Interobserver variability in the International Study Group for Pancreatic Surgery (ISGPS)-defined complications after pancreatoduodenectomy

Hendriks, T.E.; Balduzzi, A.; Dieren, S. van; Suurmeijer, J.A.; Salvia, R.; Stoop, T.F.; ... ; Int Study Grp Pancreatic Surg ISGPS

Interobserver Variability in the International Study Group for Pancreatic Surgery (ISGPS)-Defined Complications After Pancreatoduodenectomy

An International Cross-Sectional Multicenter Study

Tessa E. Hendriks, MD,†‡§ Alberto Balduzzi, MD,‖ Susan van Dieren,*†
J. Annelie Suurmeijer, MD, PhD,*† Roberto Salvia, MD, PhD,‖
Thomas F. Stoop, MD,*†¶ Marco Del Chiaro, MD, PhD,¶
Sven D. Mieog, MD, PhD,§ Mark Nielen, PhD,‡ Sabino Zani Jr, MD,#
Daniel Nussbaum, MD,# Thilo Hackert,** Jakob R. Izbicki,**
Ammar A. Javed, MD, PhD,†† D. Brock Hewitt, MD, PhD,††
Bas Groot Koerkamp, MD, PhD,‡‡ Roeland F. de Wilde, MD, PhD,‡‡
Yi Miao, MD, PhD,§§ Kuirong Jiang, MD, PhD,§§
Kohei Nakata, MD, PhD,‖‖ Masafumi Nakamura, MD, PhD,‖‖
Jin-Young Jang, MD, PhD,¶¶ Mirang Lee, MD, PhD,¶¶
Cristina R. Ferrone, MD, PhD,## Shailesh V. Shrikhande, MD, PhD,***
Vikram A. Chaudhari, MD, PhD,*** Olivier R. Busch, MD, PhD,*†
Ajith K. Siriwardena, MD, PhD,††† Oliver Strobel, MD, PhD,‡‡‡
Jens Werner, MD, PhD,§§§ Bert A. Bonsing, MD, PhD,§
Giovanni Marchegiani, MD, PhD,‖‖‖ Marc G. Besselink, MD, PhD,*†✉ and
for the International Study Group for Pancreatic Surgery (ISGPS)*

Objective: To determine the interobserver variability for complications of pancreatoduodenectomy as defined by the International Study Group for Pancreatic Surgery (ISGPS) and others.

Background: Good interobserver variability for the definitions of surgical complications is of major importance in comparing surgical outcomes between and within centers. However, data on interobserver variability for pancreatoduodenectomy-specific complications are lacking.

Methods: International cross-sectional multicenter study including 52 raters from 13 high-volume pancreatic centers in 8 countries on 3 continents. Per center, 4 experienced raters scored 30 randomly selected patients after pancreatoduodenectomy. In addition, all raters scored 6 standardized case vignettes. This variability and the "within centers" variability were calculated for 2-fold scoring (no complication/grade A vs grade B/C) and 3-fold scoring (no complication/grade A vs grade B vs grade C) of postoperative pancreatic fistula, postpancreatoduodenectomy hemorrhage, chyle leak, bile leak, and delayed gastric emptying. Interobserver variability is presented with Gwet AC-1 measure for agreement.

Results: Overall, 390 patients after pancreatoduodenectomy were included. The overall agreement rate for the standardized cases vignettes for 2-fold scoring was 68% (95% CI: 55%–81%, AC1 score: moderate agreement), and for 3-fold scoring 55% (49%–62%, AC1 score: fair agreement). The mean "within centers" agreement for 2-fold scoring was 84% (80%–87%, AC1 score; substantial agreement).

