Graëff *et al. Art Int Surg* 2023;3:147-59
DOI: 10.20517/ais.2023.12

Artificial Intelligence Surgery

Original Article

Open Access

Check for updates

The preliminary stage in developing an artificial intelligence algorithm: a study of the inter- and intra-individual variability of phase annotations in internal fixation of distal radius fracture videos

Camille Graëff[1,2], Thomas Lampert[1], Jean-Paul Mazellier[2], Nicolas Padoy[1,2], Laela El Amiri[3], Philippe Liverneaux[1,3]

[1]ICube, University of Strasbourg, Illkirch-Graffenstaden 67400, France.
[2]IHU, Strasbourg 67000, France.
[3]Department of Hand Surgery, Strasbourg University Hospitals, Strasbourg 67200, France.

Correspondence to: Dr. Camille Graëff, ICube, University of Strasbourg, CNRS, 300 boulevard Sébastien Brant, Illkirch-Graffenstaden 67400, France. E-mail: camille.graeff@etu.unistra.fr

How to cite this article: Graëff C, Lampert T, Mazellier JP, Padoy N, El Amiri L, Liverneaux P. The preliminary stage in developing an artificial intelligence algorithm: a study of the inter- and intra-individual variability of phase annotations in internal fixation of distal radius fracture videos. *Art Int Surg* 2023;3:147-59. https://dx.doi.org/10.20517/ais.2023.12

Received: 5 Apr 2023 **First Decision:** 8 Jun 2023 **Revised:** 15 Jun 2023 **Accepted:** 27 Jun 2023 **Published:** 4 Jul 2023

Academic Editor: Andrew A. Gumbs **Copy Editor:** Pei-Yun Wang **Production Editor:** Pei-Yun Wang

Abstract

Aim: As a preliminary stage in the development of an artificial intelligence (AI) algorithm for surgery, this work aimed to study the inter- and intra-individual variability of phase annotations in videos of minimally invasive plate osteosynthesis of distal radius fractures (MIPO). The main hypothesis was that the inter-individual variability was almost perfect if Cohen's kappa coefficient (k) was ≥ 81% overall; the secondary hypothesis was that the intra-individual variability was almost perfect if the F_1-score (F_1) was ≥ 81%.

Methods: The material comprised 9 annotators and three annotated MIPO videos with 5 phases and 4 sub-phases. Each video was presented 3 times to each annotator. The method involved analysing the inter-individual variability of annotations by computing k and F_1 from a reference annotator. The intra-individual variability of annotations was analysed by computing F_1.

Results: Annotation anomalies were noticed: either absences or differences in phase and sub-phase annotations. Regarding the inter-individual variability, an almost perfect agreement between annotators was observed because

k ≥ 81% for the three videos. Regarding the intra-individual variability, F_1 ≥ 81% for most phases and sub-phases with the nine annotators.

Conclusion: The homogeneity of annotations must be as high as possible to develop an AI algorithm in surgery. Therefore, it is necessary to identify the least efficient annotators (measurement of the intra-individual variability) to provide them with individual training and a personalised annotation rhythm. It is also important to optimise the definition of the phases, improve the annotation protocol and choose suitable training videos.

Keywords: Algorithm, annotation, phase detection, distal radius fracture, artificial intelligence, inter-individual variability, intra-individual variability, surgical video

INTRODUCTION

Several algorithms have been developed for phase detection in videos of soft tissue surgery such as cholecystectomy[1], sleeve gastrectomy[2], laparoscopic sigmoidectomy[3], and endoscopic myotomy[4]. Surgical phase recognition algorithms are useful both in the operating rooms to help junior surgeons and assistants locate themselves in the surgical workflow, and outside for quick and easy video review and for statistical reports. They can also serve as a building block for other applications.

Preliminary to the development of these algorithms is the precise definition of the different phases to detect. For these algorithms to be usable by the greatest number of people, a consensus between experts is essential.

In orthopaedic and traumatology surgery, some authors have recently defined by consensus, among a hundred surgeons, 5 phases and 3 sub-phases in minimally invasive osteosynthesis of distal radius fractures (MIPO)[5]. This procedure is performed for the surgical treatment of wrist fractures. It starts with a small longitudinal incision (less than 30 mm) in the volar aspect of the wrist, and a plate is then attached to the radius with several screws to reduce and maintain the fracture.

Based on this definition of phases and sub-phases, it is possible to annotate a video database. The annotation of a video with phases is a process that consists of associating each image of the video with the name of the corresponding phase of the surgery. Such annotation is necessary to have examples that will be used to train an artificial intelligence (AI) algorithm to automatically recognize those phases by itself on new unseen videos: the annotated video database is designed to act as a reference guide for the algorithm, enabling him to make informed decisions.

