AI

MDPI

Article

Feasibility of GPT-3.5 versus Machine Learning for Automated Surgical Decision-Making Determination: A Multicenter Study on Suspected Appendicitis

Sebastian Sanduleanu [1,*,†], Koray Ersahin [2,†], Johannes Bremm [3], Narmin Talibova [4], Tim Damer [2], Merve Erdogan [5], Jonathan Kottlors [3], Lukas Goertz [3], Christiane Bruns [6,7], David Maintz [3] and Nuran Abdullayev [5]

1 Department of Emergency Medicine, Vogelsbeek 5, 6001 BE Weert, The Netherlands
2 Department of General and Visceral Surgery, GFO Clinics Troisdorf, Academic Hospital of the Friedrich-Wilhelms-University Bonn, 50937 Troisdorf, Germany
3 Institute for Diagnostic and Interventional Radiology, Faculty of Medicine and University Hospital Cologne, University of Cologne, 50937 Cologne, Germany
4 Department of Internal Medicine III, University Hospital, 89081 Ulm, Germany
5 Department of Radiology and Neuroradiology, GFO Clinics Troisdorf, Academic Hospital of the Friedrich-Wilhelms-University Bonn, 53840 Troisdorf, Germany
6 Department of General, Visceral, Tumor and Transplantation Surgery, University Hospital of Cologne, Kerpener Straße 62, 50937 Cologne, Germany
7 Center for Integrated Oncology (CIO) Aachen, Bonn, Cologne and Düsseldorf, 50937 Cologne, Germany
* Correspondence: sebastian.sanduleanu1990@gmail.com
† These authors contributed equally to this work.

Abstract: Background: Nonsurgical treatment of uncomplicated appendicitis is a reasonable option in many cases despite the sparsity of robust, easy access, externally validated, and multimodally informed clinical decision support systems (CDSSs). Developed by OpenAI, the Generative Pre-trained Transformer 3.5 model (GPT-3) may provide enhanced decision support for surgeons in less certain appendicitis cases or those posing a higher risk for (relative) operative contra-indications. Our objective was to determine whether GPT-3.5, when provided high-throughput clinical, laboratory, and radiological text-based information, will come to clinical decisions similar to those of a machine learning model and a board-certified surgeon (reference standard) in decision-making for appendectomy versus conservative treatment. Methods: In this cohort study, we randomly collected patients presenting at the emergency department (ED) of two German hospitals (GFO, Troisdorf, and University Hospital Cologne) with right abdominal pain between October 2022 and October 2023. Statistical analysis was performed using R, version 3.6.2, on RStudio, version 2023.03.0 + 386. Overall agreement between the GPT-3.5 output and the reference standard was assessed by means of inter-observer kappa values as well as accuracy, sensitivity, specificity, and positive and negative predictive values with the "Caret" and "irr" packages. Statistical significance was defined as $p < 0.05$. Results: There was agreement between the surgeon's decision and GPT-3.5 in 102 of 113 cases, and all cases where the surgeon decided upon conservative treatment were correctly classified by GPT-3.5. The estimated model training accuracy was 83.3% (95% CI: 74.0, 90.4), while the validation accuracy for the model was 87.0% (95% CI: 66.4, 97.2). This is in comparison to the GPT-3.5 accuracy of 90.3% (95% CI: 83.2, 95.0), which did not perform significantly better in comparison to the machine learning model ($p = 0.21$). Conclusions: This study, the first study of the "intended use" of GPT-3.5 for surgical treatment to our knowledge, comparing surgical decision-making versus an algorithm found a high degree of agreement between board-certified surgeons and GPT-3.5 for surgical decision-making in patients presenting to the emergency department with lower abdominal pain.

Keywords: appendectomy; artificial intelligence; surgical decision-making

check for updates

Citation: Sanduleanu, S.; Ersahin, K.; Bremm, J.; Talibova, N.; Damer, T.; Erdogan, M.; Kottlors, J.; Goertz, L.; Bruns, C.; Maintz, D.; et al. Feasibility of GPT-3.5 versus Machine Learning for Automated Surgical Decision-Making Determination: A Multicenter Study on Suspected Appendicitis. *AI* **2024**, *5*, 1942–1954. https://doi.org/10.3390/ai5040096

Academic Editors: Zhongheng Zhang, Yucai Hong and Wei Shao

Received: 20 August 2024
Revised: 2 October 2024
Accepted: 14 October 2024
Published: 16 October 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Acute appendicitis (AA) is among the most common causes of lower abdominal pain leading to emergency department visits and often to urgent abdominal surgery [1]. As many as 95% of patients with uncomplicated acute appendicitis eventually undergo surgical treatment [2].

The incidence of acute appendicitis (AA) has shown a steady decline worldwide since the late 1940s. In developed nations, the occurrence of AA ranges from 5.7 to 50 cases per 100,000 inhabitants annually, with the highest incidence observed in individuals between the ages of 10 and 30 years old [3,4]. Regional differences play a significant role in the lifetime risk of developing AA, with reported rates of 9% in the United States, 8% in Europe, and much lower than 2% in Africa [5]. Furthermore, there is considerable variation in the clinical presentation of AA upon presentation at the doctor, the severity of the disease and the time it takes from the first onset of symptoms to the acute phase, the approach to radiological diagnosis, and the surgical management of patients, which is influenced by the economic status of the country, among other factors [6].

The rate of appendiceal perforation, a serious complication of AA, varies widely, ranging from 16% to 40%. This complication is more frequently seen in younger patients, with perforation rates between 40% and 57%, and in those over 50 years of age, where rates range from 55% to 70% [7]. Appendiceal perforation due to, e.g., delayed presentation is particularly concerning as it is linked to significantly higher morbidity and mortality compared to nonperforated cases of AA.

In one cohort [8], perforation was found in 13.8% of the cases of acute appendicitis and presented mostly in the age group of 21–30 years. Patients presented with abdominal pain in 100% of cases, followed by vomiting (64.3%) and fever (38.9%). Patients with perforated appendicitis had a very high (72.2%) complication rate (mostly intestinal obstruction, intra-abdominal abscess, and incisional hernia). The mortality rate in this cohort with perforated appendicitis was 4.8%.

