

2022

Training non-surgical experts to annotate open-source surgical videos for machine learning

<https://hdl.handle.net/2144/45660>

"Downloaded from OpenBU. Boston University's institutional repository."

BOSTON UNIVERSITY
SCHOOL OF MEDICINE

Thesis

**TRAINING NON-SURGICAL EXPERTS TO ANNOTATE OPEN-SOURCE
SURGICAL VIDEOS FOR MACHINE LEARNING**

by

FAADIL MOHAMMED SHARIFF

B.S., Indiana University, 2020

Submitted in partial fulfillment of the
requirements for the degree of
Master of Science

2022

Approved by

First Reader

Gwynneth Offner, Ph.D.
Director of the MS in Medical Sciences Program
Associate Professor of Medicine

Second Reader

Gabriel Brat, M.D.
Assistant Professor of Surgery
Beth Israel Deaconess Medical Center

DEDICATION

I would like to dedicate this work to my family who have always supported me in my academic endeavors.

ACKNOWLEDGMENTS

I would like to thank my mentor, Dr. Gabriel Brat, and Jevin Clark for their help in conducting this research and guiding me throughout the writing process.

**TRAINING NON-SURGICAL EXPERTS TO ANNOTATE OPEN-SOURCE
SURGICAL VIDEOS FOR MACHINE LEARNING**

FAADIL MOHAMMED SHARIFF

ABSTRACT

The use of video annotation for utilization in machine learning computer programs is an area of medicine that has shown increased demand for research in recent years. The limiting factor for the use of video annotation in surgery is the scale and efficiency in which videos can be labelled. The challenge in surgical contexts is the current notion that only surgical experts can provide accurate video annotations. To challenge this notion, we have conceived a survey to test non-surgical experts' abilities to accurately annotate open-source surgical videos. This test has been published on the crowdsourcing platform Amazon mTurk. A learning module was created to provide relevant and concise information necessary to accurately annotate the surgical video and complete the survey. This learning module illustrates important instructions on differentiating between three surgical activities of focus: cutting, suturing, and tying. The survey includes free-response and multiple-choice questions that test the accuracy of respondent's video annotation. Analyzing the results from 50 participants, more data from larger scale studies must be acquired, greater data validation systems must be implemented, and instructions in the survey and learning module must be adapted. These changes are due to high rates of inaccurate annotation for all three surgical activities. The data showed no clear indication that cutting, suturing, or tying could be accurately identified but further investigation would be prudent in the future.

TABLE OF CONTENTS

DEDICATION	iv
ACKNOWLEDGMENTS	v
ABSTRACT.....	vi
TABLE OF CONTENTS.....	vii
LIST OF TABLES	viii
LIST OF FIGURES	ix
LIST OF ABBREVIATIONS.....	x
INTRODUCTION	1
Background of Machine Learning and Artificial Intelligence.....	1
Machine Learning in Medicine.....	4
Video Annotation and Medical Applications	6
METHODS	10
RESULTS	23
DISCUSSION.....	42
Review of Findings and Problems Encountered.....	42
Review of Potential Solutions and Improvements.....	49
BIBLIOGRAPHY.....	56
CURRICULUM VITAE.....	60

LIST OF TABLES

Table 1: Survey questions with answer choices and answer key.....	15-17
Table 2: Demographic survey questions and answer choices.....	19-20
Table 3: mTurk survey scheduling details and experimental parameters.....	21
Table 4: Subset of respondent data for Q1 that included 2 and 4 timestamps.....	24
Table 5: The answer key for the suturing timestamp data.....	24
Table 6: Suturing respondent data subset for 2 timestamp calculations.....	25
Table 7: Suturing respondent data subset for 4 timestamp calculations.....	26
Table 8: Summary of data presented in Fig. 8. Measures of Central Tendency.....	34-35
Table 9: Summary of Observations/Trends discussed.....	47-49
Table 10: Summary of Issues and Potential Solutions discussed.....	53-55

LIST OF FIGURES

Fig. 1a-1f: Images from learning module	12
Fig. 2: Average Aligned Difference for Cutting Activities..	27
Figure 3: Median Absolute Aligned Difference for Cutting Activities..	28
Figure 4: Average Aligned Difference for Suturing Activities..	29
Figure 5: Median Absolute Aligned Difference for Cutting Activities.	30
Figure 6: Average Aligned Difference for Suturing Activities	31
Figure 1a: Incision made with scalpel (Thakkar, 2019).	12

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
AGI	Artificial General Intelligence
CV	Computer Vision
fps.....	Frames per second
HIT	Human Intelligence Tasks
ML.....	Machine Learning
mTurk.....	(Amazon) mechanical Turk
OS	Operating System
sec	seconds

INTRODUCTION

Background of Machine Learning and Artificial Intelligence

Machine learning and artificial intelligence are growing fields that are showing promise to revolutionize industries with continued innovation and research. Artificial intelligence (AI) is an area of study that meshes computer science expertise and engineering principles. The origin of the concept of artificial intelligence is varied but one common origin is from the father of computing, Alan Turing. Turing, who is famous for his work during World War II for the allied forces fighting the Nazis and solving the encryption machine known as Enigma, posed a question on whether machines can think. Being a scientist at his core he created a test known as the Turing Test in 1950, which was adapted from a Victorian style game, to determine if a machine could successfully replicate the role of a man in a question-and-answer format guessing game. This methodology was profoundly advanced thinking for an era when the concept of artificial intelligence had not even been coined. The term was later coined by John McCarthy, one of the founding fathers of artificial intelligence (Rajaraman, 2014). It is important to recognize the origins of this concept when considering the overarching goal or question regarding whether machines/computers can replicate or mimic human thinking. The question that must be posed once understanding that paradigm is what is the current landscape of artificial intelligence machines?

The current field of artificial intelligence can be categorized in two important manners, the first is the level of intelligence and thinking the machine can exhibit and the current forms of utilization for said machines today. The first type of artificial intelligence

machine is known as a reactive machine. The basics of such a machine are that there is no memory storage and thus no capacity for storing information to learn from previous experience. The machine in this case must react to specific input and complete a specific output and thus these machines have limited utility as they can only accomplish specialized, repetitive tasks such as competing in chess competitions. The second type of artificial intelligence machine is known as limited memory machines. These are AI machines capable of storing data from previous experience as a data gathering tool that can inform decision making for present-time tasks. These machines can be trained to analyze data sets and make predictions based on that information. Limited memory AI lends itself to more complex tasks than reactive machines and exhibit more “human-like” thinking in practice through iteration and trial-and-error. Theory of Mind is the next evolution of AI machines that at this point is strictly theoretical. This concept is based on the premise that true ‘human thinking’ requires an understanding and internalization of emotions, specifically the emotions of others and how those emotions impact people’s behaviors and their relationships. To accomplish this level of intelligence an AI machine would need to be capable of understanding abstract thoughts, emotions, and react to such phenomena. The most advanced form of AI is self-awareness. This would require machines to exhibit their own form of consciousness and to understand their own existence in the world. Both theory of mind and self-awareness forms of AI are strictly theoretical at this stage. All current uses of AI in society are forms of reactive machines or limited memory machines.

The scope of use for AI machines currently is an important topic to consider as the field is growing year to year with one report from a research firm, Markets and Markets, predicting the AI industry value growing from 58.3 billion to 309.6 billion between the years of 2021 and 2026. Narrow AI (also known as weak AI) is the commonplace form of AI that has gained prevalence in society. Narrow AI perform limited-scope tasks, examples include smartphone operating systems (OS) personal assistants such as Alexa and Siri, Google Search algorithms, and autonomous car technology from companies such as Tesla. The more complex form of AI that can be used is known as ‘strong AI’ or artificial general intelligence (AGI). Strong AI is theoretical and would exhibit the capabilities of intelligence equivalent to a human being and could utilize such intelligence in numerous businesses, technological and medical applications.

Machine learning is a concept that enables the utilization of narrow AI machines. Narrow AI is made possible through machine learning algorithms. Machine learning (ML) can be best described as computer programs and algorithms that use data sets as ‘experience’ to learn from and help computers make predictions or complete tasks based on that knowledge. ML algorithms are often referred to as ‘models’ and these models are fed large data sets with the goal of detecting patterns or predicting conclusions from new datasets.

A key distinction to make to fully understand ML programming is to compare and contrast the variables and processes involved with traditional computer programming. Traditional programming requires manual creation of a program that has logic or rules programmed by a person that can create a desired output or solution. ML programming

does not require the use of a manually constructed computer program and is an automated process. A simplified explanation of the process involves using an input, a pre-existing output, which allows the computer algorithm to run an automated process that will formulate a model. This could be utilized to predict future outcomes with unknown data sets. ML occurs in two types of formats: supervised and unsupervised learning.

Supervised learning requires data sets for training that act as an input as well as providing the desired results data set. By providing both data sets, the model can recognize patterns or relationships in the data that can be utilized in the future with new input data to predict new output data. Unsupervised learning requires data sets for training that are unlabeled or correlated with output data. The model utilizes algorithms to find features and unknown patterns in data. Supervised learning works from collecting data and producing data output through experience or known data sets to recognize relationships and model patterns. Unsupervised learning has a much broader scope that can analyze all potential patterns as there is no 'experience' or data set that it is referencing to learn from.

Machine Learning in Medicine

ML algorithms can be a powerful tool and aid projects in many industrial applications. Likewise, the use of ML technology has extreme promise in the field of medicine. While computers and technology have long been evolving with and aiding in the transformation of medicine, the use of ML has not similarly caught on. ML technology has been theoretically discussed as a promising tool with the potential to radically change and potentially improve the inefficiencies in medicine going back to 1970 (Schwartz, 1970).

The lack of adoption and integration of the power of ML technology in medical education, clinical patient care, and medical research is a reality of the current age. The power of ML in medicine can be most aptly understood when understanding the labor-intensive process of something as common as a differential diagnosis. When encountering a list of symptoms for a patient, a primary care physician must consult their knowledge and experience, their team's knowledge and experience, and assess the possible explanations. This can often be an arduous process for conditions that present similar sets of symptoms but require unique treatments. The power of ML can be harnessed in these situations and enable more efficient analysis of the facts and empower physicians to make more informed decisions (Salvatore et al., 2014). The function of ML technology is akin to a physician being able to consult with a hundred, if not thousands, of physicians at once, combining their experiences and memories for patterns that inform decision making. This technology can help mitigate issues that plague our healthcare system, such as early diagnosis of cancers in patients, helping to identify patients at risk of cardiovascular disease in earlier stages of life, patients with complicated sets of symptoms, and can even play a role in recommending the most optimal prescriptions. This last example is particularly important as many studies have noted the indisputable and unfortunate factor that human error and patterns of human behavior have played in prescribing incorrect or more familiar medications resulting in fatal consequences (Kohn et al., 2000).

Video Annotation and Medical Applications

Video annotation is a tool under the umbrella of AI in a field known as computer vision (CV). The premise of computer vision is that ML algorithms can train computers to ‘think’ like humans, particularly in terms of the sensation of vision. The goal with CV is to try and create a system that can perceive and label objects in videos, such as the eye and brain do in humans. Video annotation done manually requires a frame-by-frame analysis of videos with labelling tools to identify objects and actions of interest. This is a laborious and time-consuming process, as a 30 frames-per-second (fps) video that is 60 seconds long would have 1,800 frames to analyze. Video annotation differs from image annotation as video offers greater context thus further insight to teach AI and CV systems. Video annotation allows for labelling of objects and actions across hundreds and thousands of frames that can overcome object obstruction and categorize actions based on object movement. Research conducted on video annotation and the creation of systems or programs to enable annotation to have focused on improving annotation accuracy and efficiency by transitioning systems away from manual, user-intensive annotation (Bianco et al., 2015). These tools and programs show promise in improving the amount of data for ML that can be annotated and organized. Video annotation is an important element of CV programs that exhibit narrow AI capabilities in fields such as transportation and medicine.

The increasing rates of technological advances in medicine have been critical to improving the quality and speed of care. The field of surgery has been impacted through revolutionary advancements involving robotics and tools to perform minimally invasive

surgeries. No matter the setting, medicine and technology are forever intertwined and there are mountains of data and information being gathered in the process. The increased collection of data from medical devices, medical records, surgical tools, etc. is an area ripe for the implementation of AI and ML. In medicine, specifically in surgery, data for video annotation and CV is important to store, collect, and analyze. Understanding how to leverage this data can have profound impacts on medicine and surgery in areas such as creating automated systems for real-time guidance for surgery in operating rooms and post-surgery feedback for surgical training. The complexity of analyzing and processing the data gathered from different sources such as surgical videos is a major hurdle for widespread implementation of these technologies.

What makes surgical video annotation a particularly difficult task is the complexity and frame-by-frame time-sensitivity of actions occurring that are important to label and note. These challenges have been met with increased demand for research promoting the advancement of video annotation for these purposes. Studies have shown that video ML technology has promise in areas of surgery such as surgical phase recognition (Garrow et al., 2020). Unfortunately, a limiting factor acknowledged in this burgeoning field of medicine is the lack of qualified annotators. The current system of surgical video annotation implements relies on the use of surgical experts who have the requisite knowledge to label and tag actions in these videos. This bottleneck in data collection presents a major problem when considering the vast amounts of surgical videos that is raw data waiting to be annotated. Thus, the concept of utilizing crowdsourcing

technology to scale up the data collection/video annotation work offers a potential solution to this problem.