As this annotation is performed by human subjects, it is possible that there is an inter- and intra-individual variability of annotations. It is only when this variability has been shown to be low that it will be possible to use this annotated database to develop a phase detection algorithm. Indeed, we need consistent and accurate annotations if we want the algorithm to be able to learn properly to automatically detect the surgical phases.

In this context, the aim of this work was to study the inter- and intra-individual variability of phase annotations in MIPO videos.

The main study hypothesis was that the inter-individual variability was almost perfect if Cohen's kappa coefficient was greater than or equal to 81% overall. The secondary hypothesis was that the intra-individual variability was almost perfect if the F_1-score was greater than or equal to 81%.

METHODS

Written consent from 4 patients was obtained to use the videos of their distal radius fracture internal fixation and the local ethics committee gave its approval for this study (CE 2021-159).

The material comprised 9 annotators, 4 videos, 1 annotation software, and 1 computer code developed specifically for the study.

The 9 annotators were 1 AI researcher and 8 hand surgeons. A surgeon's expertise in distal radius fracture internal fixation is usually measured using five levels: from I (beginner) to V (expert), depending on their training and experience[6]. The AI researcher, with no surgical experience yet familiar with distal radius fracture internal fixation videos, was considered level I. The level of the 8 surgeons varied, with 2 level II, 4 level III, 1 level IV, and 1 level V.

Four MIPO[7] videos (V0, V1, V2, V3) were downloaded from a dedicated software platform (qvident®, Caresyntax™, Boston, Massachusetts, United States). The videos were recorded using an HD video camera built into the surgical lighting (TruVidia™ Wireless®, TRUMPF Medizin Systeme GmbH + Co. KG™, Saalfeld, Germany).

Software (MOSaiC, IHU Strasbourg) was used to annotate video data with labels, i.e., associate the name of the corresponding phase with each frame of the video. The labels were defined according to a previous study[5]. They consisted of 5 phases and 3 sub-phases 3a-3b-3c (V0, V1, V3) or 4 sub-phases 3a-3b-3c-3opt (V2) [Table 1].

A computer code in Python™ was developed by the AI researcher to extract the annotations and then analyse them.

The method involved four steps: annotator training, video annotation (V1, V2, V3), extraction, and analysis of the annotations.

For their training, the annotators, informed of the objective of the study, had to read three documents. These documents consisted of a tutorial on the use of the annotation software [Supplementary Material 1], an annotation protocol for distal radius fracture internal fixation videos [Supplementary Material 2], and a table summarising the different annotation labels with visual cues defining their beginning and end [Supplementary Material 3]. To become familiar with the protocol and the annotation software, each annotator first had to practice on a test video (V0), whose annotations were then analysed with the AI researcher.

For annotating of the videos, a non-collaborative group was created on the MOSaiC software with the three videos downloaded 3 times and then numbered from 1 to 9 in an order that did not respect an arithmetical progression to avoid any bias of habit (V1, V2, V3, V1, V3, V2, V1, V3, V2). Annotators did not have access to the annotations of the other annotators. Once each video annotation was completed and validated, it was no longer possible to modify its annotation.

The annotations were retrieved from the MOSaiC software using software developed in Python™. Each annotation included several associated pieces of information from which only the following were used in this study: the identifier of the annotator, the type and name of the label, the time corresponding to the

Table 1. Phases and sub-phases labels with their beginning and their end for the three videos of distal radius fracture fixation

Label		Image	Beginning	End
PHASES				
1	Installation		Beginning of the video	Beginning of the approach phase
2	Approach		Apparition of the scalpel	Beginning of the fixation phase
3	Fixation		Beginning of the introduction of the plate sub-phase	Beginning of the verification phase
4	Verification		First image, after the last moment when the surgeon touched the proximal screws, where neither surgical instruments nor the operator's hands are visible	Beginning of the closure phase
5	Closure		Apparition of the adson-forceps and the needle with the needle holder	End of the video
SUB-PHASES (of the fixation phase)				
3a	Introduction of the plate		Apparition of the plate	Beginning of the distal fixation sub-phase
3b	Distal fixation		Apparition of the K-wire	Beginning of the proximal fixation sub-phase
3c	Proximal fixation		First image, after the last moment when the surgeon touched the distal screws, where neither surgical instruments nor the operator's hands are visible	Beginning of the verification phase

3opt	Modification of the plate positioning (not present in all surgeries)		Plate grabbed to be removed from the wrist	Beginning of the distal fixation sub-phase

beginning of the annotation and its duration in milliseconds [Figure 1].

