Article

Not peer-reviewed version

Modification and Validation of the System Causability Scale Using AI-Based Therapeutic Recommendations for Urological Cancer Patients: A Basis for the Development of a Prospective Comparative Study

Emily Rinderknecht , Dominik von Winning , Anton Kravchuk , Christof Schäfer , Marco J. Schnabel , Stephan Siepmann , Roman Mayr , Jochen Grassinger , Christopher Goßler , Fabian Pöhl , Peter J. Siska , Florian Zeman , Johannes Breyer , Anna Magdalena Schmelzer , Christian Gilfrich , Sabine. D. Brookman-May , Maximilian Burger , Maximilian Haas , Matthias May *

Posted Date: 16 October 2024

doi: 10.20944/preprints202410.1290.v1

Keywords: Artificial Intelligence Integration; Large Language Models; Multidisciplinary Tumor Boards; Non-inferiority CONCORDIA Trial; SCS; Urological Cancer Treatment; Validation Study; Clinical Decision Support; Artificial Neural Network

Article

Modification and Validation of the System Causability Scale Using AI-Based Therapeutic Recommendations for Urological Cancer Patients: A Basis for the Development of a Prospective Comparative Study

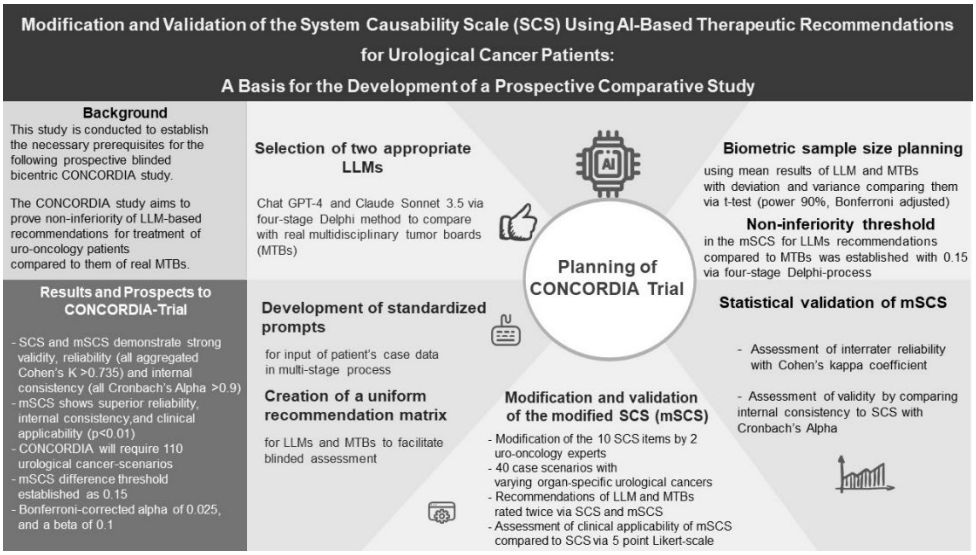
Emily Rinderknecht [1], Dominik von Winning [2], Anton Kravchuk [2], Christof Schäfer [3], Marco J. Schnabel [1], Stephan Siepmann [2], Roman Mayr [1], Jochen Grassinger [4], Christopher Goßler [1], Fabian Pohl [5], Peter J. Siska [6], Florian Zeman [7], Johannes Breyer [1], Anna Schmelzer [2], Christian P. Gilfrich [2], Sabine D. Brookman-May [8], Maximilian Burger [1], Maximilian Haas [1,#] and Matthias May [2,*,#]

[1] Department of Urology, Caritas St. Josef Hospital, University of Regensburg, Regensburg, Germany.
[2] Department of Urology, St. Elisabeth Hospital Straubing, Straubing, Germany.
[3] Department of Radiotherapy, Straubing Hospital Medical Care Centre, Straubing, Germany.
[4] Department of Hematology and Oncology, Straubing Hospital Medical Care Centre, Straubing, Germany.
[5] Department of Radiotherapy, University Hospital Regensburg, Regensburg, Germany.
[6] Department of Internal Medicine III, University Hospital Regensburg, Regensburg, Germany.
[7] Center for Clinical Studies, University Hospital Regensburg, Regensburg, Germany.
[8] Department of Urology, Ludwig-Maximilians-University (LMU), Munich, Germany.
* Correspondence: matthias.may@klinikum-straubing.de
[#] shared senior author position.

Abstract: The integration of artificial intelligence, particularly Large Language Models (LLMs), has the potential to significantly enhance therapeutic decision-making in clinical oncology. Initial studies across various disciplines have demonstrated that LLM based treatment recommendations can rival those of multidisciplinary tumor boards (MTBs); however, such data are currently lacking for urological cancers. This study provides the methodological foundation for the prospective CONCORDIA trial, which will generate this data for the first time. In this preliminary work, we evaluated the proposed measurement tool for the CONCORDIA study -the System-Causability-Scale (SCS) and its modified version (mSCS) - based on recommendations from ChatGPT-4 and an MTB for 40 urological cancer-scenarios. Both scales demonstrated strong validity, reliability (all aggregated Cohen's K > 0.74), and internal consistency (all Cronbach's Alpha > 0.9), with the mSCS showing superior reliability, internal consistency, and clinical applicability ($p<0.01$). Two Delphi processes were used to define the LLMs to be tested in the CONCORDIA study (ChatGPT-4 and Claude 3.5 Sonnet) and to establish the acceptable non-inferiority margin for LLM recommendations compared to MTB recommendations. The forthcoming ethics-approved and registered CONCORDIA non-inferiority trial will require 110 urological cancer scenarios, with an mSCS difference threshold of 0.15, a Bonferroni corrected alpha of 0.025, and a beta of 0.1. Blinded mSCS assessments of MTB recommendations will then be compared to those of the LLMs. In summary, this work establishes the necessary prerequisites prior to initiating the CONCORDIA study and validates a modified score with high applicability and reliability for this and future trials.

