Ethical implications of AI in robotic surgical training: A Delphi consensus statement

Justin W Collins, Hani J Marcus, Ahmed Ghazi, Ashwin Sridhar, Daniel Hashimoto, Gregory Hager, Alberto Arezzo, Pierre Jannin, Lena Maier-Hein, Keno Marz, et al.

Ethical implications of AI in robotic surgical training: A Delphi consensus statement

Authors: Justin W. Collins[1,2,3]*, Hani J Marcus[2], Ahmed Ghazi[4], Ashwin Sridhar[1,3], Daniel Hashimoto[5], Gregory Hager[6], Alberto Arezzo[7], Pierre Jannin[8], Lena Maier-Hein[9], Keno Marz[9], Pietro Valdastri[10], Kensaku Mori[11], Daniel Elson[12], Stamatia Giannarou[12], Mark Slack[13,14], Luke Hares[15], Yanick Beaulieu[16], Jeff Levy[17], Guy Laplante[18], Arvind Ramadorai[19], Anthony Jarc[20], Ben Andrews[21], Pablo Garcia[22], Huzefa Neemuchwala[23], Alina Andrusaite[22], Tom Kimpe[24], David Hawkes[2], John D. Kelly[1,2,3], Danail Stoyanov[2]

Affiliations:

1. University College London, Division of Surgery and Interventional Science, Research Department of Targeted Intervention.
2. Wellcome/ESPRC Centre for Interventional and Surgical Sciences (WEISS), University College London
3. University College London Hospital, Division of Uro-oncology.
4. Simulation Innovation Laboratory, University of Rochester, USA
5. Surgical Artificial Intelligence and Innovation Laboratory, Massachusetts General Hospital, USA.
6. Malone Center for engineering in healthcare, Department of Computer Science, John Hopkins University, Baltimore, USA
7. Department of Surgical Sciences, University of Torino, Italy
8. University Rennes, Inserm, Rennes, France
9. Deutsches Krebsforschungszentrum, Division of Computer Assisted Medical Interventions, Heidelberg Germany
10. STORM Lab, School of Electronic and Electrical Engineering, University of Leeds, Leeds, UK
11. Director of Information Technology Center, Nagoya University, Japan
12. Hamlyn Centre for robotic surgery, Department of Surgery and cancer, Imperial College London, UK
13. Honorary Senior Lecturer, University of Cambridge, Cambridge UK
14. CMO CMR Surgical, Cambridge, UK
15. Chief technology director, CMR Surgical, Cambridge, UK
16. Division of Cardiology and Critical Care, Sacré-Coeur Hospital, University of Montreal, Montreal, Canada
17. Institute for Surgical Excellence, Philadelphia, USA
18. Director, Global Medical Affairs at Medtronic Minimally Invasive Therapies, Brampton, Canada
19. Director, Digital-Assisted Surgery (DAS), Medtronic Surgical Robotics, North Haven, CT, USA
20. Applied Research, Intuitive Surgical, Inc., Sunnyvale, CA, USA
21. Strategy, Intuitive Surgical, Inc., Sunnyvale, CA, USA
22. Johnson and Johnson Medical Devices, CA, USA
23. C-SATS Inc., a Johnson & Johnson Medical Devices Company, CA, USA
24. BARCO NV – Healthcare division, Kortrijk, Belgium

Acknowledgements: We would like to thank David Hayward, a BBC journalist, lecturer on ethics and patient advocate who contributed to the meetings and Delphi process.

Key Words: Artificial intelligence, machine learning, narrow AI, GDPR, data protection, privacy, transparency, predictive analytics, biases, training, curriculum development, surgical education, Deep learning, Computer vision, Natural language processing, Learning algorithms, risk prediction

***Corresponding author**:
Dr. Justin W Collins Email: justin.collins@ucl.ac.uk Contact number: +44 7751003409

Word count:

Abstract 300	Tables 4
Text 3998	Figures 3
References 30	

Abstract:

Context: As the role of AI in healthcare continues to expand there is increasing awareness of the potential pitfalls of AI and the need for guidance to avoid them.

Objectives: To provide ethical guidance on developing narrow AI applications for surgical training curricula. We define standardised approaches to developing AI driven applications in surgical training that address current recognised ethical implications of utilising AI on surgical data. We aim to describe an ethical approach based on the current evidence, understanding of AI and available technologies, by seeking consensus from an expert committee.

Evidence acquisition: The project was carried out in 3 phases: (1) A steering group was formed to review the literature and summarize current evidence. (2) A larger expert panel convened and discussed the ethical implications of AI application based on the current evidence. A survey was created, with input from panel members. (3) Thirdly, panel-based consensus findings were determined using an online Delphi process to formulate guidance. 30 experts in AI implementation and/or training including clinicians, academics and industry contributed. The Delphi process underwent 3 rounds. Additions to the second and third-round surveys were formulated based on the answers and comments from previous rounds. Consensus opinion was defined as ≥ 80% agreement.

Evidence synthesis: There was 100% response from all 3 rounds. The resulting formulated guidance showed good internal consistency, with a Cronbach alpha of >0.8. There was 100% consensus that there is currently a lack of guidance on the utilisation of AI in the setting of robotic surgical training. Consensus was reached in multiple areas, including: 1. Data protection and privacy; 2. Reproducibility and transparency; 3. Predictive analytics; 4. Inherent biases; 5. Areas of training most likely to benefit from AI.

Conclusions: Using the Delphi methodology, we achieved international consensus among experts to develop and reach content validation for guidance on ethical implications of AI in surgical training. Providing an ethical foundation for launching narrow AI applications in surgical training. This guidance will require further validation.

Patient summary: As the role of AI in healthcare continues to expand there is increasing awareness of the potential pitfalls of AI and the need for guidance to avoid them. In this paper we provide guidance on ethical implications of AI in surgical training.

Introduction:

Healthcare is perceived as a biomedical industry, but it has also always been an information industry and with an exponentially growing number of data collection devices there are increasing opportunities for machine learning (ML) and artificial intelligence (AI) applications. Healthcare is one of the most promising areas for the integration of AI systems, but it is also one of the most complex. The healthcare sector incorporates many regulatory, privacy and ethical laws that aim to ensure innovation will 'first cause no harm'.

Robotic surgery is an example of technology that has impacted the surgical profession, with rapid adoption seen, since the first robotic system received United States (US) Food and Drug Administration (FDA) approval in the year 2000 [1]. There are currently multiple robotic systems commercially available, with more systems, both procedure specific and general surgical robots, planned to be available soon [2].

The current success of robotic surgery has largely been as an enabler of minimally invasive procedures, delivering improved visualisation, increased precision, and enhanced dexterity compared with standard laparoscopy [3]. Despite these apparent advantages, independent studies have concluded that the rapid uptake of robotic surgery has potential to lead to diminished patient safety [4,5]. It is recognised that there can be increased patient risks from complications during the introduction of new technology, including robotic surgery [6]. An important safety aspect is the training received prior to operating on patients and it is recognised that robotic surgery trainees require training in technology as well as surgical technique. Agreed standards of 'surgical proficiency' are needed with an understanding of how this is optimally taught [7].