The analysis of the annotations through the computation of different metrics focused, on the one hand, on the inter-individual variability of the annotations and, on the other hand, on the intra-individual variability of the annotations [Figure 2]. In the case of inter-individual variability, the metrics were computed between the annotations of an annotator chosen as a baseline (the AI researcher or annotator 0) and those of each of the other annotators. For intra-individual variability, the metrics were computed, for each annotator and each video, between the first annotation of this video (chosen as a baseline) and the other two annotations of this same video.

The three metrics computed were Cohen's kappa coefficient k (%) [Equation (1)], the F_1-score F_1 (%) [Equation (2)], and the difference δ (in absolute value), between annotators, of the start time of each phase (s). Several metrics exist to quantify the agreement between annotators, including Cohen's kappa coefficient[8], the percentage of agreement, the ICC intra-class correlation coefficient, or the F_1-score[9-11]. The percentage agreement and the intra-class correlation coefficient ICC were not used in our study, because the use of the former is sometimes criticised in this situation[12] and the latter is more complex to compute.

$$k = \frac{P_o - P_e}{1 - P_e} = \frac{2 \cdot (TP \cdot TN - FN \cdot FP)}{(TP + FP) \cdot (FP + TN) + (TP + FN) \cdot (FN + TN)} \quad (1)$$

$$F_1 = \frac{TP}{TP + 0.5 \cdot (FP + FN)} \quad (2)$$

TP: true positives; FP: false positives; TN: true negatives; FN: false negatives; P_o: observed relative agreement; P_e: hypothetical probability of agreement by chance.

The computation of k and F_1 involves the terms true positive (TP), false positive (FP), true negative (TN) and false negative (FN), which are commonly used in computer science to evaluate the performance of classification models (models which goal is to associate each data point to one class). Thus, TP represents the number of correctly predicted positive points, TN represents the number of correctly predicted negative points, FP represents the number of falsely predicted positive points, and FN represents the number of falsely predicted negative points.

We used k to analyse the inter-individual variability, as it allows us to measure the degree of agreement between two independent annotators, taking into account the possibility that agreement occurs by chance. However, this coefficient cannot be used for the analysis of intra-individual variability because it does not satisfy the independence hypothesis. Regarding the overall inter-individual variability for each video, the

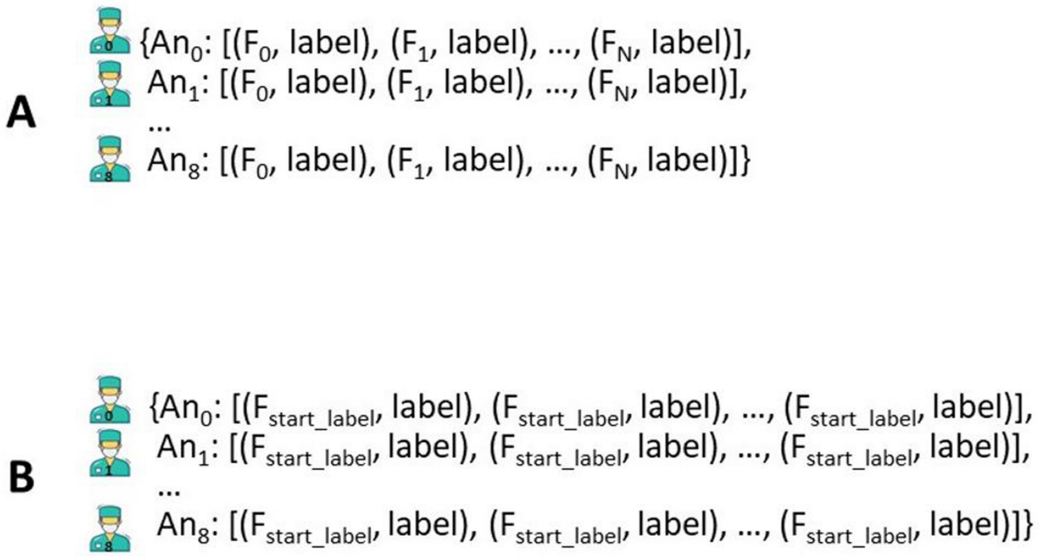

Figure 1. The data structure used to store the annotations for each video (see Table 1 for details on the labels). (A) when the data are the phase annotations; (B) when the data are the phase start time. An_i: The i^{th} annotator; F_i: the i^{th} frame of the video.