Conclusions: The interobserver variability for the ISGPS-defined complications of pancreatoduodenectomy was too high even though

From the *Amsterdam UMC, location University of Amsterdam, Department of Surgery, Amsterdam the Netherlands; †Cancer Center Amsterdam, the Netherlands; ‡Dutch institute for Clinical Auditing, Leiden, the Netherlands; §Department of Surgery, Leiden University Medical Center, Leiden, The Netherlands; ‖Unit of General and Pancreatic Surgery, Department of Surgery, Dentistry, Paediatrics and Gynaecology, The Pancreas Institute Verona, University of Verona, Verona, Italy; ¶Department of Surgery, Division of Surgical Oncology, University of Colorado Anschutz Medical Campus, Aurora, CO; #Department of surgery, Duke University, Durham, NC; **Department of General Visceral and Thoracic Surgery University Hospital Hamburg-Eppendorf, Germany; ††Department of Surgery, New York University Langone Health, New York, NY; ‡‡Department of surgery, Erasmus University Medical Center, Rotterdam, the Netherlands; §§Pancreas Center, The First Affiliated Hospital of Nanjing Medical University, China; ‖‖Department of Surgery and Oncology, Graduate School of Medical Sciences, Kyushu University, Fukuoka, Japan; ¶¶Department of Surgery, Seoul National University Center, Seoul, Korea; ##Department of Surgery, Cedars-Sinai Medical Center, Los Angeles, CA; ***Department of Gastrointestinal Surgical Oncology, Tata Memorial Hospital, Mumbai, India; †††Hepatobiliary and Pancreatic Surgery Unit, Manchester Royal Infirmary, Manchester University NHS FT, Manchester, United Kingdom; ‡‡‡Department of General Surgery, Division of Visceral Surgery, Medical University of Vienna, Vienna, Austria; §§§Department of General, Visceral and Transplant Surgery, University Hospital, LMU Munich, Munich, Germany; and ‖‖‖Department of Surgery, Oncology and Gastroenterology (DISCOG), University of Padua, Padua, Italy.
✉m.g.besselink@amsterdamUMC.nl.

G.M. and M.G.B. shared senior authorship.

The authors report no conflicts of interest.

Supplemental Digital Content is available for this article. Direct URL citations are provided in the HTML and PDF versions of this article on the journal's website, www.annalsofsurgery.com.

DOI: 10.1097/SLA.0000000000006473

the “within centers” agreement was acceptable. Since these findings will decrease the quality and validity of clinical studies, ISGPS has started efforts aimed at reducing the interobserver variability.

Keywords: interobserver variability, pancreatoduodenectomy, ISGPS, complications, morbidity, pancreatic surgery

(*Ann Surg* 2024;280:728–733)

Pancreatic surgery is associated with a high risk of major complications (26%–40%),[1,2] often requiring radiological, endoscopic, and surgical interventions leading to prolonged hospital stay and increased health care costs.[3] Therefore, accurate assessment and reporting of these complications is essential for comparing and improving health care outcomes in patients after pancreatic surgery. In the past decades, the International Study Group for Pancreatic Surgery (ISGPS) has defined and classified the most relevant pancreatic surgery-specific complications: postoperative pancreatic fistula (POPF) and delayed gastric emptying (DGE) which occur most often (11%–43%), postpancreatectomy hemorrhage (PPH), chyle leakage (CL), and postpancreatectomy acute pancreatitis, whereas bile leakage (BL) was defined by the International Study Group for Liver Surgery.[4–10] These ISGPS and International Study Group for Liver Surgery definitions are widely regarded as the standard for international research in pancreatic surgery.

In 2022, the Dutch Pancreatic Cancer Group[11] noted substantial variation in the rate of ISGPS-defined complications between 16 Dutch centers. As this variation was thought to be explained by inadequate training, online meetings were hosted for those responsible for entering the data. Unexpectedly, during these meetings, when discussing “case vignettes,” the agreement between the attendees remained poor despite the exact ISGPS definitions being shown. The consequences and reproducibility of these findings are yet unclear as international studies on the interobserver agreement of the ISGPS-defined complications of pancreatic surgery have not been performed.

This international study aims to investigate the interobserver variability for the ISGPS-defined complications of pancreatoduodenectomy.