In order to create a framework for gathering data to test this proof of concept for surgical video annotation, a large-scale crowdsourcing platform was required to outsource the work. The Amazon Mechanical Turk (mTurk) website provided such a platform for this experiment to crowdsource a requisite amount of data. The website has a few key features that enable use in gathering large amounts of data. There are two distinct users on mTurk, those who create projects or tasks for data collection are known as ‘requesters’ and those who complete projects or tasks for compensation are known as ‘workers’.

These workers are also known as mTurkers. The website considers projects on the mTurk site to be those that require human effort or intelligence thus the projects/jobs are listed as “Human Intelligence Tasks” (HITs). Common HITs on the platform include surveys, image or video annotation, transcription, academic research, data entry, writing, etc.

There are upwards of half a million registered users, with HITs varying from as low as \$0.01 dollars for quick, low-effort tasks to as much as \$11.00 - 12.00 dollars for tasks that take upwards of an hour. In recent years, the utilization of amazon mTurk as a marketplace to crowdsource data collection in academic and medical settings has increased. Reviews of the use of such methods of data collection have shown an upwards trend in academic articles in journals referencing data samples from Amazon mTurk, with one study showing a 2,117% increase over a 7-year span from 2012 to 2019 (Aguinis et al., 2020).

The promise of the Amazon mTurk site as a source for crowdsourcing the volume of data needed for surgical video annotation for ML integration is abundant. The benefits of using a service such as Amazon mTurk is the amount of research performed and data gathered regarding the utility of the site compared to more traditional methods of large-scale market research. The cost-effectiveness paired with comparatively accurate samples of data are promising as seen by a study that showed data accuracy within 10% of traditional research survey methods in a data collection period of a few hours (Bentley et al., 2017).

The goal is that utilizing Amazon mTurk to create efficient and scalable system for video annotation library enables further use with ML and AI applications particular in the surgical field of medicine. While the promise of ML and CV in surgery and medicine is more of unrealized potential than practical advancement, it remains a key field that is ripe for further study and future implementation in the healthcare industry. The ability to harness the computing power of advanced ML programs to improve the lives of millions of patients is immense. Thus, this unique system of a learning module, Qualtrics survey, and the utilization of Amazon mTurk crowdsourcing presents a unique opportunity to test the ability to gather large enough sets of data to accelerate the use of ML and CV technologies in surgical settings.

METHODS

Using Amazon mTurk as the crowdsourcing platform, we designated the HIT as a survey that linked to a Qualtrics survey. The survey includes instructive videos in a learning module that are intended to teach non-experts, people with no experience in medical or surgical training, how to recognize and annotate videos of surgery. The activity begins with learning module, it is followed by the surgical video of primary focus, then a survey on the surgical video that asks specific questions regarding annotation and timestamp markers, concluding with various demographic survey questions.

Prior to beginning the learning module, it is necessary to forewarn the respondents on mTurk about the nature of the content they will be viewing. Thus, in bold text there is a warning that states the potentially graphic nature of the surgical images and videos that are part of the learning module and the focus of the mTurk HIT. The warning stipulates that included in the images, text, and videos are surgical procedures and tools, as well as exposed anatomical structures and blood. The last portion of the warning statement indicates the length of time one might be exposed to such graphic content as the intended and approximate time to complete the HIT is 45 minutes.

The learning module begins by discussing the three surgical activities to be annotated in the open-source surgical video embedded in the HIT. The three surgical activities that are of importance are cutting, tying, and suturing. The module indicates that a unique ‘activity’ is defined by a beginning and ending timestamp for a unique surgical activity. This ‘activity’ is annotated in the following format, XX: XX – YY: YY. The mTurkers

are told in the module to be as specific as possible when differentiating between distinct activities even on a second-by-second basis.

The module defines a surgical action as only the time when a surgeon is actively using a surgical tool that is interacting with the patient. A helpful distinction provided by explaining that an ‘activity’ only begins for cutting when a surgeon utilizing a tool comes into contact with the patient’s body, not when the surgeon picks up the tool. An important instruction is dictated about the distinction between two unique activities by stipulating that a break of three seconds between two different activities or the same activity must be logged separately. An example of this process would be if a surgeon removed a tool for cutting for longer than three seconds and either resumed cutting or completed a different task. If those parameters are met, then by rule the mTurkers must annotate two distinct activities with separate timestamps.

Cutting is described in the module as an action utilizing one of two tools: a scalpel or an electrocautery device. The module continues by educating mTurkers that are not familiar with surgical tools about the use case of cutting and the specific tools to look for while annotating. The action of cutting in a surgical context is described in the module as creating an incision that helps to separate the layers of a patient’s tissue to access anatomical structures of interest. A scalpel is defined for the mTurkers as a tool for the initial stages of a surgical procedure to make a primary incision over a larger area of tissue.



Figure 1a: Incision made with scalpel (Thakkar, 2019).

An electrocautery device is a separate tool with an attached power source that the module explains is to provide an electrical current and heat to facilitate cutting in a more efficient and precise use case. The electrocautery is then explained to be used for surgical actions such as coagulating tissue that are not relevant to this HIT and video annotation. The critical final instruction in the module dictates to the mTurkers that if a cutting activity is recorded with a scalpel and an electrocautery is subsequently introduced, or vice versa, they must be annotated as two unique activities in the survey.



Figure 1b: Incision made with an electrocautery device (Thakkar, 2019).

The learning module proceeds to educate mTurkers on the specifics of the second surgical action of importance that is suturing. Suturing is defined in the module as weaving a surgical thread through a needle that is inserted into a patient's tissue to seal or stitch together a surgical incision or wound for proper healing and to diminish the risk of any complication. A helpful method illustrated in the module to identify suturing is by noting the surgical thread attached a metal hook along with pliers that are known as needle drivers that help to seal incisions.



Figure 1c: Suture needle and thread held by needle driver (Wollheim, 2021).



Figure 1d: Needle driver threading suture needle across incision (Slater, 2018).

The learning module continues with descriptions and images to illustrate how to identify and annotate the third and final surgical activity of tying. The mTurkers are provided a description of what tying accomplishes in surgical settings as it is explained that tying creates a knot that secures the suture in place. The module explains that a cutting activity is when the suturing needle is handled by the needle driver after being threaded through the patient's tissue. The module further explains that in many cases the act of tying is subsequent to suturing but it is important to distinctly identify the separation of these two actions. The module concludes with a hint that the end of a tying activity is considered when the final knot is tied, not when the thread is being cut.



Figure 1e: Completed tying activity with sutured knots (Fletcher, 2019).

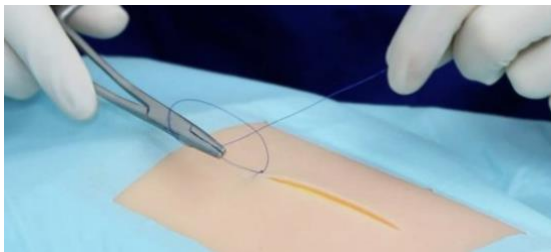


Figure 1f: Needle driver making loop for tying (Brewster, 2017).

Prior to full scale implementation and data collection on Amazon mTurk, the survey was sent to 10 users as a proof-of-concept evaluation. There are nine questions on the survey.

The survey is followed by optional demographic survey questions and a section for feedback and comments that will be analyzed for further iterations of this project.

The survey and demographic questions are listed below in table format with the logic and rationale for each question accompanied below the table.

The survey and demographic questions are listed below in table format with the logic and rationale for each question accompanied below the table.

Table 1: Survey questions with answer choices and answer key.

Question	Answer Choices	Correct Answer
<p>Please record timestamps below.</p> <p>Cutting:</p> <p>Suturing:</p> <p>Tying:</p>	<p>Open-ended response</p>	<p>Cutting: 00:11-00:23; 00:23-00:26; 00:29-00:58</p> <p>Suturing: 01:37-01:43; 01:57-02:02; 02:39-02:42; 03:11-03:22; 03:32-03:40; 03:55-4:14</p> <p>Tying: 01:43-01:52; 02:42-02:49; 02:54-03:04; 03:07-03:10; 03:23-03:32; 03:48-03:52</p>
<p>Question 1: Please select the number of distinct activities for cutting, suturing, and tying.</p>	<p>2; 3; 4; 5; 6; 7; 8; 9;</p>	<p>3, 6, 6</p>

Question 2: What is the timestamp for the second cutting activity?	00:11 – 00:23 00:23 – 00:26 01:43 – 01:52 02:54 – 03:04	00:23 – 00:26
Question 3: What is the timestamp for the third suturing activity?	01:37 – 01:43 01:43 – 01:52 02:39 – 02:42 02:54 – 03:04	02:39 – 02:42
Question 4: What is the timestamp for the fifth tying activity?	02:39 – 02:42 03:23 – 03:32 03:48 – 03:52 03:55 – 04:14	03:23 – 03:32
Question 5: At timestamp 01:34-01:50 what one distinct activity is shown?	Cutting Suturing Tying N/A	Cutting
Question 6: At timestamp 06:02-06:10 what one distinct activity is shown?	Cutting Suturing Tying N/A	Tying
Question 7: At timestamp 05:50-06:30 how many distinct activities should be logged?	2; 3; 4; 5	4

Question 8: At timestamp 01:19-01:53 how many distinct activities should be logged?	1; 2; 4; 5	2
---	------------	---

The first question is multiple parts, open-ended, and required the most thought for mTurkers to answer. The question was designed to ask for specific second-by-second timestamps of every unique surgical action. The intended answer choices would be a list of time periods with a beginning and end timestamps for cutting, tying, and suturing actions.

The second question is multiple parts and open-ended with a range of specified answer choices for the mTurkers responses. The question was designed to ask for the number of unique time periods or distinct activities in the video for cutting, tying and suturing actions.

The third question is a multiple-choice question with four distinct answer choices. This question was designed to test for the mTurkers ability to identify a unique cutting activity by prompting them to identify when the chronologically second distinct cutting activity occurred. The four answer choices are all specific time periods with a beginning and end timestamp.

The fourth question similarly is a multiple-choice question with four distinct answer choices. The question was designed to test the mTurkers ability to identify a unique suturing activity. The prompt asked to identify the third suturing activity chronologically seen in the video. The four answer choices provided specific time periods denoted with beginning and ending timestamps to select from.

The fifth question follows the format of the previous two multiple-choice questions with four distinct answer choices. This question tasks the mTurkers with identifying a suturing activity. The prompt asks to identify the time period associated with the fifth suturing activity seen chronologically in the surgical video. The four answer choices again provided specific time periods for reference with beginning and ending timestamps to select from.

The sixth question is a multiple-choice question with four distinct answer choices. The question prompt tasks mTurkers with associating a specific time period with one of the three unique surgical activities of focus. The question provides a timestamp in the early portion of the video of 01:34-01:50 and asks to identify the one distinct activity that occurs within that timeframe. The four answer choices provided are cutting, suturing, tying and n/a.

The seventh question is a multiple-choice question with four answer choices. The question is designed similarly to the previous question. The aim of this question is to test the mTurkers ability to associate a specific timeframe with one of the three unique surgical activities of focus. The question prompt provides a timestamp in the latter portion of the video of 06:02-06:10 and asks to identify the unique activity shown in that timeframe. The answer choices are cutting, tying, suturing and n/a.

The eighth question is a multiple-choice question with four answer choices. This question is designed to test the mTurkers by asking them to identify the number of unique surgical activities they annotated in a specific timeframe. The timeframe provided is 05:50-06:30 and the answer choices given are 2, 3, 4, or 5 unique surgical activities.

The ninth question is a multiple-choice question with four answer choices. The format of this question follows the design of the previous question. The intention of this question is to test the mTurkers ability to identify the number of unique surgical activities in a different specific timeframe. The question prompt provides the timeframe of 01:19-01:53. The answer choices given for this question are 1, 2, 4, or 5 unique surgical activities.

Table 2: Demographic survey questions and answer choices.

Question	Answer Choices
Question 1: Please select your age group.	18-25; 26-35; 36-45; 46-55; 56-65; 66-75; 75+
Question 2: Please select your gender identity.	Female; Male; Transgender Female/MTF; Transgender Male/FTM; Non-binary/third gender; Prefer not to say
Question 3: Please select your current education level.	High School; Trade School; College; Graduate School (Medical, Law, Dental, etc.); Post Graduate Training
Question 4: Before completing this tutorial, did you have any previous surgical annotation experience?	Yes; No; Unsure

<p>Thank you for completing our experiment. Please leave any questions, comments or feedback below.</p>	<p>Open-ended response</p>
---	----------------------------

The demographic survey appended to the end of the surgical annotation survey on the HIT comprised of four optional questions that helped provide data on our mTurkers demographic characteristics. This information would be utilized to inform improvements and iterations for further mTurk surgical annotation HIT data batches received.

The demographic survey questions were designed to get a snapshot of mTurkers characteristics and background. The questions included those asking to identify their specific age cohort, gender identity, and highest education level at the time of completing this HIT. Another important demographic data point collected was the mTurkers' previous experience with surgical annotation prior to this HIT. Lastly, an open-ended response page was presented to mTurkers for feedback on the experiment. This feedback would be catalogued and considered when iterations of new surgical annotation survey HITs were created for subsequent data collection.

