The Knee 51 (2024) 120–129

Contents lists available at ScienceDirect

The Knee

journal homepage: www.elsevier.com/locate/thekne

Can ChatGPT make surgical decisions with confidence similar to experienced knee surgeons?

Omar Musbahi [a,*], Martine Nurek [b], Kyriacos Pouris [a], Martinique Vella-Baldacchino [a], Alex Bottle [c], Caroline Hing [d], Olga Kostopoulou [b,e], Justin P. Cobb [a], Gareth G. Jones [a]

[a] MSk Lab, Sir Michael Uren Hub, Imperial College London, London, UK
[b] Department of Surgery and Cancer, Imperial College London, London, UK
[c] School of Public Health, Imperial College London, London, UK
[d] St George's University Hospitals NHS Foundation Trust, London, UK
[e] Institute of Global Health Innovation, Imperial College London, London, UK

ARTICLE INFO

Article history:
Received 12 March 2024
Revised 4 August 2024
Accepted 15 August 2024

Keywords:
Natural language processing
Knee arthroplasty
Decision making
Artificial intelligence

ABSTRACT

Background: Unicompartmental knee replacements (UKRs) have become an increasingly attractive option for end-stage single-compartment knee osteoarthritis (OA). However, there remains controversy in patient selection. Natural language processing (NLP) is a form of artificial intelligence (AI). We aimed to determine whether general-purpose open-source natural language programs can make decisions regarding a patient's suitability for a total knee replacement (TKR) or a UKR and how confident AI NLP programs are in surgical decision making.
Methods: We conducted a case-based cohort study using data from a separate study, where participants (73 surgeons and AI NLP programs) were presented with 32 fictitious clinical case scenarios that simulated patients with predominantly medial knee OA who would require surgery.

Using the overall UKR/TKR judgments of the 73 experienced knee surgeons as the gold standard reference, we calculated the sensitivity, specificity, and positive predictive value of AI NLP programs to identify whether a patient should undergo UKR.
Results: There was disagreement between the surgeons and ChatGPT in only five scenarios (15.6%). With the 73 surgeons' decision as the gold standard, the sensitivity of ChatGPT in determining whether a patient should undergo UKR was 0.91 (95% confidence interval (CI): 0.71 to 0.98). The positive predictive value for ChatGPT was 0.87 (95% CI: 0.72 to 0.94). ChatGPT was more confident in its UKR decision making (surgeon mean confidence = 1.7, ChatGPT mean confidence = 2.4).
Conclusions: It has been demonstrated that ChatGPT can make surgical decisions, and exceeded the confidence of experienced knee surgeons with substantial inter-rater agreement when deciding whether a patient was most appropriate for a UKR.

* Corresponding author at: MSk Lab, Sir Michael Uren Hub, Imperial College London, 86 Wood Ln, London W12 0BZ, UK.
E-mail address: o.musbahi19@imperial.ac.uk (O. Musbahi).

https://doi.org/10.1016/j.knee.2024.08.015
0968-0160/© 2024 IMPERIAL COLLEGE LONDON. Published by Elsevier B.V.

1. Introduction

Knee osteoarthritis (OA) affects an estimated 250 million individuals worldwide and is a leading cause of disability [1]. In 2019, OA was accountable for 18.9 million global years lived with disability (YLD), with knee OA responsible for 60.9% according to the Global Burden of Disease Study 2019 [2]. Despite optimal lifestyle modifications, analgesia, and conservative therapy, knee OA is progressive, and patients may eventually require joint replacement surgery. In the UK, the National Joint Registry of England and Wales (NJR) recorded 198,504 primary total knee replacements (TKRs) and 30,976 unicompartmental knee replacements (UKRs) from the 1 January 2020 to 31December 2022 and this is expected to increase [3].

National Institute of Clinical Excellence (NICE) have recommended UKR as a bone and ligament-sparing alternative to TKR [4,5], with a number of advantages: shorter operative time, faster recovery with shorter hospital stays and improved gait [6–8]. When using an implant with good survivorship and long-term follow up, UKR has a similar implant survival rate to TKR and has been shown to be more cost-effective [6,9].

However, there remains controversy in patient selection [4]. Since the description of the original indications by Kozinn and Scott (isolated medial or lateral compartment OA, age less than 60 years, weight less than 82 kg), various authors have proposed widening these indications, with excellent patient outcomes [10–13]. With these expanded indications, 47% of patients eligible for a TKR may be candidates for a UKR [14]. However, the procedure still only accounts for approximately 10% of all TKRs performed in the major joint registries [15–17]. This suggests that some orthopaedic surgeons, who would not be considered high-caseload unicompartmental knee surgeons (>20% practice UKA or >10 UKA/year [18]), may not be able to identify suitable patients. The decision is not without consequence; it has significant implications for both healthcare costs and patient outcomes. [6,9,19].

In the 1950 s, English mathematician and computer science pioneer Alan Turing posed the question, "can machines think?" in his paper "Computing machinery and intelligence" [20]. However, not even Turing could have predicted the extent to which artificial intelligence (AI) and natural language processing (NLP) has revolutionised the digital healthcare landscape in the last decade.

Since the landmark study by Vaswani et al., which introduced the transformer architecture for sequence-to-sequence modelling in NLP, there has been an expansion of the role of NLP in healthcare outcomes and orthopaedics [21,22]. NLP has been used to extract outcome data from unstructured electronic health record (EHR) data and to classify radiology reports in orthopaedic trauma for the presence of injuries [22,23]. Recently, there has been increased interest in open access NLP models with the release of ChatGPT (OpenAI, CA, USA) on 30 November 2022 [24]. With more than 1 million users in the first week of ChatGPT, open-access NLP models are predicted to revolutionise many industries, including healthcare [25].

In this study, we aimed to determine whether a general-purpose 'untrained' open-source natural language program can make decisions regarding a patient's suitability for a TKR or a UKR against the current standard 'experienced knee surgeon'. Furthermore, we will compare the degree of confidence that AI NLP assigned to its decisions with the confidence reported by experienced knee surgeons.

2. Materials and methods

2.1. Study design

To measure agreement between open-source general-purpose NLP programs (ChatGPT Model 3.5, Microsoft Bing, and Google Gemini) and experienced knee surgeons in determining whether a patient should receive a UKR or TKR, we re-analysed data from a separate, ongoing study of surgical decision making conducted by some of the authors (M.N., O.K., O.M., M.V., C.H., J.C.) – 'the parent study', which has completed recruitment. The 'parent study' utilised a case-based approach, where knee surgeons were presented with a series of hypothetical clinical case scenarios and were then asked to make treatment recommendations and to provide a confidence rating for the decision. We compared the responses of the study's participants (surgeons) with the NLP programs on the same case scenarios.

