Maximilian Berlet*, Jonas Fuchtmann, Lukas Bernhard, Alissa Jell, Marie-Christin Weber, Philipp-Alexander Neumann, Helmut Friess, Michael Kranzfelder, Hubertus Feussner and Dirk Wilhelm

Laparoscopic Cholecystectomy – A Proper Model Surgery for AI based Prediction of Adverse Events?

Analysis of possible predictive values on the basis of the German reimbursement statistics

https://doi.org/10.1515/cdbme-2022-0002

Abstract: Laparoscopic cholecystectomy (LCHE) is a widely employed model for surgical instrument and phase recognition in the field of machine learning (ML), with the latter being assigned to identify critical events and to avoid complications. Although ML algorithms have been proven to be effective for this instance and in selected patients, it is questionable whether patients receiving LCHE in daily clinical routine would actually benefit from adverse event prediction by ML applications. We believe, that the statistical problem of low prevalence (PREV) of potential adverse events in an unselected population and consequential low diagnostic yield was not considered adequately in recent research. Therefore, we performed a query to the G-DRG (German Diagnosis Related Groups) database of the German Federal Statistical Office with the aim to calculate prevalence of surgical and postoperative adverse events coming along with LCHE. The results enable an estimation of positive (PPV) and negative (NPV) predictive values hypothetically achievable by ML applications aiming to predict an adverse surgical course.

Keywords: Laparoscopic cholecystectomy, adverse events, prevalence, artificial intelligence

1 Introduction

Laparoscopic cholecystectomy (LCHE) is a procedure with a high degree of standardization and has replaced the open approach for most indications. In terms of machine learning (ML), it serves as a model for numerous scientific issues. LCHE is performed hundreds of thousand times per year all over industrial countries and is easy accessible for video recording in daily clinical routine. Main applications of ML-based solutions are detection of laparoscopic surgical instruments [1], anatomical structures [2], and even the prediction of surgical course [3,4]. Furthermore, LCHE serves as an established model for improvement of robotic and computer-assisted surgery. [5] Video, sensor, and clinical data represent possible input for ML applications, mostly realized in form of Convolutional Neural Networks (CNN). [6] Datasets with readily annotated video records and clinical parameters like Cholec80 or CholecSeg8k are freely available for research. [7] Although substantial advance has been achieved in the field of phase and adverse event recognition [8-10], it is unclear whether LCHE actually is a proper model surgery for postoperative outcome prediction in a real world unselected population. Numerous recent works report impressive sensitivity (SENS) and specificity (SPEC) rates exceeding the 80% mark. Taken the reported low incidence of complications, it is rather questionable, how useful those applications would come in clinical routine. SENS and SPEC do not depend on the test collective's characteristics, but are properties of the test itself. Contrarily, parameters that correlate with the prevalence of adverse events, are positive (PPV) and negative (NPV) predictive values. [11] These values stand for the probability that a prediction of an adverse event (PPV) or its denial (NPV) by a particular test is correct. In this article we deliver prevalences for a set of relevant adverse events based on the German reimbursement statistics, comprising all 1.8 million LCHE performed in Germany from 2008 to 2018. Thus, PPV and NPV achievable by hypothetical ML applications trained on these adverse events become estimatable. Figure 1 illustrates the calculation of fundamental parameters as SENS, SPEC, PPV, NPV, and PREV. The table

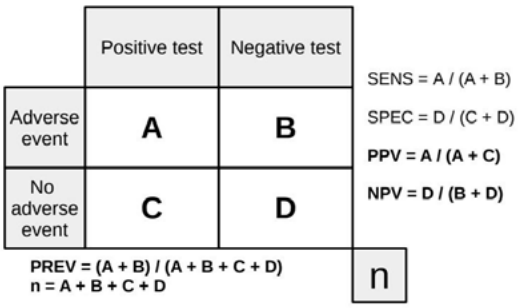

Figure 1: Contingency table of fundamental test parameters: sensitivity (SENS), specificity (SPEC), positive predictive value (PPV), negative predictive value (NPV), prevalence (PREV), and size of the whole population (n)

