Published in final edited form as:

J Anesth. 2018 April ; 32(2): 211–218. doi:10.1007/s00540-018-2463-2.

Interrater Variability in ASA Physical Status Assignment: An analysis in the pediatric cancer setting

Luis Tollinche, MD¹, Gloria Yang, BA¹, Kay-See Tan, PHD², and Ruth Borchardt, MD¹

¹Memorial Sloan Kettering Cancer Center, Department of Anesthesiology and Critical Care Medicine

²Memorial Sloan Kettering Cancer Center, Department of Epidemiology and Biostatistics

Abstract

Background—The American Society of Anesthesiologists (ASA) physical status is a universal classification system that helps clinicians categorize their patients preoperatively. However there is a lack of both inter-rater and intra-rater reliability among clinicians for the ASA physical status classification. Our study focuses on testing these reliabilities within pediatric anesthesia providers in the cancer setting.

Methods—In our retrospective observational study, a total of 1177 anesthesia records were reviewed. The cohort included all pediatric patients (< 18 years old) diagnosed with either retinoblastoma or neuroblastoma who had two or more anesthesia procedure within a 14 day time period.

Results—Overall, the ASA physical status score among two different anesthesia providers for the same patient treated at different times, had very little inter-rater reliability, $\kappa = -.042$ (95% CI: $-.17; .09$). Of the 1177 patient anesthesia records, only 25% had two or more ASA physical status score assigned by the same anesthesiologist within a 14 day time period. There was moderate intra-rater reliability $\kappa = .48$ (95% CI: $.29; .68$) for patients who were assigned an ASA physical status score by the identical anesthesia provider at different times points within a 14 day period.

Conclusion—In contrast to observations in earlier studies, findings indicate poor agreement in inter-rater reliability. Although there was moderate agreement in intra-rater reliability, one would

Corresponding Author: Luis E Tollinche, MD, 1275 York Avenue, C330F, New York, NY 10065, Office phone: (212) 639-2884, Fax: (646) 422-2146, tollinche@mskcc.org.
IRB Contact Information: Dr. Roger Wilson, Chairman, IRB, Memorial Sloan Kettering Cancer Center, 1275 York Avenue, New York, NY 10065, (212) 639-8116, wilsonr@mskcc.org

Conflicts of Interest:
None

Authors contribution:
Tollinche, Yang, Tan, Borchardt: (All authors) were:

1. substantial contributors to the conception of the manuscript
2. active participants in the drafting and revising of the manuscript
3. approved of final version of manuscript
4. agree to be accountable for all aspects of the work

EQUATOR guidelines: This manuscript adheres to the applicable EQUATOR guidelines PRISMA-P

expect to find stronger, even perfect, intra-rater reliability. These findings suggest the need to develop a specific physical status classification system directed toward patients with a systemic illness such as cancer in both young and adult patients.

Keywords

ASA physical status; reliability; pediatric; cancer

Introduction

The American Society of Anesthesiologists physical status (ASA-PS) classification system was created in 1941 by a committee of three physicians to ensure uniformity in assessing patient preoperative status. (1)

What is already known?

The controversy revolving around reliability and consistency of the ASA physical status classification system has been widely researched. Examples of this include questionnaires and surveys administered to anesthesia providers that consist of hypothetical scenarios. In these surveys, clinicians assign each hypothetical patient an ASA physical status score based on the listed comorbidities pertaining to each patient. (4–9) A majority of these papers have claimed fair to moderate inter-rater reliability among the clinician's score regardless of the clinician's years in practice, area of practice and level of education. (4, 6, 10) Hence, the assertion has been set forth that the ASA scale is a reliable tool in assessing patient preoperative status. (10, 11)

However, most of the misconception in assigning a patient his or her ASA physical status score lies within factors that are independent of the ASA scoring system. These include: age, sex, weight, possibility of pregnancy, the nature of the proposed surgical intervention, the degree or skill of the clinician providing care or any other environmental factor related or unrelated to the procedure. (1)

Goal & Hypothesis

The investigators were interested in evaluating the inter and intra-rater reliability of the ASA-PS classification system in a real-life setting. Through systematic chart review, we compared the ASA-PS assignment of actual pediatric patients at a major cancer institution who received anesthesia and were assigned an ASA-PS at different time points. We hypothesize that the study population would exhibit high inter-rater variability but very low intra-rater variability when comparing ASA PS scores assigned to the same patient within a 14 day time period.