METHODS

Study Design

This international multicenter cross-sectional retrospective clinical chart review and case-vignette study included 13 international high-volume expert centers in pancreatic surgery from 8 countries on 3 continents. Each center provided 4 raters experienced in registering complications according to the ISGPS definitions (eg, surgeons, surgical residents, researchers, trained data analysts/nurses). Each rater scored all ISGPS-defined complications in 30 randomly selected patients from their own center and in 6 standardized case vignettes (Supplementary 8, Supplemental Digital Content 1, http://links.lww.com/SLA/F237). The ISGPS definitions were mandatorily “printed on the desk” available for each reviewer during scoring. All reviewers had full access to the patient’s medical records, eg, clinical charts, laboratory results, radiology reports, correspondence, and postdischarge information. Per hospital, the scoring was performed during the same time frame. The study was reported in accordance with the STROBE guidelines.[12]

TABLE 1. Interpretation of Gwet-AC1/Fleiss’ Kappa Score

Gwet-AC1[18]/Fleiss’ Kappa[19]	
< 0	Less than chance agreement
0.01–0.20	Slight agreement
0.21–0.40	Fair agreement
0.41–0.60	Moderate agreement
0.61–0.80	Substantial agreement
0.81–0.99	Almost perfect agreement

Definitions and Data Collection

The following pancreatic surgery-specific complications were assessed: POPF,[4] PPH,[8] DGE,[7] BL,[6] and CL.[5] Postpancreatectomy acute pancreatitis[10] was only collected in the 6 standardized case vignettes as this definition was only recently introduced.

Each participating center provided the total number of pancreatoduodenectomies performed in the years 2020 and 2021 (Supplementary Table 1, Supplemental Digital Content 2, http://links.lww.com/SLA/F238). For every center, based on this total 2-year volume, a random selection of 30 patients (ie, 30 numbers) was made using an online random number generator (Appendix). The centers originated from the United States (n = 4), the Netherlands (n = 3), Italy (n = 1), China (n = 1), Germany (n = 1), Japan (n = 1), India (n = 1), and the Republic of Korea (n = 1).

The “overall” interobserver variability for scoring ISGPS definitions was determined using the 6 standardized case vignettes, which were scored by all reviewers.

Per center, an interobserver variability was calculated based on the 30 randomly selected patients. The mean per center interobserver variability was referred to as the “within center” interobserver variability.

A non-WMO declaration was obtained for this study stating that the Medical Research Involving Human Subjects Act (WMO) does not apply.

Statistical Analysis

The interobserver variability was represented as percent agreement and unweighted Gwet.AC1 score (AC1). For the purpose of this study, which is to determine the interobserver agreement adjusted for chance agreement between multiple reviewers and multiple categories, the standard Cohen Kappa is not suitable. As this, Cohen Kappa is originally designed to be used for a nominal outcome and a maximum of 2 raters. Also, it is known to have 2 types of paradoxes that problematize correctly.[13–16] AC1 was chosen as the interpretation is very similar to Cohen Kappa scoring however the scoring system is less affected by prevalence and marginal probability than the original and frequent used Cohen Kappa scoring, as it corrects for the “Kappa paradox.”[17,18] Gwet.AC1 can be interpreted as follows: Table 1, (Supplementary Table 1, Supplemental Digital Content 2, http://links.lww.com/SLA/F238.

As AC1 scores are less frequently used, we also presented the calculations as Fleiss’ Kappa score,[19] which is an extension of Cohen Kappa for easier comparison. Supplementary Tables 4–7, Supplemental Digital Content 2, http://links.lww.com/SLA/F238.

The overall interobserver variability and the mean center-variability were calculated for 2-fold scoring (ie, no complication/grade A vs grade B/C) and threefold scoring (ie, no complication/grade A vs grade B vs grade C) of the 5 common ISGPS-defined complications. Also, the interobserver variability for each ISGPS complication category was assessed. For sensitivity analyses, all patients were removed for whom all reviewers agreed on the absence of any (A/B/C) complication.

All Statistical analyses were performed using R statistical computing software version 4.2.3.

Ethics

As this study involved no individually identifiable patient data, no patients were allocated to different treatments for the study and the results are compared with existing standards it was regarded as an audit.

RESULTS

The ISGPS complications were scored for 390 patients after pancreatoduodenectomy by 52 experts. Most of the experts were pancreatic surgeons and fellows (75%), followed by surgical residents (13%) with a median clinical experience of 11 years (IQR 6–20). The remaining categories included PhD candidates and data managers. Per rater, the median experience with scoring ISGPS-defined complications was 5 years (IQR 2–10).