***Corresponding author: Maximilian Berlet:** MITI research group, Surgical department, Klinikum rechts der Isar, Technical University of Munich, Ismaninger Str. 22, 81675 München, Germany, e-mail: maximilian.berlet@tum.de, **Marie-Christin Weber, Philipp-Alexander Neumann, Helmut Friess:** Surgical department, Klinikum rechts der Isar, Technical University of Munich, **Jonas Fuchtmann, Lukas Bernhard, Alissa Jell, Michael Kranzfelder, Hubertus Feussner, Dirk Wilhelm:** MITI research group, Surgical department, Klinikum rechts der Isar, Technical University of Munich

fields (A-D) therefore have to be filled with the absolute case numbers in the particular groups.

2 Material and Methods

We performed a query to the G-DRG database of the German Federal Statistical Office (DESTATIS) [12]. The query code was written in SAS language version 9.3. All cases of LCHE and converted LCHE (OPS 5-511.1, 5-511.2) between 2008 and 2018 in Germany were included. Prevalences of relevant adverse events comprising in-house mortality, need for surgical revision, postoperative bleeding (ICD-10 T81.0), accidental violation of anatomic structures (T81.2), surgical site infection (T81.4), and anesthesiological complications (T88.2-T88.6) were calculated on this basis. Then, we estimated PPV and NPV presuming a stepwise run through the range of SENS and SPEC between 50% and 99%. The script for calculation was written in the statistical language R version 3.6. [13] Results are presented for the unselected collective of all LCHE, and additionally for the preselected group of cases with conversion to open surgery. Finally, we propose the draft of a representative test collective for ML applications and responsible sample sizes based on the characteristics of all LCHE performed during the period of observation. Rates are given in mean percentage ± standard deviation.

3 Results

The query to the G-DRG database revealed a yearly number of 164,238 ± 4,233 laparoscopic cholecystectomies. The conversion rate was at 4.65 ± 0.71%. The overall in-house mortality following LCHE and during the same hospital stay was at 0.54 ± 0.6% while the mortality rate was much higher in case of converted surgery (3.31 ± 0.66%). Of note, the mortality rate increased significantly during the period of observation. (Figure 2**B**) Nearly all adverse events were more likely to emerge in case of a conversion from laparoscopic to open surgery. Due to the data structure, it is unfortunately not possible to determine whether a complication caused the conversion or the other way around. The overall rate of revision during the same hospital stay was at 1.35 ± 0.04% and at 9.49 ± 1.6% in converted surgeries. The most common complication was a postoperative bleeding with an overall rate of 1.41 ± 0.1% and 4.05 ± 0.6% in case of conversion. Likewise, the bleeding rate increased as well during the observation time while the postoperative infection rate decreased in both, overall (0.63 ± 0.09%) and conversion group (3.61 ± 0.25%). Anesthesiological complications are rare in both collectives with 0.41 ± 0.07% for the entire cohort and 0.48 ± 0.14% for the conversion group. Accidental violation of anatomic structures occurred in 0.41 ± 0.06% overall and in 2.91 ± 0.42% in case of conversion. Based on these rates, we estimated positive and negative predictive values for hypothetical ML applications predicting these