Methods

IRB Approval

IRB exemption was approved by MSKCC under the criteria of low protocol risk level.

Retrospective Data Review

We conducted a retrospective review of 1177 individual anesthesia records of pediatric patients (< 18 years old) diagnosed with either retinoblastoma or neuroblastoma. Only patients classified as outpatient and who received at least two anesthetics within a 14 day time period between January 1, 2014 and March 1, 2016 were included. Any patient who required an inpatient admission for any reason was excluded from the study. Data points abstracted included: MRN, patient name (first and last), date of birth, sex, age at surgery, surgery date, anesthesia provider, ASA-PS score, surgery location, surgeon's name (first and last), surgical department, duration of procedure and procedure details (Table 1). Each patient was given a unique subject identification number and every procedure was assigned a unique visit number. Of the 1177 anesthesia analyzed, a total of 30 individual attending anesthesiologists had assigned at least one ASA-PS score to any of the 193 unique patients.

Given the nature of their disease, pediatric patients with neuroblastoma and retinoblastoma frequently undergo anesthesia for procedures that help characterize and treat their disease (MRI, MIBG, Bone marrow biopsies/aspirates, etc.). It is common for these patients to have multiple anesthetics in the span of a 14 day time period. All ASA-PS scores were abstracted from the patient's anesthesia electronic medical records; for consistency, only the "perioperative ASA-PS scores" were recorded (Immediate Preop Attestation). Our initial data query revealed 249 individual pediatric patients diagnosed with either retinoblastoma or neuroblastoma who had two or more outpatient anesthesia procedures between January 1, 2014 and March 1, 2016.

Patients that were excluded from the initial data query included patients who did not have more than two procedures within an allotted 14-day time period. Patients who had an inpatient stay after their procedure or who received anesthesia for anything other than a diagnostic/"extent of disease" study were excluded from our analyses. A total of 56 patients was excluded from our data analysis (Figure 1).

Statistical Analysis

Patient sex and age at procedure were summarized descriptively. The primary objective is to estimate the inter-rater reliability (IRR) of the ASA ratings, or more conventionally the agreement across raters assigning ASA ratings to each patient. We present the IRR as two chance-corrected agreement measures: Cohen's Kappa and Cohen's weighted-Kappa. Both Kappa measures represent agreement greater than that expected by chance alone.

Cohen's Kappa ('unweighted-Kappa') represents the agreement between two raters who each classify a group of patients into the 5 ASA categories, calculated based on relative observed agreement and expected probability of chance agreement. (18) However there is a disadvantage in representing the results with Cohen's Kappa. A drawback of standard Kappa statistic (unweighted-Kappa) is that it does not account for the degree of disagreement and thus treats all disagreements equally. When the variable is ordinal-such as the ASA-PS scale-it may be preferable to allow different levels of agreement to contribute to the value of Kappa. As such, the weighted-Kappa allocates different weights to the disagreement depending on the magnitude of disagreement: disagreements involving distant values are

weighted more heavily than minor disagreements involving similar values. On an ordinal scale, a 1-point difference in ASA ratings is less deleterious than a 3-point difference. An ASA-PS I versus an ASA-PS V by two raters on the same patient would be weighted heavily, whereas ratings of an ASA-PS III and an ASA-PS IV on the same patient would have low weighting in the calculation of the Kappa statistics. Weighted-Kappa considers disagreements differently from unweighted-Kappa and is particularly useful for ordinal ratings. Unlike the unweighted-Kappa that gives credit to complete agreement only, weighted-Kappa gives credit for both complete and partial agreement; however, off-diagonal cells contain weights indicating the severity of the disagreement. The penalization increases with increasing distance between the two ASA ratings. (19) Although both unweighted-kappa and weighted-kappa measures are presented, we hang more significance in our weighted-kappa as it is more appropriate for the ordinal ASA-PS scaling system. Parallel to our beliefs, the disagreement in an ASA-PS score of I and IV should be weighed more heavily than a disagreement in an ASA-PS score of II and III.