A future success of robotic surgery may be facilitated by the data robotic networks collect and for narrow AI algorithms to deliver real-time guidance [8]. However, the deployment of AI in surgery brings challenges related to the safeguarding of patient and surgeon data, the ethical boundaries of innovation, as well as the actual impact of AI on patient outcomes and the surgical team. A potentially less complex route to evaluating AI in surgery would be to assess the risks and benefits in surgical training, ideally in training labs, before direct patient involvement. Standardised training curricula could potentially reduce performance variables for input data and agreed benchmarks of performance and other defined outcome metrics would provide output data. At the core of robotic surgery technology are data processing and control algorithms that translate the surgeon's hand, wrist, and finger movements into precise movements of miniaturised surgical instruments inside the patient's body, with automated tremor cancellation. In addition, there are multiple sources of data available from video that can be interpreted with computer vision [9]. There is increasing potential to develop ML algorithms to give automated performance feedback from these data rich devices [9,10]. Automated performance metrics may also avoid certain biases that can occur in subjective trainer/trainee assessment. Robotic surgery training data can be collected both from the robotic system and from supplementary training technologies [11].

We should not underestimate the ethical and regulatory challenges that surround AI in MedTech devices that collect and store data from both patients and trainees [12,13]. As well as privacy and data issues, other identified potential 'pitfalls' include biases, accountability, explainability, transparency, and liability. Successfully addressing all these aspects will support and cultivate ML approaches and stimulate its further integration in robotic surgery systems. Whilst there is significant potential to harness this data to better understand how to improve surgical training and patient outcomes, there remains a lack of guidance on the ethical implications. This Delphi process aims to define the ethical and regulatory concerns about AI in robotic surgery training.

2. Materials and methods

The project was carried out in 3 phases: (1) A steering group was formed to review the literature and summarize the current evidence for the AI in robotic surgical training. (2) A larger expert panel convened and discussed the important aspects of ethical implications of AI application in training based on the current evidence. Following presentations and open discussion, a survey was created, with the input from the panel members. (3) Panel-based consensus findings were determined using an online Delphi process to formulate guidance and to provide recommendations for future research.

2.1 Review of the literature

The systematic review was performed in accordance with the PRISMA statement [14]. In November 2019 we undertook a comprehensive computerized search using PubMed and Medline databases. We systematically searched using the Boolean free-text search term (robot OR robotic OR "robot-assisted") AND (surgeon OR surgical OR surgery) AND (training OR course OR simulation OR curriculum) AND (AI OR "artificial intelligence" or ML OR "machine learning" OR "machine-learning". The literature review was updated in January 2020.

Articles of interest included prospective studies on the impact of robotic training utilizing novel technologies that enable data collection, such as telemetry, eye-tracking, teleprescence, video labelling and objective metric development and systematic reviews on robotic training published between July 2000, when the first robotic systems received FDA approval in the US [15] - and January 2020. Other significant studies cited in the reference list of selected papers were evaluated, as well as studies of interest published after the systematic search.

Two reviewers independently selected papers for detailed review (J.C. and H.M) evaluating the abstract and, if necessary, the full-text manuscript. Potential discrepancies were resolved by open discussion. The electronic search yielded a total of 105 potential articles. Fig. 1 summarizes the selection process. Overall, the quality of available studies was found to be low. Available evidence consists largely of expert opinion, consensus statements and small qualitative studies. There were no publications identified that focused specifically on the ethical implications of AI application for robotic surgery.

2.2 Expert panel conference meeting

An advisory panel was formed that was comprised of 30 key opinion leaders with a specialist interest in robotic surgery training and/or AI applications within healthcare. The panel was chaired by Professors Danial Stoyanov, John Kelly and Dr Justin Collins. In total 17 experts from the United Kingdom and Europe, 10 experts from the United States, 2 from Canada, and 1 from Japan were brought together to develop these consensus views. 18 were surgeons with an interest in the application of AI in robotic surgery training, one patient advocate and the remaining 11 were made up of members of the healthcare industry or academics with expertise in AI and machine learning. The median (range) for published panel members h-index and i10-idex were 25 (7-79) and 49 (7-616) respectively. The meeting comprised presentations on the subject matter and reviews of the literature findings (see appendix 1).

An overview of various potential implications for AI in training were discussed including:

- E-Learning
- Predictive analytics
- ML in surgical training
- Eye tracking
- Telemetry
- Teleprescence, video performance analytics and computer vision

- Network development
- Data protection and privacy issues
- Reproducibility and transparency issues
- Inherent biases in analytics

Participants were then divided into three discussion groups to identify the ethical implications:

- Group 1 - DATA: address issues with privacy, biases, standardisation, sensors, wearables, labelling
- Group 2 - DOMAIN: tasks, technique, instrument, other variations
- Group 3 - ACCOUNTABILITY: privacy, practical issues, network development, ownership, accountability, licenses, patient involvement?

Finally, the groups reconvened with a focus group discussion to summarise the 3 groups conclusions and a first draft survey divided into six overarching categories (see section 3.1), was generated by the panel members.

2.3 Internet survey and Delphi process

Following the consensus conference, the Delphi process was conducted to drive consensus of the experts' opinions. An Internet survey (Google forms) was generated and sent to the 30 members of the panel that comprised the surgeons and healthcare industry experts involved in robotic training. Supplementary file 1 shows a full list of the survey questions. An e-consensus reaching exercise using the Delphi methodology was then applied. The Delphi method structures group communications so that the process is effective in allowing a group of individuals to deal with a complex problem. Questions in which there was ≥80% consensus were removed from the next round of the survey. Repeated iterations of anonymous voting continued over three rounds, where an individual's vote in the next round was informed by knowledge of the entire group's results in the previous round. To be included in the final recommendations each survey item had to have reached group consensus (≥80% agreement) by the end of the 3 survey rounds. In the Delphi process the finding of 'consensus' is more relevant than the level of consensus. Levels of consensus are reported in the supplementary files (appendix 2).

Results:

3.1. Formulation of guidance

We had 100% (30/30) response rate in all three rounds. There was high inter-rater reliability which was >0.80. After three rounds of Delphi surveys, consensus was obtained in 109 elements in 6 different categories. The categories included:

- Section 1: Consensus on Terminology (define biases and data labelling terminology etc)
- Section 2: Data and privacy issues
- Section 3: Transparency and reproducibility
- Section 4: Potential biases from AI
- Section 5: Accountability and liability
- Section 6: Application of AI algorithms in robotic surgical training

There was 100% agreement within the panel that there are potential benefits to the utilisation of AI in the setting of robotic surgical training curricula, that there is currently a lack of guidelines (or guidance) on the utilisation of AI in the setting of surgical training and that the future success of AI in

surgical training will require its ethical deployment within healthcare organisations. There was also strong agreement that that there are potential risks to the utilisation of AI in the setting of robotic surgical training and that the aim of this group is to identify the ethical implications and to formulate guidance on AI in surgical training. Appendix 2 comprises a full list of the questions on the various elements of a consensus views and the levels of agreement reached.

Section 1: Terminology

Uniform communication language is important for understanding and evaluating the application of AI in surgical training. If there is ambiguity in the terminology, it may have implications in various clinical settings. At the meeting, we presented and discussed important terminologies related to AI in the setting of robotic surgical training. A summary of these agreed terms can be seen in Table 2.

Section 2: Data and privacy issues

There was consensus that data should be labelled in a standardised and reliable way to optimise datasets for AI algorithms. There was also recognition that standardised data labelling aids reporting of complications and has potential to improve patient safety to the point that there is an ethical obligation to standardise data labelling. The group identified hospital organisations, medical societies, and industry as all having a role in delivering this standardisation.

The group concluded there were several areas of data collection related to robotic surgery that can be automated. These included kinematic data (APMs) operative times, video performance analysis (computer vision), instrument usage and via the utilisation of novel technologies including eye tracking data [10, 11, 16]. There was also consensus that automated data should be confirmed to correlate with patient outcome data (e.g. PROMs) before it is used for training datasets [17].

