged the influence factors and the validation metrics into three groups: extraneous, independent, and dependent variables.

Extraneous Variables

Surgeon-specific factors $s_i \in S$ that influence a surgical process are mainly the human factors of surgeons (21) and the staff in the OR. Two actors performed the simulations of our study: one played the role of the surgeon, and the other played a combined role of assistant and scrub nurse. The surgeon-specific factors were not considered separately because the actors were directed to follow the work steps of the audio representations of the Gold Standards closely.

Surgical Processes vary due to the use of different surgical tools, instruments, and devices. The technology factors $t_i \in T$ were also considered as extraneous variables, not separated, and constant for the study due to the predefinition of instrument names, usage times, and order by the simulation scripts.

We introduced the patient-specific factor group $p_i \in P$ to indicate the patient's current situation, his or her history or future, and his/her specific anatomical and pathological circumstances. We considered the patient-specific factors group as an extraneous variable and constant because the simulations were performed on 3D-Rapid Prototyping models, which all use the same template.

For the study, we focused on data acquisition by human observers, supported by the Surgical Workflow Editor. We classified the measurement system into influence factors $m_i \in M$. We considered m_1 as structural ontology, m_2 as content ontology, and the Surgical Workflow Editor as observation support software m_3. For the observer, we opted for the factors m_4 as the observation workload and m_5 as the knowledge level of the observer. We considered $m_1,\ldots,m_5$ as extraneous variables, assuming them to be constant.

Independent Variables

The focus of this study was the validation of accuracy differences in SPM resulting from different data acquisition strategies as factor m_6, and different observer populations as factor m_7. Data acquisition by observers may be performed intra-operatively as live observation or post-operatively from videos. The observer populations (m_7) consisted of ten individuals: five medical students (4th-6th semester) and five technical students. None of them was experienced in SPM recording. Each of them performed nine observations each for video and live situations (3 observations for each of the 3 Gold Standards) in random order. Live simulations were performed 18 times because only 5 observers could observe simultaneously due to space limitations. Each Live simulation was recorded on video to serve as the Bronze standard for observations recorded in that particular session.

Dependent Variables

We defined six different metrics for validation within the context of Surgical Process Modeling: $vm_i \in VM$. The six metrics were designed to cover the facets that characterized the quality of data acquisition for SPM and were complementary to each other. For an overview of the computational order of the validation metrics, the reader is referred to Figure 2.