An intra-abdominal abscess (IAA) is another potentially severe complication occurring in 3% to 25% of patients following appendectomy [9], the risk being the highest following complicated appendicitis. Risk factors for developing postoperative IAA remain controversial and poorly defined with no evidence for differences between open and laparoscopic surgery or between aspiration and peritoneal lavage [9,10].

The clinical diagnosis of AA is often challenging and involves a combination of clinical (e.g., physical examination findings such as a positive psoas, Rovsign sign, and McBurney sign that may indicate peritonitis), age, vital signs such as temperature and blood pressure, laboratory findings (e.g., CRP, leucocytes), and radiological findings (ultrasound as well as computed tomography, depending on patient constitution and clinician's preference) [11]. In the emergency department, when a patient is suspected of having appendicitis, a thorough workup is essential to make an accurate diagnosis and determine the appropriate treatment plan. As mentioned, time is of the essence as appendiceal perforation is associated with a high complication rate.

Appendectomy has been the standard treatment for appendicitis for a long time, even though the successful use of antibiotic therapy as an alternative was reported as early as 65 years ago [12].

Evidence for antibiotics-first treatment has had renewed interest, with several randomized controlled trials concluding that a majority of patients with acute, uncomplicated (nonperforated) appendicitis (AUA) can be treated safely with an antibiotics-first strategy (conservatively), with rescue appendectomy if indicated [13–18].

With the recent worldwide coronavirus pandemic (COVID-19), health systems and professional societies, e.g., the American College of Surgeons [16], have proposed reconsideration of many aspects of care delivery, including the role of antibiotics in the treatment of appendicitis without signs indicative of high risk for perforation, in individuals unfit for surgery (e.g., immunosuppressed patients) or those having concerns to undergo operation (choice to be made through shared decision-making between patient and clinician).

The ultimate decision between explorative laparoscopy/appendectomy and conservative treatment should be made on a case-by-case basis, and while simple and user-friendly scoring systems such as the Alvarado score have been used by clinicians as a structured algorithm to aid in predicting the risk stratum of AA [1], such scoring systems are often unreliable, confusing, and not widely adopted by clinicians. In this light, algorithms that rely on high-throughput real-world data may be of current interest.

In recent years, the field of artificial intelligence (AI) has witnessed remarkable advancements, for instance, in the field of natural language processing (NLP), with the most prominent applications including chatbots, text classification, speech recognition, language translation, and the generation or summarization of texts.

In 2017, Vaswani et al. [19] introduced the Transformer deep learning model architecture, replacing previously widely used recurrent neural networks (RNNs) [20], deep learning models that are trained to process and convert a sequential data input into a specific sequential data output.

Transformers, characterized by their feedforward networks and specialized attention blocks, represent a significant advancement in neural network architecture, particularly in overcoming the limitations of recurrent neural networks (RNNs). Unlike RNNs, where each computation step depends on the previous one, Transformers can process input sequences in parallel, significantly improving computational efficiency. Additionally, the attention blocks within Transformers enable the model to learn long-term dependencies by selectively focusing on different segments of the input data [21]. A basic Transformer network comprises an encoder and a decoder stack, each consisting of several identical feed-forward neural blocks [19]. The encoder processes an input sequence to produce a set of context vectors, which are then used by the decoder to generate an output sequence. In the case of a Transformer, both the input and output are text sequences, where the words are tokenized (broken down into smaller units called tokens) and represented as elements in a high-dimensional vector [21].

Large language models (LLMs) are large Transformer models trained on extensive datasets [21].

ChatGPT-3.5 (Generative Pre-trained Transformer), an LLM, is one of the NLP architectures developed by OpenAI to output an AI chatbot that has been pre-trained on online journals, Wikipedia, and books [22]. It is a so-called large language model (LLM) that uses deep learning techniques to achieve general-purpose language understanding and generation that has gained widespread attention for its ability to generate human-like text based on a given input. The technology has shown promise in various applications, including language translation, content generation, and summarization.

One of the primary challenges in the management of hospital medical records is the need to maintain the accuracy and consistency of information. Healthcare providers must be able to quickly access and update patient records, ensuring that the data is both accurate and up-to-date. GPTs can assist in this process by automatically generating summaries of medical records, allowing healthcare professionals to quickly review and update the information as needed. Moreover, GPTs can be utilized to improve the interoperability of medical records. As healthcare systems become more interconnected, the need for seamless data exchange between different providers and institutions becomes crucial. GPTs can help bridge the gap between disparate electronic health record systems by translating medical records into a standardized format, facilitating smoother data exchange and reducing the risk of miscommunication.

Clinical decision support systems (DSSs), continuously learning artificial intelligence platforms, can integrate all available data—clinical, imaging, biologic, genetic, and validated predictive models—and may help doctors by providing patient-specific recommendations. GPTs may be able to assist by interpreting these recommendations, explaining the rationale behind them, and answering related clinical questions, thereby enhancing the decision-making process.

There are several promising results in the current literature as of August 2024 with the use of GPTs in the high-data-throughput environment of a radiology department, for instance, in helping the radiologist with choosing the appropriate radiologic study and scanning protocol, with adequate differential diagnosis, and potentially even with automated reporting [23–27]. ChatGPT nevertheless often faces criticism for its inaccuracies, limited functionality, lack of transparency in citation sources, and the need for thorough verification by the end-user. These limitations pose several potential risks, including plagiarism, hallucinations (where the model fabricates or misrepresents information), academic misconduct, and various other ethical concerns [28–30]. Therefore, ChatGPT is in our opinion better suited as a supplementary tool in the medical field rather than a primary information resource as errors in the information generated by ChatGPT could have serious implications for an individual's health. In our opinion, research should be focused on providing the algorithm with abundant real-world data, providing the algorithm with proper context, and seeing how it performs in comparison to individual healthcare domain experts.

Our hypothesis in this study is that GPT-3.5 as well as a machine learning model, when provided high-throughput clinical, laboratory, and radiological text-based information, will come to clinical decisions similar to those of a board-certified surgeon on the requirement of explorative laparoscopic investigation/appendectomy or conservative treatment in patients presenting with acute abdominal pain at the emergency department.

2. Materials and Methods

This study received ethical approval (file number 23–1061-retro) from the Institutional Review Board (IRB) of GFO Kliniken Troisdorf, and informed consent was waived due to the retrospective design of this study. No patient-identifying information was provided to the artificial intelligence.

