The Robustness Of Probabilistic Objective Performance Indicators

Max Berniker, Ye Li, Kiran Bhattacharyya, Anthony Jarc, Intuitive Surgical Inc.

INTUITIVE SAGES 2023

Introduction

Can we compute accurate objective performance indicators (OPIs) in the face of annotator variability? OPIs computed during key surgical tasks correlate with surgeon skill and clinical outcomes, making them valuable for surgical data science. Yet the initiation and termination of surgical tasks is often subject to interpretation, rendering OPIs variable and uncertain. OPIs should be robust to this variability to remain informative. Can a probabilistic approach to task identification mitigate these effects? Specifically, are probabilistic OPIs more accurate and precise than hard-boundary OPIs?

Background

In a prior study (Berniker et al., 2021) we proposed a probabilistic approach to OPI computation, presenting theoretical benefits and proof of concept on a limited data set. Here our aim is to determine if this approach is robust on a rich and extensive data set.

Convention:

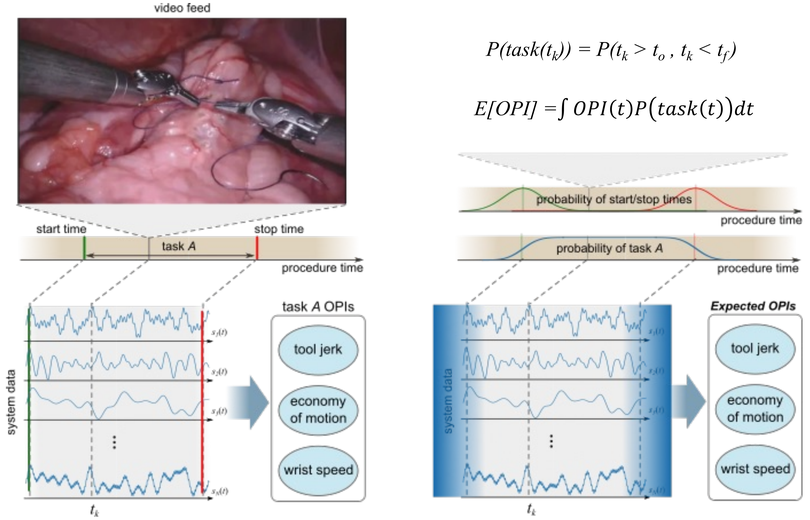
Start and stop times are annotated (by a human or a machine), demarking hard temporal boundaries within which data is used to compute OPIs.

Proposal:

Start and stop times are treated as random variables with probabilities (green and red curves). The probable duration of a task is computed (blue curve). OPIs are also random, and their expected values can be computed.

$$P(task(t_k)) = P(t_k > t_o\ ,\ t_k < t_f)$$

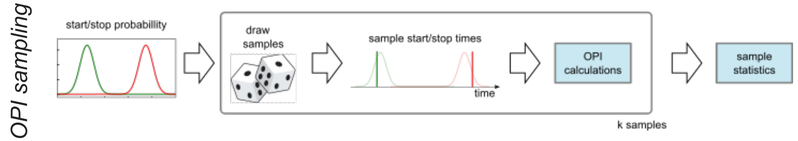
$$E[OPI] = \int OPI(t)P\big(task(t)\big)dt$$



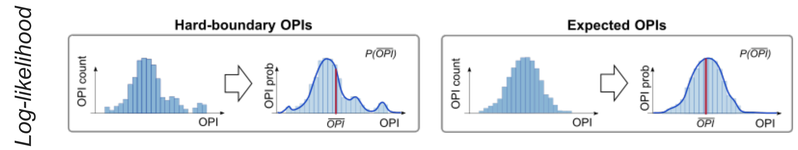
Berniker, Max, Kiran D. Bhattacharyya, Kristen C. Brown, and Anthony Jarc. "A Probabilistic Approach to Surgical Tasks and Skill Metrics." IEEE Transactions on Biomedical Engineering 69, no. 7 (2021): 2212-2219.

Methods

We compared OPIs using these two approaches on a large clinical data set. For each annotated task, we affixed a normal distribution to the start and stop times. The standard deviation (2.0 seconds, found empirically) represents annotator uncertainty. 300 sample start and stop times were randomly drawn. For each random annotation pair, six OPIs (i.e., duration, path length, angular path length, camera clutch rate, energy rate and hand clutch rate) were computed using the hard-boundary and probabilistic methods (referred as Expected OPIs). All data was obtained from the da Vinci® Xi robotic system.