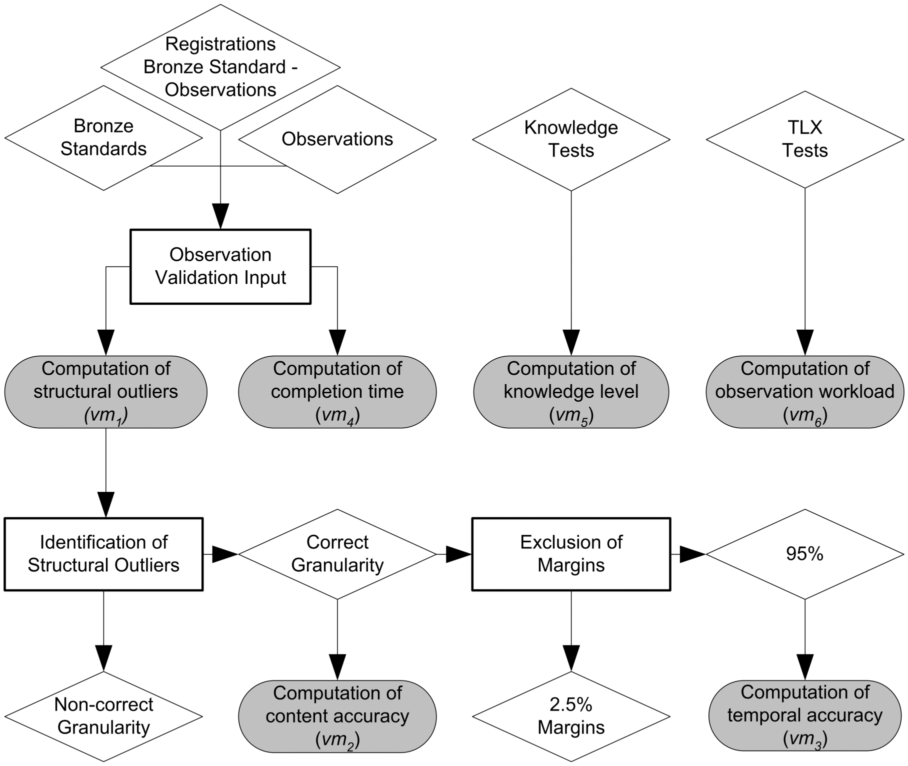


Figure 2: Computation of Validation Metrics.

- The measurement of the structural outliers of an observation (vm_1) focused on the compliance of granularity guidelines of an observation compared to its reference. Structural outliers are measured as the percentage of outlier flow objects to all flow objects in the observation. Structural outliers are defined in the context of the study as *structural parts of a Surgical Process Model,* spm_1*, contradicting the structural parts of the reference Surgical Process Model,* spm_2*, by assuming that both* $spm_i \in SPM$*, are triggered by the same structural ontology and represent the same instance of a Surgical Process* $sp_i \in SP$. In Figure 3, the various interpretations for structural outliers are shown. Flow objects registered in a 1:1 relationship to their referential flow objects reflected the *correct granularity*. Flow objects that appeared in the observation, but not in the reference, were denoted as *additional observations*. Flow objects in the reference that were not recorded were *missing observations*. Flow objects represented as multiple activities in the reference, but represented as one activity in the observations, were denoted as *decreased granularity*. One flow object of the reference represented as multiple flow objects in the observations represented *increased granularity*. A mixture of multiple flow objects in the reference and multiple flow objects in the observations represented *mixed granularity*. Before applying the validation metrics vm_2 and vm_3, all flow objects not representing the *correct* granularity were removed because only similar granularities may be compared.

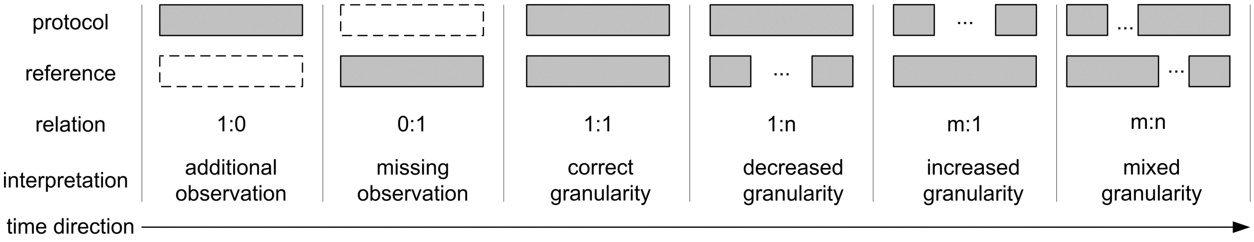


Figure 3: Types of structural outliers.

- The validation metric vm_2 estimates the validation criterion of content accuracy of an observation. Content accuracy was defined as the *distance of conceptual instances in a Surgical Process Model,* spm_1*, compared to the conceptual instances in the reference Surgical Process*

Model, spm_2*, assuming that both* $spm_i \in SPM$*, are triggered by the same content ontology and represent the same instance of a Surgical Process* $sp_i \in SP$. The metric vm_2 calculates a similarity measure for the content accuracy of perspective attributes of an observation compared to the corresponding perspective attributes of the reference. Based on the perspective attributes, a content accuracy value for each perspective was calculated. Subsequently, this procedure was repeated for activities, state transitions, and events as flow objects of the overall protocol.

- The measurement of the temporal accuracy of an observation (vm_3) indicates the temporal distance between durations of activities that was calculated based on the start timestamps and stop timestamps of registered flow objects. The calculation has been done after the rejection of temporal outliers corresponding to abnormal excessively large time deviations due to hardware failures.
- The measurement of the completion time of an observation (vm_4) is calculated as a ratio based on the time needed to create the observation protocol with respect to the duration of the reference. It begins when the first activity is set and ends when the final protocol is saved after review by the observer.
- Experimental conditions were controlled by the knowledge level of an observer (vm_5*)* and the assessment of workload observation from observer feedback (vm_6). vm_5 was used to check the learning curve of the observers. This parameter is expressed as the percentage of correct answers on the knowledge tests. The software feedback vm_6 assessed the workload of the observation task by subjective ratings of the criteria of the NASA Task Load Index (20).