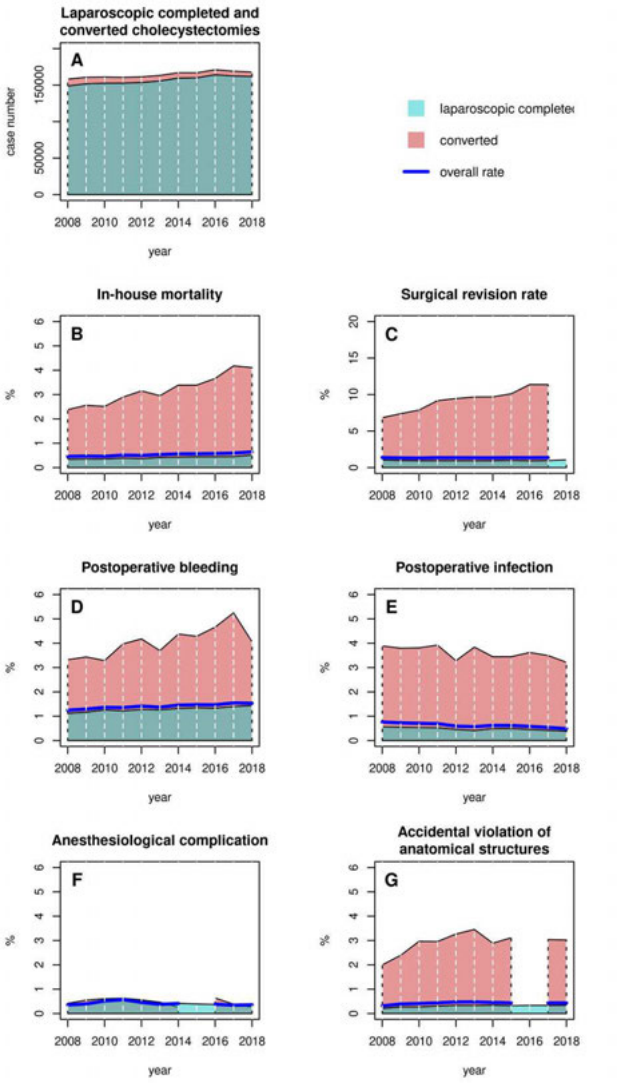

Figure 2: Adverse event rates of laparoscopic cholecystectomy in Germany between 2008 and 2018, gaps in the plots represent a lack of data for the particular year. **A** y-axis: absolute case numbers **B-G** y-axis: percentage

particular adverse events. Figure 3 illustrates the calculation based on the rates of 2018 (Panel **A-C** and **F-M**) and 2017 (Panel **D** and **E**). As expected, the maximum achievable PPV presuming all combinations of SENS and SPEC depends on the prevalence of the specific complication.

As can be seen in these illustrations, a hypothetical ML application, predicting the death of a patient during the hospital stay after LCHE with a presumed SENS of 1.0 and a SPEC of 0.99 would still just reach a maximum PPV of less than 0.4. (Figure 3**B**) This illustrates, that if the ML applications predicts the patient's death, the probability that this event will really occur will not exceed 40%. The same theoretical ML approach applied to the collective of converted LCHE presuming same SENS and SPEC would contrarily reach a PPV of about 0.8 implying that a prediction of death would actually indicate in-hospital mortality in 80% of cases. (Figure 3**C**) This fact emphasizes the necessity of a prior assessment of the pre-test probability of specific adverse events. In case of low prevalence, even a very sensitive and

specific test applied to an unselected population of LCHE patients appears to be rather useless. On the other hand, the data obtained from our query can be the basis to estimate sample size and characteristics of a representative collective used for inference of a hypothetical ML application. Table **1** depicts the composition of an unselected test collective with different sample sizes, regarding the queried complications and their probabilities. Based on the complication rates of the year 2018, 122 or more samples would be necessary to include at least one case of each adverse event. Due to the low prevalence of most complications, relatively high sample size numbers are necessary to achieve a distribution equal to that of the real population. Obviously, lower sample sizes would be necessary, when focusing only on converted cases, as adverse events show a higher prevalence in this group.