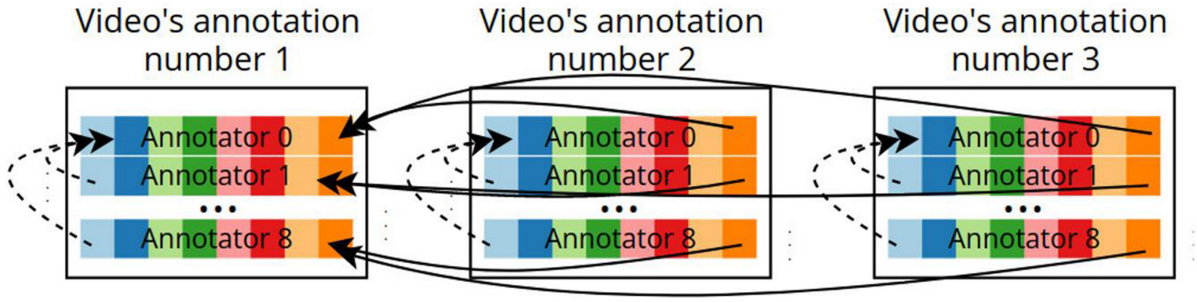

Figure 2. Visualisation of the annotations that are compared for the analysis of the inter-individual (dotted arrows) and intra-individual (full arrows) annotation variability (only one video among the three is shown here, but it is the same principle for the two others).

level of agreement between the annotators was considered almost perfect if $k \geq 81\%$, substantial if $61\% \leq k \leq 80\%$, moderate if $41\% \leq k \leq 60\%$, fair if $21\% \leq k \leq 40\%$, slight if $1\% \leq k \leq 20\%$, and less than chance if $k \leq 0\%$[8].

We used F_1 to analyse both the inter- and intra-individual variability. To quantify the agreement between annotators, it is usually considered that the higher the F_1 score, the higher the agreement, but no threshold value is proposed in the literature. We therefore decided to take, as for the k coefficient, the value of 81% as the threshold defining an almost perfect agreement.

We used δ to analyse both the inter- and intra-individual variability. The videos were recorded at a rate of 60 frames per second and the annotation in the MOSaiC software was done every ten frames (default setting). As a result, it was only possible to annotate six frames per second. As the human eye does not perceive a major change between three consecutive frames in the three videos of the study, the difference δ,

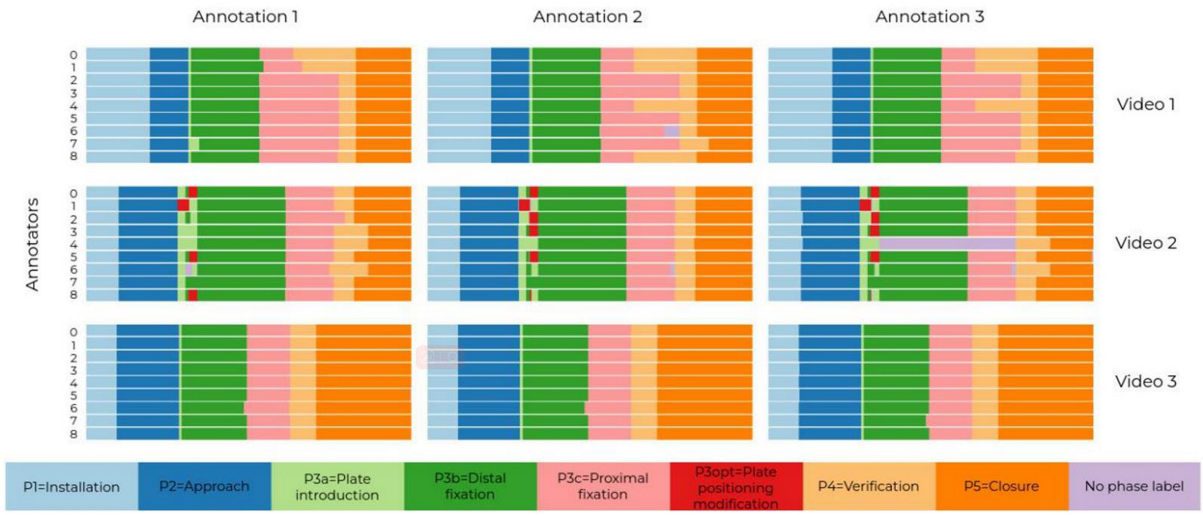

Figure 3. Overview of all the phase and sub-phase annotations made by the nine annotators.

between annotators, of the start time of each phase was arbitrarily chosen as insignificant if the median of $\delta < 0.5$ s. The threshold of 5 s, corresponding to 30 frames, was arbitrarily chosen as an upper bound to consider δ as significant if its median was superior to this value. If 0.5 s ≤ median of $\delta \leq 5$ s, δ was considered as not very significant.