The Amazon mTurk HIT was published on December 28th, 2021. The title of the HIT as visible to mTurkers was "Recognize and Record Actions in Surgical Videos by Completing a Learning Module and Answering a Survey". The title was intended to be specific and complete to help mTurkers understand the task. The description explained that the tasks involved a learning module and survey and that the goal was to test the ability to annotate surgical videos. This description elaborated on the title and gave an

idea of the commitment for effort and time required to complete the HIT. The keywords that were chosen were learning, annotation, video, surgery, and training. The keywords were selected to guide mTurkers searching for HITs and to help attract the desired number of respondents.

Table 3: mTurk survey scheduling details and experimental parameters.

Survey options	Selection
Reward per response	\$11.00
Number of respondents	50
Time allotted per Worker	5 hours
Survey expiration	14 days
Auto-approve and pay Worker deadline	3 days

The mTurk site requires requesters to specify payment and scheduling details for the HIT prior to publication. The experimental parameters, listed in Table 3, were selected were advised by Dr. Gabriel Brat, and research associate, Jevin Clark. The number of respondents and reward per response were selected via researching mTurk HITs and acknowledged the time commitment of the HIT. The time allotted per mTurker was selected upon researching average response times for HITs from mTurk forums. The survey expiration date was set at 14 days if 50 respondents had not yet been approved. The auto-approve period was selected as three days, which would pay mTurkers if not manually verified in that time. The mTurk site also provides options for worker requirements that can be customized. The only notable qualification selected for the first

batch of this HIT was that the United States was there designated location. Lastly, the task was set to be visible to only those mTurkers/workers that qualified for the HIT.

RESULTS

The results gathered from the mTurk survey HIT included 70 completed assignments with 50 assignments approved upon manual verification of “good faith” efforts to complete the annotation survey. The remaining 20 assignments completed by various Amazon mTurkers were rejected for a myriad of reasons including but not limited to: incomplete answers to the annotation survey, incoherent answers to the annotation survey or automated answers completed by computer bots. Thus, the approval percentage for assignments was 71.43 % and the rejection percentage was 28.57 %.

The time elapsed on the mTurk project page from the date of publication to verification of 50 approved assignments was 27 days. The average time to complete the HIT was 83.27 minutes amongst the 50 approved assignments. The results from the mTurk project page allowed for the download of .csv files that listed every completed assignment in a spreadsheet.

Completion of the mTurk survey assignment required mTurkers to input the unique survey code they received at the end of the Qualtrics survey. This code enabled the manual verification of the assignments done by mTurkers as the results were cross-checked with a spreadsheet downloaded from the Qualtrics account.

The first question on the survey, as seen in Table 1, asked for mTurkers to list the specific time periods for each distinct surgical activity cutting, suturing and tying. The following analysis was done based on the timestamp data acquired from the responses to said question. In order to analyze the timestamp data, there was a need to create a system to align the various timestamp responses provided by mTurkers.

The primary tool used was aligning the timestamps between the survey answer key and the mTurkers responses. This allowed for measures of central tendency to be calculated against the answer key while accounting for variability in mTurkers responses for the number of timestamps identified. Due to the open-ended response formatting of the first question on the survey, mTurkers who submitted accepted assignments exhibited a wide variation in the number of timestamps identified and listed. By aligning the timestamps, measurements and data analysis were more easily calculated. A sample calculation is shown below in table format.

Table 4(Left): Subset of respondent data for Question 1 of survey that included 2 and 4 timestamps. Table 5(Right): The answer key for the suturing timestamp data.

SUTURING RESPONSES(min:sec)		
2 Timestamp Answers		
03:17-03:24		
03:05-03:24		
03:05-03:24		
03:05-03:12		
03:17-03:24		
03:05-03:12		
00:25-00:28		
01:43-01:52		
03:05-03:12		
02:54-03:54		
03:05-03:12		
4 Timestamp Answers		
03:05-03:12, 03:17-03:24		
01:35-02:19, 02:37-03:18		
01:40-02:10, 02:40-02:44		
03:05-03:12, 03:17-03:24		
02:39-02:42, 02:54-03:04		
	ANSWER KEY(sec)	
	97	
	103	
	117	
	122	
	159	
	162	
	191	
	202	
	212	
	220	
	235	
	254	

As seen in Table 4 (Left), there is a portion of the responses from mTurkers. The data is in “minute: second” format. The responses listed under “2 Timestamp Answers” are a subcategory of the suturing timestamp data, strictly displaying the mTurkers with responses that included just one beginning and end times. Table 5 (Right) is displaying the answers for the correct timestamps for suturing activities. There are 12 times listed in seconds format, with every 2 times considered a grouping of timestamps. Thus 97 seconds and 103 seconds are in minute: second format listed as 01:37 – 01:43.

This data structure is repeated for all variations of the responses gathered from the mTurkers survey results. The data is in a similar format for the cutting and tying datasets. The mTurkers respondent data included responses from 2 timestamp answers up to 16 timestamp answers. The process of aligning the timestamps allowed for analysis and comparison of the timestamp data to the answer key between all variations of responses.

Table 6: Suturing respondent data subset for 2 timestamp calculations relative to answer key.

2 Timestamps (sec)		2 Timestamps Difference (sec)	
197	204	-100	50
185	204	-88	50
185	204	-88	50
185	192	-88	62
197	204	-100	50
185	192	-88	62
25	28	72	226
103	112	-6	142
185	192	-88	62
174	234	-77	20
185	192	-88	62

Table 6 displays a sample portion of the suturing dataset, with calculations done to align the timestamps to the answer key. The columns listed under the heading “2 Timestamps (sec)” are the beginning and end times in second format previously seen in Table 4. For example, 197 seconds and 204 seconds correspond to 03:17 – 03:24. In order to calculate the aligned difference, the beginning and end times for each response are subtracted from the beginning and end times of the answer key in Table 5. Thus, for the values of 197 seconds and 204 seconds, the aligned timestamp difference in seconds is -100 and 50 respectively. This is calculated by subtracting the answer key value, 97, from the respondent value, 197, to get -100. A similar calculation is done by subtracting the end answer value of 254 from the respondent value of 204, resulting in an aligned difference of 50.

Table 7: Suturing respondent data subset for 4 timestamp calculations relative to answer key.

4 Timestamps (sec)				4 Timestamps Difference (sec)			
185	192	197	204	-88	-89	38	50
95	139	157	198	2	-36	78	56
100	130	160	164	-3	-27	75	90
185	192	197	204	-88	-89	38	50
159	162	174	184	-62	-59	61	70

Table 7 displays a sample portion of the suturing dataset, with calculations done to align the timestamps for mTurkers responses that included 4 timestamps. The four columns under the heading “4 Timestamps (sec)” list the various responses for the 4 timestamps provided in the responses. One set of responses includes the values, 185, 192, 197, and

204 seconds. These values match the 4 timestamp answers in minute: second format seen previously in Table 4. The values in minute: second format are 03:05 – 03:12, 03:17 – 03:24.

In order to calculate the aligned difference with 4 timestamps it requires comparing the first two timestamp values of 185 and 192 to the first two answer key values while comparing the last two timestamps' values of 197 and 204 to the last two answer key values. The first two answer key values, seen in Table 5, are 97 and 103 seconds. The last two answer key values are 235 and 254 seconds. Thus, the aligned differences are calculated as followed: $(97 - 185) = -88$, $(103 - 192) = -89$, $(235 - 197) = 38$, and $(254 - 204) = 50$.

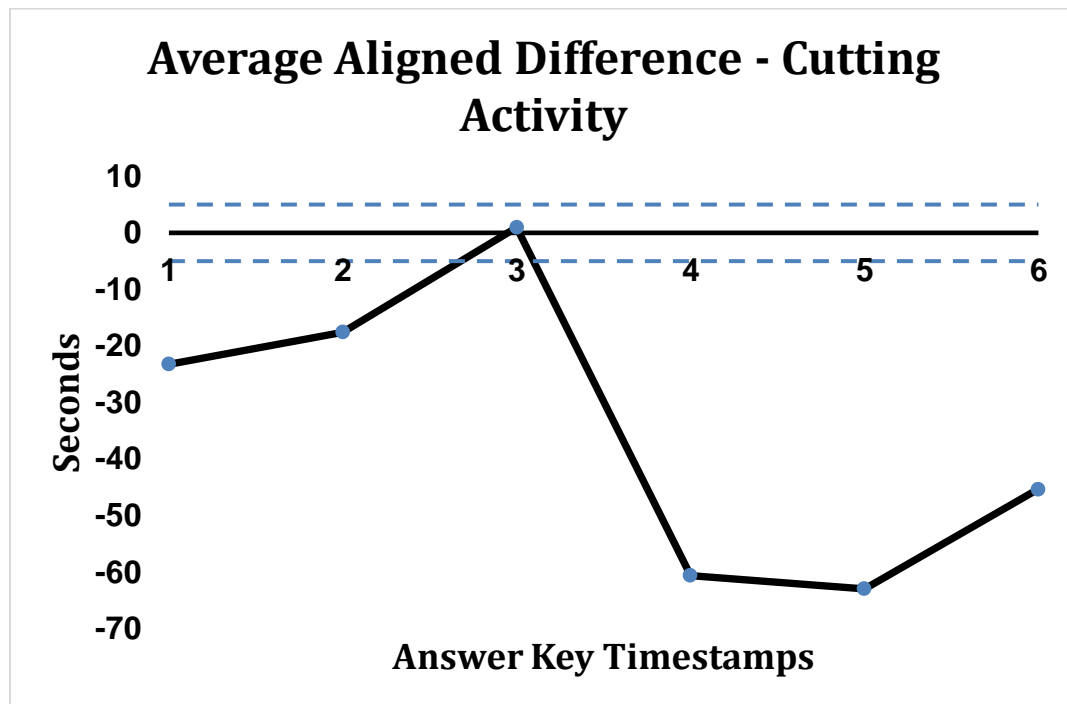


Figure 2: Average Aligned Difference for Cutting Activities. Error range was set at +/- 5 seconds.

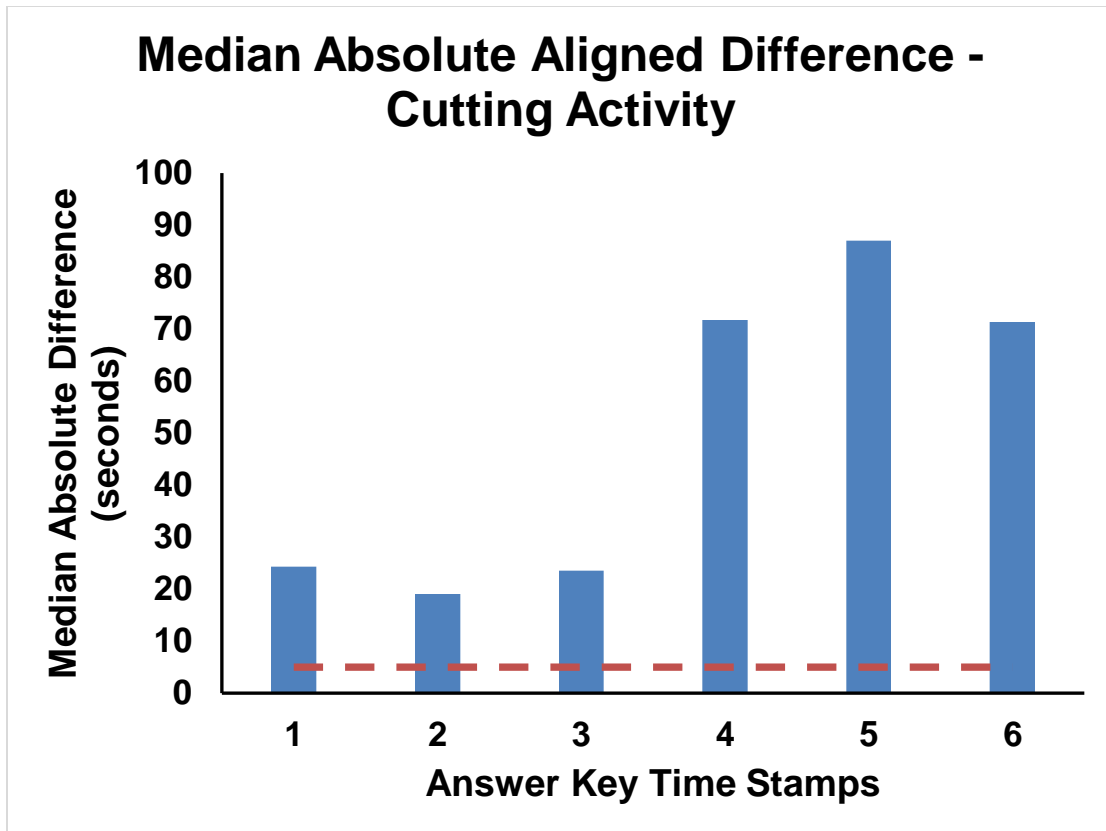


Figure 3: Median Absolute Aligned Difference for Cutting Activities. Error range was set at +5 seconds.

