Check for updates

ORIGINAL ARTICLE

# Interobserver variation in the Parkland scale. Are we seeing the same thing?

## *Variación interobservador en la escala de Parkland. ¿Estamos viendo lo mismo?*

*José L. Maldonado-Calderón[1], Lydia E. Nava-Rivera[2], Antonio Urbina-Zeglen[1], Marco V. Herrera-Santos[1], Pilar Carranza-Rosales[3], Javier Morán-Martínez[2], and Nadia D. Betancourt-Martínez[2]**

*[1]Departamento de Cirugía, Hospital General Universitario Dr. Joaquín Del Valle Sánchez, Facultad de Medicina, Universidad Autónoma de Coahuila (UAdeC), Torreón, Coahuila; [2]Departamento de Biología Celular y Ultraestructura, Facultad de Medicina, UAdeC, Torreón, Coahuila; [3]Departamento de Biología Celular y Molecular, Centro de Investigación Biomédica del Noreste, Instituto Mexicano del Seguro Social. Monterrey, Nuevo León. México*

## Abstract

**Objective:** *The aim of this study was to analyze the reliability of agreement between surgeons when using the Parkland Grading Scale for Acute Cholecystitis (PGS-AC).* **Methods:** *A total of 43 images taken out of videos of laparoscopic cholecystectomies (LCs) were collected, they were used to frame an online questionnaire that was sent to 18 surgeons and resident doctors who classified the images according to the Parkland scale criteria, followed by the evaluation of concordance between observers applying the Fleiss κ test.* **Results:** *A global Fleiss' κ value of 0.213 was obtained, which corresponds to a low interobserver concordance. Factors such as being a surgical resident, having more than 10 years of experience performing this type of procedure, or performing more than 2 LCs per week, were related to greater concordance in diagnosis.* **Conclusions:** *The low concordance found when using the Parkland grading scale, translates into a high interobserver variation related to multiple variables, which is why, we are not seeing the same.*

***Keywords:*** *Cholecystectomy. Interobserver variation. Acute. Parkland scale.*

## Resumen

**Objetivo:** *Analizar la fiabilidad de la concordancia entre cirujanos al utilizar la PGS-AC (Parkland Grading Scale for Acute Cholecystitis).* **Métodos:** *Se recolectaron 43 imágenes extraídas de vídeos de colecistectomías laparoscópicas y se realizó un cuestionario en línea que se envió a 18 cirujanos y médicos residentes, quienes clasificaron las imágenes según los criterios de la PGS-AC. Se evaluó la concordancia entre observadores aplicando la prueba kappa de Fleiss.* **Resultados:** *Se obtuvo un valor kappa de Fleiss global de 0.213, lo que corresponde a una baja concordancia interobservador. Factores como ser residente de cirugía, tener más de 10 años de experiencia realizando este tipo de procedimientos o realizar más de dos colecistectomías laparoscópicas por semana se relacionaron con una mayor concordancia en el diagnóstico.* **Conclusiones:** *La baja concordancia encontrada al utilizar la PGS-AC se traduce en una alta variación interobservador relacionada con múltiples variables, por lo que no estamos viendo lo mismo.*

***Palabras clave:*** *Colecistectomía. Variación interobservador. Aguda. Escala de Parkland.*

***Correspondence:***
Nadia D. Betancourt-Martínez
E-mail: nabetancourtm@uadec.edu.mx

Date of reception: 18-07-2023
Date of acceptance: 25-08-2023
DOI: 10.24875/CIRU.23000362

Cir Cir. 2024;92(6):709-714
Contents available at PubMed
www.cirugiaycirujanos.com

0009-7411/© 2023 Academia Mexicana de Cirugía. Published by Permanyer. This is an open access article under the terms of the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

# Introduction

Laparoscopic cholecystectomy (LC) has been proposed as the gold standard for the treatment of symptomatic cholelithiasis[1]. More than 750,000 LCs are performed each year in the United States[2], making this one of the most common elective surgeries in the world. Among the indications for LC is acute cholecystitis (AC), for which precise diagnostic criteria and management algorithms based on its severity have been established[3,4]. There are several classifications for AC, such as that of the American Association for the Surgery of Trauma, which includes clinical, imaging, intraoperative, and pathology description[5]. However, due to its complexity and the large number of variables involved, it has not yet been adopted by the surgical community on a wider scale. Recently, a classification of intraoperative findings, called the Parkland grading scale for AC (PGS-AC), has been proposed to determine and grade the difficulty of LC[6]. This scale (validated in 2019) consists of five grades, of which, grade 1 represents a normal gallbladder (GB) and grade 5 the most severe grade[7]. However, due to the short time, it has been established, the PGS-AC is not exempt from being evaluated for possible biases in its application through concordance studies[8]. The aim of this study was to analyze the evaluations of surgeons and their reliability of agreement in rating 43 cholecystectomies using the PGS-AS.