Table 1: Composition of representative inference groups for hypothetical ML applications, predicting complicated course of laparoscopic cholecystectomy based on the G-DRG data of 2018

n	Mort. (0.54%)	Conv. (4.65%)	Rev. (1.35%)	Bleed. (1.41%)	Infect. (0.63%)	Anesth. compl. (0.41%)	Accid. violation (0.41%)
122	1	6	2	2	1	1	1
366	2	17	5	5	2	2	2
610	3	28	8	9	4	3	3
1098	6	51	15	15	7	5	5
1342	7	62	18	19	8	6	6

4 Discussion

With the data obtained from our query, we were able to calculate the prevalence of conversion to open surgery during LCHE, the rate of revision surgery during the same hospital stay, and the rate of four adverse event types, as defined in the G-ICD10 system. Data samples used for training and inference seem often rather small, as there exists no generally accepted minimal sample size in the evaluation process of ML applications. Our findings reveal, that the highly standardized LCHE which is not only one of the most frequently performed surgeries in Germany, but also the most preferred model for the training and establishing of ML based algorithms, comes along with a low general average rate of complications. This raises the problem, that ML applications even with high SENS and SPEC applied to a test collective without previous filtering and densification, will achieve low PPV simply due to a low probability of actual adverse event occurrence. In our study, solely the conversion from laparoscopic to open surgery reaches responsible PPV of up to 0.8 in an unselected collective presuming the characteristics of the German population. Thus, the key to deal with low rates of adverse events may be a stepwise approach that first predicts the probability of conversion and then estimates the probability of an additive adverse event. Another strategy would be to create a representative test collective with the same characteristics as

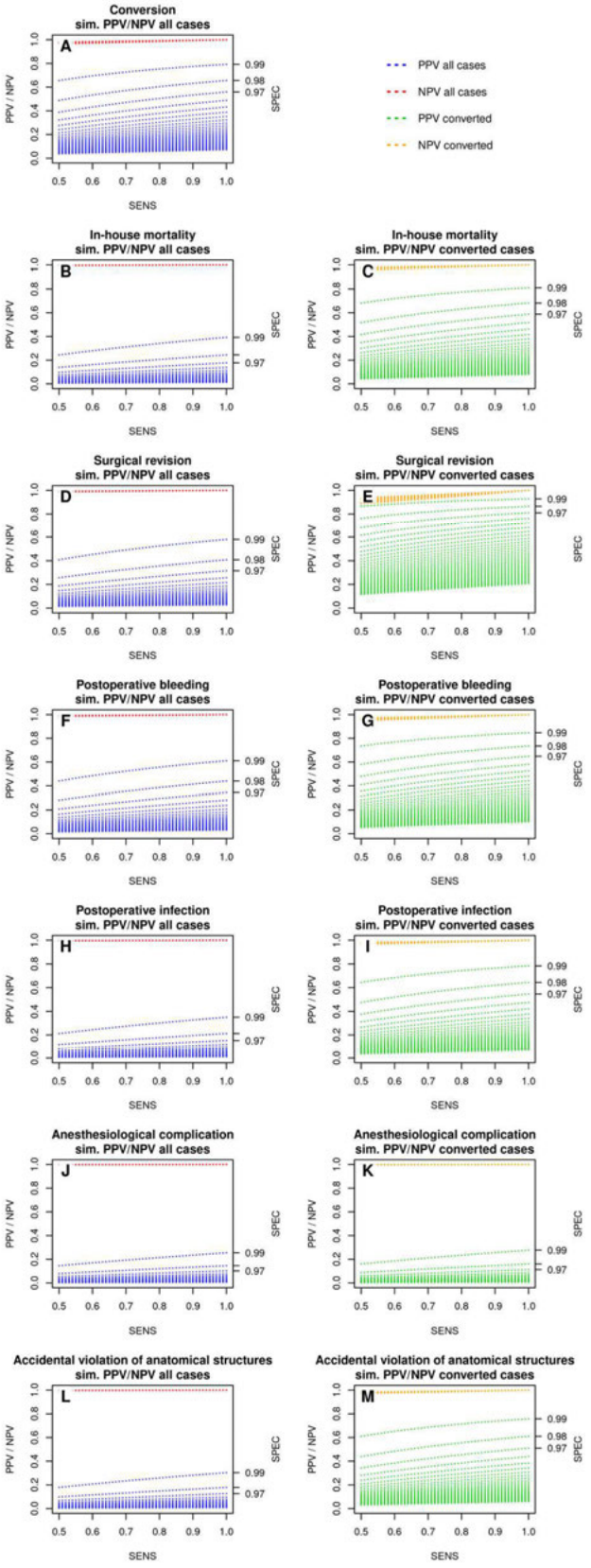

Figure 3: Simulated positive predictive values (PPV) and negative predictive values (NPV) for possible AI applications predicting the particular complications presuming each, sensitivity (SENS) and specificity (SPEC) between 0.5 and 0.99. **Left panels:** overall collective, **Right panels:** collective with necessity of conversion to open surgery, simulations except D and E are based on the data of 2018 Panels **D** and **E** base on 2017.