RESULTS

Qualitatively, annotation anomalies were noticed through absences or differences in annotation labels [Figure 3].

On five occasions during phase 3, one or more sub-phase labels were totally or partially absent from one or more videos. This label absence occurred once with annotator n°4 and four times with the annotator n°6. Concerning annotator n°4, two labels (sub-phases 3b and 3c) were totally absent in one of the three V2. Concerning annotator n°6, a label was partially absent once in one of the three V1 annotation runs and once in each of the three V2 annotation runs. In three cases, the annotation anomalies concerned the limit between the end of sub-phase 3c and the beginning of phase 4, and once the middle of sub-phase 3a. For the computation of k and F_1, these missing annotations were compensated by extending the duration of the previous label.

Video V2 resulted in the most annotation differences. It was the only video that contained sub-phase 3opt. The sub-phase (3opt) was annotated by only six out of nine annotators. It was systematically annotated in all three V2s by four annotators, in two V2s by two annotators, and in none of the V2s by three annotators. Sub-phase 3opt was annotated by four annotators in the first annotation of V2, and by six annotators in the last two annotations of V2. Apart from annotator 7, all others started sub-phase 3b at about the same time. Between phase 2 and sub-phase 3b, the annotators annotated either sub-phase 3a once, or sub-phase 3a twice, or sub-phase 3a once and sub-phase 3opt once, or sub-phase 3a twice and sub-phase 3opt once.

The quantitative analysis is shown in Tables 2-5.

Table 2. Cohen's kappa and F_1-score per phase/sub-phase, averaged over the different annotators and the different annotations of each video

Phase	Video		
	1	2	3
		k (%)	
	87	94	100
		F_1 (%)	
P1	100	100	100
P2	100	100	100
P3a	99.6	78.7	99.3
P3b	99.5	99.9	100
P3c	73.5	99.6	100
P3opt	/	54.8	/
P4	66.2	98.9	99.9
P5	99.9	99.9	100

F_1: F_1-score; k: Cohen's kappa; P1: installation; P2: approach; P3: fixation; P3a: introduction of the plate; P3b: distal fixation; P3c: proximal fixation; P3opt: modification of the plate positioning; P4: verification; P5: closure.

Table 3. Mean (M) and median (m) start time difference per phase/sub-phase (in seconds) of each video

Phase	Video					
	1		2		3	
	M	m	M	m	M	m
P1	0.0	0.0	0.0	0.0	0.0	0.0
P2	0.1	0.1	0.5	0.1	0.1	0.1
P3a	0.1	0.0	4.9	0.1	0.1	0.1
P3b	2.2	0.1	15.0|0.1	0.1|0.1	0.1	0.1
P3c	4.0	4.1	0.3	0.1	2.0	0.1
P3opt	/	/	8.1	1.3	/	/
P4	209.8	278.7	2.5	0.4	0.3	0.1
P5	3.2	0.3	9.4	0.3	0.2	0.1

P1: Installation; P2: approach; P3: fixation; P3a: introduction of the plate; P3b: distal fixation; P3c: proximal fixation; P3opt: modification of the plate positioning; P4: verification; P5: closure.

Table 4. F_1-score per phase/sub-phase, averaged over the different annotations, for each annotator

Phase	Annotator								
	0	1	2	3	4	5	6	7	8
					F_1 (%)				
P1	100	100	99.6	100	99.6	100	100	100	100
P2	100	100	99.8	100	99.8	100	100	100	100
P3a	99.7	96.7	85.9	93.1	86.4	99.6	84.6	79.8	89.2
P3b	99.9	98.9	99.0	99.4	83.3	99.7	97.4	97.3	99.9
P3c	99.9	93.8	96.6	100	69.9	99.8	97.4	99.1	92.6
P3opt	99.5	94.9	/	94.9	/	96.9	/	/	32.2
P4	100	97.3	87.5	91.4	76.3	99.6	93.2	95.4	88.4
P5	100	100	100	95.4	97.5	100	97.7	98.1	100

F_1: F_1-score; P1: installation; P2: approach; P3: fixation; P3a: introduction of the plate; P3b: distal fixation; P3c: proximal fixation; P3opt: modification of the plate positioning; P4: verification; P5: closure.