2.1. Workflow

We randomly collected $n = 63$ consecutive histopathologically confirmed appendicitis patients and $n = 50$ control patients presenting with right abdominal pain at the emergency department of two German hospitals (GFO, Troisdorf, and University Hospital Cologne) between October 2022 and October 2023.

For both groups, the following exclusion criteria were applied: (a) incomplete vital signs upon admission to the emergency department (temperature, blood pressure, and respiratory rate); (b) missing physical examination findings; (c) missing CRP and leucocyte count; (d) missing ultrasound examination findings for the surgically confirmed appendicitis cases that did not undergo an abdominal CT examination; (e) patient having contra-indications for surgery (e.g., inability to tolerate general anesthesia).

Physical examination signs taken into account were as follows [11]: (a) McBurney sign (maximum pain in the middle of the imaginary connecting Monro line between the navel and the anterior superior dextra iliac spine); (b) Blumberg sign (contralateral release pain, e.g., pain on the right when releasing the compressed abdominal wall in the left lower abdomen); (c) right lower quadrant release pain; (d) Rovsign sign (pain in the right lower abdomen when extending the colon against the cecal pole); (e) Psoas sign (pain in the right lower abdomen when lifting the straight right leg against resistance).

Based on each patient's clinical, laboratory, and radiological findings (full reports), GPT-3.5 was accessed via ChatGPT (https://chat.openai.com/) (accessed on 24 October 2023) and asked to determine the optimal course of treatment, namely laparoscopic exploration/appendectomy or conservative treatment with antibiotics using zero-shot prompting, using same dialogue box for each case to potentially enhance the context awareness of the model. The choice of GPT-3.5 instead of GPT-4 was due to the temporary unavailability of GPT-4 at the time of prompting.

Additionally, a random-forest-based machine learning classifier was trained and validated to determine the optimal course of treatment based on the same information that was provided to GPT-3.5, albeit in a more structured data format.

An example of the prompt provided to GPT-3.5 is provided in Appendix A.

It is important to mention that in all cases where GPT-3.5 did not provide a clear-cut answer, it was prompted to give its best guess estimate based on the provided information.

The results were compared with an expert decision determined by 6 board-certified surgeons with at least 2 years of experience, which was defined as the reference standard.

Figure 1 shows the study flowchart.

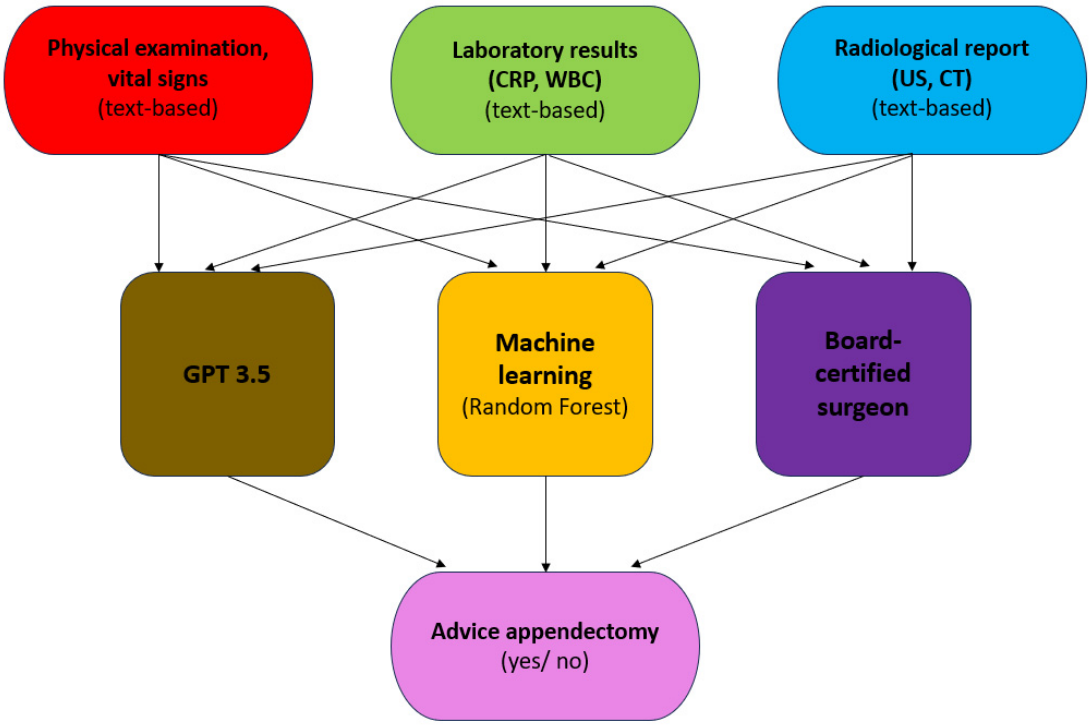


Figure 1. Study workflow.

2.2. Statistical Analysis

Statistical analysis was performed using R, version 3.6.2, on RStudio, version 2023.03.0 + 386 (https://cran.r-project.org/) (accessed on 12 November 2023). Overall agreement between the GPT-3.5 output and the reference standard was assessed by means of inter-observer kappa values as well as accuracy, sensitivity, specificity, and positive and negative predictive values with the "Caret" and "irr" packages.

Statistical significance was defined as $p < 0.05$.

2.3. Machine Learning Model Development

A random forest (RF) machine learning classifier was computed (*default settings:* 500 trees, mtry = $\sqrt{}$nr. of predictors, without internal cross-validation) and validated in an external validation cohort taking into account variables such as "age", "physical examination", "breathing rate", "systolic/diastolic blood pressure", "temperature", "CRP", "leucocyte count", "ultrasound findings", and "CT findings" indicative of appendicitis upon admission at the emergency department.

The "randomForest" package, which implements Breiman's random forest algorithm (based on Breiman and Cutler's original Fortran code) for both classification and regression tasks, was used.

The "predict" function was used to predict the label of a new set of data from the given trained model, while the "roc" function (pROC package v. 1.18.5) was used to

build an ROC curve and return a "roc" object. McNemar's Test was used to compare the predictive accuracy of the machine learning model versus the GPT-3.5 output (based on the correct/false classification according to the decision made by the board-certified surgeon.