for instance the German population. As depicted in Table 1, such a sample would need to be by far larger than any dataset currently available for the training of ML-based applications. The effort of possible solutions to the problem of low prevalences inevitably leads to the crucial question of this article: Is LCHE really an appropriate model surgery when ML scientists intent to predict critical events? All in all, LCHE seems to represent some kind of worthwhile warm-up exercise, virtually to help ML applications to find their feet. Current challenges of ML research are still to achieve a reliable instrument and phase detection. At this stage, LCHE appears practical as a rather limited selection of instruments is needed and intraoperative steps are easy to define. [14] In addition, in surgical robotics and OR management research, LCHE undoubtedly serves as a powerful model. Nonetheless, our data reveal, that a hypothetical ML application predicting adverse events after LCHE will not impact the daily clinical routine significantly. Therefore, it is mandatory to rethink the future focus of ML research in terms of surgery. Applications achieving high detection rates under controlled circumstances inside a laboratory without clinical relevance seem rather useless. A possible approach to solve this problem could be to sweep to alternative model surgeries, which are performed in high numbers too and offer a certain degree of standardization, but are related to higher rates of adverse events. For example, laparoscopic sigmoid, pancreas or oesophageal resection could become promising models as they show significantly higher complication rates and different complication profiles. [15-17] Lessons learned from LCHE regarding phase and instrument recognition could easily be transferred to such surgeries and thus be the basis for an adverse event recognition where it is actually needed. Nevertheless, our results clearly show the necessity to create larger and more representative databases with comprehensive labelling and additive medical information to foster ML research in surgery. This can be achieved only by the creation of multicenter datasets to meet characteristics similar to that of the German overall collective for instance. Realistic clinical questions demand sample sizes of far more than 100 data sets. Moreover, a responsible synthesis of new data technology in terms of ML and fundamental principles of classical statistics is essential. Pure declaration of SENS and SPEC as exclusive quality criteria seems irresponsible as well as expected occurrence rates of adverse events within the final test collective must be considered early during study design. [18] Following this strategy, a deceptive impression of the actual usability of adverse event prediction tools becomes possible. As ML applications currently are a hot topic triggering some kind of gold rush mood, it is mandatory even to address their real clinical applicability and meaningfulness as soon as possible. [19] Hence, as a first step, the advantages and disadvantages as well as the frontiers of particular model surgeries need to be analyzed thoroughly.

5 Conclusion

LCHE finds broad use as a model for training and testing of machine learning applications. As this kind of surgery is a proper choice for basic instrument and phase recognition, we do not see its strength if it comes to the actual daily clinical use in terms of outcome and adverse event prediction. Therefore, new model surgeries with higher intrinsic complication rates must be opened up. Furthermore, the creation of comprehensive multicenter datasets are mandatory to ensure reliable and representative ML research in surgery.

Author Statement

Research funding: The authors state no funding involved. Conflict of interest: Authors state no conflict of interest.