IV. RESULTS

Detailed results for structural outliers are presented in Table 4. Medical students recorded granularity correctly 92.3% (±5.7%) of all activities in the reference in live observations and 92.5% (±5.2%) in video observations, as opposed to 86.6% (±6.8%) in live observation and 91.2% (±6.7%) in video observation for technical students. The mode of data acquisition was significant. Video observations were more accurate in terms of correct granularity. Missing activities and activities with decreased granularity were more prevalent in the live observations. The observer population also had a significant influence on structural outliers. For instance, medical student observers were more likely than technical students to record granularity correctly.

Table 4: Study results for structural outliers (vm_1)

	m_6 live		**m_6 video**		**significance**	**significance**
(granularity) [%]	m_7 **medical**	m_7 **technical**	m_7 **medical**	m_7 **technical**	m_6	m_7
additional observation of activities	1.5±0.9	1.9±1.7	1.3±1.6	3.6±5.9		F=11.4, p=0.001
missing observation of activities	2.1±2.1	3.7±3.1	1.2±1.3	2.5±4.2	F=6.0, p=0.02	F=9.1, p=0.001
correct granularity of activities	92.3±5.7	86.6±6.8	92.5±5.2	91.2±6.7	F=7.4, p=0.01	F=16.2, p<0.001
decreased granularity of activities	0.4±1.1	1.3±1.9	1.7±2.8	0.8±2.2	F=14.5, p<0.001	F=7.1, p=0.01
increased granularity of activities	4.9±4.8	8.6±4.9	4.2±4.4	4.1±4.4		
mixed granularity of activities	0.03±0.02	0.1±0.3	-	0.3±1.3		
correct granularity of events	57.1±19.3	51.9±19.7	23.5±35.0	37.2±41.6	F=31.4, p<0.001	
correct granularity of state transitions	87.4±10.5	89.1±7.3	68.2±39.5	74.0±35.2	F=18.4, p<0.001	

The overall content accuracy for activities is 91.5% (±5.4%) in live and 91.5% (±5.3%) in video observation by medical observers. Content accuracy for activities was 88.9% (±2.6%) for live and 87.4% (±8.9%) for video observations by technical students (cp. Table 5). The data acquisition type had no significant influence on content accuracy for activities, but video observations produced significantly lower content accuracy for events.

Table 5: Study results for content accuracy (vm_2)

	m_6 **live**		m_6 **video**		**significance**	**significance**
(content accuracy) **[%]**	m_7 **medical**	m_7 **technical**	m_7 **medical**	m_7 **technical**	m_6	m_7
functional perspective of activities	93.1±6.7	87.1±5.6	92.9±7.8	82.7±14.1		F=36.7, p<0.001
organizational perspective of activities	97.8±3.0	97.7±2.5	98.3±2.4	96.9±6.5		
operational perspective of activities	88.3±7.3	84.5±4.4	89.2±6.2	84.8±13.2		F=11.8, p=0.001
spatial perspective of activities	70.8±9.0	70.8±8.8	70.8±10.6	68.9±9.8		F=12.5, p<0.001
total content accuracy of activities	91.5±5.4	88.9±2.6	91.5±5.3	87.4±8.9		F=15.1, p<0.001
total content accuracy of events	93.2±20.2	96.8±10.7	72.8±41.8	71.9±38.7	F=25.3, p<0.001	
total content accuracy of state transitions	98.1±5.1	99.1±3.5	99.0±4.6	98.2±8.7		

The mean absolute value for temporal accuracy was less than 2 s. for all factors. The data acquisition type had only low significant influence on temporal accuracy (cp. Table 6). The observer population had a significant influence on temporal accuracy.

Data acquisition from videos required 80 % more time than data acquisition for live observations. No significant differences were found in completion time between medical and technical observers.