3. Results

In total, n = 113 patients (n = 63 appendicitis patients confirmed by histopathology and n = 50 control patients presenting with lower abdominal pain) were included in the analysis across independent patient cohorts from two German hospitals (University Hospital Cologne and GFO Kliniken Troisdorf). Macroscopically mild, moderate, and severely inflamed appendix cases were included in the analysis.

In the first cohort from GFO Kliniken Troisdorf (n = 100), a total of n = 50 appendicitis patients confirmed by histopathology and n = 50 control patients presenting with lower abdominal pain were included (median age 35 y, 57% female). Upon admission to the emergency department, an ultrasound examination was performed for all patients, while for 29% of the patients, a CT examination was performed.

On average, 1.12 signs indicative of appendicitis were found upon physical examination (Psoas sign, Rovsign sign, McBurney/Lanz, release pain, etc.) in the appendicitis-confirmed group, while in the control group, only 0.24 physical examination signs were found on average.

The average temperature upon admission was 36.8 °C in the appendicitis-confirmed cases and 36.6 °C in the control group. The average CRP and leucocyte values were 5.85 mg/dL and 12.82/µL, respectively, in the appendicitis group and 1.19 mg/dL and 8.14/µL, respectively, in the control group.

In the second cohort from Cologne (n = 13), a total of n = 13 appendicitis patients confirmed by histopathology were included (median age 22 y, 38% female).

On average, 1.31 signs indicative of appendicitis were found upon physical examination (Psoas sign, Rovsign sign, McBurney/Lanz, release pain, etc.).

The average temperature upon admission was 36.5 °C, while the average CRP and leucocyte values were 3.51 mg/dL and 13.43/µL, respectively.

There was an agreement between the reference standard (expert decision—appendicitis confirmed by histopathology) and GPT-3.5 in 102 of 113 cases (accuracy 90.3%; 95% CI: 83.2, 95.0), with an inter-observer Cohen's kappa of 0.81 (CI: 0.70, 0.91).

All cases where the surgeon decided upon conservative treatment were correctly classified by GPT-3.5. With a specificity of 100%, a positive GPT-3.5 result tends to rule in all patients that require surgery according to the surgeon, while the sensitivity of GPT-3.5 with respect to reference standard was 83%.

Table 1 presents the individual patient characteristics per hospital cohort, while Figure 2 depicts a confusion matrix comparison constituting both cohorts between the specialist (board-certified surgeon) decision and GPT-3.5 decision on (explorative) appendectomy or conservative treatment.

Figure 3 presents training and validation ROC curves obtained by machine learning with a random forest model. The training cohort (n = 90) consisted of n = 50 appendicitis-confirmed cases and n = 40 controls from GFO Troisdorf, while the validation cohort (n = 23) consisted of all n = 13 appendicitis-confirmed cases from Cologne and n = 10 remaining controls from GFO Troisdorf.

The random forest model reached an AUC of 0.89 (CI: 0.81, 0.96) in the training cohort and an AUC of 0.91 (CI: 0.78, 1.0) in the validation cohort.

The estimated machine learning model training accuracy was 83.3% (95% CI: 74.0, 90.4), while the validation accuracy for the model was 87.0% (95% CI: 66.4, 97.2). This is in comparison to the GPT-3.5 accuracy of 90.3% (95% CI: 83.2, 95.0), which did not perform significantly better in comparison to the machine learning model (McNemar p = 0.21).

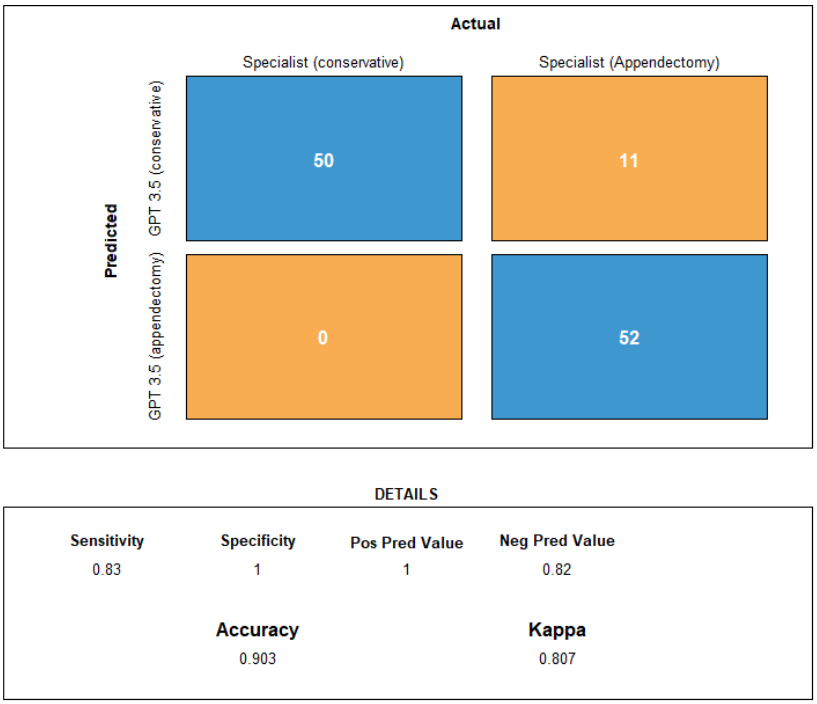


Figure 2. Confusion matrix constituting both the GFO Troisdorf and the Cologne cohort (n = 113). Comparisons are made between specialist (board-certified surgeon) decision and GPT-3.5 decision on (explorative) appendectomy or conservative treatment.