Although we present both unweighted- and weighted-Kappa measures, we will consider the weighted-Kappa as the primary result as it is more appropriate for the ordinal ASA scale. Both Kappa measures estimates the precision rather than the accuracy of the ratings (i.e., there is no gold standard for the correct rating to which to measure the accuracy of the ratings). The Kappa measures range from -1 to 1 , where -1 implies complete disagreement, 0 for agreement expected by chance, and 1 as perfect agreement. Following the suggestions from Landis and Koch (20), reliability can be described as ‘slight’ (values of $0.01 - 0.20$), ‘fair’ ($0.21 - 0.4$), ‘moderate’ ($0.41 - 0.60$), ‘substantial’ ($0.61 - 0.80$) and ‘almost perfect’ when values are above 0.80 . Values between -0.10 to 0 represent no/poor agreement; a negative kappa is an indicator of great disagreement among raters and potential necessity to retrain rater or redesign the instrument. (14)

In the 14-day periods, each patient may have more than two ASA ratings, assigned by raters from a pool of anesthesiologists. As the primary analysis, we calculate the unweighted-Kappa and weighted-Kappa using only the first two ASA ratings per patient.

As a sensitivity analysis, we repeat the calculation of the Kappa measures using all available ASA ratings from each patient. The goal of the sensitivity analysis is to show that using any other pairs of ratings would have produced the same result as using only the first two consecutive ratings. If the agreement in the sensitivity analyses were much better or worse than using only the first two, it would imply that the third (or subsequent) ratings does add valuable information and must be incorporated in the primary analyses. Cohen’s Kappa only applies to the assessment of agreement between two ratings. There is currently no conventional methodology that can handle the case where patients have different number of ratings (i.e., missing data issue), while still providing an agreement statistic that accounts for the severity of disagreement (i.e., ordered categories). Hence, we proceed with a random bootstrap sampling approach.

As each patient may contribute more than one 14-day period, each period is assumed to be independent. A representative Kappa value is then calculated for each patient-period combination based on the random sampling concept. If the patient had only two ASA

ratings, both ratings are included in the analysis. If the patient had more than two ASA ratings, two will be selected randomly without replacement to represent the patient's contribution. Cohen's unweighted-Kappa and weighted-Kappa are calculated with the resulting sample in which each patient contributes exactly two ASA ratings. This resampling procedure is repeated 1000 times, and the mean Kappa measures across the 1000 agreement statistics produce the representative Kappa values, while the 95% confidence interval is derived from the 2.5th and 97.5th percentile of the distribution of the statistics. The representative Kappa value can be considered as an average of agreement across all pairs of available ratings.

The intra-rater reliability (agreement within the same rater given a patient) is also investigated as an exploratory objective, analyzed similarly to the inter-rater reliability among patients with at least two ASA ratings assigned by the same rater. There are no formal hypothesis testing in this study. Instead, the goal is to estimate the level and variability of agreement between raters. As such, we present the potential lower bounds for a 95% confidence interval (CI) based on the fixed sample size and anticipated Kappa agreement values. Based on clinician hypothesized distribution of ASA-PS scores, the expected prevalence of the five ASA-PS categories (ASA I, II, III, IV, V) were assigned as 1%, 43%, 50%, 5%, 1%.

Given the anticipated unweighted-kappa of 0.5, a total of 193 unique subjects produces a lower limit for the 95% CI of 0.402; whereas an unweighted-kappa of 0.25 can produce a lower limit of 0.153, and the unweighted-kappa of 0.1 produces a lower limit of 0.022. All analyses were conducted in R 3.1.1 (R Development Core Team, Vienna, Austria) using the *psych* and *nlme* R packages. All statistical tests were two-sided, and $p < 0.05$ was considered statistically significant.