## Materials and methods

### *Place of study*

The study was carried out in the General Surgery service of the General University Hospital of the Faculty of Medicine of the Autonomous University of Coahuila Campus Torreón; Torreón, Coahuila, Mexico.

### *Selection of the samples*

Videos of LCs performed between 2020 and 2021 were randomly selected. From these videos, 43 images were obtained under the following criteria:

- Make a screenshot of the image that best represents the grade according to the Parkland scale (always before starting the dissection)
- Check that the instruments did not interfere with the image visualization.

### *Questionnaire framing*

With the obtained images, a questionnaire was framed as a tool to classify each image in one of the 4° according to the PGS or to classify the image as unclassifiable. This questionnaire was applied to 13 surgeons and five residents of the region who met the following inclusion criteria: being a surgeon or resident doctor, performing laparoscopic surgery, and operating in hospitals in the region.

### *Ethical considerations*

The protocol was approved by the Bioethics Committee of the School of Medicine of the Autonomous University of Durango Campus Laguna with reference number 130/20.

### *Statistical analysis*

The data obtained from the questionnaires were analyzed using the statistical package IBM SPSS statistics version 26. Descriptive statistics was used such as mean, standard deviation, and percentage frequencies. To evaluate the concordance between the evaluators (n = 18), the Fleiss' $\kappa$ statistic was used. The global and individual $\kappa$ value was determined for each classification category (unclassifiable and grade IV), stratifying by age, gender, grade, subspecialty, years of experience, number of surgeries performed per week, and work sector. Comparisons of $\kappa$ values between strata were made using the z-test for two samples.

## Results

In this study, a total of 18 evaluators were included for the rating of 43 images, obtaining a total of 774 data points. The mean age of the participating evaluators was 38.50 ± 11.66 years and 58.8% of the sample were male. Most of the observers have a medical specialty (72.2%) and 27.8% complete their medical residency. Of the total number of specialists, 38.5% have a subspecialty such as advanced laparoscopy and endoscopy. The average years of experience of the evaluators was 10.17 ± 10.93 with an average number of surgeries per week of 2.22 ± 1.35. 22.2% of the evaluators sample work in public hospitals, whereas 61.1% work in public and private hospitals (Table 1).

The global values of Fleiss' $\kappa$ in which all the degrees of the Parkland scale are included, as well

**Table 1. Characteristics of the evaluators**

| Variable | n | Half | OF | Minimum | Maximum |
|---|---|---|---|---|---|
| Age (years) | 18 | 38.50 | 11.66 | 26.00 | 63.00 |
| Gender | | | | | |
| Feminine | 8 (41.2) | | | | |
| Male | 10 (58.8) | | | | |
| Residency | 5 (27.8) | | | | |
| Residence grade (year) | 5 | 2.20 | 1.30 | 1.00 | 4.00 |
| Specialty | 13 (72.2) | | | | |
| Subspecialty | 5 (38.5) | | | | |
| Years of experience | 18 | 10.17 | 10.93 | 0.00 | 31.00 |
| Number of surgeries per week | 18 | 2.22 | 1.35 | 0.00 | 5.00 |
| Work sector | | | | | |
| Private | 3 (16.37) | | | | |
| Public | 4 (22.2) | | | | |
| Both | 11 (61.1) | | | | |

Values are presented as mean and SD (standard deviation) or frequencies (%).