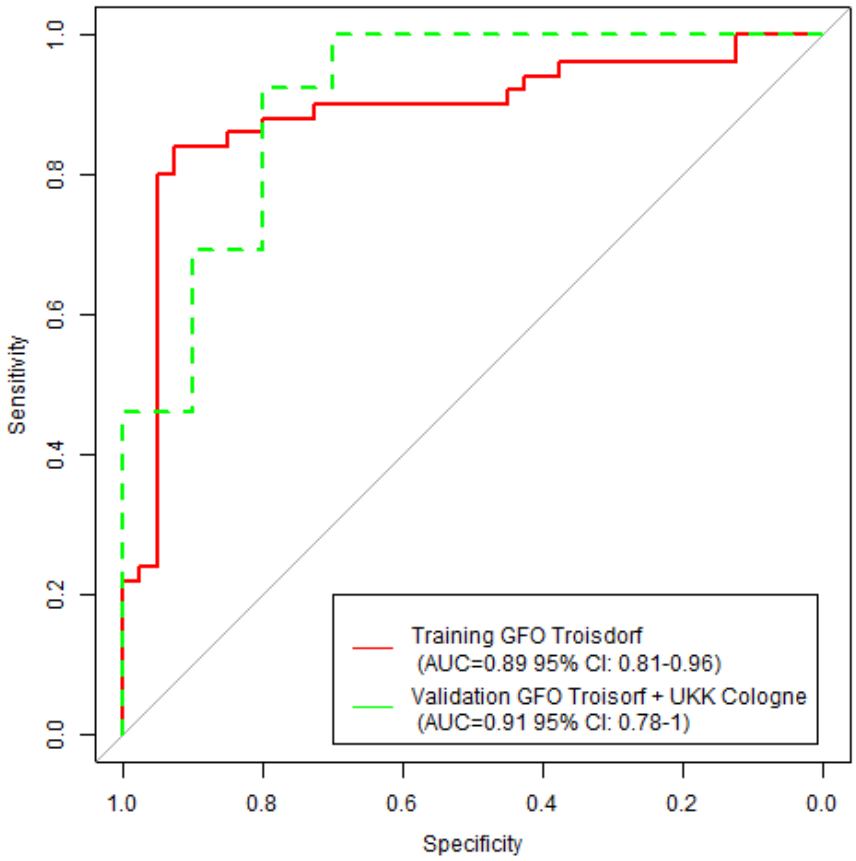


Figure 3. Training (red) and validation (green) random forest machine learning ROC curves including area-under-the-curve metrics and confidence intervals.

Table 1. Patient characteristics per hospital cohort.

	GFO Troisdorf (*n* = 100)	Cologne (*n* = 13)
Board-certified specialist decision		
Appendectomy (*n*)	50	13
Conservative (*n*)	50	0
Median age (years)	35	22
Gender		
Male (*n*)	43	8
Female (*n*)	57	5
Imaging upon ER admission		
Ultrasound (%)	100	100
Computed tomography (%)	29	31

4. Discussion

This multicenter study found a high degree of agreement between board-certified surgeons and GPT-3.5 in the clinical-, laboratory-, and radiological-parameter-informed decision for laparoscopic explorative surgery/appendectomy versus conservative treatment in patients presenting at the emergency department with lower abdominal pain.

Several medical studies were performed previously prompting GPT-3.5/4 to evaluate its performance in selecting correct imaging studies and protocols based on medical history and corresponding clinical questions extracted from Radiology Request Forms (RRFs) [24], determining top differential diagnoses based on imaging patterns [25], generating accurate differential diagnoses in undifferentiated patients based on physician notes recorded at initial ED presentation [26], and acting as a chatbot-based symptom checker [27].

In the emergency department, another study [28] conducted an analysis to evaluate the effectiveness of ChatGPT in assisting healthcare providers with triage decisions for patients with metastatic prostate cancer in the emergency room. ChatGPT was found to have a high sensitivity of 95.7% in correctly identifying patients who needed to be admitted to the hospital. However, its specificity was much lower, at 18.2%, in identifying patients who could be safely discharged. Despite the low specificity, the authors concluded that ChatGPT's high sensitivity indicates a strong ability to correctly identify patients requiring admission, accurately diagnose conditions, and offer additional treatment recommendations. As a result, the study suggests that ChatGPT could potentially improve patient classification, leading to more efficient and higher-quality care in emergency settings.

In the field of general surgery, a recent study [29] compared ChatGPT-4 with junior and senior residents as well as attendings in identifying the correct operation to perform and recommending additional workup for postoperative complications in five clinical scenarios. Each clinical scenario was run through ChatGPT-4 and sent electronically to all general surgery residents and attendings at a single institution. The authors found that GPT-4 was significantly better than junior residents ($p = 0.009$) but was not significantly different from senior residents or attendings.

Another study [30] evaluated the performance of ChatGPT-4 on surgical questions, finding a near- or above-human-level performance. Performance was evaluated on the Surgical Council on Resident Education question bank and a second commonly used surgical knowledge assessment. This study revealed that the GPT model correctly answered 71.3% and 67.9% of multiple choice and 47.9% and 66.1% of open-ended questions for the Surgical Council on Resident Education question bank and the second surgical knowledge assessment, respectively. Common reasons for incorrect responses by the model included inaccurate information in a complex question ($n = 16$, 36.4%), inaccurate information in a fact-based question ($n = 11$, 25.0%), and accurate information with circumstantial discrepancy ($n = 6$, 13.6%). The study highlights the need for further refinement of large

language models to ensure safe and consistent application in healthcare settings. Despite its strong performance, the suitability of ChatGPT for assisting clinicians remains uncertain. A significant aspect of the ChatGPT model's development is that its training primarily depends on general medical knowledge that is widely available on the internet. This approach is necessitated by the difficulty of integrating large datasets of patient-specific information into the model's training process. The challenge arises from the stringent requirements to protect patient privacy and adhere to ethical standards, which limit access to detailed, real-world clinical data. As a result, ChatGPT's responses to medical queries may lack the depth and specificity that come from direct exposure to extensive patient data. This reliance on publicly available information introduces a degree of non-scientific specificity to the model's medical-related outputs. Consequently, while ChatGPT can provide general guidance and information, it may not always offer the precise or nuanced insights that are crucial in clinical decision-making, underscoring the importance of human oversight and verification when using the tool in a healthcare context.

In light of this current understanding, we have attempted to provide GPT with highly structured and comprehensive real-world patient data. Several findings are noteworthy in our own current study.

For instance, the relatively high AUC values in the machine learning validation cohort (higher than the training AUC) indicate that the machine learning model is generalizable and not likely to overfit.

In our cohort, GPT-3.5 outperforms machine learning in terms of accuracy, highlighting the possibility that when provided with full-text data on relevant clinical findings such as physical examination and medical imaging with specific prompts, it might be able to better understand the context and generate more relevant responses in comparison to the more traditional machine learning models.