Overall Interobserver Variability

The mean overall rate of agreement for the standardized cases vignettes for 2-fold scoring was 68% (95% CI: 55%–81%, AC1 score: moderate agreement) (Supplementary Table 1, Supplemental Digital Content 2, http://links.lww.com/SLA/F238) and 55% for 3-fold scoring (49%–62%, AC1 score: fair agreement) (Supplementary Table 3, Supplemental Digital Content 2, http://links.lww.com/SLA/F238). The overall lowest percent agreement was for POPF 61% and CL 61%.

Interobserver Variability "Within Centers"

In 27% (18%–35%) of the 30 randomly selected patients per center, all reviewers agreed on the absence of any type (A/B/C) of complication. In 49% (42%–57%) of patients all reviewers agreed on the absence of any grade B/C complication. The mean "within centers" agreement for 2-fold scoring was 84% (80%–87%, AC1 score: substantial agreement) (Table 2). Per complication type, the mean agreement was the lowest for 2-fold scoring of POPF (89%) and DGE (92%) (Table 3).

Sensitivity Analysis

When excluding all patients in whom all reviewers agreed that no type (A/B/C) of ISGPS-defined complication was present, the mean center-specific agreement for twofold scoring lowered to 75.0% (70.9%–79.1%). The "within centers" agreement was the lowest for 3-fold scoring (Table 3).

The Fleiss Kappa scores for all calculations presented a slightly lower interobserver agreement (Supplementary Table 4–7, Supplemental Digital Content 2, http://links.lww.com/SLA/F238).

DISCUSSION

This first international study evaluating the interobserver variability for the 5 most relevant pancreas-specific complications following a pancreatoduodenectomy (POPF, PPH, CL, BL, and DGE) in 390 patients by 52 experts found a high overall interobserver variability based on the assessment of standardized case vignettes. The "within centers" variability seemed to be acceptable based on the assessment of 30 randomly selected patients in each participating center. When taken together, this demonstrates that improvements are needed to facilitate the use of these definitions with better interobserver agreement.

Previous studies outside the ISGPS-defined complications have reported that poor interobserver variability is a common problem for clinical classification systems.[20–22] Such classifications are, however, crucial as they provide a framework for surgical outcomes research. This ensures that health care professionals across specialties use a common language when assessing complications and comparing/benchmarking outcomes between centers.

The ISGPS-defined complications are internationally recognized by experts as the standard for scoring complications after pancreatoduodenectomy and are commonly combined with the Clavien-Dindo classification.[22] Consequently, the observed poor interobserver variability has several important implications for clinical practice and research. First, it might result in considerable hospital variation, which in turn can affect the ability to benchmark and compare complications following pancreatoduodenectomy.[23,24] Second, it decreases the quality and validity of clinical studies. Especially when trials use the ISGPS-defined complications as endpoints, the results should be interpreted with caution. Increased consistency in complication assessment enables the identification of specific patterns and areas for improvement of quality of care. Also, it might hinder correct sample size calculations for randomized controlled trials.

How to explain the observed high interobserver variability? First, differences in training and attitude may influence our judgment when scoring complications. Second, the definitions themselves may need clarification. When asking the reviewers' explanation for the high interobserver variability, a frequently heard response was "lack of clarity" in the definitions and "overlap" of particular grades. As an example, in the POPF definition "clinically relevant change of management" differentiates a grade A fistula from a grade B. However, there is no clear definition given what this "clinically relevant change of management" exactly encompasses. A more anecdotal example was the response of a surgeon, when asked why he had scored a grade C complication while his colleagues unanimously scored a

TABLE 2. "Within centers" interobserver variability

Within centers interobserver variability (2-fold scoring) Any ISGPS complication "Yes" (ISGPS grade B/C) vs "No" (none or ISGPS grade A)														
Center	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**	**L**	**M**	**mean (95%-CI)**
Percent agreement	71.7	77.8	78.9	80.6	80.6	80.6	83.9	85.0	86.7	87.8	91.1	93.3	93.3	83.9 (80.4–87.4)
gwet AC1	0.452	0.563	0.662	0.678	0.735	0.640	0.678	0.700	0.733	0.772	0.825	0.880	0.882	0.708 (0.642–0.773)