Results

In the group of pediatric patients consisting of 1177 anesthesia records, 0.2% ($n=2$) were assigned to ASA-PS I, 37% ($n=440$) were assigned to ASA-PS II, 57% ($n=676$) were assigned to ASA-PS III, 4.9% ($n=58$) assigned to ASA-PS IV and $<0.1\%$ ($n=1$) assigned to ASA-PS V.

Inter-rater reliability: Agreement of two distinct anesthesiologists on the ASA-PS assignment of a unique patient

Based on the sample using only two ratings per patient, 193 pairs of ASA ratings for 193 patients were analyzed. In the primary analysis, the ASA-PS score in the intra-operative setting exhibited little to no inter-rater reliability. The unweighted-kappa was -0.14 (95% CI, -0.26 to 0.015) and the weighted-kappa was -0.042 (95% CI, -0.17 to 0.09), indicating no agreement between the two different raters. The 193 pairs of the ASA-PS scores were distributed according to the first and second rater (Table 2). Only 77 pairs (40%) of ASA-PS scores showed exact agreement among the two different raters, while the others disagreed (Table 2). Of the pairs of ASA-PS scores that were in disagreement, 112 pairs (58%) of scores were rated exactly 1 score apart by the two raters (Table 2). From the 193 pairs, there

were a total of 4 pairs (2%) where the ASA-PS scores were 2 scores apart by the two raters (Table 2).

In a sensitivity analysis, when all ASA-PS scores were considered in a random sampling approach, the unweighted-kappa was -0.15 (95% CI, -0.24 to -0.063). The weighted-kappa was -0.009 (95% CI, -0.11 to 0.088), indicating poor to no inter-rater reliability. This result was similar to that observed when each patient contributed only the first two ASA-PS scores to the analysis (Figure 2).

Intra-rater reliability: Agreement of ASA-PS assignment of a unique patient by the same anesthesiologist

Of the 1177 individual anesthesia records, only 103 unique cases (25%) had two or more ASA-PS scores assigned by the same anesthesia provider (Table 3). The unweighted-kappa was 0.50 (95% CI: 0.34 , 0.66) and the weighted-kappa was 0.48 (95% CI: 0.29 , 0.68) indicating reasonable to moderate intra-rater reliability. The 103 pairs were distributed according to the first and second episode in which the same anesthesia provider assigned the same ASA-PS score to the same patient at two different encounters within a 14-day time period (Table 4). Of those that were assigned to the same anesthesia provider at two different episodes, 76 pairs (73%) of ASA-PS scores showed exact agreement among the two different episodes. There was minimal intra-rater disagreement when assigning the same ASA-PS scores to the same patient at two different time points. Of the 103 pairs, only 27 pairs (26%) were in disagreement.

From the ASA-PS scores that were in discord at two distinct episodes, 25 pairs (24%) of scores were rated exactly 1 score apart by the same rater (Table 4). From the 103 pairs observed for intra-rater reliability, there were a total of 2 pairs (2%) where the ASA-PS scores were 2 scores apart by the same rater (Table 4).

Discussion

The results of this retrospective observational study demonstrate a striking lack of inter-rater reliability in the ASA-PS system. In keeping with our hypothesis, we saw little to no inter-rater agreement, despite knowing the patient's diagnosis prior to their first procedure. Our results indicate the undermined validity of the ASA-PS system and imply that its intended purpose—to assess the preoperative status of a patient—has gone awry. Although there was moderate intra-rater reliability, one would expect near perfect agreement and little to no variability in intra-rater ASA-PS assignments. In contrast to past studies whose results of both unweighted Cohen's kappa (κ) and weighted Cohen's (κ_w) implied moderate to high inter-rater reliability, (4, 10, 11, 13) our data exhibited negative values. A negative kappa is generally represented as “no agreement” (14) with an even larger negative value representing an even greater disagreement among the raters. These negative values describe a scenario wherein there is worse agreement than what we would expect by chance alone in arbitrarily picking a set of numbers and determining the degree of variance.