Regarding privacy issues there was strong consensus that data should be anonymised using non-identifiable information whenever possible. Whilst recognising that with electronic patient records and increased connectivity of data points, it is increasingly difficult to anonymise data. The group agreed that strategies to anonymise data should include a combination of automated and manual approaches and that data minimisation can be best achieved by: (1) Identifying and capturing the minimum amount of data needed; (2) regular review of what data is stored and why; and (3) to only use personal data that is highly relevant and necessary for evaluation or machine learning algorithm.

Section 3: Transparency and reproducibility

Transparency is the ability of the person handling AI output data, to determine how an algorithm reaches its conclusion [18]. The committee agreed there are ethical issues with reproducibility of AI directly related to the opacity of AI thinking in deep neural networks. Regarding the implications in surgical training, the group concluded that the potential ethical issues regarding the transparency and reproducibility of AI are reduced in a dry or wet-laboratory training environment, as there is no direct patient involvement.

Section 4: Biases

Transparency also has implications for biases, as AI technologies have the potential for algorithmic bias, reinforcing discriminatory practices based on data points such as position, experience or race, sex, or other features [19]. Transparency of training data and of model interpretability would enable evaluation for potential biases. Ideally, machine learning could be a solution to resolve recognised biases. [19]. The committee reached a strong consensus view that AI can avoid certain biases that

may occur in human assessments with 100% consensus that both confirmation bias and interpretation bias would be better or at least the same with AI. Whereas there was concern that both prediction bias and information bias could be worse or equivalent with AI.

Identified elements that could affect biases in training datasets included historical datasets that are no longer representative of training outcomes, data from different geographical regions and automated data, where there is an assumption of clinical relevance, without direct evidence of impact. There was also a consensus view that cumulated data from different institutions with different curricula had potential to introduce biases if it was not adequately standardised. Another example of potential bias was data from trainees that are not representative of the intended population e.g. algorithm applied to academic centers when trainee trained on community programs. Historically there are many neural network models that were pre-trained by using open datasets. However, there is no guarantee that such data were collected in compliance with current or historical ethical regulations. The committee agreed there are additional ethical considerations regarding the utilization of pre-trained models in the development of surgical training AI.

Section 5: Accountability and liability

The committee reached consensus that clinicians have responsibility to our patients to improve their care whilst respecting their privacy, therefore systems need to be developed to protect privacy, whilst allowing AI advancements to improve healthcare. There was agreement that there should be guidelines around anonymising the data to protect privacy. The consensus views on this guidance are summarised in Table 3.

The committee agreed that the development of a network of experts related to the relevant surgical procedure would aid standardisation of data collection and data labelling and that key opinion leaders should be from multidisciplinary backgrounds to include clinicians, computer engineers, researchers, patients, patient advocates, industry, academia, and ethicists. There was also agreement that patients and the public be involved (consulted) in the development of AI systems.

Once approval is given for data collection, the committee recommended that the patient and surgeon should both give consent before collecting data.

Section 6: Application of AI algorithms in robotic surgical training

The committee agreed that to further enable machine learning of outcome data, we need data normalisation, and that full procedural decision making/strategy is also important to trace. There was consensus that we should ideally detect and record both activities and events. For example, bleeding or smoke, as a micro-activity label.

As an example of data normalisation, there was 100% consensus that we need standardised objectively defined metrics to train, test and measure surgical performance and that this should be achieved by utilising task-deconstruction on surgical procedures to identify key tasks to complete and errors to avoid. This in turn would enable surgical training to be completed with a modular approach [7] (step by step, bench marked progression), at the same time as reducing some of the variables to enhance ML algorithms.

The group agreed that training on singular skills tasks is useful in assessing proficiency. There was no agreement on whether we need AI on the whole procedure to evaluate outcomes or efficiency, or whether there are likely to be key indicative steps in any given procedure. However, there was 97% consensus that there is potential for AI to identify steps of a procedure, that are predictive of

outcome, that were not previously thought to be important. This area will benefit from further research. There was also no consensus on the need to collect data on non-technical skills.

Important elements of a surgical performance that can be used for labelling data are summarised in table 4.

The group identified various potential confounding variables that can impact outcome data. These important elements include: patient co-morbidities, disease and tumour staging, socio-economic factors, Geography/Culture, MedTech Device, instruments as well as hospital factors (size, nursing staff, nursing quality, etc.) and the provider regarding experience, training, etc.

There was agreement that metrics that define optimised surgical performance will change over time with machine learning algorithms and will themselves be impacted with advancements in knowledge, device development or other technological advances.

Regarding the technology, there were recommendations that we should measure instrument variations and compare instrument behaviour with 'capabilities' associated with better outcomes. With multi-instrument training datasets, there are opportunities for training AI systems to be robust to different tools or tool combinations.

There was agreement that practical data/findings derived from AI should, as much as possible, be open label and made available for the benefit of society, as well as being available to the public domain to aid research and development. The other areas of collaboration that were recommended, were the development of generic consent forms that ask patients to consent to the use of anonymised images, video and data for audit and research, including AI research. There was also agreement that common guidance on data aggregation strategies will help guide organisations to combine data.

The group agreed that real-time automated performance feedback in robotic surgery training, driven by AI, is ethical. Moreover, they advised that before embarking on developing an educational training platform, it should be clear from the onset who will be responsible in case harm is caused by the platform. The group also recommended that if AI utilised in training identifies a gap in knowledge or skills, there should be a remediation program developed and available for the trainee. See figure 3 for a summary of the consensus views.

Discussion:

The application of predictive AI to robotic surgical training has potential to gain value in the areas of data, human judgement, and actions. Whilst there are huge potential benefits, there are also risks. The ethical and regulatory concerns about AI in robotic surgery training can be grouped into three broad categories: the sources of data needed for ML algorithms in supervised learning; the development of algorithms; and the deployment of algorithms in surgical training.

Issues related to data privacy need to be addressed with hospitals having agreed protocols and guidance established by the DPO for trainees, trainers and patients. Established sources of data can be improved by collecting data in a standardised way, to reduce variables and enable ML algorithms. In robotic surgical training this can be achieved with standardised robotic curricula, train-the-trainer courses and agreed protocols that define surgical training at a granular level that can be aligned with telemetry data and more easily interpreted in computer vision analysis [7,10,21,22]. Further adoption of the performance metrics for identified key index procedures will be enabled by familiar definitions, that are open source and culminated together in data registries in established robotic

surgery networks [8]. The development of research networks that share open-source material is already established in areas of healthcare such as diagnostic imaging [23]. With big data, AI could contribute to intra-operative decision-support systems and post-operative risk prediction in surgical training [24].

Development of AI in surgical training should also consider the ethical implications of transparency and the ability to reproduce results in different training environments. Lack of transparency is recognised to be an important consideration in deep learning and the opacity associated with some AI algorithms is a major concern. Transparency in robotic surgical training is relevant at multiple levels. Scientific method is based on the ability to reproduce the same results when repeated by different researchers. Therefore, an inability to reproduce results affects the trustworthiness of scientific conclusions [25]. Transparency relates to model interpretability and the ease to understand or interpret how a given technology reaches a certain decision or prediction.

There are ethical implications for credentialling organisations who reach decisions on proficiency of the surgeon utilising an 'opaque' or 'black box' approach [5, 26]. It is recognised that if the ML system's reasoning can be explained, then humans can verify whether the reasoning is sound. But, if the system's reasoning cannot be explained, such evaluation is no longer feasible, and it is difficult to perform analyses that could potentially uncover new training insights. However, there are potential trade-offs with increased transparency and interpretability resulting in decreased accuracy or predictive performance of predictive models [18].