TABLE 3. "Within Centers" Interobserver Variability, Per Complication Type

	Twofold scoring* per complication type		Threefold scoring†	
	Percent agreement	gwet-AC1	percent agreement	gwet-AC1
	mean (95%-CI)	mean (95%-CI)	mean (95%-CI)	mean (95%-CI)
POPF	89.2 (87.5–90.9)	0.839 (0.812–0.866)	88.5 (87.0–90.0%)	0.845 (0.820–0.870)
PPH	94.9 (93.7–96.2)	0,938 (0.921–0.955)	94.4 (93.0–95.7)	0.936 (0.920–0.953)
DGE	92.1 (90.6–93.6)	0.892 (0.870–0.914)	89.9 (88.0–91.8)	0.882 (0.858–0.905)
CL	94.9 (93.2–96.5)	0.933 (0.909–0.964)	94.9 (93.2–96.5)	0.933 (0.909–0.957)
BL	96.2 (95.4–96.9)	0.955 (0.947–0.964)	96.1 (95.3–96.9)	0.957 (0.948 – 0.965)

*No complication/ ISGPS grade A vs grade ISGPS grade B/C.

†No complication/ ISGPS grade A vs grade ISGPS grade B vs ISGPS grade C.

BL iindicates bile leakage; CL, chyle leakage; DGE, delayed gastric emptying; POPive pancreatic F, postoperatfistula; PPH = post-pancreatectomy hemorrhage; PPAP, post-pancreatectomy acute pancreatitis.

grade B, was "the whole course of events in this case felt more like a type C complication."

The better "within centers" interobserver variabilty compared to the considerably worse "overall" interobserver variability based on the standardized cases vignettes might be explained by the lower rate of complications in the randomly selected patients. Also, the better scores could reflect better alignment regarding complication scoring within the "own" team, as within a center, the reviewers a more likely to be trained in the same manner. Moreover, the process of defining and scoring a given complication takes place only at the end of the clinical trajectory. The rater must, therefore, go back and analyze the entire clinical history to depict whether a complication has indeed occurred or not and, if present, assess its severity. This is a time-consuming process and, indeed, not completely intuitive. The interobserver variability may, therefore, be improved by using digital technologies that are nowadays widely available although this requires further study.

Some limitations of this study must be taken into account. First, the selection process for the 30 patients per center. Due to the random selection process, aimed to prevent systemic selection bias, complications were less frequently present in the total sample. This might have resulted in inflated (ie, better) interobserver variability. Also, this selection process may have introduced an uneven distribution of complications per center. Second, due to the retrospective nature of clinical chart reviews, the reviewers' interpretation could have been influenced by missing data. Although, per center, all 4 reviewers had access to the same information, this cannot be said when comparing centers. By choosing expert centers in pancreatic surgery, we expected an overall good quality of patient information in the patient record and a good representation of the daily clinical practice, registering complications. However, there may be differences between centers in the way patient information is recorded. Therefore, a comparison of the observed agreements between centers should be done with caution. Third, we did not specify a certain postoperative period which could have influenced outcomes. However, the ISGPS definitions specify postoperative periods; the POPF definition, for example, includes a 28-day period for the presence of a surgical drain, which will require a follow-up definition. Fourth, in contrast, the 6 case vignettes were based on complex "real" patients, of which some parts were hypothetical to shape the situation toward specific types of complications. As a result, these vignettes could lead to an underestimation of the interobserver variability. Also, a short vignette description will never completely imitate the "real" situation of scoring complications based on a medical record wherein a certain major event (eg, a relaparotomy) will be highlighted in several locations in a medical record. The main strengths of this study include the structured process of scoring complications by a large group of international experts. Also, the study accounted both for patients from the observer's own center and for more complex patients (ie, the case vignettes), which shapes a broad context of the interobserver variability.