On the other hand, machine learning, albeit being more time-consuming to train, offers a clearer insight into feature importance, making it easier to understand which variables contribute more to the predictions of the model and which features do less.

The results from the machine learning part of the analysis are in line with previous findings in the literature on the detection of individuals with acute appendicitis [31–34].

To our knowledge, this is the first study of the "intended use" for surgical treatment decisions in the literature that compares the decision-making of board-certified surgeons versus the GPT algorithm and machine learning based on comprehensive clinical, biochemical, and radiological information.

Certainly, there are a few limitations to our current study. The limitations of this study include the following: (1) The output of GPT-3.5 is not always straightforward, but is rather a piece of advice or recommendation to consult an external source of data. We have noticed that to achieve more precise responses, it is very important to prompt GPT-3.5 to provide the user with a resolute answer, in other words, to make a decision despite the uncertainties based on the data that were provided to the algorithm. (2) Another limitation is related to inherent biases, the inaccurate results of the LLM algorithm, and the inability of the current GPT-3.5 version to differentiate between reliable and unreliable sources. GPT-3.5 is only trained on content up to September 2021 from a limited number of online sources, which limits its accuracy on queries related to more recent events. GPT-4 is trained on data up through April 2023 or December 2023 (depending on the model version) and can browse the internet if it is prompted to do so. (3) There are significant legal, technological, and ethical concerns surrounding the use of ChatGPT in healthcare decision-making in general [35–39]. Improper utilization of this technology could lead to violations of copyright laws, health regulations, and other legal frameworks. For instance, text generated by ChatGPT may include instances of plagiarism and can contribute to the creation of hallucinations, as previously mentioned, content produced by the model that is not grounded in reality, often fabricating narratives or data. These issues may arise due to biases in the training data, insufficient information, a limited understanding of real-world contexts, or other inherent algorithmic limitations. It is important to further recognize that

ChatGPT is unable to discern the significance of information and can only replicate existing research, lacking the capability to generate novel insights like human scientists. Therefore, a thorough investigation into the ethical implications of ChatGPT is necessary, and there is a pressing need to establish global ethical standards for its use [36], particularly as a medical chatbot, on an international scale.

While GPT-3.5′s role in the decision to perform an appendectomy should in our opinion be as a decision support tool rather than a replacement for clinical judgment, it has the potential to streamline the decision-making process, improve patient outcomes, and reduce the risk of unnecessary surgeries. We acknowledge that decision-making for appendectomy encompasses surgical judgment alongside patient preference. In cases where fast decisions must be made under time pressure and uncertainty (i.e., high risks for surgical complications, lack of patient cooperation), GPT-3.5 and later versions can in our opinion be a valuable aid in the decision-making process.

As with any medical application of AI, it is important to use GPT-3.5 and GPT-4 in conjunction with the expertise of trained healthcare professionals who can make the final decisions based on both the AI's guidance and their clinical judgment [40].

In our opinion, this study merely serves as a proof of concept, and clinical adoption possibilities of the proposed approach to use GPT-3.5 as well as more commonly used supervised machine learning algorithms as a clinical decision support system (CDSS) are still subject to regulatory review and approval (although the FDA and international regulatory authorities have already issued initial guideline documents for the development and approval of tools based on machine learning (ML)/artificial intelligence (AI)) [41]. Such a clinical decision support tool, if used in a routine clinical setting in the EU, would very likely require certification as a Class II medical device under the MDR—Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC [42].

The Internet of Things (IoT), referring to devices equipped with sensors, processing capabilities, software, and various technologies that communicate and share data with other devices and systems via the internet, has made a breakthrough within the surgical practice. One literature review [43] revealed that telesurgical networks are routinely incorporated in many surgical centers and may encompass complex AI machine learning applications that aid in medical decision-making, such as ChatGPT. The IoT may play a role in suspected appendicitis in patient monitoring (monitoring patients' vital signs during surgery and recovery, allowing for continuous assessment and quicker response to complications), data-driven decision-making (tracking patient recovery more efficiently and developing personalized treatment plans), and providing real-time intra-operative feedback for the surgeon by means of IoT-enabled instruments.

With the advent of newer versions such as GPT-4 that are pre-trained on ever larger amounts of information and that can accept images as input and pull text from web pages when you share a URL in the prompt, but also grant the user the possibility to provide the LLM with additional domain-specific and unbiased information (e.g., retrieval augmented generation (RAG) and fine-tuning), such tools hold the potential to improve clinical workflows, resource allocation, and cost-effectiveness.

Author Contributions: Conceptualization, S.S., K.E., J.K. and N.A.; methodology, S.S., K.E., J.K. and N.A.; software, S.S.; validation, S.S.; formal analysis, S.S.; investigation, S.S.; resources K.E., N.T., C.B., D.M. and N.A.; data curation, K.E., T.D., M.E. and N.A.; writing—original draft preparation, S.S., J.K., L.G. and N.A.; writing—review and editing, S.S., K.E., J.K. and L.G.; visualization, S.S.; supervision, N.A. and K.E.; project administration, K.E., J.B., N.T., C.B., D.M. and N.A. All authors have read and agreed to the published version of the manuscript.

Funding: This study received partial funding from NUM 2 (Netzwerk Universitätsmedizin, Berlin, Germany) (FKZ: 01KX2121).

Institutional Review Board Statement: This study was conducted in accordance with the Declaration of Helsinki and was approved by the Institutional Review Board (IRB) of GFO Kliniken Troisdorf (file number 23–1061-retro).

Informed Consent Statement: Informed consent was waived due to the retrospective nature of this study, and no patient-identifying information was provided to the artificial intelligence.

Data Availability Statement: The data that support the findings of this study are available on request from the corresponding author, S.S.

Conflicts of Interest: The authors of this manuscript declare no relationships with any companies whose products or services may be related to the subject matter of the article.

Appendix A

Example: The GPT-3.5 prompt was as follows:

"A (XX) year old XX presents on an Emergency Department of a hospital with:

- Physical examination: . . . .
- Temperature: . . . . °C
- Tension . . . . mmHg
- Respiratory rate . . . . /min
- CRP . . . . mg/dL
- Leucocytes . . . . /μL
- (XX) contra-indications for surgery
- Ultrasound examination findings: . . . .
- CT examination: . . . .