Keywords: artificial intelligence integration; large language models; multidisciplinary tumor boards; non-inferiority CONCORDIA trial; SCS; urological cancer treatment; validation study; clinical decision support; artificial neural network

 doi:10.20944/preprints202410.1290.v1



1. Introduction

The increasing integration of artificial intelligence (AI) into healthcare has recently gained considerable attention, sparking widespread discussions, research, and early adoption. This momentum is driven by its significant potential across multiple domains within the medical sector [1–3]. Among the most groundbreaking developments are Generative Pre-trained Transformers (GPTs) and, more broadly, Large Language Models (LLMs), a critical branch of AI and machine learning (ML) These models, powered by advanced ML algorithms, can generate and interpret text without the need for prior linguistic preprocessing, such as traditional Natural Language Processing. Consequently, they hold vast potential for applications in both clinical practice and academic research within the medical field [4,5].

While GPT-2 was introduced in 2018, it was the launch of OpenAI's ChatGPT on November 30, 2022, that catapulted LLMs into mainstream consciousness, driving global interest and adoption [6]. In healthcare, the impact of AI-generated decisions, recommendations, or outcomes is profound, depending on the specific context and application. However, for these models to be safely and effectively integrated into routine clinical practice, several challenges must be addressed. These include ensuring accuracy, transparency, and accountability, alongside managing ethical concerns and maintaining the central role of the human physician, who bears ultimate responsibility for clinical decisions. As a result, the implementation of these models in medicine remains in its early stages, with significant validation still required for widespread adoption.

A recent international survey of 456 urologists revealed that 53% perceived limitations in LLM usability within clinical and academic settings. The most cited concerns included inaccurate responses (45%), lack of specificity (42%), and inconsistent answers (26%) [7]. Nonetheless, 56% of respondents believed that ChatGPT and other LLMs hold potential value for clinical decision-making, with approximately 20% having already incorporated ChatGPT into their workflows [7].

To assess the quality of AI-generated explanations, especially in the context of scientific model development, Holzinger et al. introduced the System Causability Scale (SCS) in 2020 [8]. The SCS quantifies explainability based on responses to 10 questions, each rated on a 5-point Likert scale. Its simplicity and status as a standardized tool make it highly useful for evaluating AI- and LLM-generated explanations [8].

In the survey by Eppler et al., 30% of urologists indicated potential applications for LLMs in selecting appropriate treatment options [7]. Multidisciplinary tumor boards (MTBs) serve as a critical component in delivering high-quality, guideline-based care for urological cancer patients through consensus-driven therapy recommendations [9,10]. However, these often-weekly meetings pose

 doi:10.20944/preprints202410.1290.v1

substantial time management challenges for clinicians, who are already intended to meet both clinical and academic needs. [9–11]. Although a few studies have examined the integration of LLMs into MTBs for non-urological cancers, these investigations also highlight current limitations regarding the safe and effective use of such models [12–21].

The present study serves as a preparatory investigation to lay the groundwork for a prospective trial aimed at determining whether LLM-generated treatment recommendations for genitourinary cancers (GUCs) can match those of an interdisciplinary MTB comprising specialists from urology, oncology, radiotherapy, and nuclear medicine/radiology. This preparatory study has several key objectives: (i) to adapt the SCS to meet the specific needs of the forthcoming study and validate this modified score (mSCS); (ii) to generate treatment recommendations for a representative cohort of GUC patients using both ChatGPT-4 and a real MTB, thereby enabling the design of a robust study protocol and sample size calculation for the planned trial. Ultimately, the primary objective of this study is to establish a newly validated scoring system (mSCS) tailored to the requirements of this and future similar trials.

2. Materials and Methods

2.1. Planned Prospective Trial

The proposed study is a prospective investigation designed to examine whether real MTBs can be equivalently replaced by LLMs. The study, titled "Concordance Study on Urological Tumor Boards and Large-Language-Model Substitutes" (CONCORDIA Study), will seek ethical approval and formal registration with the German Clinical Trials Register (Deutsches Register Klinischer Studien, DRKS). It will be conducted at two German hospitals: St. Josef Medical Center (University of Regensburg) and St. Elisabeth Hospital Straubing. These institutions will provide case scenarios of GUC patients, reflecting the typical distribution of tumor entities handled by their respective MTBs.

The primary aim of the study is to compare the blinded therapeutic recommendations of a real MTB – comprising specialists in urology, oncology, radiation therapy, and nuclear medicine – with the recommendations from two selected, publicly available LLMs. The study follows a non-inferiority design, evaluating the performance of LLMs against the real MTB. Study outcomes will be assessed using a modified score (mSCS).