Strengths of our study

Our study improved on many of the deficiencies cited in previous studies that investigated the inter-rater reliability of the ASA-PS classification system. First, *actual* patient records were reviewed and both the degree of inter-rater and intra-rater reliability were measured. Many previous studies rely on hypothetical patients whose descriptions are limited to a few sentences. Furthermore, abstracting “real” retrospective data helped us minimize potential bias inherent in constructing and conducting questionnaires of hypothetical patients.

Second, in contrast to other studies where the ASA-PS score is rated at the same time, our study incorporates a defined, albeit narrow, timeline to compare the different ASA-PS scores that were assigned to the same patient. The pair of ASA-PS scores for each unique patient is calculated as a dichotomous rating assigned by two different raters. This allows us to assume that each encounter for the same patient within the 14-day time period can be interchangeably paired to test for not only inter rater reliability but more importantly intra rater reliability; previous studies have not been able to analyze intra rater reliability.

A third strength in our study arises because we did not depend on a “gold standard” for ASA assignment. Past studies have taken to setting an arbitrary “correct” ASA-PS score to a hypothetical patient scenario and comparing the gathered answers to their pre-set standard. (4, 8, 11, 15, 21) It is impossible to arrive at a correct ASA score for any given patient because of the inherent properties of ASA scoring; the “correct” answer may simply arise from a pooled, majority opinion. There is disagreement among clinicians about the correct ASA-PS score for any given patient and therefore, any study that compares responses against a “correct” score is potentially flawed. Our study did not rely on a “correct” ASA score at any point. Instead, we compared respondent’s responses to each other and not to an arbitrary “correct” answer. (15, 17)

The final strength in our study design is that it holds reproducibility if the study was to be extended to other institutions that are not in a cancer setting. Our methodology can be easily reproduced at any institution capable of reviewing patient records.

In the cancer setting

The poor to moderate degree of agreement in both inter-rater and intra-rater reliability may be ascribed to a single confounding factor-cancer. Cancer is viewed as a systemic illness because it is defined as a disease in which abnormal cells divide uncontrollably, potentially invade nearby tissue and also have the ability to spread to other parts of the body through the blood and lymphatic system. Therefore, cancer should be factored into stratifying the ASA-PS scaling system similar to other underlying health conditions.

A patient with active lung cancer who is otherwise healthy should be classified as an ASA 3 given the multisystem involvement of lung cancer (e.g. effects on coagulation cascade, lung parenchyma, toxicity from therapies, etc.) But a patient with metastatic lung cancer who suffers from hypertension, coronary artery disease, renal failure and history of cerebrovascular accidents would similarly be classified an ASA 3. The current ASA scoring paradigm suggests little utility for describing the cancer patient. Because patients with cancer can range vastly in their health conditions depending on their stage and level of

treatment, it is imperative to define and stratify the current ASA-PS classification system by incorporating specific cancer criteria.

Previous studies have suggested adding examples to the currently used ASA-PS scaling system or further refining the examples. (15) Existing examples for each ASA-PS category include smoking history, body mass index, heart disease, alcohol consumption and other various illnesses. (15) There is, however, a conspicuous absence of cancer related examples (e.g stage/grade of cancer, chemo, radiation, immunotherapy, etc). These shortfalls of the current ASA-PS classification system call for a novel form of the ASA-PS scaling system with examples that can guide anesthesia providers who are practicing in the cancer setting.

Unblinded ASA score assignments

In the instance a previous anesthesia record is available, any and every anesthesia provider is able to view the patient's previous ASA-PS score assignment before designating a score. We would expect to see an improvement in inter and intra rater reliability given that the anesthesia provider is unblinded to previous ASA-PS scores; however, our results demonstrate otherwise.