Table 6: Study results for temporal accuracy (vm_3) and completion time (vm_4)

	m_6 **live**		m_6 **video**		**significance**	**significance**
	m_7 **medical**	m_7 **technical**	m_7 **medical**	m_7 **technical**	m_6	m_7
temporal accuracy [s]	1.7±0.4	1.9±0.5	1.5±0.3	1.8±0.8	F=4.9, p=0.03	F=12.1, p=0.001
completion time	1.6±0.1	1.2±0.2	2.3±0.5	2.1±0.4	F=389.7, p<0.001	

Nearly all workload criteria, and also the estimation of one's own performance, were rated higher for live observations (cp. Table 7). All workload criteria were rated more demanding by the technical observer population.

The Gold Standards had a significant influence only on the number of structural outliers. Medical students scored 94.1 % correct answers on the knowledge tests, while technical students scored 78.3 %.

Table 7: Study results for observation workloads (vm_5)

	m_6 **live**		m_6 **video**		**significance**	**significance**
(criteria)	m_7 **medical**	m_7 **technical**	m_7 **medical**	m_7 **technical**	m_6	m_7
Effort	60.0±21.2	69.6±11.3	51.6±24.5	67.2±11.5	F=6.5, p=0.01	F=19.7, p<0.001
Frustration	49.0±18.0	46.8±11.7	38.2±19.2	42.3±14.0	F=14.6, p<0.001	
Mental Demand	58.7±19.6	69.6±15.0	54.9±20.6	70.0±16.9		F=18.9, p<0.001
Performance	47.1±16.0	48.8±13.4	37.8±17.0	46.5±16.8	F=8.7, p=0.004	F=5.3, p=0.02
Physical Demand	41.5±23.7	61.0±12.7	35.4±20.4	57.6±10.1		F=52.3, p<0.001
Temporal Demand	73.7±17.2	77.4±14.0	44.0±19.0	56.8±16.5	F=39.2, p<0.001	F=11.0, p=0.001
Total Workload	62.4±14.2	67.1±10.3	49.1±15.3	60.3±10.0	F=39.2, p<0.001	F=17.8, p<0.001

V. DISCUSSION

A. Significance of the Work

To the best of our knowledge, this is the first extensive validation of knowledge acquisition for Surgical Process Models in the medical domain; no previous comparisons of live and video observations were found in the literature. Based on a rigorous control of influence factors (22) affecting Surgical Processes and the definition of validation metrics, a complex and rigorous study has been designed and conducted. Former studies validated observations based on inter-observer agreements (23; 24) and used correlations as indirect metrics to quantify the agreements. For valid observations, a threshold of 85% inter-observer agreement is reported (24). Our results were calculated based on direct comparison of observation results with the observed process as a reference.

We found that observers generally record accurately, robustly, and reproducibly. The accuracy of data acquisition for live or video observation was comparable.

The results for structural outliers give a measurement for the assessment of the granularity of an SPM. Nearly all of the activities were observed with correct granularity. In contrast to the observer population, the influence of the data acquisition type had low significance. We may conclude that differences between video and live observations of activities regarding the validation criterion of structural outliers are not statistically significant.

The observations for state transitions and events were unacceptable. Seemingly, the concentration of the observers was focused on the interventional site and on the monitor displaying the endoscope view, not on the monitor displaying the state transitions and the events. This might be compensated by introducing acoustic signals that highlight them for the observers or perhaps even for the surgeons themselves in the operating room.

Content accuracy showed no significant differences between the data acquisition strategies. Thus, we conclude that live and video observations may be considered similar regarding the validation criterion of content accuracy. The medical observers recorded the activity content significantly better than the technical observers. Low accuracy occurred mainly because students could not properly assess the spatial perspective. None of the perspectives showed a significant difference by data acquisition strategy. However, there is still work to do to develop a method for direct global content accuracy comparison that accounts for the positive and negative variation in granularity.

The completion time was far longer when recording from videos than from live simulations. This result is especially interesting when considering the comparable outcomes of live and video observations for granularity, content and temporal accuracy.