2.2. Knee surgeon cohort

The parent study recruited practicing UK-based knee consultants or knee fellowship-trained surgeons who had experience in performing both TKR and UKR procedures. Participants were recruited via an e-mail distributed to members of the British Association for Surgery of the Knee (BASK) to ensure that only orthopaedic surgeons with a special interest in knee surgery participated. From the parent study data, we sampled the data from 73 knee surgeons who were experienced in both TKR and UKR procedures from the UK. They all met the criteria for a medium/high caseload UKR (>10 per year) [18].

2.3. Open-Source NLP programs

We trialled three commonly used open-source general-purpose NLP programs: ChatGPT model 3.5, Microsoft Bing, Google Gemini. ChatGPT is an general-purpose open-source NLP model, which employs a variant of the transformer architecture introduced by Vaswani et al. called the Generative Pre-trained Transformer (GPT) [21,26]. The transformer model is a neural network architecture that utilises self-attention mechanisms to process input sequences [26]. Self-attention allows the model to consider how each word relates to others in the sentence [26].

2.4. Clinical case scenarios

Thirty-two unique and fictitious clinical case scenarios were created for the ‘parent study’, aiming to simulate patients with knee OA who would potentially require surgery. The scenarios were designed to resemble patients seen in clinical practice by an orthopaedic knee surgeon. Each scenario comprised subject demographics, medical history, physical examination findings, and radiographic images. The following patient characteristics were held constant across scenarios: history findings (activity demand, no preference between UKR and TKR, no history of rheumatoid inflammatory disease), examination findings (normal medial and laral collateral ligaments, normal range of movement), radiographic findings (extent of wear patch on anteroposterior and Rosenberg views showing Ahlback grade III–IV medial tibiofemoral joint, equivalent to Kellgren–Lawrence grade 3–4, no loss of space in the lateral tibiofemoral compartment, no further information can be determined from skyline views), and MRI findings (intact posterior cruciate ligament, severe chondral damage on the medial femoral condyle and medial tibial plateau, mild cartilage fibrillation on the lateral femoral condyle and medial patellar facet).

The following variables were manipulated in each scenario: age, body mass index (BMI), location of pain, American Society of Anesthesiologists (ASA) grade, and anterior cruciate ligament (ACL) status. These variables were selected for manipulation as they are significant factors in the decision-making process for surgeons when choosing between UKR and TKR, as identified in the literature [27–41]. Figue 1 illustrates an example scenario provided to the surgeons.

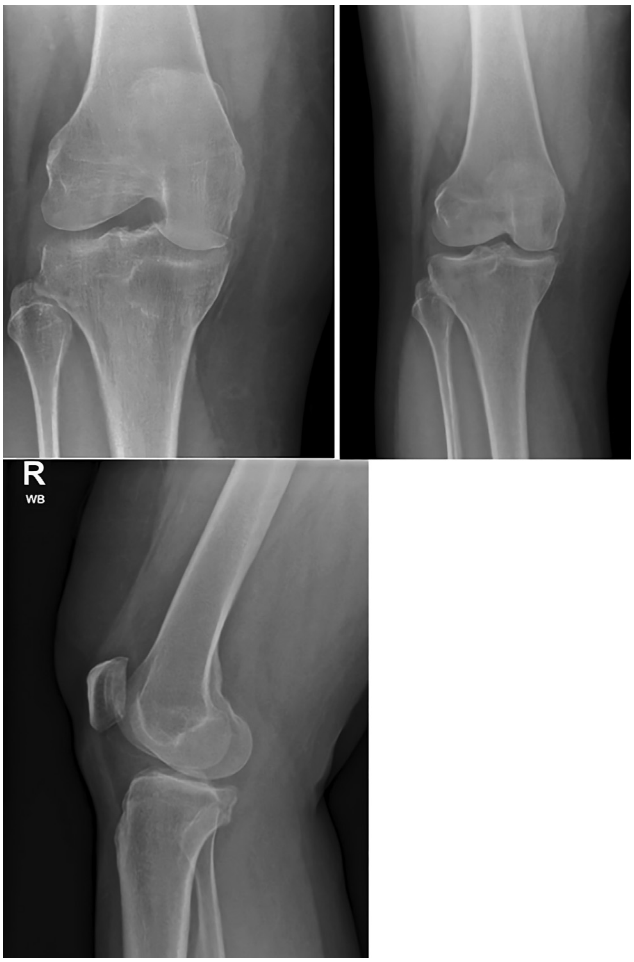

Figure 1. Scenario provided to the surgeon cohort. Patient Y (aged 48 years, body mass index 19 kg/m^2) presents with generalised knee pain. Their ASA Physical Classification Score is 3 (severe systemic disease). On examination, anterior cruciate ligament is abnormal.

2.5. Procedure

Each surgeon was presented with 32 clinical scenarios in two different sessions (example Figure 1) in a random order via an online survey hosted by Qualtrics^{TM} (SAP, UT, USA);. At the end of each scenario, surgeons were asked: “Which surgery would you recommend for this patient?”, responding on an 11-point scale (range −5 to +5), with −5 = “definitely TKR” to +5 = “definitely medial UKR”, and 0 = “undecided”.

To evaluate the decision making and confidence of open-source NLP programs (ChatGPT model 3.5, Google Gemini, Microsoft Bing) compared with the human surgeons (the ground truth), in the same 32 fictitious scenarios, these scenarios were input into the open-source NLP programs.

As with the surgeons, each NLP open-source program was asked: *“In this scenario, would you recommend Total Knee Replacement or Unicompartmental Knee Replacement, and with what degree of confidence on an 11-point scale, with −5 representing ’Definitely TKR,’ +5 representing ’Definitely UKR,’ and 0 representing ’Undecided’?”*. An example ChatGPT scenario and response can be seen in Figure 2.

After trialling, two of the three open-source NLP programs (Google Gemini and Bing) were consistently unable to provide a recommendation between a UKR and TKR for each scenario. Similarly, in the scenarios where the recommendation was provided, the confidence rating was not provided, making it challenging to compare the results with the orthopaedic knee surgeon results. Microsoft Bing was not able to provide either a recommendation or a confidence rating in any of the 32 scenarios. Google Gemini was able to provide a recommendation in decision and confidence rating in three of the 32 scenarios only. Hence, only ChatGPT (model 3.5) was used for the following study.

The scenarios were input into ChatGPT three times by two members of the study group (O.M. and K.P.) on different days, yielding six ChatGPT ‘runs’/trials per scenario. ChatGPT data collection was conducted without prior knowledge of the surgeon cohort results and recorded separately. ChatGPT cannot process radiographic images and thus the radiographic report, determined via consensus of two FRCR (Fellowship of Royal College of Radiology Examination) qualified radiologists, was used.