With supervised learning, the accuracy of predictions relies heavily on the accuracy of the underlying annotations used to train the model. Poorly labelled data will yield poor results in surgical assessments [27]. Standardised labelling and transparency of data labelling used in the ML training process for a supervised learning algorithm is imperative to ensure consistency and ultimately accuracy. Therefore, the need to have shared open-source metrics for labelling data is paramount to the success of AI in robotic surgery training that can be critically evaluated.

Whilst there was much consensus on guidance for the collection of performance metrics, there was no consensus on the need to collect data on non-technical skills (NTS), which may reflect a lack of current focus on this area or the difficulties in accurately labelling the data of NTS [28].

The deployment of AI in surgical training is still early in its development. Although significant advances are being made, these advances are focused on narrow applications of the technology to specific aspects of performance such as surgical instrument tracking [29]. One of the largest gaps in current knowledge is understanding the scope of the domain of data and the variability that it may contain. Systems to collect large scale, quantitative information across different institutional and geographical boundaries are still not widely used. This means that AI development, which may require such data, is still developed in silo and hence needs international consortia and resources. It is therefore critical to establish agreed criteria for structuring such data and defining what information is key to be extracted from it.

The application of AI in surgical training is in the phase of discovery and development, and critical appraisal of new devices, publications and software is necessary to appropriately evaluate their impact on robotic surgical training. However, data on potential applications of AI to surgery have been promising [30]. There are also opportunities to evaluate predictive AI in robotic surgical training in laboratory settings, that do not directly involve patients and may avoid many identified ethical implications.

Limitations: future studies should acknowledge the ethical and legal implications of developing AI in surgery and the need to recognise data protection, privacy by design and transparency issues. In this study, we focus on the use of ML in surgical training in the general sense. We do not address specific considerations that may be needed for particular ML approaches, for example deep learning as opposed to classical rule-based or logical algorithms. AI applications for surgical training will require careful evaluation and validation with predefined training goals before they can be routinely deployed. In addition, variability and robustness of any technology needs extensive experimental testing and understanding which can only happen if performed at scale. Understanding of such variability should dictate how adoption can happen and be used to identify risk factors from using such technologies for the different stakeholders.

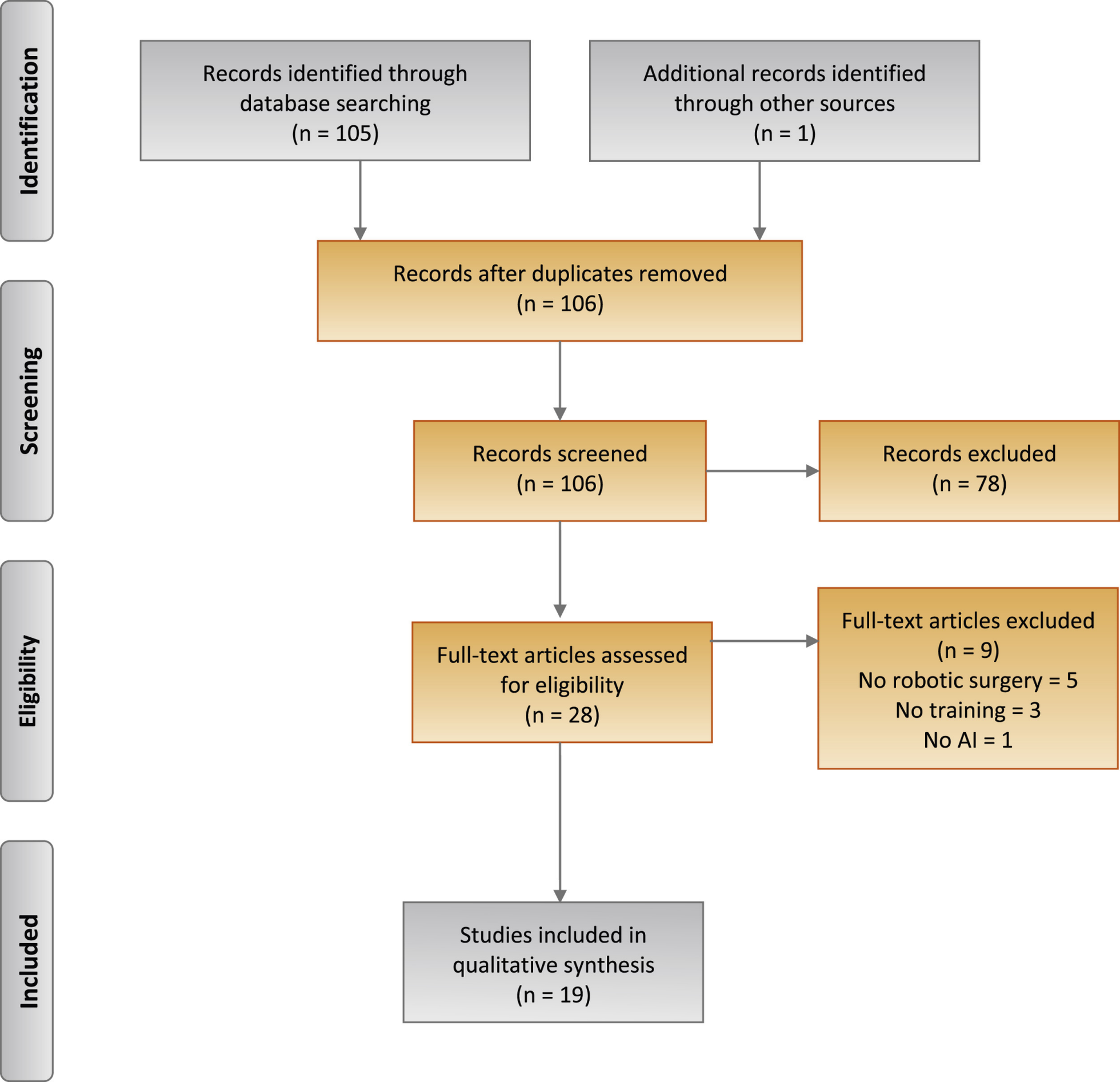
Conclusions:
The integration of AI into robotic surgical training has many opportunities but also risks. Recognised risks on the ethical application of AI include data and privacy issues, transparency, biases, accountability, and liabilities. Using the Delphi methodology, we achieved international consensus among experts to develop and reach content validation for guidance on the ethical implications of using AI in various surgical training settings. This guidance lays the foundation for launching narrow AI applications in surgical training. This guidance will require further validation.