Figures 2 and 3 are a graphical representation of analysis done on the responses given to the first question on the survey. Figure 2 shows the average aligned difference in seconds between the cutting activity timestamp responses and the survey answer key responses.

The dotted lines in both figures represent a five second margin of acceptable error. Two measures of central tendency were measured for the given timestamp data, average and median aligned difference. For the average aligned difference, the non-absolute value of the data was used, as seen in the responses.

The average aligned difference for the cutting activity timestamps was only within the accepted margin of error for the 3rd timestamp. All the other timestamps can be seen

having negative values, depicting the average response being earlier than the answer key value for the cutting activity. For the median difference in timestamps, as depicted in Figure 3, the absolute value of the aligned difference was used. This analysis resulted in all 6 cutting activity median differences being outside the accepted margin of error. The values for the median absolute difference being positive indicates that the median response was greater than 5 seconds later than the answer key timestamps for that specific cutting activity. For both average and median aligned difference, the greatest magnitude of difference occurred in the 5th cutting activity timestamps data. For the 5th cutting activity the average aligned difference was -62.96 seconds and the median absolute aligned difference was 87.04 seconds.

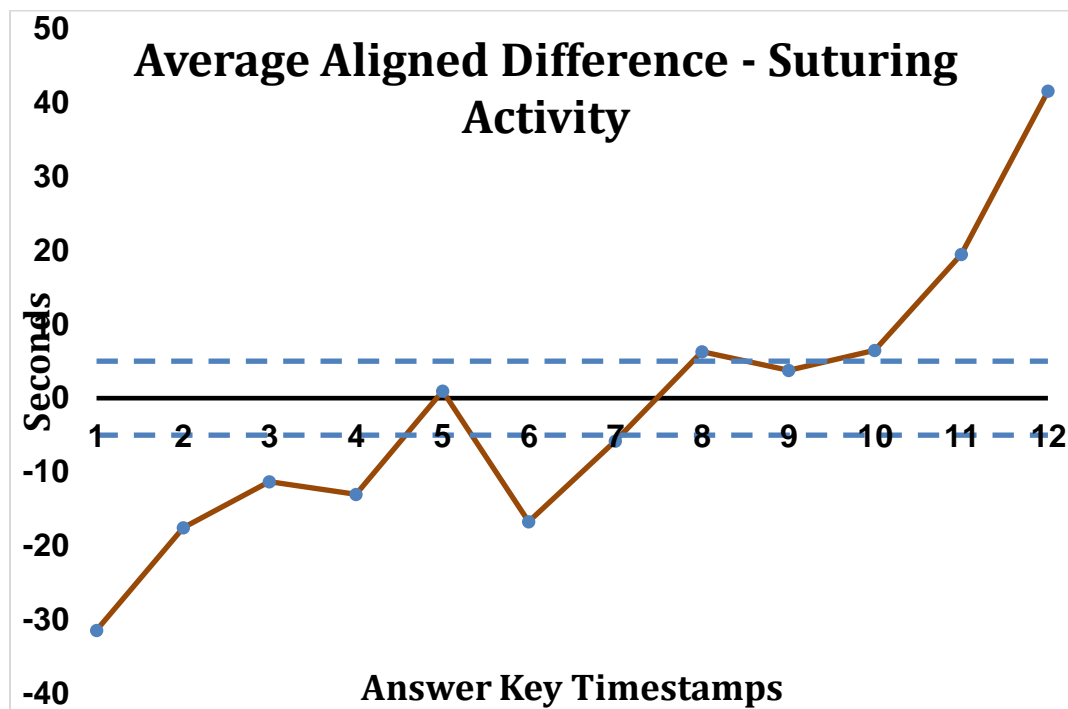


Figure 4: Average Aligned Difference for Suturing Activities. Error range was set at +/- 5 seconds.

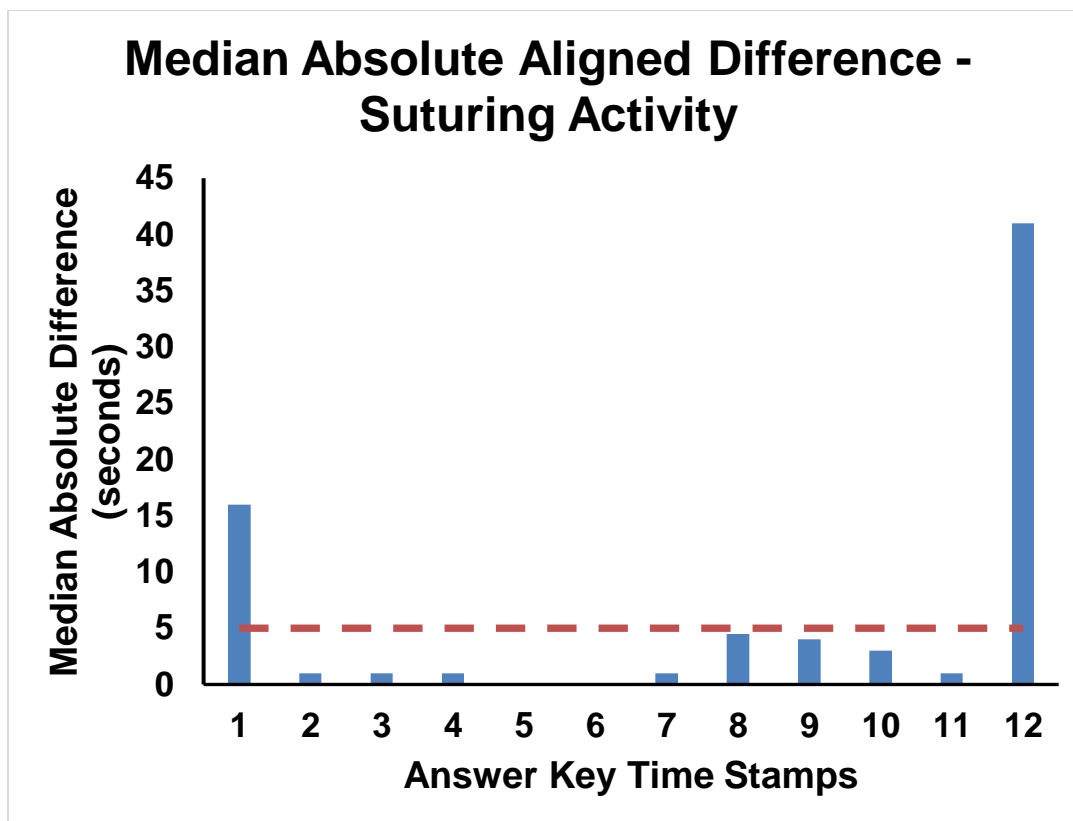


Figure 5: Median Absolute Aligned Difference for Cutting Activities. Error range was set at +5 seconds.

Figures 4 and 5 are graphs that show the analysis done for the responses to the first survey question. The focus of these two figures is to analyze the difference between the suturing activity timestamps submitted in the survey response compared to the answer key. The two measures of central tendency calculated were the average and median aligned differences. As seen in Figure 4, the data was depicted as non-absolute values. For the median differences in timestamps, the data was represented as absolute values. The accepted margin of error for these graphs was set at plus or minus 5 seconds from the answer key timestamp.

The data in Figure 4, shows that only timestamps 5 and 9 were within the accepted margin of error. The average aligned difference for timestamps 1-4, 6, and 7 were all negative values. These negative values indicate that the average response for the suturing activity annotated early or prior to the answer key timestamp. For timestamps 5, and 8-12 the timestamps were positive values, indicating that the average response for this suturing activity was annotated after the answer key timestamp. The average aligned difference is greatest in magnitude for suturing activities at the beginning and end of the video as the first and last timestamps are -31.38 seconds and 41.5 seconds respectively.

The data in Figure 5 shows that the median absolute difference for suturing activity timestamps was within the 5 second margin of error for all timestamps except for the first and last activities. The greatest difference in the median absolute difference occurred with the last activity, as seen with the 41 second median difference.

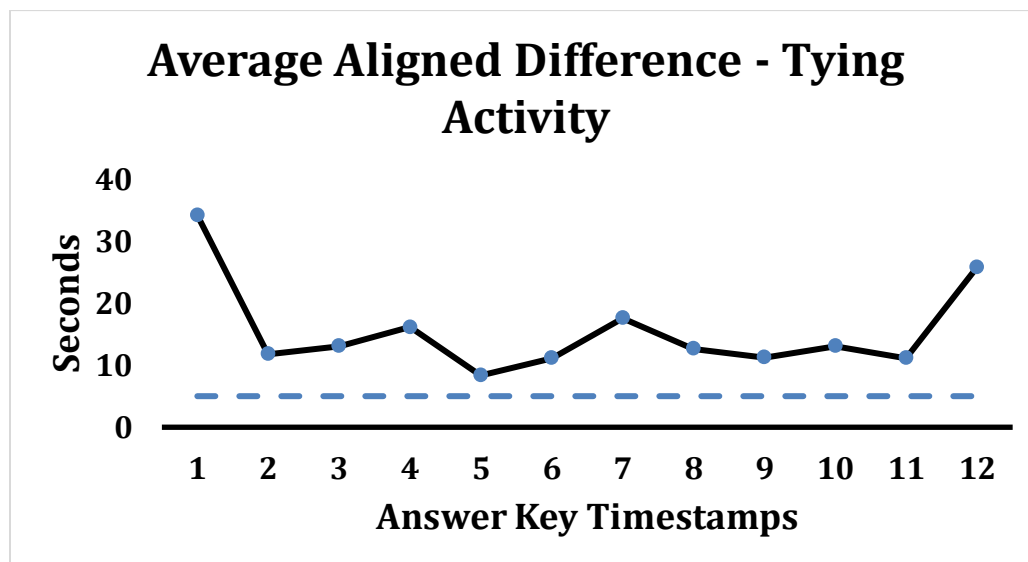


Figure 6: Average Aligned Difference for Suturing Activities. Error range was set at +/- 5 seconds.

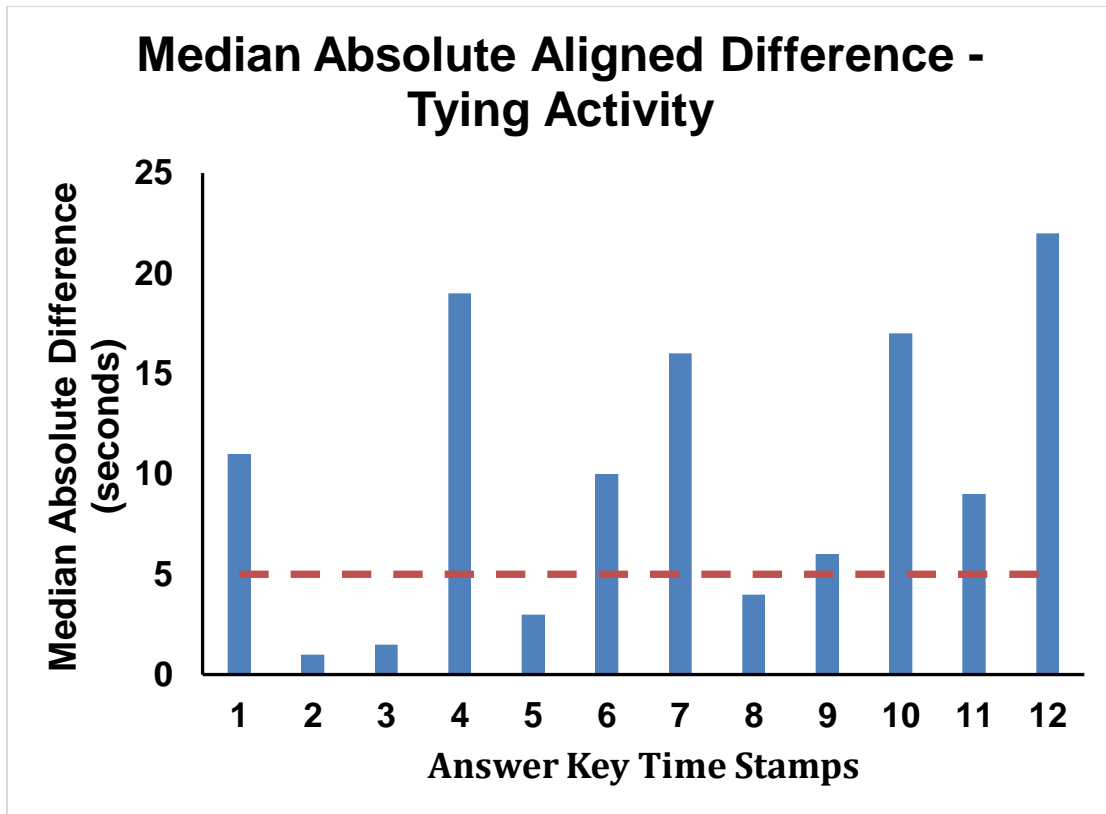


Figure 7: Median Absolute Aligned Difference for Cutting Activities. Error range was set at +5 seconds.