2.6. Statistical analysis

Confidence ratings (−5 = “Definitely TKR” and +5 = “Definitely UKR”) were averaged per scenario, for the surgeon cohort and the ChatGPT cohort, and used to determine each cohort’s overall decision per scenario (UKR if >0, TKR if <0, undecided if 0). Sensitivity, specificity, and the positive predictive value of ChatGPT to accurately predict surgeons’ decisions to perform UKR were calculated using a two by two contingency table, together with 95% confidence intervals (CIs). Cohen’s Kappa correlation coefficient was used to assess agreement in these per-scenario decisions. For the mean confidence ratings of the treatment decision, differences were compared using an independent *t*-test, whilst a two-way mixed effects average raters

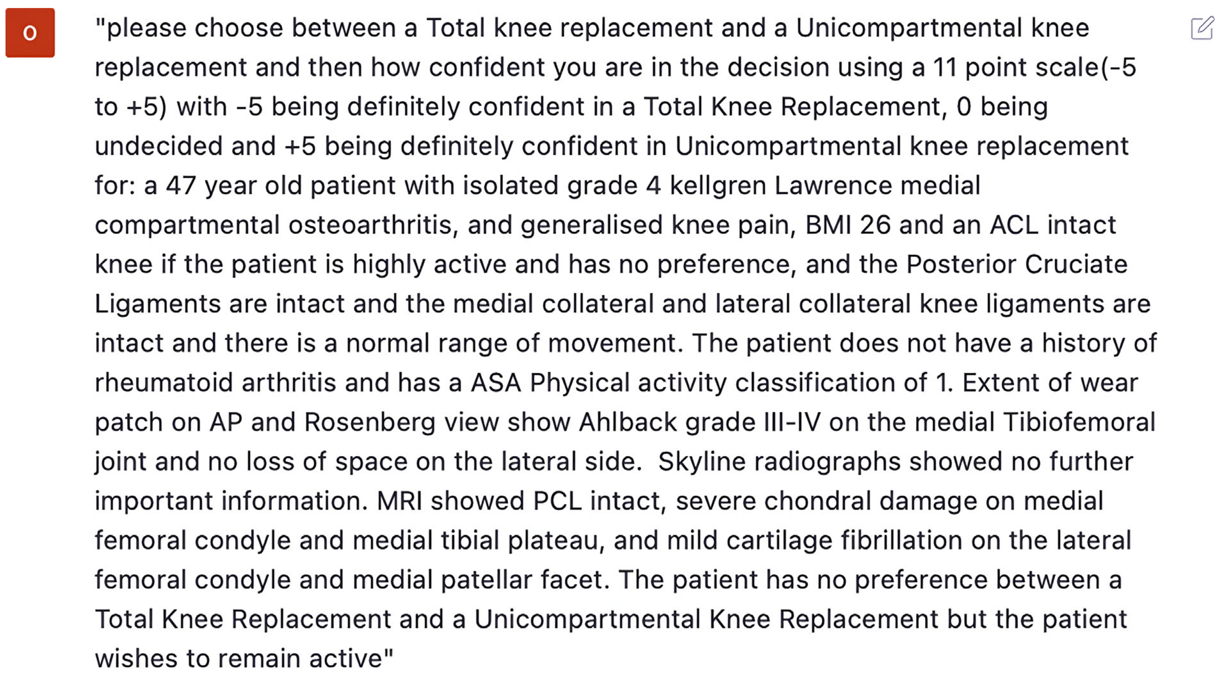

Figure 2. (a) Example scenario provided to the ChatGPT. (b) Answer provided by ChatGPT for scenario question in (a).

