Supplementary material

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.

Section 1: General questions (consensus levels in brackets):

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)