**Table 2. Overall values (including all grades) of Fleiss' κ, stratified by different variables**

| Variable | Overall value of Fleiss' κ | CI 95% | p-value[sa] |
|---|---|---|---|
| Global | 0.213 | 0.212-0.213 | < 0.0001 |
| Age (years) | | | |
| ≥ 38 | 0.251 | 0.250-0.252 | < 0.0001 |
| < 38 | 0.203 | 0.202-0.203 | < 0.0001 |
| Gender | | | |
| Feminine | 0.209 | 0.209-0.210 | < 0.0001 |
| Male | 0.204 | 0.204-0.205 | < 0.0001 |
| Degree | | | |
| Home | 0.162 | 0.161-0.164 | < 0.0001 |
| Specialty | 0.231* | 0.231-0.232 | < 0.0001 |
| Subspecialty | | | |
| With Subspecialty | 0.190 | 0.189-0.192 | < 0.0001 |
| No subspecialty | 0.215 | 0.214-0.215 | < 0.0001 |
| Years of experience | | | |
| ≤ 10 years | 0.186 | 0.186-0.187 | < 0.0001 |
| > 10 years | 0.274† | 0.272-0.275 | < 0.0001 |
| Number of surgeries per week | | | |
| ≤ 2 | 0.165 | 0.164-0.166 | < 0.0001 |
| > 2 | 0.249‡ | 0.248-0.250 | < 0.0001 |
| Work sector | | | |
| Private/Public | 0.190 | 0.190-0.191 | < 0.0001 |
| Both | 0.255† | 0.254-0.256 | < 0.0001 |

*p < 0.001. †p < 0.0001. ‡p < 0.00001, when comparing between categories; 95% CI: 95% confidence intervals. a: p value of the Fleiss' κ statistic.

as the "unclassifiable" category, are shown in Table 2. The global value was 0.213 (CI 95 %: 0.212-0.213) (p < 0.0001), indicating a weak strength of concordance. The global Fleiss' κ was calculated, stratifying by age, gender, degree, subspecialty, years of experience, number of surgeries performed per week, and work sector. The level of agreement for all strata ranged from slight (κ ≤ 0.20) to fair (κ = 0.21-0.40). For comparisons of the κ value between strata, the reliability of agreement was higher for specialists in contrast to residents (0.231 [0.231-0.232] vs. 0.162 [0.161-0.164], respectively), as for the category of > 10 years of experience compared to the category of ≤ 10 years (0.274 [0.272-0.275] vs. 0.186 [0.186-0.187], respectively). In addition, statistically significant differences between the global κ values were found for: number of surgeries performed per week, with this variable's results being higher for the category of > 2 surgeries in contrast to ≤ 2 (0.249 [0.248-0.250] vs. 0.165 [0.164-0.166], respectively); as well as for work sector, where those evaluators who work in private and public sectors have a significantly higher value of κ, than surgeons who work in a single sector (private or public) (0.255 [0.254-0.256] vs. 0.190 [0.190-0.191], respectively) (Table 2).

For the concordance assessment by individual categories (grade IV and unclassifiable), the highest κ values were for the categories of greater severity, grades IV and V (0.248 [0.248-0.249] vs. 0.405 [0.404-0.405], respectively); indicating weak to moderate concordance for these categories (Fig. 1A). When stratifying by age, it was found that older evaluators (≥ 38 years) tend to have greater concordance in classifying grades II, IV, and V (κ = 0.262 [0.260-0.264], 0.398 [0.396-0.4] and 0.512 [0.510-0.514], respectively), and significantly lower for unclassifiable cases (κ = 0.141 [0.140-0.143]) compared to younger evaluators (κ = 0.144 [0.254-0.257], 0.167 [0.166-0.169], 0.397 [0.396-0.399], and 0.249 [0.246-0.251]) (p < 0.0001) (Fig. 1B). According to gender, female evaluators have lower concordance for grades IV and V (κ = 0.176 [0.174-0.178] and 0.292 [0.291-0.294]) and a higher concordance when classifying in grade III as well as in unclassifiable (κ = 0.223 [0.221-0.225] and 0.287 [0.285-0.269]) compared to the male evaluators (κ = 0.282 [0.280-0.283], 0.527 [0.526-0.529], 0.109 [0.108-0.111], and 0.108 [0.106-0.109]) (p < 0.0001) (Fig. 1C). On the other hand, specialists tend to have greater agreement when evaluating grades I-II and IV (κ = 0.247 [0.246-0.248],