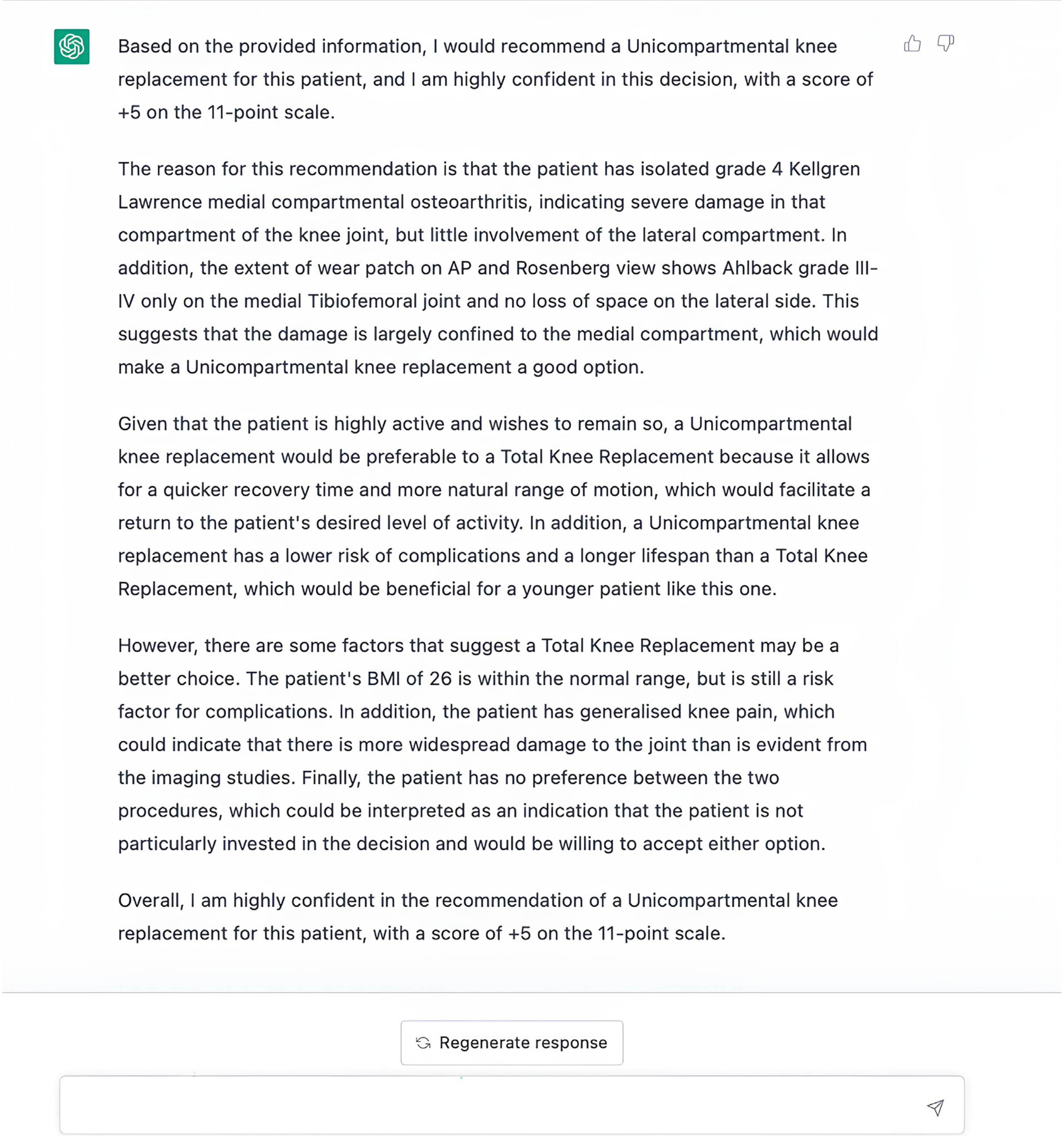

Fig. 2 (*continued*)

intra-class correlation coefficient (ICC) was used to assess agreement between the two groups (surgeons vs. AI). The Kappa Fleiss test was also calculated to determine inter-rater reliability for the surgeon group as a cohort and the inter-rater reliability of the AI cohort. All analysis was performed using R statistical software v10 [42].

2.7. Ethics

Ethical approval was granted by Imperial College Research and Ethics Committee (ICREC number 21IC7201).

3. Results

3.1. Sensitivity and specificity of ChatGPT

ChatGPT responses were compared with the surgeon's responses (reference standard), the sensitivity for ChatGPT in determining whether a patient should undergo UKR was 0.91 (95% CI: 0.71 to 0.98) and the specificity was 0.70 (95% CI: 0.39 to 0.93). The accuracy was 0.84 (95% CI: 0.67 to 0.94). The positive predictive value for ChatGPT was 0.87 (95% CI: 0.72 to 0.94).

3.2. UKR vs. TKR

Table 1 shows the number and proportion of times that TKR, UKR, and 'undecided' were chosen per scenario.

There was disagreement between the knee surgeon cohort and ChatGPT group in five scenarios (15.6%) (Table 1). There was disagreement between the moderate volume surgeons (10–29 UKRs per year) and high-volume surgeons (>30 UKRs per year) in four cases (12.5%). Overall, the surgeon cohort indicated that UKR was most appropriate in 22 scenarios (69%). Similarly, ChatGPT suggested UKR in 23 scenarios (72%). As a cohort, neither the surgeon group nor ChatGPT responses were "undecided" in the 32 scenarios. The Kappa Fleiss statistic was 0.16 ($P<0.05$) for the surgeon cohort, indicating low agreement among the surgeons' decisions. The Kappa Fleiss statistic for the AI cohort was 0.49 indicating moderate agreement among the AI response.

The surgeon cohort and ChatGPT showed significant agreement in their TKR/UKR decisions (Table 1 and Figure 3). The Cohen's kappa coefficient was 0.63 (95% CI: 0.31 to 0.83) for determining 'UKR' or 'TKR' indicating substantial agreement between the AI group and the surgeon group [43].

3.3. Confidence in decision making

Figure 3 presents the mean confidence and 95% confidence interval for each scenario for surgeon cohort vs. ChatGPT. ChatGPT was more confident when deciding on UKR (ChatGPT mean confidence = 2.4 vs. surgeon cohort mean confidence = 1.7). ChatGPT was also more confident when deciding on TKR (ChatGPT mean confidence = −2.0 vs. surgeon cohort mean confidence = −1.1).

The per-scenario mean difference and independent samples t-test result is shown in Table 1.

There was significant intra-rater correlation between the two cohorts for the 32 scenarios with average raters' intra-rater correlation coefficient of 0.66 (95% CI: 0.31 to 0.83, $P<0.05$).

4. Discussion

To our knowledge, this is the first study to investigate the ability of an untrained (non-task-specific) AI algorithm to make a therapeutic decision against the medical gold standard. In addition, this is the only study that compared the confidence of an AI model in its decision making with the confidence of human medical practitioners.

Beard et al. investigated the decision making of four surgeons in selecting TKR or UKR in 140 patients with end-stage medial knee OA [44]. Beard et al. found a variation in decision making of up to 59% for four surgeons in deciding between UKR and TKR [44]. In our study, there was an agreement of 84.4% in the 32 scenarios between the surgeon and the AI cohort.

From the three open-source general-purpose NLP programs (ChatGPT, Microsoft Bing, Google Gemini), only ChatGPT was able to provide a decision between UKR and TKR and provide a confidence rating. Overall, ChatGPT was more confident in its UKR decision making than the knee surgeon cohort (surgeon mean confidence in UKR=1.7 v.s ChatGPT mean confidence in UKR=2.4). Clinician overconfidence bias is a recognised phenomenon, in which clinicians overestimate the accuracy of their clinical judgements [45]. Importantly, high confidence in an judgement can override diagnostic support and determine both the process and outcome of the clinical diagnostic process, leading to more biased evaluations [46]. For example, in knee surgery, a TKR surgeon may have a fixed opinion towards performing a TKR. In this study, ChatGPT appeared to be more confident than the knee surgeon cohort suggesting that AI may be at an even higher risk of overconfidence bias, and clearly needs to be examined in more detail before deploying these algorithms in clinical practice.

This study is not without limitations. We compared 73 knee surgeons' responses with only six responses by ChatGPT. Similarly, ChatGPT model 3.5 cannot process radiographic images and hence we relied on the radiograph report that was determined via consensus of two FRCR (Fellowship of Royal College of Radiology Examination) qualified radiologists, and the accuracy of this report may be a key factor in the observed results. Each surgeon reviewed each scenario once, hence we were unable to calculate the intra-rater repeatability. Similarly, we only looked at a few variables that contributed to the decision making (BMI, location of knee pain, ASA score, ACL status), while keeping other important factors such as the limb alignment, physical examination, radiographic outcomes constant. Hence, these fictional cases are not fully representative of real-life cases requiring complex decision making.

With the availability of an open-source NLP such as ChatGPT, patients can ask and receive a medical decision for their symptoms and more importantly, we have shown that ChatGPT can also provide a confidence rating for its decision, simu-

Table 1
Results of 73 surgeons and the six ChatGPT responses, confidence scores and overall decision for each scenario.