References

[1] Anteby R, Horesh N, Soffer S, Zager Y, Barash Y, Amiel I, et al. (2021) Deep learning visual analysis in laparoscopic surgery: a systematic review and diagnostic test accuracy meta-analysis. Surg Endosc 35:1521–1533

[2] Altieri M, Hashimoto D, María Rivera A, Namazi B, Alseidi A, Okrainec A, et al. (2020) Using Artificial Intelligence to Identify Surgical Anatomy, Safe Zones of DISSection, and Dangerous Zones of Dissection during Laparoscopic Cholecystectomy. J Am Coll Surg 231:e21

[3] Ward TM, Hashimoto DA, Ban Y, Rosman G, Meireles OR (2022) Artificial intelligence prediction of cholecystectomy operative course from automated identification of gallbladder inflammation. Surg Endosc 1–9

[4] Di Martino M, Mora-Guzmán I, Jodra VV, Dehesa AS, García DM, Ruiz RC, Nisa FG-M, et al. (2021) How to Predict Postoperative Complications After Early Laparoscopic Cholecystectomy for Acute Cholecystitis: The Chole-Risk Score. J Gastrointest Surg 25:2814–2822

[5] D. Ranev and J. Teixeira, „History of Computer-Assisted Surgery“, Surgical Clinics, Bd. 100, Nr. 2, S. 209–218, Apr. 2020, doi: 10.1016/j.suc.2019.11.001.

[6] Tranter-Entwistle I, Eglinton T, Connor S, Hugh TJ (2022) Operative difficulty in laparoscopic cholecystectomy: considering the role of machine learning platforms in clinical practice. Artif Intell Surg 2:46–56

[7] Hong W-Y, Kao C-L, Kuo Y-H, Wang J-R, Chang W-L, Shih C-S (2020) CholecSeg8k: A Semantic Segmentation Dataset for Laparoscopic Cholecystectomy Based on Cholec80. ArXiv Prepr ArXiv201212453

[8] Czempiel T, Paschali M, Keicher M, Simson W, Feussner H, Kim ST, Navab N (2020) Tecno: Surgical phase recognition with multi-stage temporal convolutional networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp 343–352

[9] Beyersdorffer P, Kunert W, Jansen K, Miller J, Wilhelm P, Burgert O, et al. (2021) Detection of adverse events leading to inadvertent injury during laparoscopic cholecystectomy using convolutional neural networks. Biomed Eng Biomed Tech 66:413–421

[10] Mascagni P, Alapatt D, Urade T, Vardazaryan A, Mutter D, Marescaux J, et al. (2021) A computer vision platform to automatically locate critical events in surgical videos: documenting safety in laparoscopic cholecystectomy. Ann Surg 274:e93–e95

[11] Vecchio TJ (1966) Predictive Value of a Single Diagnostic Test in Unselected Populations. N Engl J Med 274:1171–1173

[12] Statistisches Bundesamt (Destatis) (2022) Fallpauschalenbezogene Krankenhausstatistik (DRG)

[13] R Core Team (2020) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria

[14] Hashimoto DA, Ward TM, Meireles OR (2020) The role of artificial intelligence in surgery. Adv Surg 54:89–101

[15] Ritz J-P, Reissfelder C, Holmer C, Buhr HJ (2008) [Results of sigma resection in acute complicated diverticulitis: method and time of surgical intervention]. Chir Z Alle Geb Oper Medizen 79:753–758

[16] Mendoza AS, Han H-S, Ahn S, Yoon Y-S, Cho JY, Choi Y (2016) Predictive factors associated with postoperative pancreatic fistula after laparoscopic distal pancreatectomy: a 10-year single-institution experience. Surg Endosc 30:649–656

[17] Junemann-Ramirez M, Awan MY, Khan ZM, Rahamim JS (2005) Anastomotic leakage post-esophagogastrectomy for esophageal carcinoma: retrospective analysis of predictive factors, management and influence on longterm survival in a high volume centre. Eur J Cardiothorac Surg 27:3–7

[18] Gholipour C, Fakhree MBA, Shalchi RA, Abbasi M (2009)
Prediction of conversion of laparoscopic cholecystectomy to open surgery with artificial neural networks. BMC Surg 9:1

[19] El Hechi M, Ward TM, An GC, Maurer LR, El Moheb M, Tsoulfas G, Kaafarani HM (2021) Artificial intelligence, machine learning, and surgical science: reality versus hype. J Surg Res 264:A1