Regarding the inter-individual variability, almost perfect agreement was observed between the different annotators, with k ≥ 81% for the three videos. For all phases and sub-phases of the three videos, $F_1 \geq 81\%$, except for sub-phase 3c and phase 5 in V1 and for sub-phase 3a and sub-phase 3opt in V2. A median of $\delta <$ 0.5 s was observed for all phases and sub-phases in all three videos, except for sub-phase 3c in V1 and sub-phase 3opt in V2 where 0.5 s ≤ median of $\delta \leq 5$ s and for phase 4 in V1 where a median of $\delta > 5$ s was observed.

Regarding the intra-individual variability, $F_1 \geq 81\%$ for all phases and sub-phases of the nine annotators, except for sub-phase 3c and phase 4 of annotator 4, for sub-phase 3a of annotator 7 and for sub-phase 3opt of annotator 8. A median of $\delta < 0.5$ s was observed for all phases and sub-phases of the nine annotators, except for phase 4 of annotators 2, 6 and 8, for sub-phase 3opt of annotators 3 and 5 and for the second annotation of sub-phase 3a of annotator 6 where 0.5 s ≤ median of $\delta \leq 5$ s.

DISCUSSION

The American Board of Surgery investigated the suitability of video-based assessment as an adjunct to certification for assessing technical skills[13]. These videos were analysed by surgeons in a subjective manner. It would be interesting to obtain an objective and automatic analysis by AI algorithms. It is in this context that our study takes all its interest by studying the annotation variability of videos later used to train an algorithm.

The AI researcher (annotator 0) was used as the reference annotator because they were more familiar with the annotation task and the annotation protocol than the other annotators. This choice was confirmed by the fact that annotator 0 presented the most homogeneous annotations of all [Table 4].

Generally speaking, as the quality of the videos interferes with the reliability of annotation, some authors have divided non-surgical videos into four levels of difficulty[14]. In our study, V3, with k = 100% (the maximum), can be considered as easy to annotate, as shown by the extreme similarity of the annotations made by the nine annotators [Figure 3]. For V1 and V2, the difficulty is greater, as the k values are not as high, although they remain above 81%. As the annotation protocol and annotators were the same throughout the study, the only difference between these three videos was the surgical procedure recorded. As V3 was the shortest of the three, the length of the surgery could be an indirect indicator of the difficulty of annotating the corresponding video. It seems logical to assume that, for similar fractures, the more fluid the surgery, the shorter its duration. A less fluid surgery could result in phase changes that are more difficult to identify.

The difference between the mean and the median of δ could be grouped into three cases [Tables 3 and 5]. In the first case, the mean and the median of δ were both low when all annotators started the phase or sub-phase at about the same time as annotator 0. In the second case, the mean and the median of δ were both high when the majority of annotators started the phase or sub-phase with a significant time delay compared to annotator 0. In the third case, the mean of δ was high, but the median of δ was low when the majority of annotators started the phase or sub-phase at about the same time as annotator 0, but a minority started it with a significant time delay compared to annotator 0.

Some of the results of this study were not above the threshold required to achieve an almost perfect agreement.

Table 5. Mean (M) and median (m) start time difference per phase/sub-phase (in seconds), for each annotator

Phase	Annotator																	
	0		1		2		3		4		5		6		7		8	
	M	m	M	m	M	m	M	m	M	m	M	m	M	m	M	m	M	m
P1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
P2	0.1	0.1	0.1	0.1	0.8	0.1	0.2	0.1	0.8	0.1	0.0	0.0	0.1	0.1	0.0	0.0	0.0	0.0
P3a	0.1	0.1	1.3	0.2	0.1	0.1	0.0	0.0	0.1	0.1	0.1	0.1	0.0 3.0	0.0 3.0	0.1	0.1	0.0	0.0
P3b	0.0 0.0	0.0 0.0	0.1	0.1	12.7	0.1	12.8	0.1	0.2	0.1	0.1 0.1	0.1 0.1	12.8	0.2	16.7	0.1	0.0 0.0	0.0 0.0
P3c	0.2	0.1	9.0	0.0	0.2	0.1	0.6	0.2	0.1	0.1	1.3	0.1	2.4	0.2	3.2	0.2	0.2	0.1
P3opt	0.2	0.2	0.1	0.1	/	/	2.8	2.8	/	/	1.6	1.6	/	/	/	/	0.5	0.5
P4	0.1	0.0	18.2	0.5	11.7	1.3	0.1	0.1	93.0	0.3	0.8	0.1	6.1	3.6	0.5	0.1	50.8	2.0
P5	0.0	0.0	0.2	0.2	0.3	0.1	14.7	0.1	7.9	0.1	0.1	0.0	7.4	0.1	11.7	0.1	0.2	0.1

P1: Installation; P2: approach; P3: fixation; P3a: introduction of the plate; P3b: distal fixation; P3c: proximal fixation; P3opt: modification of the plate positioning; P4: verification; P5: closure.