Four types of tests were performed to answer the following questions: 1) were the samples from distinct distributions?[1] 2) did the OPIs have different means?[2] 3) did the samples sets have different variances?[3] 4) were the data likelihoods different?



Log likelihood ratios were computed (4), as the mean expected OPIs to the mean hard-boundary OPIs[4]. If the probability of the mean expected OPI is larger (the ratio is larger), this implies the distribution is less variable.

Results

We compared the hard-boundary and probabilistic OPIs across over 1200 surgical cases across seven procedure types (radical prostatectomy, cholecystectomy, inguinal hernia, ventral hernia, hysterectomy, lobectomy and lower anterior resection). In total there were ~18K annotated tasks with ~700 to ~4500 tasks per procedure type.

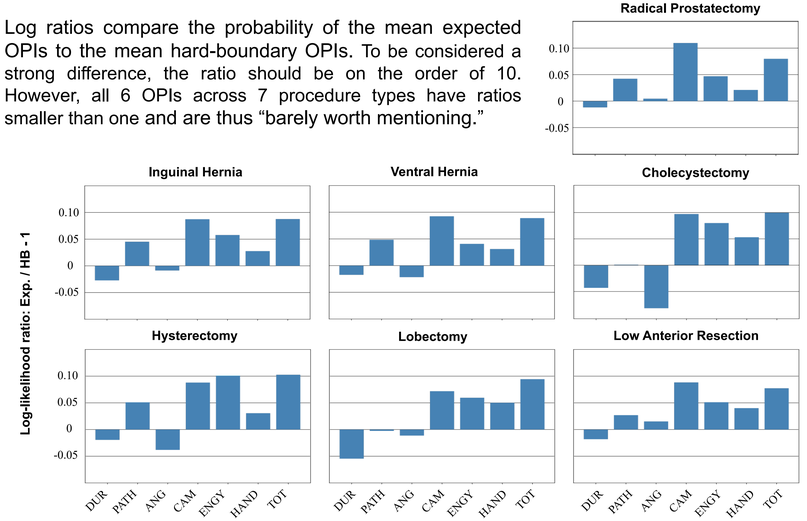
The results followed a consistent trend: most OPIs were statistically indistinguishable. Of those that were statistically distinct, the expected values had a smaller variance and narrower distributions around the mean.

1. Two-sample Kolmogorov-Smirnov tests (with means subtracted). 2. Independent sample t-tests for each task without assuming equal variances. 3. Levene test 4. The PDF for each task was approximated with a Gaussian kernel density estimator using the sample data.

Below is a summary of the tests performed to compare the OPIs computed with the two approaches (*N* is the annotated tasks sample size). Their populations were mostly indistinguishable. When distinct, the probabilistic OPIs had smaller variances.

Procedure type	K-S test		Means		Variance		Var. Exp < HB	
	N	Sig. diff.	N	Sig. diff.	N	Sig. diff.	N	Sig. diff.
Radical Prostatectomy	3277	1026 (31.3%)	1026	727 (70.9%)	1026	848 (82.7%)	848	763 (90.0 %)
Inguinal Hernia	4607	1372 (29.8%)	1372	948 (69.1%)	1372	1081 (78.8%)	1081	886 (82.0%)
Ventral Hernia	966	314 (32.5%)	314	208 (66.2%)	314	231 (73.6%)	231	188 (81.4%)
Cholecystectomy	4296	1674 (38.9%)	1674	1170 (69.9%)	1674	1285 (76.8%)	1285	960 (74.7%)
Hysterectomy	2592	1006 (38.8%)	1006	706 (70.2%)	1006	794 (78.9%)	794	648 (81.6%)
Lobectomy	732	302 (41.3%)	302	225 (74.5%)	302	236 (78.2%)	236	186 (78.8%)
Low Anterior Resection	2586	981 (37.9%)	981	688 (70.1%)	981	726 (74.0%)	726	564 (77.7%)

Log ratios compare the probability of the mean expected OPIs to the mean hard-boundary OPIs. To be considered a strong difference, the ratio should be on the order of 10. However, all 6 OPIs across 7 procedure types have ratios smaller than one and are thus “barely worth mentioning.”



Conclusion

We find that the probabilistic OPIs are more robust and less variable than their hard-boundary counterparts. The results suggest that this probabilistic approach holds promise for the computational analysis, and conceptual interpretation, of surgical data.