Based on this information alone could you advise whether the patient will have to undergo laparoscopic exploration/appendectomy or whether conservative treatment with/without antibiotics is rather warranted?".

References

1. Di Saverio, S.; Podda, M.; De Simone, B.; Ceresoli, M.; Augustin, G.; Gori, A.; Boermeester, M.; Sartelli, M.; Coccolini, F.; Tarasconi, A.; et al. Diagnosis and treatment of acute appendicitis: 2020 update of the WSES Jerusalem guidelines. *World J. Emerg. Surg.* **2020**, *15*, 27. [CrossRef] [PubMed]
2. Sceats, L.A.; Trickey, A.W.; Morris, A.M.; Kin, C.; Staudenmayer, K.L. Nonoperative management of uncomplicated appendicitis among privately insured patients. *JAMA Surg.* **2019**, *154*, 141–149. [CrossRef] [PubMed]
3. Ilves, I. Seasonal variations of acute appendicitis and nonspecific abdominal pain in Finland. *World J. Gastroenterol.* **2014**, *20*, 4037–4042. [CrossRef]
4. Viniol, A.; Keunecke, C.; Biroga, T.; Stadje, R.; Dornieden, K.; Bösner, S.; Donner-Banzhoff, N.; Haasenritter, J.; Becker, A. Studies of the symptom abdominal pain—A systematic review and meta-analysis. *Fam. Pract.* **2014**, *31*, 517–529. [CrossRef]
5. Bhangu, A.; Søreide, K.; Di Saverio, S.; Assarsson, J.H.; Drake, F.T. Acute appendicitis: Modern understanding of pathogenesis, diagnosis, and management. *Lancet* **2015**, *386*, 1278–1287. [CrossRef]
6. Gomes, C.A.; Abu-Zidan, F.M.; Sartelli, M.; Coccolini, F.; Ansaloni, L.; Baiocchi, G.L.; Kluger, Y.; Di Saverio, S.; Catena, F. Management of Appendicitis Globally Based on Income of Countries (MAGIC) Study. *World J. Surg.* **2018**, *42*, 3903–3910. [CrossRef]
7. Livingston, E.H.; Woodward, W.A.; Sarosi, G.A.; Haley, R.W. Disconnect between incidence of nonperforated and perforated appendicitis: Implications for pathophysiology and management. *Ann. Surg.* **2007**, *245*, 886–892. [CrossRef]
8. Potey, K.M.; Kandi, A.A.P.; Jadhav, S.; Gowda, V.M. Study of outcomes of perforated appendicitis in adults: A prospective cohort study. *Ann. Med. Surg.* **2023**, *85*, 694–700. [CrossRef] [PubMed] [PubMed Central]
9. Mulita, F.; Plachouri, K.-M.; Liolis, E.; Kehagias, D.; Kehagias, I. Comparison of intra-abdominal abscess formation after laparoscopic and open appendectomy for complicated and uncomplicated appendicitis: A retrospective study. *Videosurg. Other Miniinvasive Tech.* **2021**, *16*, 560–565. [CrossRef]