For example, the annotation of sub-phase 3opt in V2 was not homogeneous among all annotators. There are at least three reasons for this confusion. The first reason is that although the existence of sub-phase 3opt was mentioned in the annotation protocol, the video that was used to train the annotators (V0) did not include it. The annotators, therefore, did not have the opportunity to become familiar with it. The second reason is that sub-phases 3a and 3opt are difficult to distinguish since they both involve placing the plate with its mounted guide under the *pronator quadratus* muscle. The third reason is that the existence of the sub-phase 3opt is generally not systematic and that this sub-phase may appear in diverse ways throughout different surgical videos [Table 6]. The consequence of this confusion is that the event corresponding to sub-phase 3opt should not have been defined as a sub-phase in the annotation protocol. A solution to this confusion could be to replace sub-phase 3opt with a succession of sub-phases 3a and 3b. The detection by the algorithm of the event corresponding to sub-phase 3opt could be done either with a punctual action or by deduction from a succession of sub-phases 3a and 3b.

Another example of confusion is the boundary between sub-phase 3c and phase 4 in V1. Phase 4 consists of checking the correct position of the plate and screws by fluoroscopy, the correct mobility of the wrist and the freedom of movement of the flexor tendons[5]. The start of phase 4, which may involve screw changes, is not always clear because screw changes may also be performed during sub-phase 3c. The confusion only appeared with V1, where one of the proximal screws was changed during phase 4, because the annotators considered that this screw change necessarily corresponded to sub-phase 3c. The reason for this confusion is probably that V0 did not have a screw change during phase 4. The annotators did not have the opportunity to become familiar with this screw change. It can be seen in Figure 3 that among the annotators who did not integrate this screw change in phase 4, three of them (4, 6, 8) did not annotate this phase homogeneously. The solution to this confusion could be to use a training video containing a screw change during phase 4.

Table 6. Different scenarios that could be observed during a MIPO surgery

Phase	Different cases that could be observed during the fixation phase						
	Case 1	**Case 2**	**Case 3**	**Case 4**	**Case 5**	**Case 6**	**Case n**
P1							
P2							
P3	Introduction of the plate	Introduction of the plate	Introduction of the plate	Introduction of the plate	Introduction of the plate	Introduction of the plate	*Etc.*
	Distal fixation	Removal of the plate	Distal fixation	Distal fixation	Removal of the plate	Removal of the plate	
	Proximal fixation	Introduction of the plate	Removal of the plate	Proximal fixation	Introduction of the plate	Introduction of the plate	
		Distal fixation	Introduction of the plate	Removal of the plate	Distal fixation	Distal fixation	
		Proximal fixation	Distal fixation	Introduction of the plate	Removal of the plate	Proximal fixation	
			Proximal fixation	Distal fixation	Introduction of the plate	Removal of the plate	
				Proximal fixation	Distal fixation	Introduction of the plate	
					Proximal fixation	Distal fixation	
						Proximal fixation	
P4							
P5							

P1: Installation; P2: approach; P3: fixation; P4: verification; P5: closure.

The main study hypothesis of almost perfect inter-individual variability was verified since Cohen's kappa coefficient was globally greater than or equal to 81%. Concerning the intra-individual variability, the secondary hypothesis was verified for the majority of annotators (0, 1, 2, 3, 5, 6) since the F_1-score was greater than or equal to 81% for all phases and sub-phases. The secondary hypothesis was verified in some phases and sub-phases for a minority of annotators (4, 7, 8).

In conclusion, the homogeneity of annotations must be as high as possible to develop an AI algorithm in surgery. According to our results, this homogeneity depends on the material (annotators) and the method (protocol, *etc.*). The performance of the annotators can be influenced by their clinical experience, motivation, and fatigue. It is interesting to pinpoint the best performing annotators to be able to scale up the annotated data production of the AI algorithm development. The worst performing annotators should be identified (measurement of the intra-individual variability) to provide them with individual training and a personalised annotation rhythm. Regarding the method, the definition of the phases should be optimised (especially in the case of the sub-phase 3opt), the annotation protocol improved, and suitable training videos chosen (in order to cover the variety of situations that could happen).

DECLARATIONS

Acknowledgments

The authors thank Audrey Daiss, Antoine Martins, Laurine Cafarelli, Stéphanie Gouzou, Liliya Efremova, and Marie-Cécile Sapa for their participation in the study.