The small increases of the measured ratios of the knowledge tests during the data acquisition sessions showed the effectiveness of the training sessions. We trained the technical observers in a similar manner to the medical observers, but they were not able to attain the same level of knowledge. Values for all workload criteria were lower for video observations. Technical observers rated all workload criteria more demanding than medical observers. This may have influenced the lower granularity, content accuracy, and temporal accuracy of the technical observers (compared to the medical observers).

The validity of the simulation was checked by comparing the Bronze Standards to the Gold Standards. In the context of the study, the Gold Standards were held as the objective and unequivocal models that were, by definition, the simulation scripts. The Bronze Standards were viewed as the best results that the observers could achieve. The simulation validation was used to cross-check the validity of the simulations. For instance, the mean ratio of correct granularity of the Bronze Standards was 97.2%, and the mean content accuracy was 96.3%. Thus, the actors introduced only a very few simulation errors. The realism of the Gold Standards did not comprise each possible aspect of a FESS but worked as a robust simulation base for a continuous repetition of the Surgical Processes. Furthermore, the experimental design was used to hold constant the influences P, S, and T of the patient, the surgeon, and the technology on the Surgical Process. To study s_i, for instance, one would evaluate differences resulting from different personal ‘styles’ of surgeons or different education levels that may result in different procedure courses (if the same intervention was performed twice on the same patient by two surgeons).

We referred to the reasons for variation in Surgical Processes caused by using different surgical instruments or devices as technology-specific factors. The limited number of instruments representing the technology-factors T represents a restriction, but this was ignored to facilitate the work of the actors. Advantages and disadvantages of live and video observations are shown in table 8. The choice of the data acquisition strategy does not only depend on the objectives of clinical studies to be performed, but also on the available ressources for observation.

Table 8: Advantages and disadvantages of video and live observation

	Video observation	Live observation
Advantages	- temporal resolution can be increased by pausing the video - knowledge acquisition can be repeated, if the structural ontology or the content ontology need to be altered - workload of the observers is less than in live observations	- instantaneous access to information and possibility to ask for hidden information, e.g. surgical decisions - dynamic repositioning of the observer in the OR, e.g. if line of sight is blocked
Disadvantages	- not all information for the SPM can be captured on video - field of view can be blocked by intervention participants - high costs of time for data acquisition	- loss of information due to distraction or increased workload of the observer - limited temporal resolution

B. Limitations of the Present Study

Limitations to our work include:

- The observations were based on simulated Surgical Processes. Of course, simulations are not 100% realistic. Ideally, the study would have used real surgical cases, but that would have prevented control of many factors that could affect results.
- The validation metrics used for assessing the quality of data acquisition for Surgical Process Models need to be validated in additional studies.

C. Implications for Future Work

In this study, we proposed an innovative experimental design for the validation of knowledge acquisition for Surgical Process Models. This validation method may be extended and modified, and it may be used to validate modifications of S , T , P , or M . The actual design could be proposed as validation support for other more technical approaches such as those described in (25) or (26). We are unaware of any research results that delineate which measures of observations for SPMs are acceptable and which are not; we plan to address this topic in future work. Additionally, knowledge bases could be developed and validated to support and facilitate observations for SPMs. If a knowledge base were used that contained information about which actions can be performed with a specific surgical instrument, e.g., Blakesley, the Surgical Workflow Editor could propose the action *dissect* to the observer and ask for confirmation, as soon as *Blakesley* is chosen as instrument.

VI. CONCLUSIONS

The results of this study can provide useful guidance for the design of other studies to acquire knowledge for SPMs. We demonstrated the validity of video as well as live observations for modeling SPMs and that trained human observers generally record accurately, robustly, and reproducibly. We also outlined the areas where human observations were less accurate; future work should concentrate on these areas. Live observations of state transitions and events should be supported by a technical sensor system with intra- or post-observation synchronization to the observer protocol or an acoustic signal that draws the attention of the observer to the displaying device. For routine use, we recommend live observations due to their relative speed, flexibility, and effectiveness. If high precision is needed or SPM parameters, such as the ontologies used, are altered during the study, video observations are preferable. Trained medical students can be highly accurate observers.