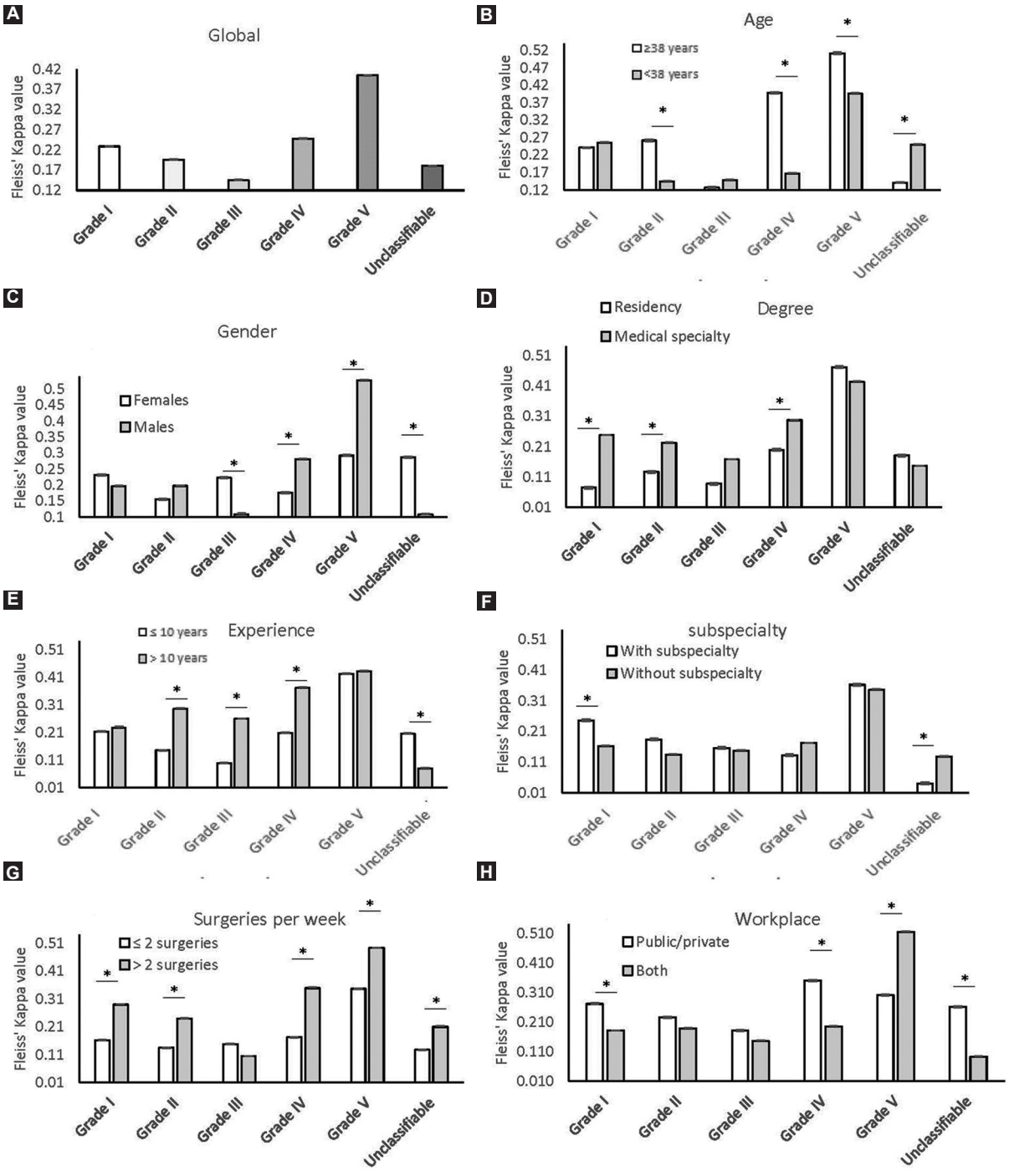


**Figure 1.** *Fleiss κ values by category of the Parkland scale and stratifying by different variables.* **A:** *κ values of the total sample.* **B:** *κ values by age strata.* **C:** *gender.* **D:** *degree.* **E:** *experience as a surgeon.* **F:** *subspecialty.* **G:** *surgeries performed per week.* **H:** *health sector (private/public). (*p < 0.0001).*

0.222 [0.221-0.224], and 0.297 [0.296-0.298]) compared to residents (κ = 0.075 [0.072-0.078], 0.127 [0.124-0.130], and 0.199 [0.196-0.202]) (p < 0.0001) (Fig. 1D).