Authors' contributions

Data collection: Graëff C, El Amiri L, Liverneaux P

Writing of the annotation protocol: Graëff C, Liverneaux P
Creation of the set-up on MOSaiC: Graëff C, Mazellier JP
Extraction of the annotations from MOSaiC: Graëff C, Mazellier JP
Analysis of the annotations: Graëff C, Lampert T, Padoy N, Liverneaux P
Writing of the paper: Graëff C, Lampert T, Padoy N, Liverneaux P

Availability of data and materials
Not applicable.

Financial support and sponsorship
This work of the Interdisciplinary Thematic Institute HealthTech, as part of the ITI 2021-2028 program of the University of Strasbourg, CNRS and Inserm, was supported by IdEx Unistra (ANR-10-IDEX-0002), SFRI (STRAT'US project, ANR-20-SFRI-0012) and IHU Strasbourg (ANR-10-IAHU-02) under the framework of the French Investments for the Future Program.

Conflicts of interest
Philippe Liverneaux has competing interests with Caresyntax. None of the other authors have competing interest.

Ethical approval and consent to participate
This study was performed in line with the principles of the Declaration of Helsinki. Approval was granted by the Ethics Committee of Strasbourg University Hospitals (CE 2021-159). Informed consent was obtained from the four patients included in the study.

Consent for publication
The authors affirm that human research participants provided informed consent for the publication of the four videos.

REFERENCES
1. Twinanda AP, Shehata S, Mutter D, Marescaux J, de Mathelin M, Padoy N. EndoNet: a deep architecture for recognition tasks on laparoscopic videos. *IEEE Trans Med Imaging* 2017;36:86-97. DOI PubMed
2. Hashimoto DA, Rosman G, Witkowski ER, et al. Computer vision analysis of intraoperative video: automated recognition of operative steps in laparoscopic sleeve gastrectomy. *Ann Surg* 2019;270:414-21. DOI PubMed PMC
3. Kitaguchi D, Takeshita N, Matsuzaki H, et al. Real-time automatic surgical phase recognition in laparoscopic sigmoidectomy using the convolutional neural network-based deep learning approach. *Surg Endosc* 2020;34:4924-31. DOI PubMed
4. Ward TM, Hashimoto DA, Ban Y, et al. Automated operative phase identification in peroral endoscopic myotomy. *Surg Endosc* 2021;35:4008-15. DOI PubMed PMC
5. Graëff C, Daiss A, Lampert T, et al. Preliminary stage in the development of an artificial intelligence algorithm: variations between 100 surgeons in phase annotation in a video of internal fixation of distal radius fracture. *Orthop Traumatol Surg Res* 2023. In Press. DOI PubMed
6. Tang JB, Giddins G. Why and how to report surgeons' levels of expertise. *J Hand Surg Eur Vol* 2016;41:365-6. DOI PubMed
7. Liverneaux PA. The minimally invasive approach for distal radius fractures and malunions. *J Hand Surg Eur Vol* 2018;43:121-30. DOI PubMed
8. Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. *Fam Med* 2005;37:360-3. Available from: https://pubmed.ncbi.nlm.nih.gov/15883903/. [Last accessed on 30 Jun 2023].
9. Bajpai S, Bajpai R, Chaturvedi HK. Evaluation of inter-rater agreement and inter-rater reliability for observational data: an overview of concepts and methods. *J Indian Acad Appl Psychol* 2015;41:20-7. Available from: https://www.researchgate.net/publication/273451591_Evaluation_of_Inter. [Last accessed on 30 Jun 2023].
10. Hripcsak G, Rothschild AS. Agreement, the f-measure, and reliability in information retrieval. *J Am Med Inform Assoc* 2005;12:296-8. DOI PubMed PMC

11. Lavanchy JL, Gonzalez C, Kassem H, Nett PC, Mutter D, Padoy N. Proposal and multicentric validation of a laparoscopic Roux-en-Y gastric bypass surgery ontology. *Surg Endosc* 2023;37:2070-7. DOI PubMed PMC
12. Hallgren KA. Computing inter-rater reliability for observational data: an overview and tutorial. *Tutor Quant Methods Psychol* 2012;8:23-34. DOI PubMed PMC
13. Pryor AD, Lendvay T, Jones A, Ibáñez B, Pugh C. An American board of surgery pilot of video assessment of surgeon technical performance in surgery. *Ann Surg* 2023;277:591-5. DOI PubMed
14. Vondrick C, Ramanan D, Patterson D. Efficiently scaling up video annotation with crowdsourced marketplaces. In: Computer Vision - ECCV 2010. Berlin: Springer Berlin Heidelberg; 2010. p. 610-23. DOI