These findings then lead to the question of how to improve the observed high interobserver variability for the ISGPS-defined complications. A more standardized manner of registering patient data would be recommended. Data should be collected prospectively using case record forms at prespecified time points. Also, the manner of scoring should be done in a more standardized manner, for example, by using a clear flowchart or facilitated by a decision tool in a smartphone/online application. Also, improving the current definitions according to the MECE-principle (mutually exclusive and collectively exhaustive) by making more clear cutoffs between the different grades would be helpful and could be integrated into an application. The ISGPS has started an improvement project to reach this goal which will clearly also require international validation.

In conclusion, the observed interobserver variability for the ISGPS-defined complications illustrates the need for improvement of the use and details of these definitions for patients undergoing pancreatic surgery. New applications should be developed and tested to reduce the interobserver variability for the ISGPS definitions and grading systems.

COLLABORATORS

Nynke Michiels, Valerie Rebattu, Fabio Casciani, Salvatore Paiella, Serena Mele, Christopher Wolfgang, Sarah Kaslow, Peter Allen, Dan Blazer, Oskar Franklin, Salvador Rodriguez Franco, Michael Kirsch, Toshitaka Sugawara, Rutger Theijse, Marie Capelle, Roel Haen, Martina Nebbia, Louisa Bolm, Zhi Ven Fong, Amit Chopde, Aditya Kunte, Kaival Gundavda, Gurudutt Varty, Naoki Ikenaga, Toshiya Abe, Zipeng Lu, Baobao Cai, Mara Götz, Faik G. Uzunoglu, Jan Bardenhagen, Fiete Gehrisch, Won-Gun Yun, Youngmin Han, Savio George A. Da P. Barreto, Horacio Asbun, Vollmer Charles, Falconi Massimo, Hartwig

Werner, MD, Prof, Adham Mustapha, Fingerhut Abe, Bockhorn Maximillian, Zyromski Nicholas, Boggi Ugo, Sato Asahi, Halloran Christopher, Butturini Giovanni, Fusai Giuseppe Kito, Friess Helmut, Lillemoe Keith D, Conlon Kevin.