This study also provided an estimate of the expected accuracy of modeling surgical processes by observation. We identified influence factors that can serve as basis for designing similar studies, in which, for example, the work of surgeons with varying levels of experience or the effect of the use of different surgical instruments might be compared. Our validation metrics can be applied to studies with comparable reference standards, but producing such references is a significant challenge.

Modeling surgical processes is undoubtably a challenge for the observers. Special advance training is required, for example, for live observations in the operating room. The study setup, of course in a narrower context, as well as the validation metrics, can be used to benchmark the level of observers in training. For instance, if it were important for the observers to achieve a certain degree of content accuracy before they can participate in clinical studies, the methods used in this study could be used to measure their proficiency.

ACKNOWLEDGEMENTS

We thank the team that supported the study at the Innovation Center Computer Assisted Surgery, University of Leipzig: T. Baumgärtner, M. Ceschia, M. Czygan, F. Eckhardt, J. Georgi, D. Goldstein, A. Hoffmeier, K. Hornoff, B. Kaschek, N. Ritter, and S. Schumann. We would also like to thank I. Hertel and M. Hofer of the Department of ENT-Surgery, University Hospital Leipzig for their support.
ICCAS is funded by the German Federal Ministry of Education and Research (BMBF) and the Saxon Ministry of Science and Fine Arts (SMWK) within the scope of the Unternehmen Region with grant numbers 03 ZIK 031 and 03 ZIK 032.

References

[1] Lemke HU, Vannier MW. The operating room and the need for an IT infrastructure and standards. Int J Comput Assist Radiol Surg. 2006;1(3):117–122.

[2] Cleary K, Kinsella A, Mun SK. OR2020 Workshop report: operating room of the future. In: Lemke HU, Inamura K, Doi K, Vannier MW, Farman AG, editors. CARS 2005: Proceedings of the 19th International Conference on Computer Assisted Radiology and Surgery; 2005 June 22-25; Berlin, Germany. Amsterdam: Elsevier; 2005. p. 832–838.

[3] Lemke HU. Summary of the white paper of DICOM WG24: DICOM in Surgery". In: Horii SC, Andriole KP, editors. SPIE Medical Imaging 2007. vol. PACS and Imaging Informatics. Bellingham, WA: SPIE; 2007. p. 651603.

[4] Burgert O, Neumuth T, Gessat M, Jacobs S, Lemke HU. Deriving DICOM Surgical Extensions from Surgical Workflows. In: Horii SC, Andriole KP, editors. SPIE Medical Imaging 2007. vol. PACS and Imaging Informatics. Bellingham, WA: SPIE; 2007. p. CID65150B.

[5] Jannin P, Raimbault M, Morandi X, Riffaud L, Gibaud B. Model of surgical procedures for multimodal image-guided neurosurgery. Comput Aided Surg. 2003;8(2):98–106.

[6] Payne PRO, Mendonca EA, Johnson SB, Starren JB. Conceptual konwledge acquisition in biomedicine: A methodological review. J Biomed Inform. 2007 March;40(5):582–602.

[7] Kromrey H. Empirical Social Sciences (in German). 11th ed. Stuttgart: Lucius & Lucius; 2006.

[8] MacKenzie CL, Ibbotson AJ, Cao CGL, Lomax A. Hierarchical decomposition of laparoscopic surgery: a human factors approach to investigating the operating room environment. Min Invas Ther All Technol. 2001;10(3):121–128.

[9] Münchenberg J, Brief J, Raczkowsky J, Wörn H, Hassfeld S, Mühling J. Operation planning of robot supported surgical interventions. In: IROS 2000: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems; 2000 Oct 30 -Nov 5; Takamatsu, Japan. vol. 1. IEEE; 2000. p. 547–552.

[10] Vawdrey DK, Gardner RM, Evans RS, Orme Jr JF, Clemmer TP, Greenway L, et al. Assessing Data Quality in Manual Entry of Ventilator Settings. J Am Medical Inform Assoc. 2007 February;14(3):295–303.

[11] Rosenbloom ST, Miller RA, Johnson KB, Elkin PL, Brown SH. Interface Terminologies: Facilitating Direct Entry of Clinical Data into Electronic Health Record Systems. J Am Medical Inform Assoc. 2006

February;13:277–288.