The following key objectives of this preparatory study will be addressed methodologically in subsequent sections: (1) Selection of two appropriate LLMs for comparison with the MTB; (2) Development of standardized prompts for data input on GUC patients and the creation of a uniform recommendation matrix to ensure blinded assessment; (3) Modification and validation of the newly developed mSCS using a cohort of 40 urological tumor patients across various organ-specific cancers; (4) Biometric sample size planning for the prospective trial, preceded by a moderated Delphi process to determine the acceptable non-inferiority threshold for LLM performance compared to the MTB; and (5) Precise documentation of statistical methods to validate the mSCS and compare results between the groups (MTB vs. LLM).

To streamline this preparatory study, the therapeutic recommendations from the real MTB were compared only with those from the premium version of the top-performing LLM. The second LLM will be tested in the main trial using the non-inferiority threshold derived from this preparatory comparison. Ethical approval for this preparatory study was obtained (UKR-EK-24-3835-104). The methodological details are presented in the following sections.

2.2. Selection of Two Appropriate LLMs for Comparison with the MTB (1)

The selection of LLMs was made through a consensus among the study authors (ER, MH, DvW, AK, CGo, and MM) using a four-stage Delphi method moderated by MS. The process included: Round 1: Identification of all available LLMs and discussion of their theoretical suitability for the study; Round 2: Review of preliminary results based on unspecific German-language prompts applied to virtual GUC case scenarios for each LLM; Round 3: Secret voting among the six panelists, with two points awarded for the most suitable LLM and one point for the second choice (a maximum

 doi:10.20944/preprints202410.1290.v1

of 12 points per LLM, with 18 points total from all panelists); Round 4: A final moderated discussion of the results, leading to consensus on the two LLMs selected for the study.

The selection criteria included: (1) availability, (2) suitability for answering medical queries, and (3) response quality. LLMs considered in this process included ChatGPT-4, ChatGPT-3.5 (both OpenAI), Claude 3.5 Sonnet (Anthropic), Copilot (Microsoft), Gemini (Google), Llama 3 (Meta), and Med-PaLM2 (Google). Panelists were also encouraged to favor LLMs utilizing different transformer-based architectures designed for natural language processing tasks.

2.3. Development of Standardized Prompts for Data Input on GUC Patients and the Creation of a Uniform Recommendation Matrix for Both LLMs and the MTB to Facilitate Blinded Assessment (2)

German-language prompts were chosen for querying the LLMs to avoid translation errors and ensure direct comparability with the German-language recommendations of the real MTB. The initial prompts were developed by MM based on previous LLM queries and relevant studies [12–17,19,20]. These prompts were then tested in a multi-stage process by the working group (ER, MH, DvW, and AK), using the two selected LLMs. Based on the results, the prompts were refined and optimized for consistency and accuracy.

The working group also discussed formal and linguistic adjustments needed to ensure that the recommendations from both the MTB and LLMs could be sufficiently blinded for evaluation. This resulted in the development of a clear recommendation matrix. The criteria for finalizing the prompts included: (a) clinical relevance of the recommendations, (b) consideration of key patient characteristics across the five main GUC types (prostate cancer, bladder cancer, kidney cancer, testicular cancer, penile cancer), (c) inclusion of a multidisciplinary treatment perspective, (d) reference to current evidence, and (e) the ability to offer alternative therapeutic strategies.

2.4. *Modification and Validation of the Newly Developed mSCS Using a Cohort of 40 Patients with Varying Organ-Specific GUCs (3)*

The SCS is a metric for evaluating LLM recommendations. It is calculated by assigning a score between 1 and 5 on a Likert scale to each of the 10 items it comprises. These items are listed in the second column of **Table 2**. The possible ratings are: 1 = strongly disagree, 2 = disagree, 3 = neutral, 4 = agree, and 5 = strongly agree. The ratings for the 10 categories are summed, and the resulting score is obtained by dividing the sum by 50. This yields scores ranging from 0.2 to 1.0, with 1.0 representing the optimal result [8].

The 10 items of the original SCS [8] were reviewed by two uro-oncology experts (MM, CGo) and assessed for their applicability to the study. The experts proposed modifications of items to improve the tool for evaluating treatment recommendations. The items were revised for the mSCS, but the number of items and the consecutive calculation of the score correspond exactly to the original SCS.

Forty case scenarios, encompassing the five main GUCs (16 prostate cancer, 9 bladder cancer, 7 kidney cancer, 4 testicular cancer, and 4 penile cancer cases), were presented to the real MTBs in Regensburg and Straubing (20 per site) and also submitted to the top-rated LLM, based on the Delphi process. Recommendations from both the MTB and the LLM were rated using the SCS by two independent uro-oncologists (ER and MH for Regensburg; DvW and AK for Straubing). The same raters applied the mSCS 14 days later to the same recommendations, with the sequence of case scenarios altered. Any discrepancies between the two raters were resolved by a third adjudicator (CGo in Regensburg; MM in Straubing).

Following the two rating rounds, all four raters (ER, MH, DvW, and AK) were asked to assess the clinical applicability of the mSCS compared to the original SCS using a 5-point Likert scale (1 = severe deterioration, 2 = deterioration, 3 = equality, 4 = improvement, 5 = strong improvement). The four ratings were combined into one rating for further analysis by determining the modal value. For the statistical analysis, the comparative SCS item was always assigned the value 3.

To comply with data protection guidelines, the case scenarios were realistic but fictitious. These scenarios were developed based on the real patient cases typically presented at the MTBs and were

 doi:10.20944/preprints202410.1290.v1

created by experienced uro-oncologists (CGo in Regensburg, MM in Straubing). Each case was formatted as a table in bullet-point form, mirroring the style used in real-life MTB presentations.