REFERENCES

1. Butturini G, Marcucci S, Molinari E, et al. Complications after pancreaticoduodenectomy: the problem of current definitions. *J Hepatobiliary Pancreat Surg*. 2006;13:207–211.
2. Harnoss JC, Ulrich AB, Harnoss JM, et al. Use and results of consensus definitions in pancreatic surgery: a systematic review. *Surgery*. 2014;155:47–57.
3. Lin JW, Cameron JL, Yeo CJ, et al. Risk factors and outcomes in postpancreaticoduodenectomy pancreaticocutaneous fistula. *J Gastrointest Surg*. 2004;8:951–959.
4. Bassi C, Marchegiani G, Dervenis C, et al. The 2016 update of the International Study Group (ISGPS) definition and grading of postoperative pancreatic fistula: 11 Years After. *Surgery*. 2017;161:584–591.
5. Besselink MG, van Rijssen LB, Bassi C, et al. Definition and classification of chyle leak after pancreatic operation: a consensus statement by the International Study Group on Pancreatic Surgery. *Surgery (United States)*. 2017;161: 365–372.
6. Koch M, Garden OJ, Padbury R, et al. Bile leakage after hepatobiliary and pancreatic surgery: a definition and grading of severity by the International Study Group of Liver Surgery. *Surgery*. 2011;149:680–688.
7. Wente MN, Bassi C, Dervenis C, et al. Delayed gastric emptying (DGE) after pancreatic surgery: a suggested definition by the International Study Group of Pancreatic Surgery (ISGPS). *Surgery*. 2007;142:761–768.
8. Wente MN, Veit JA, Bassi C, et al. Postpancreatectomy hemorrhage (PPH): an International Study Group of Pancreatic Surgery (ISGPS) definition Surgery. *Surgery*. 2007;142:20–25.
9. Bassi C, Dervenis C, Butturini G, et al. Postoperative pancreatic fistula: an international study group (ISGPF) definition. *Surgery*. 2005;138:8–13.
10. Marchegiani G, Barreto SG, Bannone E, et al. Postpancreatectomy acute pancreatitis (PPAP): definition and grading from the International Study Group for Pancreatic Surgery (ISGPS). *Ann Surg*. 2022;275:663–672.
11. Strijker M, Mackay TM, Bonsing BA, et al. Establishing and coordinating a nationwide multidisciplinary study group: lessons learned by the Dutch Pancreatic Cancer Group. *Ann Surg*. 2020;271:E102–E104.
12. von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Lancet*. 2007;370:1453–1457.
13. Lantz CA, Nebenzahl E. Behavior and interpretation of the κ statistic: resolution of the two paradoxes. *J Clin Epidemiol*. 1996;49:431–434.
14. Cicchetti DV, Feinstein AR. High agreement but low kappa: II. Resolving the paradoxes. *J Clin Epidemiol*. 1990;43:551–558.
15. Feinstein AR, Cicchetti DV. High agreement but low Kappa: I. the problems of two paradoxes. *J Clin Epidemiol*. 1990;43: 543–549.
16. Derksen BM, Bruinsma W, Goslings JC, et al. The Kappa Paradox Explained. *J Hand Surg Am*. 2024;49:482–485.
17. Wongpakaran N, Wongpakaran T, Wedding D, et al. A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. *BMC Med Res Methodol*. 2013;13:1–7.
18. Gwet KL. Computing inter-rater reliability and its variance in the presence of high agreement. *Br J Math Stat Psychol*. 2008; 61:29–48.
19. Fleiss JL. Measuring nominal scale agreement among many raters. *Psychol Bull*. 1971;76:378–382.
20. Poletajew S, Zapala L, Piotrowicz S, et al. Interobserver variability of Clavien–Dindo scoring in urology. *Int J Urol*. 2014;21:1274–1278.
21. De La Rosette, JJMCH, Opondo D, et al. Categorisation of complications and validation of the Clavien Score for percutaneous nephrolithotomy. *Eur Urol*. 2012;62:246–255.
22. Clavien PA, Sanabria JR, Strasberg SM. Proposed classification of complications of surgery with examples of utility in cholecystectomy. *Surgery*. 1992;111:518–526.
23. Raptis DA, Sánchez-Velázquez P, MacHairas N, et al. Defining benchmark outcomes for pancreatoduodenectomy with portomesenteric venous resection. *Ann Surg*. 2020;272:731–737.
24. Sánchez-Velázquez P, Muller X, Malleo G, et al. Benchmarks in pancreatic surgery: a novel tool for unbiased outcome comparisons. *Ann Surg*. 2019;270:211–218.

APPENDIX

Random Integer Generator

This form allows you to generate random integers. The randomness comes from atmospheric noise, which for many purposes is better than the pseudo-random number algorithms typically used in computer programs https://www.random.org/integers/.

DISCUSSANT

Massimo Falconi (Milano, Italy)

Thanks to the European Surgical Association for the opportunity to discuss this study, and congratulations to Dr Hendriks for her interesting presentation and to the colleagues of ISGPS for their effort in conducting this interesting study involving 13 centers worldwide. The paper reports the findings from an international multicenter collaboration by the International Study Group for Pancreas Surgery (ISGPS) on the interobserver variability of ISGPS-defined complications, including pancreatic fistula, delayed gastric emptying, and postpancreatectomy hemorrhage. The study raises very important questions about the validity of complication-specific classifications that are broadly used in pancreatic surgery studies. I believe their findings will certainly stimulate a discussion within ISGPS and the pancreatic surgeons' community about revising and improving the classifications. However, the results are not surprising, as multiple studies have been performed in the past evaluating the variability and inconsistency in postoperative complication reporting and classification. In addition, the main results of this research derive from a retrospective clinical chart review performed by multiple assessors, which is inevitably influenced by the quality of the available data.

Please Dr. Hendriks, could you address the following questions: First, in 2010, Professor Clavien's group found that surgical residents did not record up to 80% of postoperative events correctly even after specific training. This was probably related to a lack of motivation considering it is a time-consuming and unrewarded activity. In addition, there is convincing evidence that nonclinicians are better data collectors than clinicians. Why do you think interobserver variability was poor in your study, especially for case vignettes? Was it due to a lack of training, a lack of motivation?