Assessing the assumption of "no time effect"

It would be ideal if 2 ASA-PS scores for the same patient were measured, compared and contrasted on the same day. However, in an institutional setting, it is not pragmatic.

A major assumption we take when observing patients during the 14-day time period is that the patients are at a stable condition across the timeline allowing us to interchangeably pair the ASA-PS scores. This allows us to justify that any lack of agreement can be attributed to the lack of agreement between raters rather than changing patient health status over time. We assessed the potential time effect with an exploratory analysis using linear mixed-effects models for the outcome of longitudinal ASA ratings, which included patient-level random effects and time (days since first ASA rating) as fixed effect. The null hypothesis is that there is no significant relationship between ASA ratings and time since the first ASA ratings (i.e., estimate of time effect =0) Using only two ASA ratings per patient (primary analysis), there was no significant time effect ($p=0.166$). In addition, estimates calculated by considering only pairs of ratings within 7 days or within the median of 3 days, the conclusions were the same. For example, among the $N=159$ pairs of ASA ratings within 7 days, the weighted-Kappa for inter-rater reliability was -0.068 (95% CI, -0.24 to 0.11). Even though we selected a 14 day time period, our study analysis actually revealed a median time of 3 days between procedures. The median number of days between two ASA PS scores given by the same rater for the same patient was 3 days where the interquartile range (IQR) was 2 to 7 days ranging from 0 to 14 days (Figure 3). This implies that a majority of our patients returned for a second procedure within 3 days of the first procedure. That is to say that even though we selected an inclusion criteria of 14 days, half of the patients had ASA scores assigned less than 3 days apart. The median number of days between two records was 3 days. Such a narrow window of 3 days would make the possibility of a real clinical change even less likely. Finally, if patient's true physical status were to deteriorate over this 14 day time period, we would expect to see a trend of worsening ASA scores in our

analysis. Instead, we saw a small improvement in ASA scores from first to second procedure rating. (Estimate of time effect in the linear mixed-effects model= -0.012 , $p=0.136$)

ASA-PS scores are intended to reflect a patient's true medical condition and not the type of surgery that is planned. Nevertheless, previous studies have demonstrated that ASA-PS scores are erroneously influenced by the type or severity of the patient's planned procedure. (12) Details of the surgery were reviewed to control for type of procedure and to investigate the possibility that clinicians are erroneously influenced by surgery type in assigning ASA-PS scores perioperatively. It is incorrect to use the procedure type to influence ASA score assignments. High risk surgery does not carry a different ASA assignment than an MRI, for example. ASA scores are based on actual physical status of patient and are independent of the acuity of the procedure or surgery. Unfortunately, some providers are not aware of this guideline or are unconsciously influenced by the type of procedure and this may explain some of the variability in score assignment.

The effect of new information learned from first procedure

One possibility for a change in ASA score was that information learned during the first procedure could affect the rater assessment during the second procedure. That is, perhaps a new cancer diagnosis was made at the conclusion of the first procedure that affected the second procedure rater's assignment of ASA score. We would expect a trend of deteriorating ASA scores overall. Disabusing this potential shortfall, we remind that our data showed a trend, albeit slight, of IMPROVING ASA scores (Estimate of time effect= -0.012 , $p=0.136$).

Limitations

The observation of cancer patients exhibiting defined characteristics draws a limitation in generalizing our findings to other patient populations. It is likely for clinicians, when assessing patients with a systemic illness such as cancer, to skew toward one ASA-PS score. Thus, cancer alone-independent of other health conditions-inherently increases your ASA-PS score, compared to a patient who has the same health conditions but does not have cancer. Consequently, we should expect to find a stronger inter-rater reliability and even stronger intra-rater reliability. It is more likely that our results underscore the rather lackluster ASA-PS classification system. Anesthesiologists who have been working at a cancer institution for many years may "reset" their ASA-PS scale and underestimate the impact of cancer-a multi organ systemic disease-when considering the influence of cancer on ASA PS score. At our institution, almost all patients have a cancer diagnosis. And by definition ASA 1 and 2 must be rarely encountered. This is clearly not the case given our analysis. Our anesthesiologists may unconsciously skew the ASA scale and begin to compare patients to each other and not to the guidelines of ASA PS scoring. Further investigation should include anesthesia providers at non-cancer hospitals and whether there exists even more disagreement when assigning ASA-PS scores to cancer patients. While we can control for multiple variables in constructing the study design, we cannot control or neglect the human variables in the providers who have become accustomed to clinically assessing cancer patients and assigning an ASA-PS score.