Figures 6 and 7 are graphs depicting the mTurk response data for the first question on the survey. These figures illustrate the analysis done using both average and median differences from the answer key as measures of central tendency. These figures focus on the responses gathered from the survey about annotating the specific tying activities in the video as timestamps. Figure 6 displays the calculated average timestamp difference for each tying activity when comparing the responses to the answer key timestamp, the values were in non-absolute value format. Figure 7 displays the calculated median timestamp difference for each tying activity between the responses and the answer key, with values converted to absolute values.

The average aligned difference was not within the accepted margin of error of 5 seconds for any of the tying activity timestamps. The greatest average difference occurred at the first and last timestamps, with values of 34.12 seconds and 25.77 seconds respectively. A similar trend can be seen when analyzing the median absolute difference data, as there is a 22 second median absolute difference for the last activity timestamp.

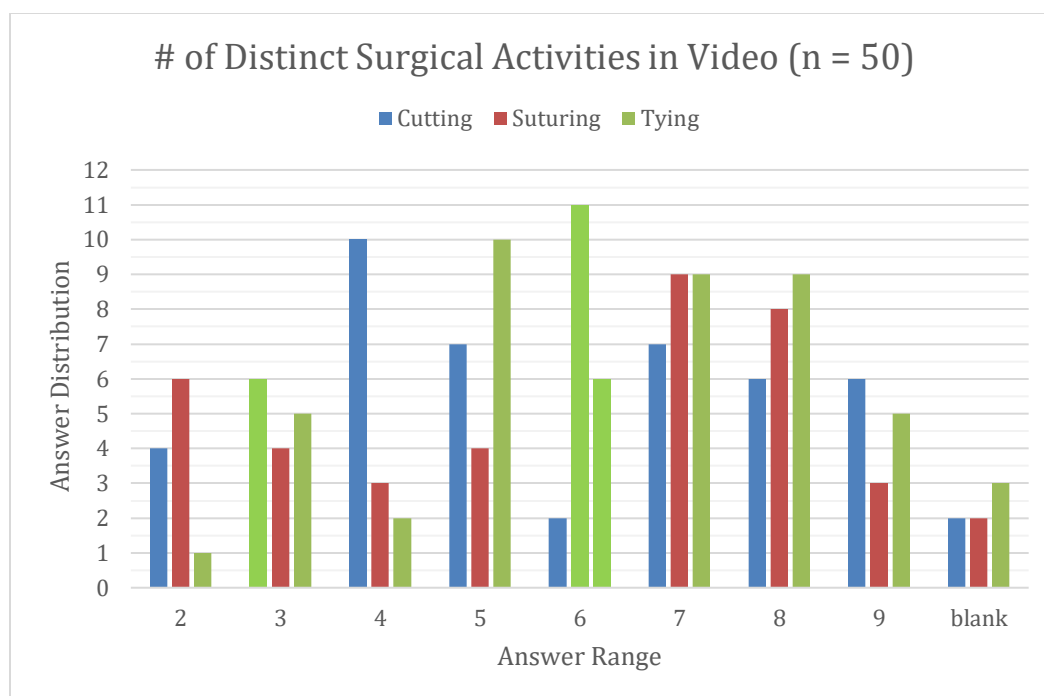


Figure 8: Question 2 on survey respondent distribution. Correct values highlighted in green: cutting = 3, suturing = 6, tying = 6.

The second question on the survey asked mTurkers to select between 2 and 9 of the number of distinct surgical activities they annotated in the video for the three actions of cutting, suturing and tying respectively. The correct number of distinct activities can be seen in the green highlighted bars of the graph. The correct number of distinct activities for cutting, suturing, and tying respectively were 3, 6, and 6. The mode value for the cutting activity subset was 4 activities with 10 out of 50 mTurkers selecting that choice.

The number of mTurkers that correctly identified the number of cutting activities as 3 activities was 6 out of 50. The mean value for the cutting activity subset, excluding the blank responses, was 5.5 activities. The median value, excluding the blank responses, was 5 cutting activities. The mode value for the suturing activity subset was 6 activities with 11 out of 50 mTurkers selecting that choice. The mode value matched the correct value for annotation of suturing activities. The mean value for the suturing activities subset, excluding the blank responses, was 5.75. The median value, excluding the blank responses, was 6 suturing activities. The mode value for the tying activity subset was 5 activities with 10 out of 50 mTurkers selecting that choice. The number of mTurkers that correctly identified the number of tying activities as 6 activities was 6 out of 50. The mean value for the cutting activity subset, excluding the blank responses, was 6.19 activities. The median value, excluding the blank responses, was 6 tying activities. The respective measures of central tendency for the specific surgical activities are listed in table format along with the correct values.

Table 8: Summary of data presented in Figure 8. Measures of Central Tendency included.

	Cutting	Suturing	Tying
Correct Value	3	6	6
(% of respondents that identified)	(12%)	(22%)	(12%)
Mean	5.5	5.75	6.19
Median	5	6	6

Mode	4	6	5
(# of respondents that identified)	(10)	(11)	(10)

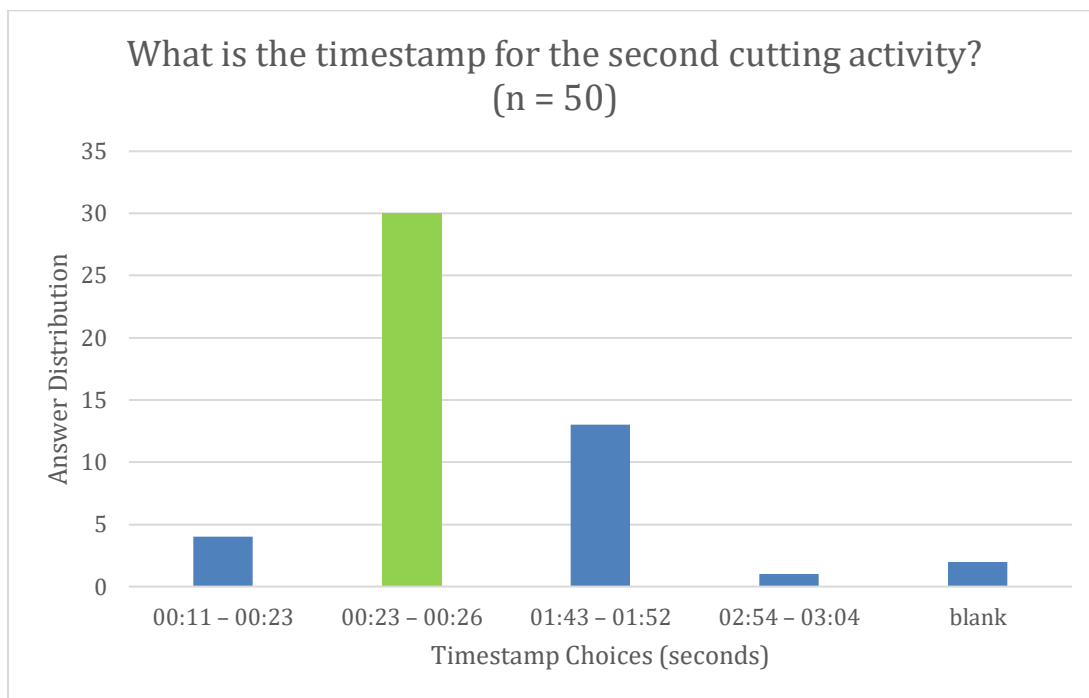


Figure 9: Question 3 on survey. Correct answer highlighted in green.

The third question asks to identify the timestamp that correlates with the second cutting activity in the surgical video. The correct answer was the timestamp of 00:23 – 00:26. Amongst the 50 accepted assignments 30 correctly identified this timestamp, a percentage of 60 %. The next most common answer with a timestamp of 01:43 – 01:52, which was a tying activity, was identified by 13 accepted assignments, or 26%.

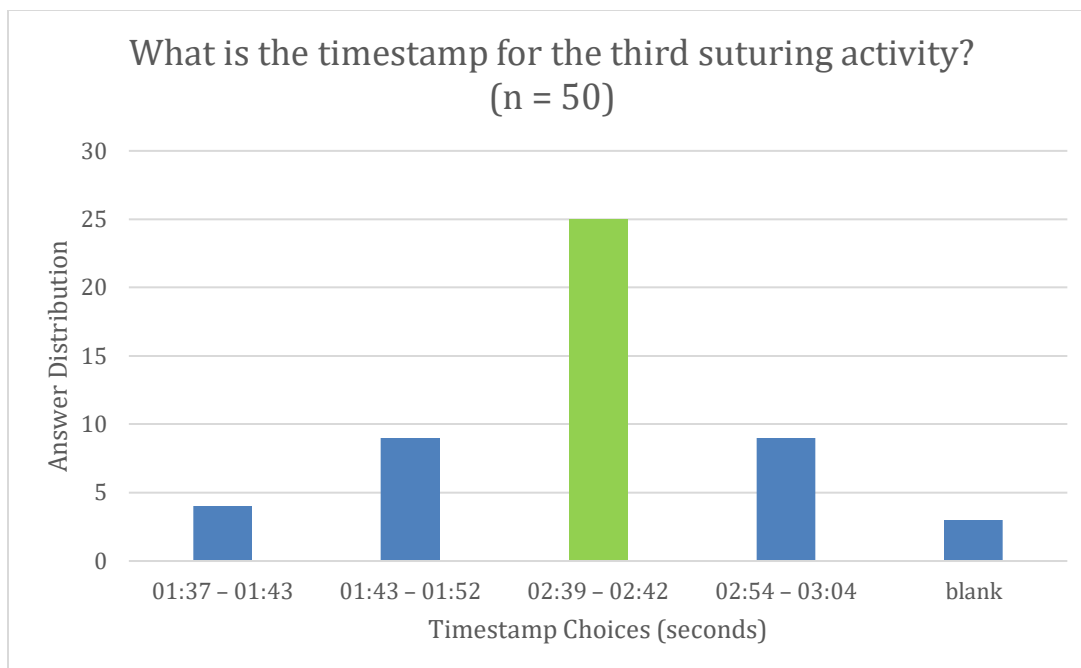


Figure 10: Question 4 on survey. Correct answer highlighted in green.

The fourth question asked respondents to identify the timestamp that correlates to the third suturing activity in the surgical video. The correct timestamp is highlighted in green as 02:39 – 02:42. The number of accepted assignments that correctly identified this timestamp was 25. The percentage of accepted assignments with the correct answer was 50%. The next most common answers, each identified by 9 accepted assignments or 18% respectively, had timestamps of 01:43 – 01:52 and 02:54 – 03:04. The timestamps of 01:43 – 01:52 and 02:54 – 03:04 were both tying activities.

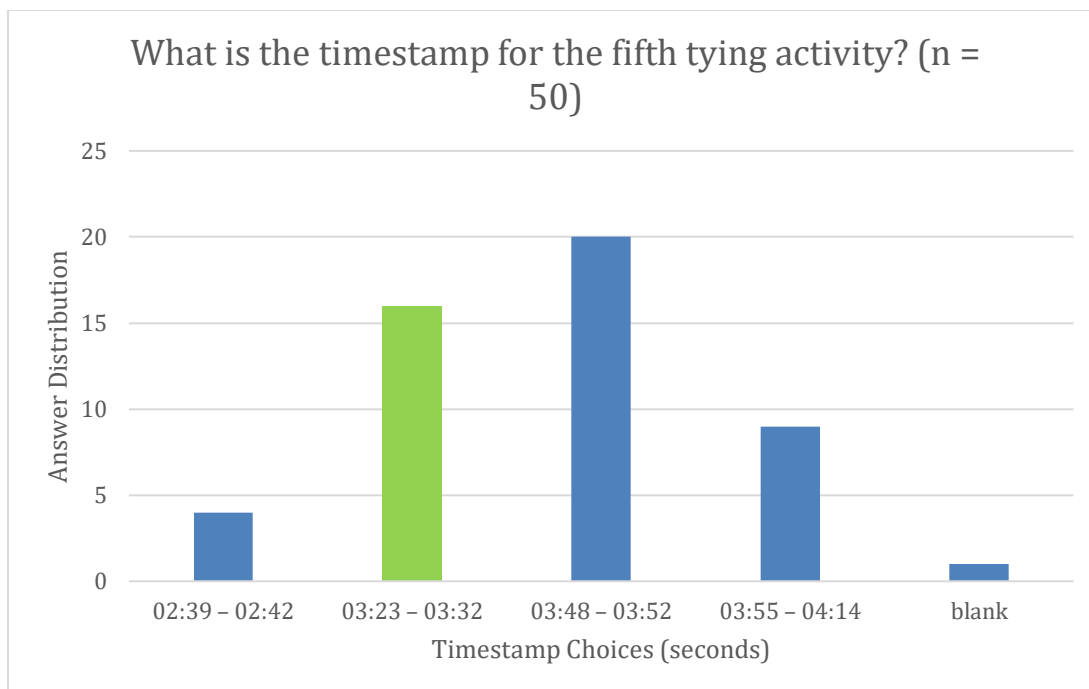


Figure 11: Question 5 on survey. Correct answer highlighted in green.

The fifth question on the survey asked respondents to identify the timestamp that correlates to the fifth tying activity in the surgical video. The correct response was the timestamp of 03:23 – 03:32. This was identified by 16 out of 50 accepted assignments or 32% correct. The answer choice that received the greatest percentage of responses was the timestamp of 03:48 – 03:52. This timestamp was a tying activity but was mistakenly identified as the fifth tying activity by 40% of the accepted assignments from mTurkers.