Second, considering the retrospective nature of chart review for interobserver variability within centers. Do you think this influenced your results? Would prospectively collecting data through standardized CRFs help reduce missing data and improve variability?

Finally, technically, how can we improve these classifications and create clearer definitions and cutoffs between complication grades?

Response From Tessa E Hendriks (Amsterdam, the Netherlands)

Thank you so much for these interesting questions. With regard to your first question about whether the issue is a lack of training or a lack of motivation; as a junior doctor, I cannot speak to the motivation of senior surgeons. However, I believe training could play a role. While all were familiar with the definitions, some nuances might be missed if not studied thoroughly. This is an important point, but it is challenging to address with the current ISGPS definitions.

Your second question concerns whether retrospective chart review influenced the poor interobserver agreement. In this study, all 4 reviewers had the same access to the data, and of course, I think it is difficult to dive into the patient file; maybe some surgeons had an extra coffee in the morning and thereby were a little sharper and did retrieve some of the missing information. So, I think, yes, that is a great nuance. On the other hand, the standardized case vignettes showed us that even if all necessary data were available, experts still couldn't agree on the complication and grade.

To answer your final question on how we can improve this in the future: I was inspired by a recent take on artificial intelligence. While AI might offer a solution eventually, it is not yet ready for application. In the meantime, we have developed an ISGPS smartphone app that sharpens the definitions by making clear cutoff points. We will validate this app in the coming year.

Christiane Bruns (Cologne, Germany)

Thank you for a nice presentation. I do have a few question about the selected case presentations validated by the reviewers. First, what exactly means a postpancreatectomy hemorrhage? Does it mean a first sentinel bleeding, or does it mean a substantial or massive bleeding based on a pancreatic fistula? How did you define that case? Second, how did you define delayed gastric emptying as a complication: severity, duration, and therapy?

Response From Tessa E Hendriks (Amsterdam, the Netherlands)

Thank you for these very important questions. With regard to your first question, the example cases were all based on real-world patients. In the definition of post-pancreatectomy hemorrhage, we noticed a problem within the definition because the definition, on one hand, wants to make the difference to early-onset and late-onset bleeding, telling us something about sentinel bleeds. Then, this definition at the same time wants to make a differentiation in severity grade: from no clinical relevancy to life-threatening complication. I think the problem with this complication definition is that it wants 2 things; tell us something about the etiology, but also wants to tell us something about the severity. Therefore, I think this definition might need some clearer cutoff points. We made a very simple case of a patient with an early-onset bleed with severe consequences and almost all surgeons scored this patient a grade C. However, it was an early onset bleed, and therefore severe complications would be in the B category.

Concerning your second question – for the delayed gastric emptying the same applies. There are so many days where different cutoff points are and that makes it much more difficult to really define the patient to the correct category, so therefore we made a similar case.

Pierre-Alain Clavien (Zurich, Switzerland)

Thank you very much for bringing this controversial topic to ESA, which is a key platform to critically talk about limitations of complication reporting. This is indeed the role of academic surgeons to remain unbiased when evaluating our results, rather than just congratulate ourselves for our "superior job." The drawback of your report – disclosing our inability to judge our work – might lead others, such as our internist colleagues, to ask for a moratorium on performing duodeno-pancreatectomy until we do better in reporting our results. What are your recommendations? Should we rely only on our validated routine complication grading system by severity or comprehensive complication index (CCI) for overall morbidity? Can we still use the ISGPS grading system to better focus on specific negative events like delayed gastric emptying and bleeding. In short, what do you suggest preventing the confusion?

Response From Tessa E Hendriks (Amsterdam, the Netherlands)

First, I am deeply honored by your question, particularly given your expertise in this field. I kind of expected a question like this. I believe we should not stop using these complication definitions, as they describe medical problems that we can act upon. The ISGPS definitions give us the handles to tackle these problems. Therefore, they are important as they tell us where to look and where to intervene. We still need to use them, but we must improve them. Second, I think there are other measures, such as "The Clavien-Dindo score," which, of course, we always report in our studies, tell us something about the severity of a complication. However, we need to improve and sharpen the current ISGPS definitions and find a way of entering the data in such a way that we reach reliable conclusions. Potentially with the support of a smartphone app.