10. Burini, G.; Cianci, M.C.; Coccetta, M.; Spizzirri, A.; Di Saverio, S.; Coletta, R.; Sapienza, P.; Mingoli, A.; Cirocchi, R.; Morabito, A. Aspiration versus peritoneal lavage in appendicitis: A meta-analysis. *World J. Emerg. Surg.* **2021**, *16*, 44. [CrossRef]
11. Moris, D.; Paulson, E.K.; Pappas, T.N. Diagnosis and Management of Acute Appendicitis in Adults. *JAMA* **2021**, *326*, 2299–2311. [CrossRef] [PubMed]
12. Ehlers, A.P.; Talan, D.A.; Moran, G.J.; Flum, D.R.; Davidson, G.H. Evidence for an Antibiotics-First Strategy for Uncomplicated Appendicitis in Adults: A Systematic Review and Gap Analysis. *J. Am. Coll. Surg.* **2016**, *222*, 309–314. [CrossRef] [PubMed] [PubMed Central]
13. Eriksson, S.; Granström, L. Randomized controlled trial of appendicectomy versus antibiotic therapy for acute appendicitis. *Br. J. Surg.* **1995**, *82*, 166–169. [CrossRef] [PubMed]
14. Styrud, J.; Eriksson, S.; Nilsson, I.; Ahlberg, G.; Haapaniemi, S.; Neovius, G.; Rex, L.; Badume, I.; Granström, L. Appendectomy versus antibiotic treatment in acute appendicitis. a prospective multi-center randomized controlled trial. *World J. Surg.* **2006**, *30*, 1033–1037. [CrossRef]
15. Turhan, A.N.; Kapan, S.; Kütükçü, E.; Yiğitbaş, H.; Hatipoğlu, S.; Aygün, E. Comparison of operative and non operative management of acute appendicitis. *Turk. J. Trauma Emerg. Surg.* **2009**, *15*, 459–462.
16. Hansson, J.; Körner, U.; Khorram-Manesh, A.; Solberg, A.; Lundholm, K. Randomized clinical trial of antibiotic therapy versus appendicectomy as primary treatment of acute appendicitis in unselected patients. *Br. J. Surg.* **2009**, *96*, 473–481. [CrossRef]
17. Vons, C.; Barry, C.; Maitre, S.; Pautrat, K.; Leconte, M.; Costaglioli, B.; Karoui, M.; Alves, A.; Dousset, B.; Valleur, P.; et al. Amoxicillin plus clavulanic acid versus appendicectomy for treatment of acute uncomplicated appendicitis: An open-label, non-inferiority, randomised controlled trial. *Lancet* **2011**, *377*, 1573–1579. [CrossRef]
18. CODA Collaborative. A Randomized Trial Comparing Antibiotics with Appendectomy for Appendicitis (CODA). *N. Engl. J. Med.* **2020**, *383*, 1907–1919. [CrossRef]
19. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008. [CrossRef]
20. Medsker, L.R.; Jain, L.C. Recurrent neural networks. *Des. Appl.* **2001**, *5*, 64–67.
21. Li, J.; Dada, A.; Puladi, B.; Kleesiek, J.; Egger, J. ChatGPT in healthcare: A taxonomy and systematic review. *Comput. Methods Programs Biomed.* **2024**, *245*, 108013. [CrossRef] [PubMed]
22. *ChatGPT*, version 3.5; OpenAI: San Francisco, CA, USA, 2023. Available online: https://openai.com/chatgpt (accessed on 22 September 2023).
23. Dave, T.; Athaluri, S.A.; Singh, S. ChatGPT in medicine: An overview of its applications, advantages, limitations, future prospects, and ethical considerations. *Front. Artif. Intell.* **2023**, *6*, 1169595. [CrossRef] [PubMed] [PubMed Central]
24. Gertz, R.J.; Bunck, A.C.; Lennartz, S.; Dratsch, T.; Iuga, A.-I.; Maintz, D.; Kottlors, J. GPT-4 for Automated Determination of Radiological Study and Protocol based on Radiology Request Forms: A Feasibility Study. *Radiology* **2023**, *307*, e230877. [CrossRef] [PubMed]
25. Kottlors, J.; Bratke, G.; Rauen, P.; Kabbasch, C.; Persigehl, T.; Schlamann, M.; Lennartz, S. Feasibility of Differential Diagnosis Based on Imaging Patterns Using a Large Language Model. *Radiology* **2023**, *308*, e231167. [CrossRef]
26. Ten Berg, H.; van Bakel, B.; van de Wouw, L.; Jie, K.E.; Schipper, A.; Jansen, H.; Kurstjens, S. ChatGPT and Generating a Differential Diagnosis Early in an Emergency Department Presentation. *Ann. Emerg. Med.* **2023**, *83*, 83–86. [CrossRef]
27. You, Y.; Gui, X. Self-diagnosis through ai-enabled chatbot-based symptom checkers: User experiences and design con-siderations. *AMIA Annu. Symp. Proc.* **2020**, *2020*, 1354–1363.
28. Gebrael, G.; Sahu, K.K.; Chigarira, B.; Tripathi, N.; Thomas, V.M.; Sayegh, N.; Maughan, B.L.; Agarwal, N.; Swami, U.; Li, H. Enhancing Triage Efficiency and Accuracy in Emergency Rooms for Patients with Metastatic Prostate Cancer: A Retrospective Analysis of Artificial Intelligence-Assisted Triage Using ChatGPT 4.0. *Cancers* **2023**, *15*, 3717. [CrossRef]
29. Palenzuela, D.L.; Mullen, J.T.; Phitayakorn, R. AI Versus MD: Evaluating the surgical decision-making accuracy of ChatGPT-4. *Surgery* **2024**, *176*, 241–245. [CrossRef] [PubMed]
30. Beaulieu-Jones, B.R.; Berrigan, M.T.; Shah, S.; Marwaha, J.S.; Lai, S.-L.; Brat, G.A. Evaluating Capabilities of Large Language Models: Performance of GPT4 on Surgical Knowledge Assessments. *medRxiv* **2023**. Update in *Surgery* **2024**, *175*, 936–942. [CrossRef] [PubMed] [PubMed Central]
31. Phan-Mai, T.-A.; Thai, T.T.; Mai, T.Q.; Vu, K.A.; Mai, C.C.; Nguyen, D.A. Validity of Machine Learning in Detecting Complicated Appendicitis in a Resource-Limited Setting: Findings from Vietnam. *BioMed Res. Int.* **2023**, *2023*, 5013812. [CrossRef]
32. Marcinkevics, R.; Wolfertstetter, P.R.; Wellmann, S.; Knorr, C.; Vogt, J.E. Using Machine Learning to Predict the Diagnosis, Management and Severity of Pediatric Appendicitis. *Front. Pediatr.* **2021**, *9*, 662183. [CrossRef] [PubMed]
33. Mijwil, M.M.; Aggarwal, K. A diagnostic testing for people with appendicitis using machine learning techniques. *Multimedia Tools Appl.* **2022**, *81*, 7011–7023. [CrossRef]
34. Akbulut, S.; Yagin, F.H.; Cicek, I.B.; Koc, C.; Colak, C.; Yilmaz, S. Prediction of Perforated and Nonperforated Acute Appendicitis Using Machine Learning-Based Explainable Artificial Intelligence. *Diagnostics* **2023**, *13*, 1173. [CrossRef] [PubMed]
35. Mu, Y.; He, D. The Potential Applications and Challenges of ChatGPT in the Medical Field. *Int. J. Gen. Med.* **2024**, *17*, 817–826. [CrossRef] [PubMed] [PubMed Central]
36. Stahl, B.C.; Eke, D. The ethics of ChatGPT—Exploring the ethical issues of an emerging technology. *Int. J. Inf. Manag.* **2024**, *74*, 102700. [CrossRef]

37. Guleria, A.; Krishan, K.; Sharma, V.; Kanchan, T. ChatGPT: Ethical concerns and challenges in academics and research. *J. Infect. Dev. Ctries.* **2023**, *17*, 1292–1299. [CrossRef] [PubMed]
38. Emsley, R. ChatGPT: These are not hallucinations—they're fabrications and falsifications. *Schizophrenia* **2023**, *9*, 52. [CrossRef]
39. Chelli, M.; Descamps, J.; Lavoué, V.; Trojani, C.; Azar, M.; Deckert, M.; Raynier, J.-L.; Clowez, G.; Boileau, P.; Ruetsch-Chelli, C. Hallucination Rates and Reference Accuracy of ChatGPT and Bard for Systematic Reviews: Comparative Analysis. *J. Med. Internet Res.* **2024**, *26*, e53164. [CrossRef]
40. Schwarz; Nicolas, T. *Allgemein- und Viszeralchirurgie Essentials*; Thieme: Stuttgart, Germany, 2017; p. 172. [CrossRef]
41. Baumgartner, C.; Baumgartner, D. A regulatory challenge for natural language processing (NLP)-based tools such as ChatGPT to be legally used for healthcare decisions. Where are we now? *Clin. Transl. Med.* **2023**, *13*, e1362. [CrossRef]
42. EUR-Lex Document 32017R0745. 2017. Available online: https://eur-lex.europa.eu/eli/reg/2017/745/oj (accessed on 28 August 2024).
43. Mulita, F.; Verras, G.-I.; Anagnostopoulos, C.-N.; Kotis, K. A Smarter Health through the Internet of Surgical Things. *Sensors* **2022**, *22*, 4577. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.