	Number of surgeons choosing TKR	Number of times AI chose TKR	Proportion of surgeons choosing TKR (%)	Proportion of AI choosing TKR (%)	Number of surgeons choosing UKR	Number of times AI chose UKR	Proportion of surgeons choosing UKR (%)	Proportion of AI choosing UKR (%)	Number of surgeons choosing undecided	Number of times AI chose undecided	Proportion of surgeons choosing undecided (%)	Proportion of AI choosing undecided (%)	Surgeon mean confidence	AI mean confidence	Difference in confidence (*P*)	Overall surgeon decision	Overall AI decision	Agreement between the two cohorts
Scenario 1	5	0	7	0	63	6	86	100	5	0	7	0	3.4	3.0	0.4 (*P*=0.70)	UKR	UKR	Yes
Scenario 2	8	0	11	0	62	6	85	100	3	0	4	0	3.2	3.0	0.3 (*P*=0.83)	UKR	UKR	Yes
Scenario 3	33	0	45	0	34	6	47	100	6	0	8	0	0.1	2.8	2.7 (*P*=0.06)	UKR	UKR	Yes
Scenario 4	39	6	53	100	33	0	45	0	1	0	1	0	-0.1	-3.0	2.9 (*P*=0.07)	TKR	TKR	Yes
Scenario 5	1	0	1	0	65	6	89	100	7	0	10	0	3.5	3.3	0.2 (*P*=0.82)	UKR	UKR	Yes
Scenario 6	6	0	8	0	63	6	86	100	4	0	5	0	3.4	3.0	0.4 (*P*=0.71)	UKR	UKR	Yes
Scenario 7	3	0	4	0	65	4	89	67	5	2	7	33	3.3	2.0	1.3 (*P*=0.15)	UKR	UKR	Yes
Scenario 8	35	1	48	17	36	0	49	0	2	5	3	83	-0.2	-0.5	0.3 (*P*=0.92)	TKR	TKR	Yes
Scenario 9	31	0	42	0	39	3	53	50	3	3	4	50	0.5	1.5	1.0 (*P*=0.52)	UKR	UKR	Yes
Scenario 10	16	6	22	100	56	0	77	0	1	0	1	0	2.6	-3.3	5.9 (*P*<0.05)	UKR	TKR	No*
Scenario 11	12	2	16	33	58	4	79	67	3	0	4	0	2.3	1.8	0.48 (*P*=0.71)	UKR	UKR	Yes
Scenario 12	14	0	19	0	58	4	79	67	1	2	1	33	2.4	2.0	0.4 (*P*=0.74)	UKR	UKR	Yes
Scenario 13	10	0	14	0	60	6	82	100	3	0	4	0	2.8	3.0	0.2 (*P*=0.87)	UKR	UKR	Yes
Scenario 14	42	1	58	17	28	4	38	67	3	1	4	17	-0.8	1.7	2.4 (*P*=0.15)	TKR	UKR	No*
Scenario 15	36	0	49	0	28	6	38	100	9	0	12	0	-0.6	3.0	3.6 (*P*<0.05)	TKR	UKR	No*
Scenario 16	5	0	7	0	62	4	85	67	6	2	8	33	3.3	2.0	1.3 (*P*=0.18)	UKR	UKR	Yes
Scenario 17	27	0	37	0	38	6	52	100	8	0	11	0	0.5	4.0	3.5 (*P*<0.05)	UKR	UKR	Yes
Scenario 18	31	0	42	0	41	6	56	100	1	0	1	0	0.5	4.0	3.5 (*P*<0.05)	UKR	UKR	Yes
Scenario 19	18	0	25	0	46	4	63	67	9	2	12	33	1.5	2.0	0.5 (*P*=0.7)	UKR	UKR	Yes
Scenario 20	33	2	45	33	39	0	53	0	1	4	1	67	0.3	-1.0	1.3 (*P*=0.43)	UKR	TKR	No*
Scenario 21	47	6	64	100	22	0	30	0	4	0	5	0	-1.5	-3.0	1.5 (*P*=0.32)	TKR	TKR	Yes
Scenario 22	50	6	68	100	21	0	29	0	2	0	3	0	-1.8	-3.3	1.5 (*P*=0.31)	TKR	TKR	Yes
Scenario 23	30	0	41	0	37	4	51	67	6	2	8	33	0.6	2.0	1.4 (*P*=0.33)	UKR	UKR	Yes
Scenario 24	30	0	41	0	41	6	56	100	2	0	3	0	0.4	3.0	2.6 (*P*=0.10)	UKR	UKR	Yes
Scenario 25	22	0	30	0	39	3	53	50	12	3	16	50	0.7	1.3	0.6 (*P*=0.62)	UKR	UKR	Yes
Scenario 26	57	4	78	67	13	2	18	33	3	0	4	0	-2.7	-1.7	1.0 (*P*=0.50)	TKR	TKR	Yes
Scenario 27	43	3	59	50	25	1	34	17	5	2	7	33	-1.1	-1.0	0.1 (*P*=0.94)	TKR	TKR	Yes
Scenario 28	31	2	42	33	41	4	56	67	1	0	1	0	0.6	1.6	1.0 (*P*=0.59)	UKR	UKR	Yes
Scenario 29	17	0	23	0	51	6	70	100	5	0	7	0	1.7	2.8	1.1 (*P*=0.39)	UKR	UKR	Yes
Scenario 30	36	4	49	67	36	2	49	33	1	0	1	0	-0.1	-1.0	0.9 (*P*=0.61)	TKR	TKR	Yes
Scenario 31	26	0	36	0	40	6	55	100	7	0	10	0	0.5	2.7	2.2 (*P*=0.15)	UKR	UKR	Yes
Scenario 32	54	2	74	33	15	3	21	50	4	1	5	17	-2.3	0.3	2.6 (*P*<0.06)	TKR	UKR	No*

Shaded columns denote surgeon data, unshaded columns describe artificial intelligence (AI) data. TKR, total knee replacement; UKR, unicompartmental knee replacement.

Shaded columns denote surgeon data, unshaded columns describe artificial intelligence (AI) data. TKR, total knee replacement; UKR, unicompartmental knee replacement.
*Scenarios where there was disagreement between overall surgeon decision and the overall AI surgeon decision.