Surgeons with more than 10 years of experience maintain higher concordance when evaluating grades II-IV (κ = 0.296 [0.293-0.298], 0.259 [0.256-0.261], and 0.372 [0.369-0.374]) and a lower concordance the unclassifiable category (κ = 0.080 [0.077-0.082]) in contrast to evaluators with less experience (κ = 0.146 [0.144-0.147], 0.099 [0.096-0.1], and 0.209 [0.208-0.210])

($p < 0.0001$) (Fig. 1E). For surgeons with a subspecialty, agreement is higher for grade I (0.246 [0.243-0.249]) and significantly lower for the unclassifiable category ($\kappa = 0.04$ [0.037-0.043]) compared to evaluators without a subspecialty ($\kappa = 0.221$ [0.220-0.222] and 0.190 [0.189-0.191]) ($p < 0.0001$) (Fig. 1F). Performing more than two surgeries a week is related to greater concordance classifying with all grades except for grade III ($\kappa = 0.289$ [0.287-0.291], 0.239 [0.237-0.240], 0.349 [0.347-0.351] 0.492 [0.490-0.494], and 0.210 [0.208-0.212]) compared to surgeons performing two or fewer surgeries per week ($\kappa = 0.162$ [0.160-0.163], 0.134 [0.133-0.136], 0.172 [0.170-0.173], 0.346 [0.345-0.348], and 0.127 [0.126-0.129]) ($p < 0.0001$) (Fig. 1G). The evaluators who work in a single hospital (public or private), have a greater concordance for grades I, IV, and for the category of not classifiable ($\kappa = 0.270$ [0.268-0.273], 0.350 [0.348-0.352], and 0.260 [0.258-0.263]) and significantly lower for grade V ($\kappa = 0.3$ [0.298-0.302]) compared to surgeons who work in public and private hospitals ($\kappa = 0.180$ [0.179-0.181], 0.194 [0.193-0.196], 0.092 [0.09-0.093], and 0.514 [0.512-0.515]) ($p < 0.0001$) (Fig. 1H).

## Discussion

The Parkland scale for AC is based on intraoperative findings during a LC and consists of five grades which predict surgical difficulty, while helping to predict the need for conversion to open surgery. Requesting help early on during the surgery from more experienced surgeons can predict better post-operative results, improve surgeon reimbursement, or have even been reported to be useful for discriminating the severity of the disease itself[8]. The scale has an interclass correlation coefficient of 0.804 (95% CI: 0.733-0.867; $p = 0.0001$)[6], as reported by Madni et al., during the imagological evaluation of severely inflamed GBs. In contrast, in this study, a low interobserver concordance was found, according to the $\kappa$ index, which resulted in 0.213 (95% CI: 0.212-0.213) ($p < 0.0001$). This last value also contrasts with that reported by Baral et al.[9], who conclude that this scale is useful for predicting possible post-operative outcomes such as white blood cell's augment, conversion to open surgery, subtotal cholecystectomy, duration of surgery, and bile leaks in patients undergoing LC in rural settings. However, in this last study, the variation between the four surgeons, all belonging to the same institution, was not evaluated, in addition to the fact that most of the images were classified by a single individual, which represents a great bias during the evaluation; even more so due to what was found in the present study, in which the factor "place of work," propitiates a significant variation for higher grade classifications. On the other hand, the potential of this scale to discriminate the severity of AC has been reported, however, as in the previous study, only two highly experienced surgeons were included, performing more than 300 surgeries per year, as reported by the authors, although concordance between both surgeons was not analyzed[10]. The reported experience agrees with what was obtained in this study, which states that the experience of observers with more than 10 years of experience maintains a greater concordance between them for the identification of intermediate categories, which are the most difficult to classify. Being consistent to Madni et al., who also reported that the surgeons who evaluated, these images belonged to the same division of burns, trauma, and intensive care of the same institution, which could have added bias to their assessment. In our work, the observers belonged to different institutions, both public and private in the region, which we consider a strength to our favor. Another aspect that strengthens the development of our work is the number of observers. Since there were 18 observers evaluating 43 images, we were able to obtain 774 data points to analyze, compared to what was done by other studies, in which 550 data points were obtained[6]. In the study conducted by Madni et al., observations were obtained from a retrospective review of intraoperative "initial views" (still images) of the GB. However, it was decided to carry it out in this format to identify the inter-observer variation among a greater number of professionals, as well as the factors involved in this subjectivity, similar to our strategy.