2.5. Biometric Sample Size Planning for the Prospective Trial, Preceded by a Moderated Delphi Process with the Entire Study Team to Establish What Level of Difference in the mSCS, Derived from Preliminary Study Results, Would Still Be Considered Non-Inferior for LLMs Compared to the MTB (4)

Biometric sample size planning was conducted by a statistician experienced in prospective study design (FZ). The mean mSCS results for the LLM and MTB, along with their respective standard deviations, were used to estimate the expected variance. A t-test was used to compare the two groups, with power set at 90% (beta = 0.1) and a one-sided 2.5% level of significance (alpha = 0.025), adjusted by Bonferroni correction since two LLMs were compared against the real MTB (adjusted p-value: 0.0125).

The non-inferiority margin for the LLMs, indicating acceptable performance compared to the MTB, was determined by the study authors (ER, MH, DvW, AK, CGo, and MM) through a second four-round Delphi process moderated by author MS. The process included: Round 1: Presentation of four differences in mSCS, proposed by the moderator, that were still associated with non-inferiority (0.05, 0.1, 0.15, and 0.2); Round 2: Group discussion on the clinical implications of these four thresholds, based on five GUC scenarios prepared by the moderator; Round 3: Secret voting among the six panelists, with two points awarded for the most suitable difference and one point for the second choice (a maximum of 12 points per cutoff, with 18 points total from all panelists); Round 4: A final moderated discussion of the results, leading to a consensus on the difference at which clinical non-inferiority of the LLMs compared to the real MTB can still be acknowledged.

2.6. Precise Documentation and Listing of Statistical Methods to validate The mSCS and Compare Results between the Groups (MTB vs. LLM) (5)

Interrater reliability was assessed using Cohen's kappa coefficient [22,23]. To simplify analysis for kappa calculation, the 5-point Likert scale ratings for the 10 SCS and mSCS items were dichotomized. Scores differing by more than 1 point were labeled as 'disagree', while scores within ±1 were labeled 'agree'. Reliability was tested for each item in both the SCS and mSCS for the MTB and LLM, with pooled analyses conducted. The interpretation of Cohen's kappa (K) can be classified based on the guidelines established by Landis and Koch, who propose the following framework for interpreting the strength of agreement: <0.00: poor agreement, 0.00–0.20: slight agreement, 0.21–0.40: fair agreement, 0.41–0.60: moderate agreement, 0.61–0.80: substantial agreement, and 0.81–1.00: almost perfect agreement [24].

The validity of the mSCS was assessed by comparing its internal consistency with that of the original SCS, using Cronbach's Alpha [25]. While there are different interpretations of Cronbach's Alpha (α) in the literature [26], we adhere to the commonly used structure as follows: <0.5: unacceptable, 0.50–0.59: poor, 0.60–0.69: questionable, 0.70–0.79: acceptable, 0.80–0.89: good, ≥0.9: excellent.

Consensus judgments were used in both systems for this analysis. Differences in clinical applicability between the SCS and mSCS were tested for significance using the Wilcoxon signed-rank test.

All *p*-values were two-tailed, and statistical significance was set at $p \leq 0.05$. Statistical analyses were performed using SPSS 29.0 (IBM Corp., Armonk, NY, USA).

3. Results

3.1. Selection of Two Appropriate LLMs for Comparison with the MTB (1)

As part of the Delphi process, ChatGPT-4 and Claude 3.5 Sonnet were selected as the most suitable LLMs.

 doi:10.20944/preprints202410.1290.v1

The distribution of points in the anonymous voting between the LLMs using the Delphi method (round 3) showed the following results: ChatGPT-4 received 11 points, Claude 3.5 Sonnet received 5 points, and ChatGPT-3.5 received 2 points. All other LLMs under consideration received no points.

The moderated discussion revealed the following reasons for the low scores achieved by the other LLMs: The response quality from Copilot as well as ChatGPT-3.5 appeared inferior in test inputs. Gemini did not sufficiently adhere to the required formal conditions of the recommendations in test inputs. Llama 3 was excluded due to its lack of availability in Europe and the frequent issuance of the error message "Sorry, I can't help you" in reference to a doctor's consultation. Med-PaLM2 was not sufficiently available.

3.2. Development of Standardized Prompts for Data Input on Urological Tumor Patients and the Creation of a Uniform Recommendation Matrix for Both LLMs and the MTB to Facilitate Blinded Assessment (2)

The developed prompt is shown in **Table 1**. The individual components of the prompt have been assigned to corresponding objectives in the table. They are color-coded according to the following scheme: Task (yellow). Information provided (green). Request for completeness and indication of preferred option (gray). Geographical categorization (blue). Formal requirements (purple).

Table 1. Representation of the standardized sequence of input commands (prompts) into the large language model in German, along with the translation of the prompts into English.