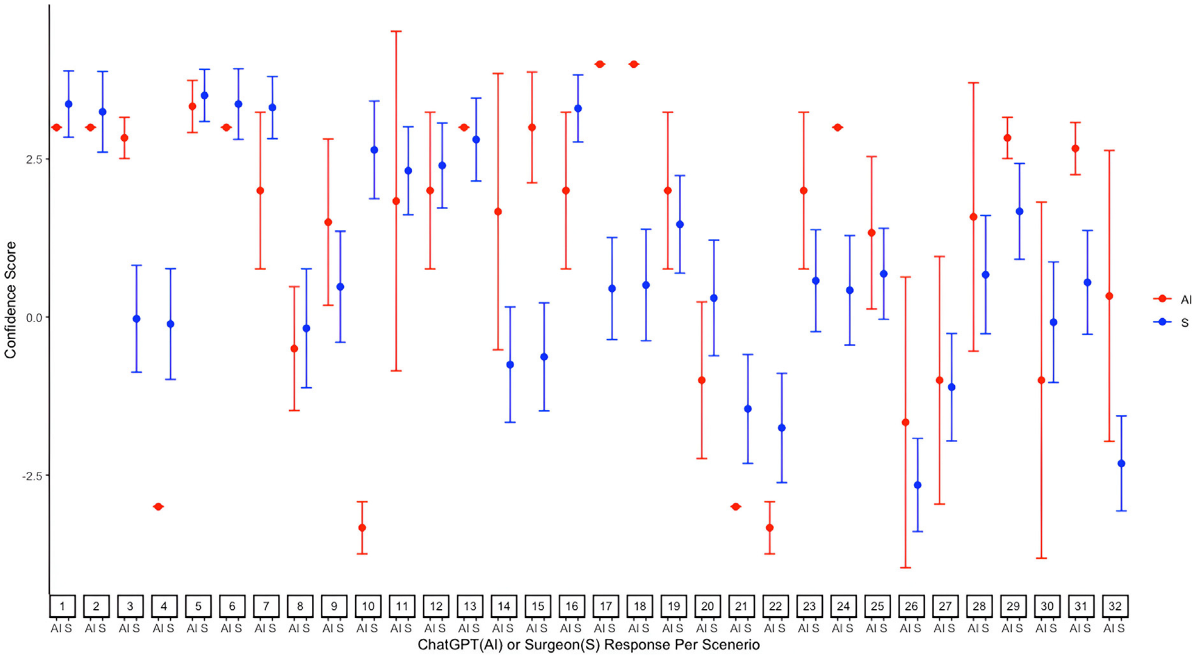

Figure 3. Graph showing the mean confidence of ChatGPT (AI) and the knee surgeon (S) with the associated 95% confidence Interval for each scenario.

lating human surgical decision making. With studies showing ChatGPT achieving a pass in the United States Medical Licensing Exams (USMLE) without specialised input from human trainers, it is not inconceivable to imagine a future healthcare model where AI models will provide a greater assistive role in the field of medical decision-making [47–51]. However, as we have demonstrated in this study, AI models are not immune to incorrect decision making and hence relying solely on open-source NLP advice can lead to inappropriate treatment recommendations. We have not investigated the direct clinical benefit and cost savings in healthcare decision making by utilising AI NLP programs, other ethical issues, or the complex role of 'confidence' in decision making by AI NLP programs, but these are fertile areas for future studies.

5. Conclusion

ChatGPT has a high positive predictive value in deciding between UKR and TKR for participants with surgery-indicated knee OA. This untrained general-purpose NLP program approximated the decision making, and exceeded the confidence, of experienced knee surgeons with acceptable inter-rater agreement. Greater transparency and publicly available information on training sources for open-source NLP models should be made available to facilitate their potential use in healthcare.

Funding

O.M. is sponsored by NIHR grant (NIHR ID 302632). M.N. was supported by an NIHR Imperial Patient Safety Translational Research Centre grant (award number PSTRC-2016-004). All surgeons who took part were reimbursed £50 for their time (NIHR Imperial PSTRC grant, award number PSTRC-2016-004). A.B.'s Unit at Imperial is affiliated with the National Institute of Health Research (NIHR) Imperial Patient Safety Translational Research Centre. The NIHR Imperial Patient Safety Translational Centre is a partnership between the Imperial College Healthcare NHS Trust and Imperial College London.

CRediT authorship contribution statement

Omar Musbahi: Writing – review & editing, Writing – original draft, Visualization, Supervision, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Martine Nurek:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Kyriacos Pouris:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Martinique Vella-Baldachino:** Writing – review & editing, Writing – original draft, Visualization, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Alex Bottle:** Writing – review & editing, Visualization, Supervision, Resources, Investigation, Funding acquisition, Data curation. **Caroline Hing:** Writing – review & editing, Visualization, Resources, Investigation, Funding acquisition, Data curation. **Olga Kostopoulou:** Writing – review & editing, Visualization, Supervision, Resources, Methodology, Investigation, Data curation, Conceptualization. **Justin P. Cobb:** Writing – review & editing, Writing – original draft, Visualization, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Gareth G. Jones:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors gratefully acknowledge infrastructure support from the NIHR Academy, NIHR Imperial Patient Safety Translational Research Centre, the NIHR Imperial Biomedical Research Centre, and the NIHR Clinical Research Network. The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care. Members of the *UNITES* Consortium are also acknowledged for supporting this research: Barry Andrews, Tony Smith, P. J. Walmsley, David W. Elson, Alex Anderson, Chloe E. H. Scott, Alasdair J. A. Santini, Benedict J. A. Lankester, J. William Tice, Reza Mansouri, Arjuna Imbuldeniya, Tarek Boutefnouchet, Nathanael Ahearn, Stefan Bajada, James R. A. Smith, Pregash Ellapparadja, Alan R. Norrish, Sean O'Leary, Jon Campion, Benjamin V. Bloch, Ricardo J. Pacheco, Simon B. Barton, Adrian Cassar-Gheiti, David Selvan, Jonathan Phillips, Sushrut Kulkarni, Randeep S. Aujla, Philip G. Turner, Amit Patel, Rahul S. Kotwal, Ashim Mannan, Rahul Bhattacharyya, Ahmed Mabrouk Aaron Biing Yann, Keshav Mathur, Muhammad Adeel Akhtar, Robert William Walker, Rakesh Kucheria, A. D. Liddle, Lebur Rohman, Ravikumar Pydisetty, Manish Divekar, Manish Kiran, David Houlihan-Burne, N. D. Rossiter, Sanjeev Agarwal, N. J. London, A. D. Toms, Zuhair Nawaz, Phil Hopgood, Ghias Bhattee, Moez Zeiton, Khalid Al-DadahHywel Davies, Oliver S. Schindler, H. B. Waterson, S. P. White, A. J. Kelly, Yuvraj Agrawal, Christopher Wilson, Nicholas E. Ohly, Andrew Lavender, Morgan Bayley, Fazal Ali, Nivraj Singh Bhamber, Tarique Parwez, Christopher Buckle, Zameer Shah, Jeremy Rushbrook, Damon Simmons, Amit Bishnoi, Reshid Berber, Richard Parkinson, D. Prakash, James R. D. Murray, Sujit Agarwal.