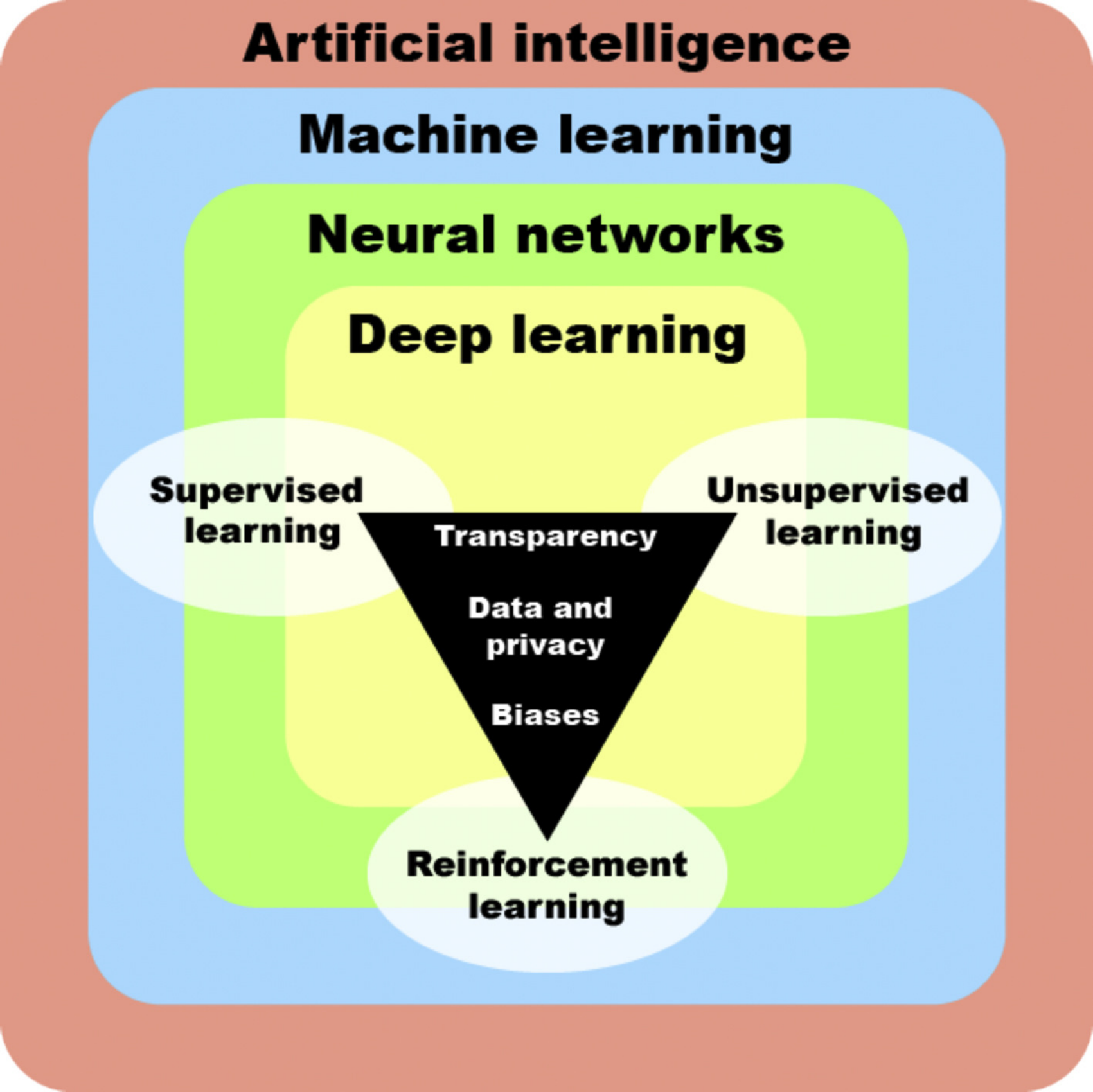
Figures

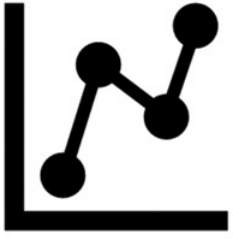
Fig. 1. PRISMA flow diagram summarising the study selection process.

Fig. 2. Relationship of terminology and link to ethical implications.

Fig. 3. Summary of the guidance for optimising the application of AI in robotic surgery training.

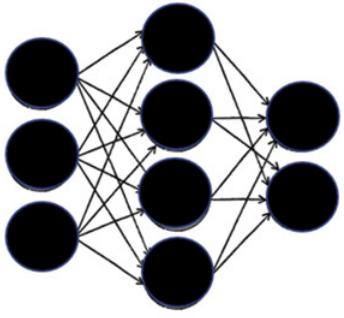

Identification
Screening
Eligibility
Included
Records identified through database searching (n = 105)
Additional records identified through other sources (n = 1)
Records after duplicates removed (n = 106)
Records screened (n = 106)
Records excluded (n = 78)
Full-text articles assessed for eligibility (n = 28)
Full-text articles excluded (n = 9)
No robotic surgery = 5
No training = 3
No AI = 1
Studies included in qualitative synthesis (n = 19)


Artificial intelligence
Machine learning
Neural networks
Deep learning
Supervised learning
Unsupervised learning
Transparency
Data and privacy
Biases
Reinforcement learning



DATA

- STANDARDISED
- VERIFIED
- COLLOBARATIVE
- OPEN SOURCE
- CONSENTED



DEVELOPMENT

- TRANSPARENT
- REPRODUCIBLE

DEPLOYMENT

- EXPLAINABLE
- SUPPORTED

References

1. Ballantyne GH, Moll F. The da Vinci telerobotic surgical system: the virtual operative field and telepresence surgery. *Surg Clin North Am* 2003; 83(6):1293-1304
2. A Gözen, J Rassweiler. Robotic surgery in Urology: New kids on the block. Urologe A. 2020 Sep;59(9):1044-1050. doi: 10.1007/s00120-020-01293-8.
3. Sheetz KH, Claflin J, Dimick JB. Trends in the Adoption of Robotic Surgery for Common Surgical Procedures. JAMA Netw Open 2020; 3(1):e1918911.
4. Sheetz KH, Dimick JB. Is It Time for Safeguards in the Adoption of Robotic Surgery? *JAMA* 2019; 321(20):1971-1972.
5. Dimitrios Stefanidis, Elizabeth M Huffman, Justin W Collins, Martin A Martino, Richard M Satava, Jeffrey S Levy. Expert Consensus Recommendations for Robotic Surgery Credentialing. Ann Surg. 2020 Nov 17. doi: 10.1097/SLA.0000000000004531. Online ahead of print.
6. Parsons JK, Messer K, Palazzi K, et al. Diffusion of surgical innovations, patient safety, and minimally invasive radical prostatectomy. JAMA Surg 2014; 149(8):845-851.
7. JW Collins, J Levy, D Stefanidis, A Gallagher, M Coleman, T Cecil et al. Utilising the Delphi Process to Develop a Proficiency-based Progression Train-the-trainer Course for Robotic Surgery Training. Eur Urol. 2019 May;75(5):775-785. doi: 10.1016/j.eururo.2018.12.044. Epub 2019 Jan 19. Review.
8. Collins J, Akre O, Challacombe B, Karim O, Wiklund P. Robotic networks: delivering empowerment through integration. BJU Int. 2015;116(2):167-8.
9. Stoyanov D. Surgical vision. Ann Biomed Eng. 2012 Feb;40(2):332-45. doi: 10.1007/s10439-011-0441-z.
10. Hung AJ, Chen J, Gill IS. Automated Performance Metrics and Machine Learning Algorithms to Measure Surgeon Performance and Anticipate Clinical Outcomes in Robotic Surgery. JAMA Surg 2018; 153(8):770-771.
11. Chen IA, Ghazi A, Sridhar A et al. Evolving robotic surgery training and improving patient safety, with the integration of novel technologies. World J Urol. 2020 Nov 6. doi: 10.1007/s00345-020-03467-7.Online ahead of print.
12. Shane O'Sullivan, Nathalie Nevejans, Colin Allen, Andrew Blyth, Simon Leonard, Ugo Pagallo et al. Legal, regulatory, and ethical frameworks for development of standards in artificial intelligence (AI) and autonomous robotic surgery. Int J Med Robot. 2019 Feb;15(1):e1968. doi: 10.1002/rcs.1968.
13. Michael J. Rigby. Ethical Dimensions of Using Artificial Intelligence in Health Care. AMA Journal of Ethics: 2019; Vol. 21, Number 2: E121-124.
14. Moher D, Liberati A, Tetzlaff J, Altman DG, Group P. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Plos Med. 2009;6(7):e1000097.
15. Leal Ghezzi T, Campos Corleta O. 30 Years of robotic surgery.World J Surg 2016;40:2550–7.
16. Tony Tien, Philip H. Pucher, Mikael H. Sodergren, Kumuthan Sriskandarajah, Guang-Zhong Yang, Ara Darzi. Eye tracking for skills assessment and training: a systematic review. JSR 2014; 191 (1): 169-178.
17. Ahern S, Ruseckaite R, Ackerman IN. Collecting patient-reported outcome measures. Intern Med J. 2017 Dec;47(12):1454-1457. doi: 10.1111/imj.13633.
18. Jianxing He, Sally L. Baxter, Jie Xu, Jiming Xu, Xingtao Zhou, Kang Zhang. The practical implementation of artificial intelligence technologies in medicine. Nat Med. 2019 January; 25(1): 30–36. doi:10.1038/s41591-018-0307-0.].