Prompt - Original input in German	Prompt - English translation
Formuliere eine stichpunktartige Therapieempfehlung für den folgenden Patientenfall. Die Vortherapien und andere relevante Befunde sind im Fall enthalten. Beschränke dich hierbei nicht nur auf Medikamente, sondern beschreibe alle möglichen Therapieoptionen. Bitte benenne Therapien und Medikamente, falls du eine medikamentöse Therapie vorschlägst, konkret. Versuche, deine Empfehlung anhand der in Deutschland zugelassenen und leitliniengerechten Therapien zu treffen. Bitte benenne zudem explizit die aus deiner Sicht beste Therapieoption für den individuellen Patienten. Begrenze mit Deinen Antworten auf maximal 80 Wörter und orientiere Dich in ihnen an folgender Struktur: 1.) Präferierte Therapie-empfehlung (falls vorhanden), 2.) Therapie-alternativen, 3.) Begründung der Empfeh-lungen, 4.) Supportivmaßnahmen / ergänzende Therapien, 5.) Weiterführende Informationen / Erklärungen. Patientenfall:	Formulate a key point-based treatment recommendation for the following patient case. The previous therapies and other relevant findings are included in the case. Do not limit yourself to medication but describe all possible treatment options. Please name therapies and medications specifically if you are suggesting drug therapy. Try to make your recommendation based on the therapies approved in Germany and in line with the guidelines. Please also explicitly state what you consider to be the best treatment option for the individual patient. Limit your answers to a maximum of 80 words and base them on the following structure: 1) Preferred therapy recommendation (if available), 2) Therapy alternatives, 3) Justification of the recommendations, 4) Supportive measures / supplementary therapies, 5) Further information / explanations. Patient case:

The prompt components are assigned to objectives: Task (yellow), Information (green), Completeness & Preferred Option (gray), Geographical Categorization (blue), Formal Requirements (purple).

3.3. Modification and Validation of the Newly Developed mSCS Using a Cohort of 40 GUC Patients with Varying Organ-Specific Cancers (3)

Table 2 shows the SCS and the mSCS. All items, except item 4, were modified.

Table 2. Items of original System Causability Scale (SCS) and modified SCS (mSCS).

 doi:10.20944/preprints202410.1290.v1

Item	SCS	mSCS
1	I found that the recommendation included all relevant known causal factors with sufficient precision and granularity	I found that the recommendation included all relevant patient-specific factors (individual patient data such as individual tumour stages, previous treatments and specific health conditions) with sufficient precision and granularity
2	I understood the explanations within the context of my work.	I found the quality and representativeness of the recommendations, particularly in relation to oncological scenarios, sufficient.
3	I could change the level of detail on demand.	I found that all reasonable treatment alternatives were specified.
4	I did not need support to understand the explanations.	I did not need support to understand the explanations.
5	I found the explanations helped me to understand causality	I found that the recommendation was explained and made transparent.
6	I was able to use the explanations with my knowledge base.	I found the recommendation to be consistent with current clinical guidelines.
7	I did not find inconsistencies between explanations.	I did not find inconsistencies between explanations/recommendations.
8	I think that most people would learn to understand the explanations very quickly.	I think that most health care professionals would learn to understand the explanations very quickly.
9	I did not need more references in the explanations: e.g., medical guidelines, regulations.	I found the recommendation demonstrates access to the latest research and clinical guidelines.
10	I received the explanations in a timely and efficient manner	I found the quality of interaction (ease of use and accessibility) sufficient.

3.4. Biometric Sample size planning for the Prospective Trial, Preceded by a Moderated Delphi Process with the Entire Study Team to Establish What Level of Difference in the mSCS, Derived from Preliminary Study Results, Would Still Be Considered Non-Inferior for LLMs Compared to the MTB (4)

After evaluating the LLM and MTB recommendations for the 40 sample tumor cases using the mSCS, the recommendations were compared with the consecutive ratings. Detailed discussions were held to determine which differences in the mSCS corresponded to which differences in the content of recommendations, especially regarding clinical implications. Based on these discussions, barriers that could generally be considered as meaningful non-inferiority measures were assessed. This resulted in the considered non-inferiority thresholds of 0.05, 0.1, 0.15 and 0.2.

In the following anonymous voting as part of the Delphi process, the non-inferiority threshold of 0.15 difference in mSCS received the highest score of 9 points. The thresholds 0.1, 0.05 and 0.2 received 5, 3 and one point respectively. Finally, another moderated discussion was held regarding the best-scoring non-inferiority threshold of 0.15, in which it was jointly agreed that this maximum difference in mSCS clinically represents a non-inferiority of the recommendation quality.

The mean value of the mSCS scores obtained was 0.992 ± 0.013 for the MTB recommendations. There was a slight inferiority in the mean mSCS of the recommendations of the LLM, which was 0.897 ± 0.144. Using a two-sided t-test, the non-inferiority threshold of 0.15 difference previously established in the Delphi process, and the previously established alpha (0.05) and beta (0.1) levels, the required sample size was 87 cases.

To account for potential dropouts, we increased the sample size by 25%, targeting 109 participants. One additional case was included to achieve an even number for the bicentric study, ensuring equal distribution across centers. This resulted in a final sample size of 110 cases for the planned prospective study.

3.5. Validation of the mSCS and Comparation between the Groups (MTB vs. LLM) (5)

 doi:10.20944/preprints202410.1290.v1

3.5.1. Interrater Reliability

To assess the agreement between the two independent raters, Cohen's Kappa was calculated. The kappa (K) values are shown in **Table 3**.

Regarding the SCS rating of the MTB recommendations, kappa values of 0.7 to 1.0 ($p < 0.001$) were obtained for the individual items. The pooled analysis resulted in K = 0.90 ($p < 0.001$). With regard to the SCS rating of the LLM recommendations, kappa values of 0.65 to 0.90 ($p < 0.001$) were obtained for the individual items. The pooled analysis resulted in K = 0.74 ($p < 0.001$). In summary, substantial to almost perfect interrater reliability was shown for the SCS across all items.