[12] Rosenbloom ST, Miller RA, Johnson KB, Elkin PL, Brown SH. A Model for Evaluating Interface Terminologies. J Am Medical Inform Assoc. 2007 October;15:65–76.

[13] Workflow Management Coalition. Terminology & Glossary. Winchester, UK: Workflow Management Coalition; 1999. Document Number WFMC-TC-1011, Document Status -Issue 3.0.

[14] Neumuth T, Trantakis C, Eckhardt F, Dengl M, Meixensberger J, Burgert O. Supporting the analysis of intervention courses with surgical process models on the example of fourteen microsurgical lumbar discectomies. In: Lemke HU, Inamura K, Doi K, Vannier MW, Farman AG, editors. CARS 2007: Proceedings of 21st International Conference on Computer Assisted Radiology and Surgery; 2007 June 27-30; Berlin, Germany. Berlin: Springer; 2007. p. 436–438.

[15] Jannin P, Morandi X. Surgical models for computer-assisted neurosurgery. NeuroImage. 2007;37(3):783–91.

[16] Neumuth T, Strauß G, Meixensberger J, Lemke HU, Burgert O. Acquisition of Process Descriptions from Surgical Interventions. In: Bressan S, Küng J, Wagner R, editors. DEXA 2006: Proceedings of 17th International Conference on Database and Expert Systems Applications; 2006 Sep 4-8; Krakow, Poland. vol. LNCS 4080. Berlin, Heidelberg: Springer; 2006. p. 602–611.

[17] Neumuth T, Durstewitz N, Fischer M, Strauß G, Dietz A, Meixensberger J, et al. Structured Recording of Intraoperative Surgical Workflows. In: Horii SC, Ratib OM, editors. SPIE Medical Imaging 2006. vol. PACS and Imaging Informatics. Bellingham, WA: SPIE; 2006. p. CID61450A.

[18] White SA. Introduction to BPMN; BPTrends, www.bptrends.com, July 2004.

[19] Jablonski S, Bussler C. Workflow Management -Modeling Concepts, Architecture and Implementation. International Thomson Computer Press; 1996.

[20] Hart SG, Staveland LE. Development of NASA-TLX (task load index): results of empirical and theoretical research. In: Hancock PA, Meshkati N, editors. Human Mental Workload. North-Holland: Elsevier Science Publishers B.V.; 1988. p. 53–71.

[21] Shah J, Darzi A. The impact of inherent and environmental factors on surgical performance in laparoscopy: a review. Min Invas Ther & Allied Technol. 2003;12(1-2):69–75.

[22] Jannin P, C G, Maurer Jr CR. Model for defining and reporting reference-based validation protocols in

medical image processing. Int J Comput Assist Radiol Surg. 2006;1(2):63–73.

[23] Reneman MF, Fokkens AS, Dijkstra PU, Geertzen JHB, Groothoff JW. Testing lifting capacity: validity of determining effort level by means of observation. Spine. 2005;30(2):E40–6.

[24] Baglio ML, Baxter SD, Guinn CH, Thompson WO, Shaffer NM, Frye FHA. Assessment of interobserver reliability in nutrition studies that use direct observation of school meals. J Am Diet Assoc. 2004;104(9):1385–92.

[25] Sudra G, Speidel S, Müller-Stich BP, Becker A, Ott M, Dillmann R. Situation modelling and situation recognition for a context-aware augmented reality system. In: Lemke HU, Inamura K, Doi K, Vannier MW, Farman AG, editors. CARS 2007: Proceedings of 21st International Conference on Computer Assisted Radiology and Surgery; 2007 June 27-30; Berlin, Germany. Berlin: Springer; 2007. p. 434–436.

[26] Padoy N, Horn M, Feussner H, Berger MO, Navab N. Recovery of surgical workflow: a model-based approach. In: Lemke HU, Inamura K, Doi K, Vannier MW, Farman AG, editors. CARS 2007: Proceedings of 21st International Conference on Computer Assisted Radiology and Surgery; 2007 June 27-30; Berlin, Germany. Berlin: Springer; 2007. p. 481–482.