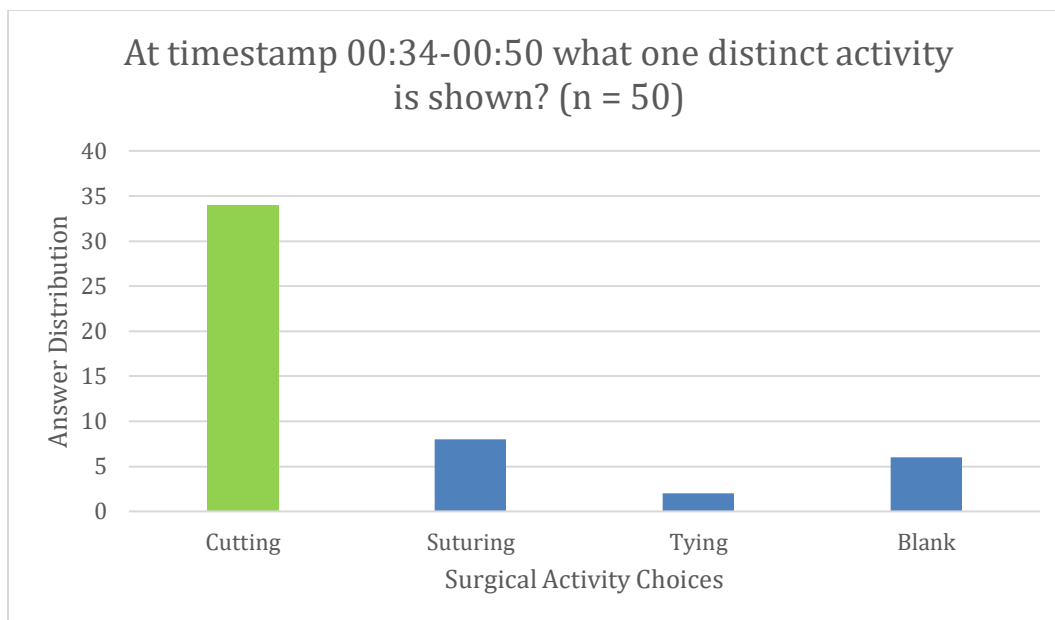


Figure 12: Question 6 on survey. Correct answer highlighted in green.

The sixth question on the survey asked respondents to identify the specific surgical activity that occurred within the timestamp of 00:34 – 00:50. The correct response, highlighted in green, was a cutting activity. Of the 50 accepted assignments 34 selected the correct choice. The percentage correct was 78%. The answer choice that received the next greatest percentage of responses was for a suturing activity, with 8 out of 50 or 16%.

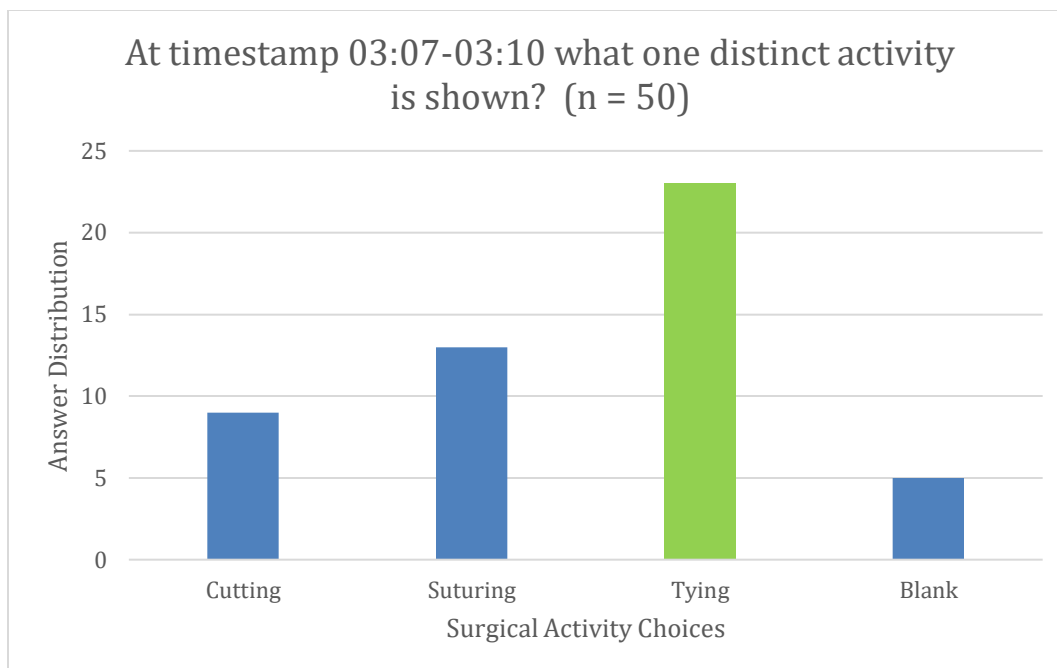


Figure 13: Question 7 on survey. Correct answer highlighted in green.

The seventh question on the survey asked respondents to identify the specific surgical activity that occurred within the timestamp of 03:07 – 03:10. The correct response, highlighted in green, is shown to be a tying activity. This answer choice was selected by 23 of the 50 accepted assignments, a percentage of 46% correct. The second most common answer choice seen with this question was the selection of a suturing activity occurring in this timestamp. There were 13 out of 50 or 26% of accepted assignments that selected this incorrect choice.

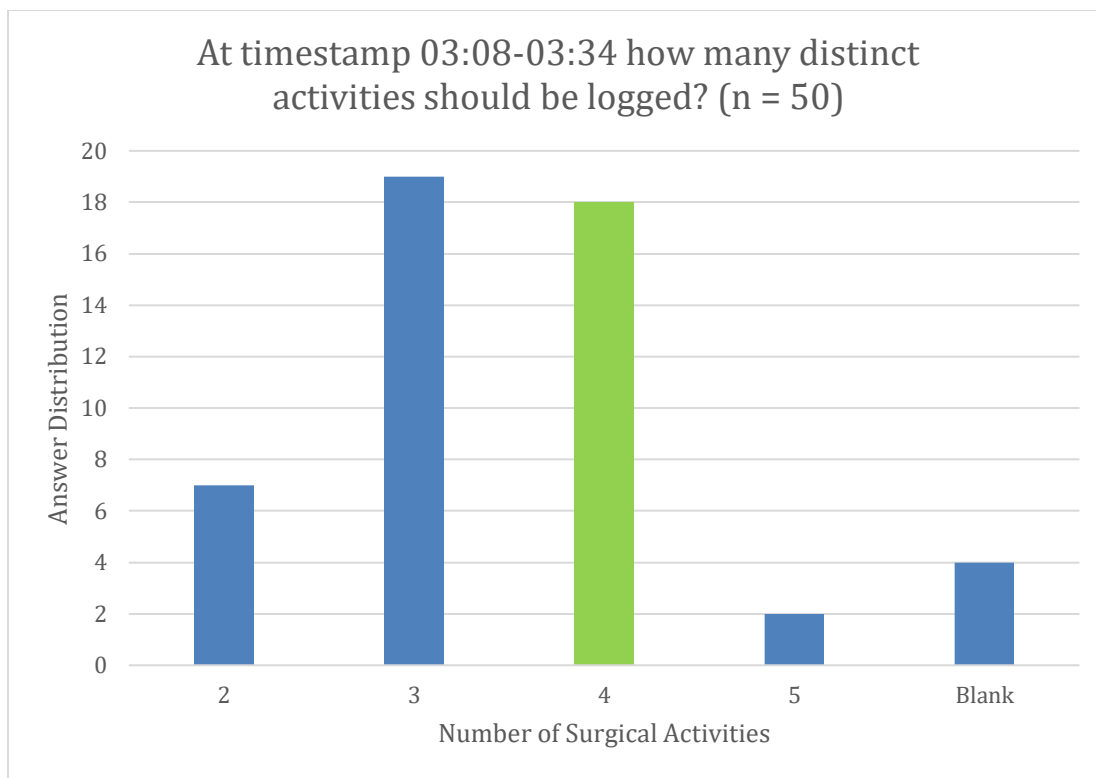


Figure 14: Question 8 on survey. Correct answer highlighted in green.

The eighth question on the survey asked respondents to specify the number of distinct activities, as defined in the learning module, that occurred within the timestamp of 03:08 – 03:34. The correct response, highlighted in green, is that 4 distinct activities occurred in that timeframe. The number of accepted assignments out of 50 that correctly chose 4 activities was 18 or 36%, the second most common answer selected. The most common answer choice selected was that 3 distinct activities occurred in that timeframe, with 19 out of 50 or 38% of accepted assignments selecting that choice.

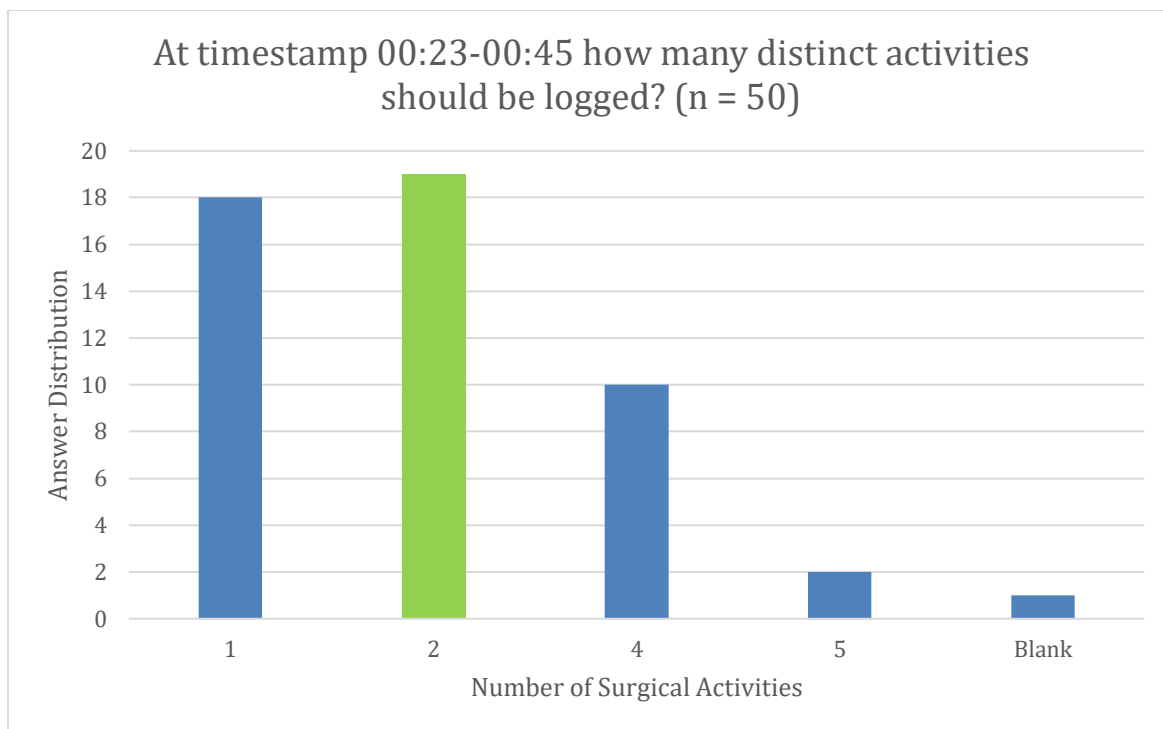


Figure 15: Question 9 on survey. Correct answer highlighted in green.

The ninth question on the survey asked respondents to specify the number of distinct activities that occurred in the timestamp of 00:23 – 00:45. The correct response, as shown above in green, is that 2 distinct activities occurred within that timeframe. The number of accepted assignments that correctly chose that value was 19 out of 50. The 38% that correctly identified the answer choice was the most common selection. The second most selected answer choice was that 1 activity occurred within that timeframe, with 18 out of 50 or 36% of accepted assignments selecting that choice.

DISCUSSION

Review of Findings and Problems Encountered

The first iteration of the mTurk surgical annotation survey was a learning experience that shed light on numerous initial trends that can be explored further in subsequent batches of data collection and iteration. The utilization of Amazon mTurk site offered interesting insight into its benefits, drawbacks, and limitations that also would benefit from further exploration and scrutiny. Additionally, the formatting and structure of the Qualtrics survey that was linked within the mTurk HIT elicited certain questions about how to structure the survey to help non-surgical users to complete the survey in eliminating unwanted confusion while aiding them to provide an optimal learning and testing environment for better results. Understanding and exploring these three main areas related to this initial iteration of this experiment can shed light on the potential for this experimental framework to accomplish the task of scaling up the speed and effectiveness of surgical video annotation for use in machine learning algorithms.

Utilization of Amazon mTurk as the host for gathering data in an open-source format resulted in a handful of observations throughout the course of data collection and analysis. There were clear benefits that were a result of the access to such a large population of potential mTurkers that would be useful for larger sample sizes in future iterations of this project. There were also noticeable issues and problems that caused concern. Focusing on the issues that were noticed, one primary issue discovered was the quality of the data received and the systems for verification of data prior to payment of participant mTurkers.