For the mSCS ratings regarding the MTB recommendation, for all Items the Kappa values were at least K = 0.75, indicating at least substantial agreement. For the mSCS ratings regarding the LLM recommendation, slightly more dispersion was observed. The lowest kappa value obtained was K = 0.65, which also indicates a substantial agreement. In the pooled analysis of interrater reliability across all items of the mSCS, K = 0.95 ($p < 0.001$) was obtained for the MTB recommendations and K = 0.81 ($p < 0.001$) for the LLM recommendations (Table 3).

Table 3. Interrater realiability (K): SCS/mSCS Ratings concerning MTB vs. LLM recommendations.

	Interrater realiability SCS		**Interrater reliability mSCS**	
	MTB	**LLM**	**MTB**	**LLM**
	K (*p*-value)	**K (*p*-value)**	**K (*p*-value)**	**K (*p*-value)**
Item 1	0.80 (<.001)	0.70 (<.001)	0.90 (<.001)	0.70 (<.001)
Item 2	0.95 (<.001)	0.75 (<.001)	1.00 (<.001)	0.85 (<.001)
Item 3	0.70 (<.001)	0.75 (<.001)	0.80 (<.001)	0.65 (<.001)
Item 4	1.00 (<.001)	0.90 (<.001)	1.00 (<.001)	0.95 (<.001)
Item 5	0.90 (<.001)	0.70 (<.001)	0.75 (<.001)	0.70 (<.001)
Item 6	0.95 (<.001)	0.65 (<.001)	1.00 (<.001)	0.80 (<.001)
Item 7	0.90 (<.001)	0.70 (<.001)	1.00 (<.001)	0.80 (<.001)
Item 8	0.85 (<.001)	0.80 (<.001)	1.00 (<.001)	0.85 (<.001)
Item 9	0.90 (<.001)	0.85 (<.001)	1.00 (<.001)	0.95 (<.001)
Item 10	1.00 (<.001)	0.75 (<.001)	1.00 (<.001)	0.80 (<.001)
All Items	**0.90 (<.001)**	**0.74 (<.001)**	**0.95 (<.001)**	**0.81 (<.001)**

3.5.2. Agreement between SCS and mSCS

Agreement of the consensus ratings between SCS and mSCS in dichotomized form was calculated using Cohen's kappa (K). The ratings of the MTB recommendations exhibited an almost perfect agreement of K = 0.96 ($p < 0.001$). With regard to the ratings of the LLM recommendations, there was an almost perfect agreement of K = 0.88 ($p < 0.001$). In the pooled analysis of all ratings, the agreement between SCS and mSCS was K = 0.93 ($p < 0.001$). Overall, this shows an almost perfect agreement.

3.5.3. Internal Consistency

The internal consistency of the ratings in the mSCS compared to the SCS (dichotomized) was tested using Cronbach's alpha. An excellent internal consistency was found with Cronbach's alpha values of 0.992 for the ratings of the MTB recommendations, 0.934 for the ratings of the LLM recommendations and 0.964 in the pooled analysis of the ratings of the MTB and LLM recommendations.

After excluding item 4, the only non-modified item, Cronbach's alpha values of 0.989 were obtained for the ratings of the MTB recommendations, 0.926 for the ratings of the LLM recommendations and 0.957 in the pooled analysis of the ratings of the MTB and LLM recommendations.

3.5.4. Evaluation of Clinical Applicability of the mSCS Compared to the SCS

 doi:10.20944/preprints202410.1290.v1

The median Likert score for the clinical applicability of the mSCS items was 4.5 (IQR: 4-5), while the score for the SCS items was 3 based on upfront determination. There was a statistically significant increase in the Likert scores after the modification of the SCS ($Z = -2.739$, $p = 0.006$, $n = 10$), suggesting that the modification had a positive effect.

4. Discussion

The current study is intended as a preparatory investigation for the prospective, bicentric CONCORDIA Study. Hence, the specific LLMs to be used, the optimal input prompts for the LLMs, a sufficient measurement tool adapted to the specific research question, and the sample size calculation for the main study were developed based on 40 case scenarios, that were discussed in a real MTB and subsequently compared with treatment recommendations from an LLM (ChatGPT-4).

MTBs consist of regular meetings of representatives of various clinical specialties, who discuss patient management and provide evidence-based and individual therapy decisions [27]. One can easily imagine that interdisciplinary exchange and various perspectives on patient cases ultimately lead to a more profound and higher-quality therapy decision. This effect seems to be particularly pronounced in tumor entities where established treatment options involve multiple specialties, such as surgery, medical oncology, radiation therapy, or nuclear medicine (as is often the case with GUCs). On the other hand, MTBs consume substantial personnel and financial resources to facilitate interdisciplinary exchange, which poses a genuine challenge in a world where both resources represent a true scarcity [9–11,28,29]. To investigate the effect of MTBs on patient outcomes, Huang et al. conducted a Meta-Analysis including 134,287 patients with various cancer entities from 59 studies. The authors found a significantly prolonged survival time (median survival time 30.2 months vs. 19 months) in patients managed by an MTB, suggesting that their implementation is likely worthwhile whenever possible [28].