References

[1] Vos T, Flaxman AD, Naghavi M, Lozano R, Michaud C, Ezzati M, et al. Years lived with disability (YLDs) for 1160 sequelae of 289 diseases and injuries 1990–2010: A systematic analysis for the global burden of disease study 2010. Lancet 2012;380:2163–96.
[2] Long H, Liu Q, Yin H, Wang K, Diao N, Zhang Y, et al. Prevalence trends of site-specific osteoarthritis from 1990 to 2019: Findings from the global burden of disease study 2019. Arthritis Rheumatol 2022;74:1172–83. doi: https://doi.org/10.1002/art.42089.
[3] National Joint Registry for England W, Northern Ireland and Wales. NJR 20th Annual Report 2023 Secondary NJR 20th Annual Report 2023 Available at: https://reports.njrcentre.org.uk/Portals/0/PDFdownloads/NJR%2020th%20Annual%20Report%202023.pdf [last accessed March 2024].
[4] Mittal A, Meshram P, Kim WH, Kim TK. Unicompartmental knee arthroplasty, an enigma, and the ten enigmas of medial UKA. J Orthop Traumatol 2020;21:15. doi: https://doi.org/10.1186/s10195-020-00551-x.
[5] Excellence NNifHaC. Joint Replacement (primary): hip, knee and shoulder. Secondary Joint Replacement (primary): Hip, knee and shoulder; 2020. Available at: https://www.nice.org.uk/guidance/ng157/resources/joint-replacement-primary-hip-knee-and-shoulder-pdf-66141845322181 [last accessed April 2024].
[6] Liddle AD, Judge A, Pandit H, Murray DW. Adverse outcomes after total and unicompartmental knee replacement in 101,330 matched patients: A study of data from the National Joint Registry for England and Wales. Lancet 2014;384:1437–45. doi: https://doi.org/10.1016/s0140-6736(14)60419-0.
[7] Zuiderbaan HA, van der List JP, Khamaisy S, Nawabi DH, Thein R, Ishmael C. Unicompartmental knee arthroplasty versus total knee arthroplasty: Which type of artificial joint do patients forget? Knee Surg Sports Traumatol Arthrosc 2017;25:681–766. doi: https://doi.org/10.1007/s00167-015-3868-1.
[8] Wiik AV, Manning V, Strachan RK, Amis AA, Cobb JP. Unicompartmental knee arthroplasty enables near normal gait at higher speeds, unlike total knee arthroplasty. J Arthroplasty 2013;28:176–8. doi: https://doi.org/10.1016/j.arth.2013.07.036.
[9] Beard DJ, Davies LJ, Cook JA, MacLennan G, Price A, Kent S. The clinical and cost-effectiveness of total versus partial knee replacement in patients with medial compartment osteoarthritis (TOPKAT): 5-year outcomes of a randomised controlled trial. Lancet 2019;394:746–56. doi: https://doi.org/10.1016/s0140-6736(19)31281-4.
[10] Kozinn SC, Scott R. Unicondylar knee arthroplasty. J Bone Joint Surg Am 1989;71:145–50.
[11] Pandit H, Jenkins C, Gill H, Barker K, Dodd C, Murray D. Minimally invasive Oxford phase 3 unicompartmental knee replacement: Results of 1000 cases. J Bone Joint Surg Br 2011;93:198–204.
[12] Campi S, Tibrewal S, Cuthbert R, Tibrewal SB. Unicompartmental knee replacement – Current perspectives. J Clin Orthop Trauma 2018;9:17–23. doi: https://doi.org/10.1016/j.jcot.2017.11.013.
[13] Rodríguez-Merchán EC, Gómez-Cardero P. Unicompartmental knee arthroplasty: Current indications, technical issues and results. EFORT Open Rev 2018;3:363–73. doi: https://doi.org/10.1302/2058-5241.3.170048.
[14] Hamilton TW, Rizkalla JM, Kontochristos L, Marks BE, Mellon SJ, et al. The interaction of caseload and usage in determining outcomes of unicompartmental knee arthroplasty: A meta-analysis. J Arthroplasty 2017;32:3228–3237.e2.
[15] Ben-Shlomo Y BA, Boulton C. The National Joint Registry 19th Annual Report 2022. Available at: https://www.ncbi.nlm.nih.gov/books/NBK587514/ [last accessed March 2024].
[16] National Joint Registry for England W, Northern Ireland and Wales. 16th Annual Report 2019 for the National Joint Registry for England, Wales, Northern Ireland and Wales and the Isle of Man. 2020.
[17] Henkel C, Mikkelsen M, Pedersen AB, Rasmussen LE, Gromov K, Price A. Medial unicompartmental knee arthroplasty: increasingly uniform patient demographics despite differences in surgical volume and usage-a descriptive study of 8,501 cases from the Danish Knee Arthroplasty Registry. Acta Orthop 2019;90:354–439. doi: https://doi.org/10.1080/17453674.2019.1601834.
[18] Mohammad HR, Judge A, Murray DW. A matched comparison of cementless unicompartmental and total knee replacement outcomes based on the National Joint Registry for England, Wales, Northern Ireland and the Isle of Man. Acta Orthop 2022;93:478–87. doi: https://doi.org/10.2340/17453674.2022.2743.