19. Char DS, Shah NH & Magnus D. Implementing machine learning in health care—addressing ethical challenges. N. Engl. J. Med 2018; 378: 981–983.
20. Chen J, Oh PJ, Cheng N, et al. Use of automated performance metrics to measure surgeon performance during robotic vesicourethral anastomosis and methodical development of a training tutorial. J Urol 2018;200:895–902.
21. JW Collins, A Ghazi, D Stoyanov et al. Utilising an accelerated delphi process to develop guidance and protocols for telepresence applications in remote robotic surgery training. European Urology Open Science 22, 23-33. https://doi.org/10.1016/j.euros.2020.09.005
22. Daniel Hashimoto, Guy Rosman, Elan Witkowski et al. Computer Vision Analysis of Intraoperative Video. Automated Recognition of Operative Steps in Laparoscopic Sleeve Gastrectomy. Ann Surg. 2019 Sep;270(3):414-421. doi: 10.1097/SLA.0000000000003460.
23. https://monai.io/ Last reviewed 24/12/2020
24. Thomas M Ward, Daniel A Hashimoto, Yutong Ban. Automated operative phase identification in peroral endoscopic myotomy. Surg Endosc. 2020 Jul 27. doi: 10.1007/s00464-020-07833-9.
25. Stodden, V. (2010). The scientific method in practice: Reproducibility in the computational sciences (MIT Sloan Research Paper No. 4773-10). Cambridge, MA: Massachusetts Institute of Technology. doi:10.2139/ssrn.1550193.
26. Piotr M Patrzyk, Daniela Link, Julian N Marewski. Human-like machines: Transparency and comprehensibility. Behav Brain Sci. 2017 Jan;40:e276. doi: 10.1017/S0140525X17000255.
27. Hashimoto DA, Rosman G, Rus D & Meireles OR Artificial intelligence in surgery: promises and perils. Ann. Surg. 2018; 268: 70–76.
28. JW Collins, P Dell'Oglio, AJ Hung, NR Brook. The Importance of Technical and Non-technical Skills in Robotic Surgery Training. Eur Urol Focus 2018 Sep;4(5):674-676. doi: 10.1016/j.euf.2018.08.018.
29. Maria Robu, Abdolrahim Kadkhodamohammadi, Imanol Luengo, Danail Stoyanov. Towards real-time multiple surgical tool tracking, Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization 2020, DOI: 10.1080/21681163.2020.1835553.
30. Daniel A. Hashimoto, Thomas M. Ward, Ozanan R. Meireles. The Role of Artificial Intelligence in Surgery. Advances in Surgery. Sep 2020: REVIEW; 54, 89-101. DOI:https://doi.org/10.1016/j.yasu.2020.05.010

Appendix 1:

List of presentations and presenter

Overview of AI in surgical intervention and training	Danail Stoyanov
Online training: what are the needs and opportunities?	Ashwin Sridhar
Potential pitfalls of AI	Justin Collins
Meeting the challenges of GDPR and data analysis	Justin Collins
Intuitive Surgical: Intelligent surgery	Anthony Jarc
Medtronic: learning from shared datasets	Arvind Ramadorai/Guy Laplante
CMRSurgical: Implementation of complex technologies. The use of the IDEAL-D framework	Mark Slack
CSATS: Vide performance analysis metrics in surgical training	Huzefa Neemuch
Eye tracking metrics in remote training	Ahmed Ghazi
Using predictive analytics to improve patient selection	Kensaku Mori
Interactive live remote virtual proctoring in the AI era	Yanick Beaulieu
Developing feedback and early warning systems in surgery	Pierre Jannin
Knowledge transfer, learning and community network development	Justin Collins
Introduction to an accelerated Delphi process and next steps	Justin Collins

List of expert panel member with h-index and i10-index

Panel member	Country	h-index	i10-index
Justin Collins	UK	21	36
Ahmed Ghazi	USA	13	15
Danail Stoyanov	UK	35	99
Yanick Beaulieu	Canada	15	19
Ashwin Sridhar	UK	15	19
Hani Marcus	UK	25	49
Jeffrey Levy	USA	7	7
John. D. Kelly	UK	39	92
Daniel Hashimoto	USA	17	21
Daniel Elson	UK	40	91
Pierre Jannin	France	36	88
Alberto Arezzo	Italy	47	137
Gregory Hager	USA	76	337
Keno Marz	Germany	11	11
Stamatia Giannarou	UK	17	23
Kensaku Mori	Japan	43	616
Lena Maier Hein	Germany	33	102
Pietro Valdastri	UK	40	101
David Hawkes	UK	79	342
Tom Kimpe	Belgium	19	33
Mark Slack	UK	24	34
Luke Hares	UK	NR	NR
Guy Laplante	Canada	NR	NR
Arvind Ramadorai	USA	8	8
Pablo Garcia	USA	24	45
Huzefa Neemuchwala	USA	12	17
Alina Andrusaite	USA	NR	NR
Ben Andrew	USA	NR	NR
Anthony Jarc	USA	12	15

Appendix 2:

Results of the Delphi process completed over 3 rounds and levels of agreement are summarised below. Questionnaires were completed by the committee members independently and anonymously using online google forms.

Questionnaire:

The aim of this Delphi process survey is to reach consensus views on the ethical implications of AI in surgical training. We aim to achieve this by reaching agreement on important aspects of the guidance. The questionnaire will be completed over 3 rounds and any questions that reaches 80% consensus will be removed from the following round(s). For remaining questions, contributors will be informed of the current percentage of agreement from the previous round.

Many questions have sections at the end to add comments or suggest alternative answers if you do not agree with the statement. There is also space at the end of each section to add additional questions or comments to aid clarification.

All answers/comments from the questionnaire will be anonymized both in evaluation and reporting. We are collecting emails to identify who has responded to the questionnaire and who requires reminders. Many thanks!

In round 1 there were 5 general questions, 30 questions on data, 26 questions on domain issues and 41 questions on accountability, total of 102 questions. 71/102 reached consensus of >80% in the first round, including all 5 general questions.

In round two there were 21 questions on data (including 5 new questions and one duplicate question removed), 10 questions (including 7 new questions) on domain issues and 16 questions (including 5 new questions) on accountability. 27/47 reached consensus of >80% in the second round.

In the final round there were 11 questions (including 2 new questions) on data, 6 questions (including 2 new questions) on domain issues and 7 questions on accountability. 11/24 reached consensus of >80% in the final round.

We had 100% response rate to all three rounds. Consensus was reached in 109 out of 173 questions overall.