LLMs are poised to take the scientific and clinical medical world by storm with their abilities in natural language processing, data analysis, predictive modeling, and generating evidence-based recommendations [1–5]. A particularly advantageous feature of LLMs is the ability to provide logical, coherent, and scientifically correct answers to various text questions, which is facilitated by deep learning algorithms and the access to large-scale and up-to-date databases. And it is precisely from this feature of LLMs that the research question of the CONCORDIA study is derived, namely whether LLMs can replace the complex, resource intensive decision-making process of an MTB and ultimately generating a recommendation that is not inferior to that of the MTB.

Currently, there is limited evidence on the use of LLMs as auxiliary tools in MTBs for other cancer entities, and, to the best of our knowledge, no studies have been conducted in the context of GUC or compared the blinded recommendations of LLMs with those of an actual MTB [12,14,15,30]. One study investigated ChatGPT 3.5 and ChatGPT-4 as decision-making tools for 30 primary head and neck cancer cases [15]. Although the LLMs performed exceptionally well in providing clinical recommendations, explanations, and summaries, they suggested significantly more treatment options than the MTB and occasionally recommended incorrect guidelines. The authors concluded that, while ChatGPT may support the MTB process, it is not capable of replacing it [15]. Another study by Stalp et al. evaluated ChatGPT 3.5s performance in suggesting treatments in 30 breast cancer cases [14]. While the therapy recommendations were judged to be mostly accurate, the quality of the recommendations was higher in primary cases, and complex patient histories posed a particular challenge for the LLM [14]. The study also demonstrated that the quality of recommendations is directly influenced by the prompt [14]. These findings align with another study by Griewing et al., which showed that using an extended input model further improved the quality of the LLMs' recommendations [18]. In the current study, this issue was addressed by refining and optimizing the initial prompts for consistency and accuracy in a multi-stage process by the working group (ER, MH, DvW, and AK).

Another critical step in the design of the CONCORDIA study was the determination of the optimal sample size. Based on the results of the current study, we took several factors into account: 1.) Power and effect size: In accordance with available recommendations on power analysis for

 doi:10.20944/preprints202410.1290.v1

clinical research studies, a statistician experienced in prospective study design (FZ) conducted the sample calculation based on a desired power set at 90% (beta = 0.1) and Bonferroni corrected alpha at 0.0125 (since a one-sided 2.5% level of significance is assumed and two LLMs will be compared against the real MTB) [31]. The expected variance was estimated based on the mean mSCS results (±STD) for the LLM and MTB and the non-inferiority-margin was set to 0.15, based on a four-round Delphi process, as described above. This led to a minimal required sample size of 87 patients. 2.) Adjustment for dropouts or missing information: To compensate potential dropouts or missing data, the targeted patient number was increased by 25%, corresponding to 109 patients. To achieve an equal case distribution between the two study centers, the final patient number was set at 110. 3.) Sample representativeness: The current study reflects the real-world care in an actual MTB and encompasses the full spectrum of GUCs in their respective frequencies (16 prostate cancer, 9 urothelial cancer, 7 renal cell cancer, 4 testicular cancer, and 4 penile cancer cases). In the planned CONCORDIA study, case scenarios across different GUC entities will be distributed as accurately as possible by analyzing the actual frequency distribution of the real MTB cases for the two study centers.

Limitations

A limitation of both the current study and the upcoming CONCORDIA study is that, due to data privacy regulations from the local ethics committee, we are unable to discuss and compare real patient cases. To address this, we will create realistic case scenarios that are not based on actual patients. As previously mentioned, our goal is to align the distribution of these scenarios with the actual frequency distribution of the two MTBs across different GUC entities, ensuring a representative cohort for the CONCORDIA study. Additionally, when designing the cases, we are carefully preserving the structure of the original MTB cases to facilitate comparability between the two centers and ensure greater consistency in the case vignettes.

5. Conclusions

This study demonstrates the successful modification and validation of the SCS for evaluating AI-based therapeutic recommendations, specifically for patients with urological cancer. The modified version (mSCS) exhibited improved reliability, internal consistency, and clinical applicability when compared to the original SCS. Results from this preliminary work provide a robust methodological basis for the forthcoming non-inferiority trial CONCORDIA, which will compare treatment recommendations derived from LLMs and MTBs. The study's design ensures a comprehensive assessment of AI's role in clinical decision-making, with significant implications for future integration of AI in clinical oncology. The validated mSCS offers a valuable tool for evaluating LLM recommendations in this and similar trials. Ultimately, this work paves the way for advancing AI-supported cancer care.

Author Contributions: Conceptualization, M.M.; methodology, E.R., M.H. and M.M.; software, E.R., F.Z. and M.H.; validation, E.R., M.H. and M.M.; formal analysis, E.R., M.H. and M.H.; investigation, E.R., D.vW., A.K., M.H., C.S., M.J.S,, S.S., R.M., J.G., C.G. (Christopher Goßler), F.P., P.J.S., J.B., A.S., S.D.B.-M. and M.M.; resources, C.G. (Christian Gilfrich) and M.B.; data curation, E.R., M.H. and M.M.; writing—original draft preparation, E.R., M.H. and M.M.; writing—review and editing, E.R., M.H., S.D.B.-M and M.M.; visualization, M.M.; supervision, C.G. (Christian Gilfrich) and M.B.; project administration, M.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Ethics Committee of the University of Regensburg (protocol code 24-3835-104; date of approval July 16, 2024).