[19] Cobb JP. Patient safety after partial and total knee replacement. Lancet 2014;384:1405–2147.
[20] Turing AM. Computing Machinery and Intelligence: Springer, 2009.
[21] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A. Attention is all you need. Adv Neural Inf Process Syst 2017;30.
[22] Bovonratwet P, Shen TS, Islam W, Ast MP, Haas SB, Su EP. Natural language processing of patient-experience comments after primary total knee arthroplasty. J Arthroplasty 2021;36:927–34. doi: https://doi.org/10.1016/j.arth.2020.09.055.
[23] Olthof AW, Shouche P, Fennema EM, Ijpma FFA, Koolstra RHC, Stirler VMA, et al. Machine learning based natural language processing of radiology reports in orthopaedic trauma. Comput Methods Programs Biomed 2021;208:106304. doi: https://doi.org/10.1016/j.cmpb.2021.106304.
[24] OpenAI. ChatGPT: Optimizing Language Models for Dialogue. 2022.
[25] C V. ChatGPT: New AI chat has everyone talking to it. Secondary ChatGPT: New AI chat has everyone talking to it; 2022. Available at: https://www.bbc.com/news/technology-63861322 [last accessed February 2024].
[26] Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners. OpenAI blog 2019;1:9.
[27] Salman LA, Abudalou A, Khatkar H, Ahmed G, Dakin SG, Kendrick B. Impact of age on unicompartmental knee arthroplasty outcomes: a systematic review and meta-analysis. Knee Surg Sports Traumatol Arthrosc 2023;31:986–97. doi: https://doi.org/10.1007/s00167-022-07132-x.
[28] Kyriakidis T, Asopa V, Baums M, Verdonk R, Totlis T. Unicompartmental knee arthroplasty in patients under the age of 60 years provides excellent clinical outcomes and 10-year implant survival: A systematic review. A study performed by the Early Osteoarthritis group of ESSKA-European Knee Associates section. Knee Surg Sports Traumatol Arthrosc 2023;31:922–32. doi: https://doi.org/10.1007/s00167-022-07029-9.
[29] Walter N, Weber J, Kerschbaum M, Lau E, Kurtz S, Alt V. Revision arthroplasty after unicompartimental knee arthroplasty. J Orthop Surg Res 2021;16:666. doi: https://doi.org/10.1186/s13018-021-02767-x.
[30] Hansen EN, Ong KL, Lau E, Kurtz SM, Lonner JH. Unicondylar knee arthroplasty has fewer complications but higher revision rates than total knee arthroplasty in a study of large United States databases. J Arthroplasty 2019;34:1617–25. doi: https://doi.org/10.1016/j.arth.2019.04.004.
[31] Parratte S, Argenson JN, Pearce O, Pauly V, Auquier P, Aubaniac JM. Medial unicompartmental knee replacement in the under-50s. J Bone Joint Surg Br 2009;91:351–6. doi: https://doi.org/10.1302/0301-620x.91b3.21588.
[32] Musbahi O, Hamilton TW, Crellin AJ, Mellon SJ, Kendrick B, Murray DW. The effect of obesity on revision rate in unicompartmental knee arthroplasty: A systematic review and meta-analysis. Knee Surg Sports Traumatol Arthrosc 2021;29:3467–77. doi: https://doi.org/10.1007/s00167-020-06297-7.
[33] Ellis RT, Nettrour JF, Keeney JA. TKA is more durable than UKA for morbidly obese patients: A two-year minimum follow-up study. J Arthroplasty 2021;36:1933–41. doi: https://doi.org/10.1016/j.arth.2020.12.039.
[34] Bonutti PM, Goddard MS, Zywiel MG, Khanuja HS, Johnson AJ, Mont MA. Outcomes of unicompartmental knee arthroplasty stratified by body mass index. J Arthroplasty 2011;26:1149–53. doi: https://doi.org/10.1016/j.arth.2010.11.001.
[35] Kandil A, Werner BC, Gwathmey WF, Browne JA. Obesity, morbid obesity and their related medical comorbidities are associated with increased complications and revision rates after unicompartmental knee arthroplasty. J Arthroplasty 2015;30:456–60. doi: https://doi.org/10.1016/j.arth.2014.10.016.
[36] Emerson Jr RH, Higgins LL. Unicompartmental knee arthroplasty with the oxford prosthesis in patients with medial compartment arthritis. J Bone Joint Surg Am 2008;90:118–22.
[37] Kleeblad LJ, Zuiderbaan HA, Hooper GJ, Pearle AD. Unicompartmental knee arthroplasty: State of the art. J ISAKOS 2017;2:97–107. doi: https://doi.org/10.1136/jisakos-2016-000102.
[38] Pascual-Leone N, Sloan M, Shah VM, Sheth NP. Patients with significant medical comorbidities may benefit from unicompartmental knee arthroplasty over total knee arthroplasty. J Knee Surg 2024;37:79–83. doi: https://doi.org/10.1055/s-0042-1760388.
[39] Goodfellow JW, Kershaw CJ, Benson MK, O'Connor JJ. The Oxford Knee for unicompartmental osteoarthritis. The first 103 cases. J Bone Joint Surg Br 1988;70:692–701. doi: https://doi.org/10.1302/0301-620x.70b5.3192563.
[40] Boissonneault A, Pandit H, Pegg E, Jenkins C, Gill HS, Dodd CA. No difference in survivorship after unicompartmental knee arthroplasty with or without an intact anterior cruciate ligament. Knee Surg Sports Traumatol Arthrosc 2013;21:2480–3246. doi: https://doi.org/10.1007/s00167-012-2101-8.
[41] Cartier P, Sanouiller JL, Grelsamer RP. Unicompartmental knee arthroplasty surgery. 10-year minimum follow-up period. J Arthroplasty 1996;11:782–8. doi: https://doi.org/10.1016/s0883-5403(96)80177-x.
[42] Das BK, Jha DN, Sahu SK, Yadav AK, Raman RK, Kartikeyan M. Introduction to R Software. Concept Building in Fisheries Data Analysis: Springer, 2022. https://doi.org/10.1007/978-981-19-4411-6.
[43] McHugh ML. Interrater reliability: The kappa statistic. Biochem Med (Zagreb) 2012;22:276–82.
[44] Beard DJ, Holt MD, Mullins MM, Malek S, Massa E, Price AJ. Decision making for knee replacement: variation in treatment choice for late stage medial compartment osteoarthritis. Knee 2012;19:886–9. doi: https://doi.org/10.1016/j.knee.2012.05.005.
[45] Croskerry P, Norman G. Overconfidence in clinical decision making. Am J Med 2008;121:S24–9. doi: https://doi.org/10.1016/j.amjmed.2008.02.001.
[46] Kourtidis P, Nurek M, Delaney B, Kostopoulou O. Influences of early diagnostic suggestions on clinical reasoning. Cogn Res Princ Implic 2022;7:103. doi: https://doi.org/10.1186/s41235-022-00453-y.
[47] Abedi V, Razavi SM, Khan A, Avula V, Tompe A, Poursoroush A. Artificial intelligence: A shifting paradigm in cardio-cerebrovascular medicine. J Clin Med 2021;10:5710. doi: https://doi.org/10.3390/jcm10235710.
[48] Saboury B, Morris M, Siegel E. Future directions in artificial intelligence. Radiol Clin North Am 2021;59:1085–95. doi: https://doi.org/10.1016/j.rcl.2021.07.008.
[49] Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol 2017;2:230–43. doi: https://doi.org/10.1136/svn-2017-000101.
[50] Pantanowitz L, Quiroga-Garza GM, Bien L, Heled R, Laifenfeld D, Linhart C, et al. An artificial intelligence algorithm for prostate cancer diagnosis in whole slide images of core needle biopsies: A blinded clinical validation and deployment study. Lancet Digit Health 2020;2:e407–16. doi: https://doi.org/10.1016/s2589-7500(20)30159-x.
[51] Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepano C. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health 2023;2:e0000198. doi: https://doi.org/10.1371/journal.pdig.0000198.