There were some participant mTurkers who did not fill out the survey at all. These participants utilized less time relative to the recommended HIT duration of 45-60 minutes. These results were either blank or illogical in the context of survey questions. This was a major issue in the critical open-ended question that asked respondents/mTurkers to list out the specific timestamps for the three surgical activities of cutting, suturing and tying. This question was the most critical portion of the results and of the 50 accepted HITs there was a wide variety of data, indicating confusion or lack of effort, with nearly 10% of respondents leaving these questions blank. Overall, there was a pattern of ambiguity in the data, perhaps caused by the format of the survey or possibly due to the fundamental culture of Amazon mTurkers. Additionally, the lack of an established system for data verification led to a wide variety of data and limited usable data.

There is a large percentage of unusable data that was financially rewarded and the system for verification of data to approve payment is a manual process. This requires downloading two separate files from the Qualtrics survey site and from the Amazon mTurk project management page. The simplified rule agreed upon by the research team was to accept good faith efforts done to accomplish the task.

Analyzing the data received from the 50 accepted assignments or HITs from the mTurk survey reveals numerous observations and patterns that are promising and ripe for further exploration and experimentation. One pattern that was noted was the consistently large margin of error for the cutting activity aligned differences for both average and median

absolute values as can be seen in [Figures 2](#) and [3](#). The mTurkers inability to correctly annotate the number of activities for cutting is exhibited in [Table 8](#) as well.

When asked to identify the number of distinct surgical activities for each action, the mTurkers were unable to do so specifically regarding identifying the cutting activity.

This stands in contrast to greater accuracy in identifying the number of distinct suturing and tying activities. As seen in [Table 8](#), the correct number of cutting activities in the survey video was 3, yet the mean (5.5), median (5), and mode (4) were all incorrect and not +/- 1 activity from being correct. One can conclude, by looking at the data for suturing and tying activities, that respondents had a greater ability to distinguish the correct number of activities, which was 6 for both. For the suturing activity data, the mean (5.75), median (6), and mode (6) were all within +/- 1 activity of the correct answer. Similar accuracy in identifying the correct number of activities for tying, can be seen when noting the mean (6.19), median (6), and mode (5), as all 3 values are within +/- 1 activity of the correct answer.

The difficulty in annotation of cutting activities in the video are more pronounced than with suturing and tying. The reason for this requires further experimentation to concretely assess however a hypothesis that can be considered is that mTurkers are less familiar with the tools involved with cutting and how the actions present themselves in video. The tools used to cut that were depicted in the learning module of this experiment were a scalpel and an electrocautery device. The scalpel has a more identifiable blade appearance and has conventionally been used as a surgical instrument to make incisions because of its practicality, low-cost and accuracy for procedures (Vore et al., 2002). The

prototypical cutting instrument being analogous to a scalpel contrasts with the other major cutting tool discussed in the learning module, an electrocautery device.

Electrocautery devices are tools that utilize an electric current to create incisions, destroy tissue, and coagulate blood in surgical procedures. They are one of the staple tools for surgeons and many studies have reported on its usefulness and the upward trend in utilization of electrocauteries due to its flexibility in surgical settings (Massarweh et al., 2006). Since electrocauteries can perform various actions that are often used consecutively in surgical procedures, it is possible that the source of confusion for cutting annotation in our experiment is a result of this unfamiliarity for mTurkers. In future iterations of the mTurk surgical video annotation survey, testing for this distinction between the recognition of a scalpel versus an electrocautery device could provide useful insights into the complexities required in surgical video annotation.

Analysis of the suturing and cutting activity data from both the open-ended timestamp question and the multiple-choice questions hints at a few notable patterns and developments that could be explored and scrutinized further. The first noteworthy pattern for suturing activity was that the first and last activities were annotated the most incorrectly. These trends can be seen in [Figures 4](#) and [5](#) with the magnitudes for average aligned difference and median absolute aligned difference being greatest for timestamps 1 and 12. In contrast, the middle activities for suturing were identified correctly, within the error margin of 5 seconds, for timestamps 2-11 in [Figure 5](#) depicting the median absolute aligned difference. This seems to indicate that mTurkers were able to identify suturing activities as it is possible that the sources of error for timestamps 1 and 12 were

due to extreme outliers in the data. The median absolute aligned difference for timestamp 12 being approximately 41 seconds is the largest margin of error exhibited in all the data analyzed. Looking at the surgical video, this indicates that the last timestamp for a suturing activity was incorrectly annotated to a large degree, with many respondents annotating an entirely different clip occurring in the surgical video.

Another possible variable that could be influencing the levels of error seen in those timestamps is a misunderstanding of suturing activities as depicted in the learning module. The learning module stipulates that suturing begins when the needle driver is guiding the suturing needle and surgical thread into the skin adjacent to an incision. Subsequently, the end of a suturing action occurs when the needle driver begins to handle the surgical thread to begin a tying action. If the learning module was not read correctly or was read correctly but confusion remained, mTurkers would be left having a misunderstanding of the specific parameters that guided the answer key timestamps for each distinct suturing activity.

The hypothesis mentioned above, regarding a misunderstanding of how the learning module classifies the beginning and end of suturing can also be a potential influence in the errors seen for tying activity data. As previously mentioned, suturing and tying activities are closely related as both involve the use of surgical threads to patch up incisions made. The key distinction that the learning module makes as a training tip for mTurkers completing the survey is that once the surgical thread is being handled by the needle driver, that is the beginning of a tying activity.

Another trend noticed is that the median difference for tying activity timestamps as seen in [Figure 7](#) was often greater than 10 seconds late, which could be due in part to a confusion or misinterpretation of when tying activities are completed. The learning module again stipulates that the end of the tying activity is classified when the last knot is pulled tight, not when the surgical thread is being cut. Since the surgical thread is only cut after the last knot is generally pulled tight, this error could be contributing to a positive median absolute aligned difference for the tying activity data.

A potential contrast in the ability to accurately recognize and annotate the two types of surgical activities, suturing and tying, has potential for further exploration and consideration in further studies. When utilizing the data depicting the median absolute aligned difference, there were 3 out of the 12 timestamps that were within the accepted margin of error. Looking at this data, there was a clear misunderstanding or inability to accurately annotate the timestamps for tying activities. This becomes more evident and curious when a comparison is made with suturing activities. As seen in the graph depicting median absolute aligned difference, all 10 of the middle timestamps of 12 overall fell within the accepted 5 second margin of error.

In order to grasp the patterns and trends discussed above these findings and observations have been listed below in Table 9.

Table 9: Summary of Observations/Trends discussed.

Observation/Trend	Brief Explanation
-------------------	-------------------

<p>Scalability of mTurk for surgical video annotation</p>	<p>The ability to expand the experiment and increase data collection with little effort in the set-up phase bodes well for large-scale surgical video annotation studies</p>
<p>Incomplete data</p>	<p>10% of HITs accepted left some portion of survey blank, even greater percentage of HITs rejected left portions or entire survey blank</p>
<p>Low yield of coherent data for Q1</p>	<p>Of the HITs accepted many failed to answer Question 1, the open-ended timestamp annotation question, which was of paramount importance</p>
<p>Large margin of error for cutting activity data</p>	<p>In Q1 and Q2, there was a clear inability to accurately annotate or identify the times for cutting activities or even the # of activities</p>
<p>Measures of Central Tendency in Q2 for suturing and tying activities</p>	<p>This question found that respondents were mostly able to identify the correct # of activities for suturing and tying when ignoring outlying data</p>

<p>Misunderstanding of difference between suturing and tying activities in learning module</p>	<p>It is possible that a variable that impacted levels of error for annotating suturing and tying occurred due to confusion about when a suturing activity ends and tying activity begins</p>
<p>Misunderstanding of when tying activities end in learning module</p>	<p>Trend seen with great positive errors for many of the tying activity timestamps. Perhaps due to not knowing when tying ends. This occurs when knot is pulled taught not when thread is cut.</p>

Review of Potential Solutions and Improvements

After analyzing the data there are questions that can be considered and used to inform further exploration into surgical video annotation in crowdsourcing settings. First, when debating the use of an accepted margin of error, to what extent does the margin of error matter for metrics such as median absolute aligned difference when comparing data for activities such as suturing and tying if a majority of the timestamps are well outside the accepted margin of error. In such a situation both points of data are fairly considered as missed annotations with just the magnitude of error being the difference. Perhaps a comparison can be made as it is a median value and thus is significant because it speaks to how many respondents near median were closer to the actual answer.

There are some potential issues to consider as well when evaluating the data and considering improvements to this concept of utilizing non-experts for surgical video

annotation. Firstly, there is wide variability in the source material for annotation. Many of these surgical educational videos are posted on websites such as YouTube for educational purposes thus there is a lack of control for factors such as video quality, video length, and video editing. One factor that could have impacted the inaccuracy for annotating surgical activities in the video associated with the mTurk survey was the manner in which the video clips were stitched together to provide a cohesive understanding for the surgical procedure. Previous studies in surgical video annotation have highlighted this potential hurdle as many studies and projects have resorted to editing existing surgical video clips with the help of surgeons and medical professionals. These surgeons and medical personnel have re-edited videos in order to limit the scope of non-surgical actions occurring on screen to limit distractions and background actions. This has been done in the hope of seeing more accurate video annotation accomplished by non-surgical experts (Ward et al., 2021). Understanding how video clips are spliced together in this amateur, point-of-view videos can impact the accuracy of surgical video annotation and is a variable that must be considered when further exploring this field.

Another important source of variability that has been noted that stands to complicate tasks to standardize surgical video annotation is surgical hand coordination and dexterity variability. The experiment published on Amazon mTurk with the learning module and survey importantly had two different video sources. Considering the artistic parallels when conducting surgery using hand tools such as scalpels, electrocauteries, and sutures there is undoubtedly variability in technique, pacing and hand coordination among surgeons. This has been studied at length in studies on surgical technical ability

evaluations utilizing computer vision and ML technology. The conclusions from these studies have been that there is high variance in the dexterity in situations where surgeons require the use of both hands, classified as “bimanual” (Law, 2017). This variability in surgical procedure can impact how a surgeon maneuvers their hands and tools to create incisions, the method and pace in which they use an electrocautery to cut or dissect tissue, and how they suture and close repaired surgical wounds. Additionally, there is similar variability in the actions of right-hand and left-hand dominant surgeons that could impact surgical techniques and thus the ability for non-surgical experts to analyze and annotate surgical videos. One additional step in the experimental process that could improve variance in annotation results that accounts for hand coordination and dexterity variability is a categorization process that accounts for surgeon variability by labelling videos with common variables.

There are some interesting potential improvements or solutions that could be suggested and be worthy of further consideration to make surgical video annotation a more scalable process to gather greater sums of data for ML. From a broad perspective, the utilization of a multi-layered system that integrates the current process of strictly relying on surgeons and medical expertise to annotate videos with larger more robust mechanisms such as crowdsourcing to process greater amounts of surgical video data. This process has been explored using the same Amazon mTurk crowdsourcing capabilities yet variations of the order in which to verify the video annotations still requires further consideration (Deal et al., 2017).

On a more micro scale, there are potential alterations to the system utilized to conduct this experiment, between survey alterations with formatting and new questions, a more robust and structured system to verify responses for payment approval to mTurkers, and more to consider when ideating on future improvements to the survey and experiment. Overall, there was a 71.43% approval rate amongst the 70 completed assignments, resulting in the goal of 50 approved assignments/HITs. It took 27 days to arrive at that number of approved assignments. The core issues that plagued this low percentage and lengthy period of data collection were the number of automated computer bots that was deployed to complete the mTurk HIT, large volumes of incoherent uncompleted HITs that lacked good faith efforts, and a manual system for payment verification of completed assignments. These types of problems have long been documented as hurdles for data gathering in research and business contexts on Amazon mTurk, with the two problematic cohorts being labelled as “cheaters” and “speeders” (Smith et al., 2016).

There is a lot of research data collection being done on the Amazon mTurk site and through the utilization of crowdsourcing initiatives in general and there are helpful considerations and data that inform how to best improve upon the first iteration of this surgical video annotation survey. An overarching concept that seeks to improve mTurker data selection is to create layers of screening that can help eliminate unwanted and unusable data while also limiting the costs for data collection. One such example is to utilize the Amazon mTurk tool that provides approval rates for mTurkers. An example use case for this tool would be setting the requirements to accomplish the HIT at a level above 90% (Cobanoglu et al., 2021). Another consideration for improvements within the

survey is the concept of anchor questions or validity-checking mechanisms that are questions embedded nonconsecutively amongst the research questions. A potential format for such validity-checking mechanisms could be paired questions at the beginning and end of the survey that ask related questions in different formats. One possible example of this could be, assuming you are seeking U.S college graduates, to identify the name of the university that the respondent attended at the beginning of the survey and asking which state the university is located in. Paired validity-checking mechanism questions such as these have been found to eliminate 20% of crowdsourcing tasks that were unable to correctly answer such questions (Cavusoglu, 2019). Other methods of control that could increase data quality are utilizing smart survey features that exist through programs such as Qualtrics that can enable restrictions on minimum time per question on a survey, thus mitigating the impacts of the aforementioned “speeders” (Smith et al., 2016). There is also the utilization of systems such as Captcha that can mitigate the prevalence of computer bots designed to automate the completion of mTurk HITs. Additionally, while not a cost cutting mechanism, for research sample size concerns it may be prudent to standardize the practice of seeking 10-15% more accepted assignments that intended to buffer the number of unusable respondent data that is received.