<u>Section 1: General questions (consensus levels in brackets):</u>

1.1. Do you agree that there are potential benefits to the utilisation of AI in the setting of surgical training curricula? – 100% consensus (round 1)
1.2. Do you agree that there are potential risks to the utilisation of AI in the setting of surgical training curricula? – 97% consensus (round 1)
1.3. Do you agree that there is currently a lack of guidelines (or guidance) on the utilisation of AI in the setting of surgical training? – 100% consensus (round 1)
1.4. Do you agree that the future success of AI in surgical training will require its ethical deployment within healthcare organisations? – 100% consensus (round 1)
1.5. Do you agree that an aim of this group is to identify the ethical implications and to formulate guidance on AI in surgical training? – 97% consensus (round 1)

Level of agreement	General statements about the potential for telepresence in robotic surgery training
100%	• There are potential benefits to the utilisation of AI in the setting of surgical training curricula • There is currently a lack of guidelines (or guidance) on the utilisation of AI in the setting of surgical training • The future success of AI in surgical training will require its ethical deployment within healthcare organisations
95%	• There are potential risks to the utilisation of AI in the setting of surgical training curricula • An aim of this group is to identify the ethical implications and to formulate guidance on AI in surgical training

Table a: General questions on the ethical implications of AI in surgical training

Section 2: Data issues

2.1 Do you agree that data should be labelled in a standardised and reliable way to optimise AI? – 97% consensus (round 1)

2.2 Do you agree that data should ideally be anonymised using non-identifiable information whenever possible? – 90% consensus (round 1)

2.3 Do you agree that with electronic patient records and increased connectivity of data points, that it is increasingly difficult to anonymise data? – 87% consensus (round 1)

2.4. Data minimisation can be achieved by which of the following elements (can tick multiple answers as required)

- Limit amount of data used – 73% consensus (no consensus)
- Regularly review what data is stored and why – 93% consensus (round 2)
- Identify and capture the minimum amount of data needed – 87% consensus (round 2)
- Only use personal data that is highly relevant and necessary for evaluation or machine learning algorithm – 80% consensus (round 2)

2.5. Strategies to anonymise data can be (can tick multiple answers as required)

- Manual
- Automated
- Combination of automated and manual approaches - 97% consensus (round 1)

2.6. Is there an ethical obligation to standardise data labelling? - 80% consensus (round 1)

2.6a Does standardised data labelling aid reporting of complications? – 93% consensus (round 1)

2.6b Does standardised data labelling have potential to improve patient safety? – 93% consensus (round 2)

2.7. Who has an ethical obligation to standardise data and reporting mechanisms?

- Hospital organisations – 80% consensus (round 1)
- Medical societies – 93% consensus (round 2)
- Patient led organisations – 37% consensus (no consensus)
- Industry – 87% consensus (round 3)

- None of the above

2.8. With robotic surgery what data collection can be automated? (can tick multiple boxes)

- Kinematic data (APMs) – 97% consensus (round 1)
- Operative time – 90% consensus (round 1)
- Video performance analysis – 90% consensus (round 1)
- Haptic feedback – 63% consensus (no consensus)
- Eye tracking data – 87% consensus (round 1)
- Instrument usage (new response) – 97% consensus (round 2)
- Operative phases (new response) – 77% consensus (round 3 - no consensus)
- Workflow information (new response) – 67% consensus (round 3 - no consensus)

2.9. Should automated data be correlated with patient outcome data (e.g. PROMs) before it is used for training datasets? – 80% consensus (round 2)

2.10. Do you agree that AI can avoid certain biases that may occur in human assessments? – 93% consensus (round 1)

2.11. Confirmation bias occurs when researchers use data/answers to confirm their hypothesis or beliefs. Do you think this will be better or worse with AI?

- Better – 47% (round 3)
- Worse – 0% (round 3)
- The same – 53% (round 3)

Overall, 100% consensus the same or better

2.12. Interpretation bias occurs when researchers' conclusions or assessments are affected by descriptions e.g. what speed was the car going when it hit the fence, compared with what speed was the car doing when it smashed into the fence. Do you think interpretation bias will be better or worse with AI?

- Better – 73% (round 3)
- Worse – 0% (round 3)
- The same – 27% (round 3)

Overall, 100% consensus the same or better

2.13. Prediction bias occurs when data points are too focused so that association is concluded. For example, predicting where crimes will most likely occur and deploying police to that area can result in more arrests being made in that area. Poor programming can result in prediction bias. Do you think this will be better or worse with AI?

- Better – 10% (round 3)
- Worse – 47% (round 3)
- The same – 43% (round 3)

Overall, 90% consensus the same or worse

2.14. Information bias occurs when researchers use information they believe is already linked to their outcome. For example, Google used algorithms to predict flu epidemics in the US related to

searches for flu medicines, which resulted in inaccurate conclusions. Do you think this will be better or worse with AI?

- Better – 3% (round 3)
- Worse – 50% (round 3)
- The same – 47% (round 3)

Overall, 97% consensus the same or worse

2.15. What elements could affect biases in training datasets? (can tick multiple boxes).

- Historical datasets that are no longer representative of training outcomes -87% consensus (round 1)
- Data from different geographical regions – 97% consensus (round 2)
- Automated data, where there is an assumption of clinical relevance, without direct evidence of impact – 90% consensus (round 2)
- Data from institutions with different curricula (new response) – 87% consensus (round 2)
- Data from trainees that are not representative of the intended population (e.g., algorithm applied to academic centers when trained on community programs) (new response) – 93% consensus (round 3)

New questions round 2

2.16 Do you agree that a lack of education around use of AI can impact misuse, misinterpretation, misapplication of AI techniques and results? – 97% consensus (round 2)

2.17 Historically there are many neural network models that were pre-trained by using open datasets. However, there is no guarantee that such data were collected in compliance with current or historical ethical regulations. Do you agree there are additional ethical considerations regarding the utilization of pre-trained models in the development of surgical AI? – 83% consensus (round 2)

Section 3: Domain issues

3.1. Do you agree that we need standardised objectively defined metrics to train, test and measure surgical performance? – 100% consensus (round 1)

3.2. Should surgical procedures undergo task-deconstruction to identify key tasks to complete and errors to avoid? – 100% consensus (round 1)

3.3. Should training be completed with a modular approach (step by step, bench marked progression)? – 100% consensus (round 1)

3.4. Is training on singular skills tasks useful in assessing proficiency? – 97% consensus (round 2)

3.5a. Do we need AI on the whole procedure or are there likely to be key steps in any given procedure?

- Needs to be whole procedure – 13% consensus (round 3)
- Only needs key (selected) steps – 50% consensus (round 3)
- Need to look at whole procedure, in the first instance, as AI may be able to identify steps that are predictive of outcome that were not previously thought to be important – 37% consensus (round 3)

3.5b. Do you agree that there is potential for AI to identify steps that are predictive of outcome that were not previously thought to be important? (new question round 2) – 97% consensus (round 2)

3.5c. Do we need AI on the whole procedure to assess efficiency? (new question round 3) – 70% consensus (round 3)

3.6. Important elements of a surgical performance that can be used for labelling data include: (can tick multiple boxes)

- Phases (of the procedure) - 97% consensus (round 1)
- Visual cues (anatomical landmarks, areas of interest etc) - 97% consensus (round 1)
- Error event (error that results in harm to tissue or patient) - 93% consensus (round 1)
- Technical errors (not necessarily associated with an event) e.g. using wrong instrument to grasp bowel (traumatic grasper) - 87% consensus (round 1)
- Mechanical error (device malfunction) - 90% consensus (round 2)
- Automated performance metrics (kinematic data) - 90% consensus (round 1)
- Non-technical skills e.g. communication (new response round 2) - 73% consensus (round 3)
- Tissue characteristics e.g., inflamed, fibrotic etc (new response round 2) - 97% consensus (round 3)
- Anatomical variations (new response round 2) - 93% consensus (round 3)
- Disease staging (new response round 3) - 67% consensus (round 3)

3.7. For errors related to device failure, do you prefer the term?

- Device error - 83% consensus (round 1)
- Mechanical error – 7%
- Other suggestions – 10%

3.8. Do you agree that metrics that define optimised surgical performance will change over time with machine learning algorithms? – 97% consensus (round 1)

3.9. Do you agree that metrics that define optimised surgical performance will change over time with advancements in knowledge, device development or other technological advances? – 97% consensus (round 1)

3.10. For outcome data, elements that can affect outcome include: (can tick multiple boxes). Note: These outcomes could be linked to both macro or micro granularity of tasks and events during a procedure.