Acknowledgments

Disclosure of Funding:

This research was funded in part through the NIH/NCI Cancer Center Support Grant P30 CA008748

References

1. Daabiss M. American Society of Anaesthesiologists physical status classification. *Indian J Anaesth.* 2011; 55(2):111–5. [PubMed: 21712864]
2. Jacqueline R, Malviya S, Burke C, Reynolds P. An assessment of interrater reliability of the ASA physical status classification in pediatric surgical patients. *Paediatr Anaesth.* 2006; 16(9):928–31. [PubMed: 16918653]
3. Aplin S, Baines D, J DEL. Use of the ASA Physical Status Grading System in pediatric practice. *Paediatr Anaesth.* 2007; 17(3):216–22. [PubMed: 17263735]
4. Burgoyne LL, Smeltzer MP, Pereiras LA, Norris AL, De Armendi AJ. How well do pediatric anesthesiologists agree when assigning ASA physical status classifications to their patients? *Paediatr Anaesth.* 2007; 17(10):956–62. [PubMed: 17767631]
5. Ranta S, Hynynen M, Tammisto T. A survey of the ASA physical status classification: significant variation in allocation among Finnish anaesthesiologists. *Acta Anaesthesiol Scand.* 1997; 41(5): 629–32. [PubMed: 9181166]
6. Riley R, Holman C, Fletcher D. Inter-rater reliability of the ASA physical status classification in a sample of anaesthetists in Western Australia. *Anaesth Intensive Care.* 2014; 42(5):614–8. [PubMed: 25233175]
7. Mak PH, Campbell RC, Irwin MG, American Society of A. The ASA Physical Status Classification: inter-observer consistency. *American Society of Anesthesiologists. Anaesth Intensive Care.* 2002; 30(5):633–40. [PubMed: 12413266]
8. Ringdal KG, Skaga NO, Steen PA, Hestnes M, Laake P, Jones JM, et al. Classification of comorbidity in trauma: the reliability of pre-injury ASA physical status classification. *Injury.* 2013; 44(1):29–35. [PubMed: 22277107]
9. Ihejirika RC, Thakore RV, Sathiyakumar V, Ehrenfeld JM, Obremskey WT, Sethi MK. An assessment of the inter-rater reliability of the ASA physical status score in the orthopaedic trauma population. *Injury.* 2015; 46(4):542–6. [PubMed: 24656923]
10. Cohen J. A Coefficient of Agreement for Nominal Scales. *Educ Psychol Meas.* 1960; 20(1):37–46.
11. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol Bull.* 1968; 70(4):213–20. [PubMed: 19673146]
12. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics.* 1977; 33(1):159–74. [PubMed: 843571]
13. McHugh ML. Interrater reliability: the kappa statistic. *Biochem Med (Zagreb).* 2012; 22(3):276–82. [PubMed: 23092060]
14. Sankar A, Johnson SR, Beattie WS, Tait G, Wijesundera DN. Reliability of the American Society of Anesthesiologists physical status scale in clinical practice. *Br J Anaesth.* 2014; 113(3):424–32. [PubMed: 24727705]
15. Hurwitz EE, Simon M, Vinta SR, Zehm CF, Shabot SM, Minhajuddin A, et al. Adding Examples to the ASA-Physical Status Classification Improves Correct Assignment to Patients. *Anesthesiology.* 2017; 126(4):614–22. [PubMed: 28212203]
16. Reed JF 3rd. Homogeneity of kappa statistics in multiple samples. *Comput Methods Programs Biomed.* 2000; 63(1):43–6. [PubMed: 10927153]
17. Sweitzer B. Three Wise Men (x2) and the ASA-Physical Status Classification System. *Anesthesiology.* 2017; 126(4):577–8. [PubMed: 28212206]
18. Cohen J. A Coefficient of Agreement for Nominal Scales. *Educ Psychol Meas.* 1960; 20(1):37–46.
19. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol Bull.* 1968; 70(4):213–20. [PubMed: 19673146]

20. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977; 33(1):159–74. [PubMed: 843571]
21. Goodhart IM, Andrzejowski JC, Jones GL, Berthoud M, Dennis A, Mills GH, et al. Patient-completed, preoperative web-based anaesthetic assessment.
22. Nasr VG, DiNardo JA, Faraoni D. Development of a Pediatric Risk Assessment Score to Predict Perioperative Mortality in Children Undergoing Noncardiac Surgery. *Anesthesia and analgesia*. 2017; 124(5):1514–9. [PubMed: 27575562]
23. Faraoni D, Vo D, Nasr VG, DiNardo JA. Development and Validation of a Risk Stratification Score for Children With Congenital Heart Disease Undergoing Noncardiac Surgery. *Anesthesia and analgesia*. 2016; 123(4):824–30. [PubMed: 27529321]

Figure 1. Overview of the study design

Consort diagram exhibiting study overview; including the inclusion and exclusion criteria as well as the primary and secondary aims that were being studied.

Figure 2. Overview of the distribution of ASA ratings given by rater 1 and rater 2 using only the first 2 ASA ratings per patient

Using only the first two ASA ratings per patient (ignore any additional ASA ratings/other periods), this bar graph represents the distribution of the 193 pairs of ASA ratings.

Figure 3. Distribution of the number of days between two procedures within the same patient
Based on the sample using only two ratings per patient, 193 pairs of ASA ratings for each of the 193 patients were analyzed. The bar graph represents the distribution of the days between the first and second ASA rating. The median number of days between two ASA ratings was 3 days (IQR 1 to 7, range 0 to 14).

Table 1

Demographical characteristics of the cohort of pediatric patients defined by their respective inter-rater and intra-rater agreement for the ASA-PS score.

Patient Level	Study Population [n = 193]
Parameter	
Gender, n (%)	
Male	114 (59%)
Female	79 (41%)
Age	
Median (min, max)	4.0 (2.0, 5.0)
Procedure Level	Service [n = 507]
Parameter	
Service, n (%)	
Anesthesiology & Critical Care	232 (46%)
Dental Oncology	1 (0.2%)
Gastroenterology (GI)	1 (0.2%)
Interventional Radiology (IR)	2 (0.4%)
Neurology	14 (2.8%)
Ophthalmology	210 (41%)
Pediatric	46 (9.1%)
Radiation Oncology	1 (0.2%)

Table 2

Distribution of agreement between the first and second rater using only the first and second ASA-PS score per patient

Rater 1	Rater 2			
	ASA 1	ASA 2	ASA 3	ASA 4
Total	ASA 1	ASA 2	ASA 3	ASA 4
ASA 1	0	0	1	0
ASA 2	0	27	51	1
ASA 3	1	49	50	7
ASA 4	0	1	5	0
Total	1	77	107	8

Weighted kappa = -0.042 (95% CI, -0.17 to 0.09) Un-weighted kappa = -0.14 (95% CI, -0.26 to -0.015)

Table 3

Distribution of agreement between the same rater at two different occurrences for the 103 pairs of ASA-PS scores

Procedure 1	Procedure 2				Total
	ASA 1	ASA 2	ASA 3	ASA 4	
ASA 1	0	0	1	0	1
ASA 2	0	31	14	1	46
ASA 3	0	10	43	1	55
ASA 4	0	0	0	2	2
Total	0	41	58	4	103

Weighted kappa = 0.48 (95% CI, 0.29 to 0.68) Un-weighted kappa = 0.50 (95% CI, 0.34 to 0.66)