Modifications to the scale have been proposed to make each grade more objective and specific. In concordance with Sugrue et al., we propose a modification to the PGS based on the "type" quality of the adhesions, since sometimes loose adhesions that are easy to remove decrease in grade as they are removed without much effort at the beginning of the dissection, which completely changes the perspective of having an initial view with a high Parkland when performing a CL with a low Parkland after a shallow dissection. Thus, we propose the subclassification of each grade, starting from grade II in firm or loose adhesions[8]. In addition, an issue we consider important is the name, since this tool was proposed as the Parkland Scale for cholecystitis, however, as

surgeons we know that a significant portion of LCs is due to scheduled surgeries for symptomatic cholelithiasis, not only due to AC, and at the time of LCs, there can be changes due to recurrent biliary colic and/or previous AC (adhesions) and not precisely acute inflammatory changes. This is confirmed by histopathological analysis since most of the LCs performed are reported as chronic cholecystitis. Due to this, we propose to rename the Parkland scale for cholecystitis to the Parkland scale for LC. We do not question the usefulness of the Parkland scale as a tool for the intraoperative stage of LC; however, it should be used with more caution and be subjected to public scrutiny to the framing of similar studies that test its accuracy before using it in an indiscriminate and universal way. To the best of our knowledge, this is the first study that tests the interobserver variation of the Parkland scale, and the consideration of variables that could influence this variation; however, we consider the important to carry out a multicenter study, with a greater number of observers considering previous experience using the scale, to be able to extrapolate this data.

## Conclusion

There is little concordance between observers when the PGS is applied, which translates into a high inter-observer variation related to variables such as their age, gender, professional degree, experience, and place of work. In response, we are not seeing the same.

## Acknowledgments

The authors thank the Department of Surgery of the Hospital General Universitario Joaquín del Valle Sánchez in Torreón, Coahuila, Mexico, for their willingness and collaboration in the development of this study.

## Funding

The authors declare that there was no financial support for this study.

## Conflicts of interest

The authors declare that they have no conflicts of interest.

## Ethical disclosures

**Protection of human and animal subjects.** The authors declare that no experiments were performed on humans or animals for this study.

**Confidentiality of data.** The authors declare that they have followed the protocols of their work center on the publication of patient data.

**Right to privacy and informed consent.** The authors have obtained approval from the Ethics Committee for analysis and publication of routinely acquired clinical data and informed consent was not required for this retrospective observational study.

**Use of artificial intelligence for generating text.** The authors declare that they have not used any type of generative artificial intelligence for the writing of this manuscript nor for the creation of images, graphics, tables, or their corresponding captions.

## References

1. Sanford DE. An update on technical aspects of cholecystectomy. Surg Clin North Am. 2019;99:245-58.
2. Moghul F, Kashyap S. Bile duct injury. In: StatPearls. Treasure Island, FL: StatPearls Publishing; 2021.
3. Gutt C, Schläfer S, Lammert F. The treatment of gallstone disease. Dtsch Arztebl Int. 2020;117:148-58.
4. Yokoe M, Hata J, Takada T, Strasberg SM, Asbun HJ, Wakabayashi G, et al. Tokyo guidelines 2018: diagnostic criteria and severity grading of acute cholecystitis (with videos). J Hepatobiliary Pancreat Sci. 2018;25:41-54.
5. Tominaga GT, Staudenmayer KL, Shafi S, Schuster KM, Savage SA, Ross S, et al. The American association for the surgery of trauma grading scale for 16 emergency general surgery conditions: disease-specific criteria characterizing anatomic severity grading. J Trauma Acute Care Surg. 2016;81:593-602.
6. Madni TD, Leshikar DE, Minshall CT, Nakonezny PA, Cornelius CC, Imran JB, et al. The Parkland grading scale for cholecystitis. Am J Surg. 2018;215:625-30.
7. Madni TD, Nakonezny PA, Barrios E, Imran JB, Clark AT, Taveras L, et al. Prospective validation of the Parkland grading scale for cholecystitis. Am J Surg. 2019;217:90-7.
8. Elkbuli A, Meneses E, Kinslow K, Boneva D, McKenney M. Current grading of gall bladder cholecystitis and management guidelines: is it sufficient? Ann Med Surg (Lond). 2020;60:304-7.
9. Baral S, Chhetri RK, Thapa N. Utilization of an intraoperative grading scale in laparoscopic cholecystectomy: a Nepalese perspective. Gastroenterol Res Pract. 2020;2020:8954572.
10. Lee W, Jang JY, Cho JK, Hong SC, Jeong CY. Does surgical difficulty relate to severity of acute cholecystitis? Validation of the parkland grading scale based on intraoperative findings. Am J Surg. 2020;219:637-41.