To help reflect on the various problems and theorized solutions to issues seen during the implementation of this study the below Table provides a comprehensive review.

Table 10: Summary of Issues and Potential Solutions discussed.

Issue Noted	Explanation	Potential Solution
-------------	-------------	--------------------

Source material variability	Lack of control for factors such as video quality, video length, and video editing	Explore a multi-tiered approach of crowdsourcing and use of medical expertise so video “noise” reduced to improve annotation
Surgical hand coordination and dexterity variability	Different video sources for teaching and testing annotation utilize different surgeons with different habits	Create a system to categorize surgical videos for video annotation by surgeon characteristics to test for improved annotation
Users implementing computer bots that complete survey in automated fashion	This phenomenon heavily decreased the approval rate for HITs	Create layers of screening that can help eliminate unwanted and unusable data; filter mTurkers by approval rate, anchor questions or validity-checking mechanisms
Manual system for verification of completed HITs paired with large amounts of poor data	This issue increased the period required for data collection and created bottleneck prior to analysis and would	Utilize screening tools such as smart survey features that decrease “cheating” and “speeding”; minimum time per question, anchor

	likewise increase experimental costs	questions, filter mTurkers by approval rate
--	---	--

The data gathered from the first iteration of the surgical video annotation survey published on Amazon mTurk has elicited valuable data that can inform future iterations and areas of research related to surgical video annotation, crowdsourced research data collection, etc. There are a handful of valuable insights gathered that are both promising developments and questions that remain unanswered about the role that mTurk can have for surgical video annotation data collection for ML and AI applications. The next logical steps seem to be a more robust sample size of data collected from an improved mTurk published survey to confirm trends seen in this phase of experimentation.

BIBLIOGRAPHY

1. Rajaraman, V. (2014). JohnMcCarthy — Father of artificial intelligence. *Resonance*, 19(3), 198–207. <https://doi.org/10.1007/s12045-014-0027-9>
2. Schwartz, W. B. (1970). Medicine and the Computer. *New England Journal of Medicine*, 283(23), 1257–1264. <https://doi.org/10.1056/nejm197012032832305>
3. Salvatore, C., Cerasa, A., Castiglioni, I., Gallivanone, F., Augimeri, A., Lopez, M., Arabia, G., Morelli, M., Gilardi, M. C., & Quattrone, A. (2014). Machine learning on brain MRI data for differential diagnosis of Parkinson’s disease and Progressive Supranuclear Palsy. *Journal of Neuroscience Methods*, 222, 230–237. <https://doi.org/10.1016/j.jneumeth.2013.11.016>
4. Kohn, L., Corrigan, J., & Donaldson, M. (2000). *To Err Is Human*. National Academies Press. <https://doi.org/10.17226/9728>
5. Bianco, S., Ciocca, G., Napoletano, P., & Schettini, R. (2015). An interactive tool for manual, semi-automatic and automatic video annotation. *Computer Vision and Image Understanding*, 131, 88–99. <https://doi.org/10.1016/j.cviu.2014.06.015>
6. Garrow, C. R., Kowalewski, K.-F., Li, L., Wagner, M., Schmidt, M. W., Engelhardt, S., Hashimoto, D. A., Kenngott, H. G., Bodenstedt, S., Speidel, S., Müller-Stich, B. P., & Nickel, F. (2020). Machine Learning for Surgical Phase Recognition. *Annals of Surgery*, 273(4), 684–693. <https://doi.org/10.1097/sla.0000000000004425>

7. Aguinis, H., Villamor, I., & Ramani, R. S. (2020). MTurk Research: Review and Recommendations. *Journal of Management*, 47(4), 823–837.
<https://doi.org/10.1177/0149206320969787>
8. Bentley, F. R., Daskalova, N., & White, B. (2017). Comparing the Reliability of Amazon Mechanical Turk and Survey Monkey to Traditional Market Research Surveys. *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems - CHI EA '17*.
<https://doi.org/10.1145/3027063.3053335>
9. Thakkar, N. (2019, May 21). *Figure 1: Skin incision by conventional scalpel*. ResearchGate; ResearchGate. https://www.researchgate.net/figure/Skin-incision-by-conventional-scalpel_fig1_333255064
10. Wollheim, D. (2021, May 20). *3 Types of Wound Closure: Primary, Secondary, and Tertiary* /. WCEI. <https://blog.wcei.net/3-types-of-wound-closure>
11. Slater, G. (2018, December 27). *Figure 1: Surgical Repair of Lateral Ligament. A. Minimally invasive...* ResearchGate; ResearchGate.
https://www.researchgate.net/figure/Surgical-Repair-of-Lateral-Ligament-A-Minimally-invasive-incision-over-the-lateral_fig1_331246700
12. Fletcher, J. (2019, May 28). *What to know about dissolvable stitches*. Medicalnewstoday.com; Medical News Today.
<https://www.medicalnewstoday.com/articles/325297#about>

13. Brewster, C. (2017, July 29). *Simple Interrupted Suture - OSCE guide | Wound Suturing | Geeky Medics*. Geeky Medics. <https://geekymedics.com/simple-interrupted-suture-osce-guide/>
14. Vore, S. J., Wooden, W. A., Bradfield, J. F., Aycock, E. D., Vore, P. L., Lalikos, J. F., & Hudson, S. S. (2002). Comparative Healing of Surgical Incisions Created by a Standard “Bovie,” The Utah Medical Epitome Electrode, and a Bard-Parker Cold Scalpel Blade in a Porcine Model: A Pilot Study. *Annals of Plastic Surgery*, 49(6), 635–645. <https://doi.org/10.1097/00000637-200212000-00014>
15. Massarweh, N. N., Cosgriff, N., & Slakey, D. P. (2006). Electrosurgery: History, Principles, and Current and Future Uses. *Journal of the American College of Surgeons*, 202(3), 520–530. <https://doi.org/10.1016/j.jamcollsurg.2005.11.017>
16. Ward, T. M., Fer, D. M., Ban, Y., Rosman, G., Meireles, O. R., & Hashimoto, D. A. (2021). Challenges in surgical video annotation. *Computer Assisted Surgery*, 26(1), 58–68. <https://doi.org/10.1080/24699322.2021.1937320>
17. Law, H. (2017). *Skill Assessment using Computer Vision based Analysis*. <https://www.semanticscholar.org/paper/Skill-Assessment-using-Computer-Vision-based-Law/d23bf3200adece389d6e7c866ca9105d999b23fa>
18. Deal, S. B., Stefanidis, D., Brunt, L. M., & Alseidi, A. (2017). Development of a multimedia tutorial to educate how to assess the critical view of safety in laparoscopic cholecystectomy using expert review and crowd-sourcing. *The*

American Journal of Surgery, 213(5), 988–990.

<https://doi.org/10.1016/j.amjsurg.2017.03.023>

19. Smith, S. M., Roster, C. A., Golden, L. L., & Albaum, G. S. (2016). A multi-group analysis of online survey respondent data quality: Comparing a regular USA consumer panel to MTurk samples. *Journal of Business Research*, 69(8), 3139–3148. <https://doi.org/10.1016/j.jbusres.2015.12.002>
20. Cobanoglu, C., Cavusoglu, M., & Turktarhan, G. (2021, March). *A beginner's guide and best practices for using crowdsourcing platforms for survey research: The Case of...* ResearchGate; unknown.
https://www.researchgate.net/publication/350206089_A_beginner's_guide_and_best_practices_for_using_crowdsourcing_platforms_for_survey_research_The_Case_of_Amazon_Mechanical_Turk_MTurk
21. Cavusoglu, M. (2019). *Online and Self-Directed Learning Readiness Among Hospitality and Tourism College Students and Industry Professionals*. Digital Commons @ University of South Florida.
<https://digitalcommons.usf.edu/etd/7763/>

CURRICULUM VITAE

Name: Faadil Mohammed Shariff

Contact Information:

Email: faadilms@bu.edu

Education:

B.S., Informatics, Indiana University Bloomington, May 2020

M.S., Medical Sciences, Boston University, Candidate May 2022

Research and Professional Experience:

Surgical Informatics Lab – Harvard Medical School

November 2021 – Present

Boston, MA

Research Intern, Master's Thesis

- Working on master's thesis under the supervision of Dr. Gabriel Brat, Assistant Professor of Surgery at Beth Israel Deaconess Medical Center and Harvard Medical School.
- Concurrently conducting research for publication on the prospects of training non-surgical experts to annotate open-source surgical videos for machine learning.
- Gaining familiarity with data analysis on Microsoft Excel and the Amazon mTurk crowdsourcing platform.

EYESPOT Medical Optical Boutique

July 2021 – January 2022*Ophthalmic and Optical Assistant**Chestnut Hill, MA*

- Gained hands-on experience working up primary eye care, contact lens and dry eye patients
- Worked alongside clinicians and opticians.
- Experienced taking patient histories, measuring lens power, performing automated keratometry readings, taking automated refractions with pupillary distance
- Performed retinal optical coherence tomography (OCT) scans and fundus photos for optical patients
- Learned how to care for patients and customers seeking to purchase eyewear and/or contact lenses

Beth Israel Deaconess Medical Center – Longwood Medical Eye Clinic**June 2021 – November 2021***Clinical Assistant**Boston, MA*

- Performing preliminary eye exams and tests, acting as a physician's scribe
- Assisting in clinical activities such as intravitreal injections and laser treatments
- Developing and maintaining close relationships with patients, families, staff, and faculty

- Shadowing ophthalmologists during clinical hours

St. Joseph Regional Medical Center

June 2015 – August 2021

Research Intern

Mishawaka, IN

- Collected and organized data from trauma patient files into Excel spreadsheets
- Created and registered patient information via an Access Database to track voluntary participation in the study
- Conceptualized statistical trends in blood transfusions for trauma patients

Publications:

- Bunch CM, Thomas AV, Stillson JE, Gillespie L, Khan RZ, Zackariya N, Shariff F, Al-Fadhli M, Mjaess N, Miller PD, McCurdy MT, Fulkerson DH, Miller JB, Kwaan HC, Moore EE, Moore HB, Neal MD, Martin PL, Kricheff ML, Walsh MM. Preventing Thrombohemorrhagic Complications of Heparinized COVID-19 Patients Using Adjunctive Thromboelastography: A Retrospective Study. *Journal of Clinical Medicine*. 2021; 10(14):3097. <https://doi.org/10.3390/jcm10143097>
- Speybroeck, J., Marsee, M., Shariff, F., Zackariya, N., Grisoli, A., Lune, S. V., Larson, E. E., Hatch, J., McCauley, R., Shariff, F., Aversa, J. G., Son, M., Agostini, V., Campello, E., Simioni, P., Scărlătescu, E., Kwaan, H., Hartmann, J., Fries, D., & Walsh, M. (2020). Viscoelastic testing in benign hematologic disorders: Clinical perspectives and future implications of point-of-care testing to assess hemostatic competence. *Transfusion*, 60 Suppl 6, S101–S121. <https://doi.org/10.1111/trf.16088>

- Walsh, M., Kwaan, H., McCauley, R., Marsee, M., Speybroeck, J., Thomas, S., Hatch, J., Vande Lune, S., Grisoli, A., Wadsworth, S., Shariff, F., Aversa, J. G., Shariff, F., Zackariya, N., Khan, R., Agostini, V., Campello, E., Simioni, P., Scărlătescu, E., & Hartmann, J. (2020). Viscoelastic testing in oncology patients (including for the diagnosis of fibrinolysis): Review of existing evidence, technology comparison, and clinical utility. *Transfusion*, 60 Suppl 6, S86–S100. <https://doi.org/10.1111/trf.16102>
- R, S., J, S., M, M., FS, S., D, C., A, C., M, T., S, T., S, W., G, W., D, F., M, W., N, Z., F, S., & H, A.-F. (2020). Emergency Department Cesarean Section for Placental Abruptio; Anticipation from Prehospital History with Preparation for Immediate Delivery. *Current Opinion in Gynecology and Obstetrics*, 3(1), 379–383. <https://doi.org/10.18314/cogo.v3i1.2037>
- Grisoli, A., Dynako, J., Zimmer, D., Zackariya, N., Shariff, F., Walsh, M., Mamczak, C. N., Peterson, C., Boyer, B., Hurwich, M., & Duprat, G. (2019). Management of a Pediatric Type 3C Open Femoral Fracture Following a High-Velocity Gunshot Wound at an Adult Level II Trauma Center. *Pediatric Emergency Care*, Publish Ahead of Print. <https://doi.org/10.1097/pec.0000000000001736>

Presentations:

Informatics Capstone Fair - May 2019

- Capstone Project: Created demo iOS app to function as a digital journal for assisting documentation for patients to manage their physical therapy and personal fitness and aid healthcare providers.
- Coded application using SQL, Python, and Swift

Volunteering:

Memorial Hospital - Beacon Health Systems

November 2013 – July 2014

Outpatient Surgery Waiting Room Clerk

- Summarized pre-op and post-op procedures to patient's families and fielded any further inquiries.