- Patient co-morbidities - 100% consensus (round 1)
- Tumour staging - 90% consensus (round 1)
- Socio-economic factors - 87% consensus (round 1)
- Geography/Culture - 83% consensus (round 1)
- MedTech Device - 87% consensus (round 1)
- Instruments - 83% consensus (round 1)
- Hospital factors (size, nursing staff, nursing quality, etc.) (new response round 2) - 90% consensus (round 2)
- Provider (e.g. experience, training, etc.) (new response round 2) - 97% consensus (round 2)
- Disease process and stage (new response round 2) - 87% consensus (round 2)

3.11. For machine learning of outcome data, do we need data normalisation? Note: Attempting to normalise to some commonly agreed criteria. – 90% consensus (round 1)

3.12. Is full procedural decision making/strategy important to trace? – 90% consensus (round 1)

3.13. Should we detect activities or events? Note: For example, bleeding or smoke, as a micro-activity label.

- Activities
- Events
- Both – 90% consensus (round 1)

3.14. Should we measure instrument variations? Note: This relates to different manufacturer or different instrument features/capabilities – 97% consensus (round 1)

3.15. If we prove outcome, should we compare instrument behaviour with 'capabilities'? Note: This relates to instrument dexterity for example, which could be associated with a better outcome – 90% consensus (round 1)

3.16. Do we need multi-instrument training datasets? Note: This refers to training AI systems to be robust to different tools or tool combinations – 90% consensus (round 1)

Section 3: Accountability issues

4.1. Do you agree that as clinicians we have responsibility to our patients to improve their care. Therefore, systems need to be developed to protect privacy, whilst allowing AI advancement to improve care – 83% agreement (round 1).

4.2. Should there be agreed standards around anonymising the data to protect privacy? – 97% consensus (round 1).

4.3. Organisational accountability under GDPR, is legally obliged to put into place which of the following: (can tick multiple boxes)

- Comprehensive governance issues – 90% consensus (round 1)
- Privacy impact assessments – 93% consensus (round 2)
- Privacy measures by design – 93% consensus (round 3)

4.4. Who has responsibility to anonymise data? (MCQ)

- The data processor
- The data controller
- The data protection officer
- Everyone handling the data has shared responsibility - 87% consensus (round 1)

4.5. Should all data that is planned to be collected, be proactively approved, and stored according to guidelines from the organisation's data protection office? – 100% agreement (round 1)

4.6. Should the data protection officer (DPO) have overall responsibility for data protection compliance matters? – 97% agreement (round 1)

4.6a. The responsibilities of a DPO include: (can tick multiple boxes)

- Informing and advising organisations of their obligations under data protection law - 90% consensus (round 1)

- Monitoring compliance with the regulations and related policies, including raising of awareness and training of staff - 97% consensus (round 1)
- Providing procedures, guidance and advice in support of this policy e.g., for Data Protection Impact Assessment (DPIAs) - 100% consensus (round 1)
- Acting as the organisations first point of contact with the Information Commissioner's Office (ICO) - 80% consensus (round 1)
- Handling subject access requests and official requests for personal data from third parties - 93% consensus (round 2)
- Investigating losses and unauthorised disclosures of personal data - 83% consensus (round 1)

4.7. Lawfulness: processing data has to be done for a specific purpose that the user has agreed to and has to match up with how it is described - 87% consensus (round 1)

4.8. Purpose limitations: data is to be used for a specific purpose that the user has been made aware of through explicit consent – 83% consensus (round 3)

4.9. Data minimisation: review what data you have and why. Only capture the minimum amount of data you need – 87% consensus (round 2)

4.10. Data accuracy: make sure that the data is accurate and ideally stored in a way that allows the user to update or delete the data themselves (securely) – 80% consensus (round 1)

4.11. Storage limitations: data that is no longer required should be removed. If kept for longer than needed data should be pseudonymised to protect user's identity – 97% consensus (round 1)

4.12. Integrity: processors should protect user data against unlawful processing or loss. Ideally having encryption of user data and privacy by design processes – 100% consensus (round 1)

4.13. Would the development of a network of expert's aid standardisation of data collection and data labelling? – 100% consensus (round 1)

4.14. Do we need key opinion leaders to be from multidisciplinary backgrounds to include: (can tick multiple boxes)

- Clinicians – 100% consensus (round 1)
- Computer engineers – 100% consensus (round 1)
- Researchers – 97% consensus (round 1)
- Patients – 83% consensus (round 1)
- Patient advocates (new response round 2) – 87% consensus (round 2)
- Industry (new response round 2) – 97% consensus (round 2)
- Academia (new response round 2) – 93% consensus (round 2)
- Ethicists (new response round 2) – 93% consensus (round 2)
- Politicians (new response round 2) – 33% consensus (round 3)

4.15. Should patients and the public be involved (consulted) in the development of AI systems? – 93% consensus (round 2)

4.16. Should patients and the public be involved in the conception stage of AI systems for surgical training? 30% agreement (round 3)

4.17. Should practical data/findings derived from AI be, as much as possible, open label and made available for research teams, to the benefit of society? – 100% consensus (round 1)

4.18. Should practical data/findings derived from AI be, as much as possible, open label and made available to the public domain? – 90% consensus (round 1)

4.19. Do you agree that organisations would benefit from generic consent forms that ask patients to consent to the use of anonymised images, video and data for audit and research, including AI research? – 97% consensus (round 1)

4.20 Once ethics approval is given, who should give consent for collecting data? (Can tick multiple boxes)

- The hospital/Trust organisation – 43% consensus (round 3)
- The patient – 93% consensus (round 1)
- The surgeon – 83% consensus (round 3)
- The NHS/equivalent 'higher' regulatory body – 13% consensus (round 3)

4.21 Do you agree there should be guidance on data aggregation strategies when data is combined? – 93% consensus (round 1)

4.22 Are there ethical issues with reproducibility of AI (opacity of AI thinking in deep neural networks)? – 87% consensus (round 1)

4.23 Do you agree that ethical issues regarding the reproducibility of AI are reduced in a laboratory training environment, as there is no direct patient involvement in a lab training environment? – 93% consensus (round 2)

4.24. Do you agree that before embarking on developing an educational training platform, it should be clear from the onset who will be responsible in case harm is caused by the platform? – 97% consensus (round 1)

4.25. Do you agree that real-time automated performance feedback in training, driven by AI, is viable? – 93% consensus (round 1)

4.26. Do you agree that real-time automated performance feedback in training, driven by AI, is ethical? – 93% consensus (round 1)

4.27. If AI utilised in training identifies a gap in knowledge or skills, should there be a remediation program developed and available for the trainee? – 97% consensus (round 1)