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Acknowledgments: We extend our gratitude to the inventors of various large language models, the companies that make them available for public use, and the many users worldwide who, through their collective intelligence and the input of scientific databases, contribute to optimizing the training algorithms. This optimization holds the potential to enhance decision-making for selecting the most appropriate therapies for complex carcinoma cases in the near future.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell. 2023;6:1169595. doi:10.3389/frai.2023.1169595.
2. Ray PP. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet of Things and Cyber-Physical Systems. 2023;3:121–54. doi:10.1016/j.iotcps.2023.04.003.
3. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med. 2022;28:31–8. doi:10.1038/s41591-021-01614-0.
4. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. 2023;29:1930–40. doi:10.1038/s41591-023-02448-8.
5. Kowalewski K-F, Rodler S. Large Language Models in der Wissenschaft. [Large language models in science]. Urologie. 2024;63:860–6. doi:10.1007/s00120-024-02396-2.
6. OpenAI. Introducing ChatGPT: 2022 Nov 30. https://openai.com/blog/chatgpt.
7. Eppler M, Ganjavi C, Ramacciotti LS, Piazza P, Rodler S, Checcucci E, et al. Awareness and Use of ChatGPT and Large Language Models: A Prospective Cross-sectional Global Survey in Urology. Eur Urol. 2024;85:146–53. doi:10.1016/j.eururo.2023.10.014.
8. Holzinger A, Carrington A, Müller H. Measuring the Quality of Explanations: The System Causability Scale (SCS): Comparing Human and Machine Explanations. Kunstliche Intell (Oldenbourg). 2020;34:193–8. doi:10.1007/s13218-020-00636-z.
9. Pillay B, Wootten AC, Crowe H, Corcoran N, Tran B, Bowden P, et al. The impact of multidisciplinary team meetings on patient assessment, management and outcomes in oncology settings: A systematic review of the literature. Cancer Treat Rev. 2016;42:56–72. doi:10.1016/j.ctrv.2015.11.007.
10. Taylor C, Munro AJ, Glynne-Jones R, Griffith C, Trevatt P, Richards M, Ramirez AJ. Multidisciplinary team working in cancer: what is the evidence? BMJ. 2010;340:c951. doi:10.1136/bmj.c951.
11. Perez-Gracia JL, Awada A, Calvo E, Amaral T, Arkenau H-T, Gruenwald V, et al. ESMO Clinical Research Observatory (ECRO): improving the efficiency of clinical research through rationalisation of bureaucracy. ESMO Open. 2020;5:e000662. doi:10.1136/esmoopen-2019-000662.
12. Levin G, Gotlieb W, Ramirez P, Meyer R, Brezinov Y. ChatGPT in a gynaecologic oncology multidisciplinary team tumour board: A feasibility study. BJOG 2024. doi:10.1111/1471-0528.17929.
13. Schmidl B, Hütten T, Pigorsch S, Stögbauer F, Hoch CC, Hussain T, et al. Assessing the use of the novel tool Claude 3 in comparison to ChatGPT 4.0 as an artificial intelligence tool in the diagnosis and therapy of primary head and neck cancer cases. Eur Arch Otorhinolaryngol 2024. doi:10.1007/s00405-024-08828-1.
14. Stalp JL, Denecke A, Jentschke M, Hillemanns P, Klapdor R. Quality of ChatGPT-Generated Therapy Recommendations for Breast Cancer Treatment in Gynecology. Curr Oncol. 2024;31:3845–54. doi:10.3390/curroncol31070284.
15. Schmidl B, Hütten T, Pigorsch S, Stögbauer F, Hoch CC, Hussain T, et al. Assessing the role of advanced artificial intelligence as a tool in multidisciplinary tumor board decision-making for primary head and neck cancer cases. Front Oncol. 2024;14:1353031. doi:10.3389/fonc.2024.1353031.
16. Aghamaliyev U, Karimbayli J, Giessen-Jung C, Matthias I, Unger K, Andrade D, et al. ChatGPT's Gastrointestinal Tumor Board Tango: A limping dance partner? Eur J Cancer. 2024;205:114100. doi:10.1016/j.ejca.2024.114100.
17. Benary M, Wang XD, Schmidt M, Soll D, Hilfenhaus G, Nassir M, et al. Leveraging Large Language Models for Decision Support in Personalized Oncology. JAMA Netw Open. 2023;6:e2343689. doi:10.1001/jamanetworkopen.2023.43689.
18. Griewing S, Gremke N, Wagner U, Lingenfelder M, Kuhn S, Boekhoff J. Challenging ChatGPT 3.5 in Senology-An Assessment of Concordance with Breast Cancer Tumor Board Decision Making. J Pers Med 2023. doi:10.3390/jpm13101502.
19. Vela Ulloa J, King Valenzuela S, Riquoir Altamirano C, Urrejola Schmied G. Artificial intelligence-based decision-making: can ChatGPT replace a multidisciplinary tumour board? Br J Surg. 2023;110:1543–4. doi:10.1093/bjs/znad264.

20. Lukac S, Dayan D, Fink V, Leinert E, Hartkopf A, Veselinovic K, et al. Evaluating ChatGPT as an adjunct for the multidisciplinary tumor board decision-making in primary breast cancer cases. Arch Gynecol Obstet. 2023;308:1831–44. doi:10.1007/s00404-023-07130-5.
21. Delourme S, Redjdal A, Bouaud J, Seroussi B. Measured Performance and Healthcare Professional Perception of Large Language Models Used as Clinical Decision Support Systems: A Scoping Review. Stud Health Technol Inform. 2024;316:841–5. doi:10.3233/SHTI240543.
22. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull. 1968;70:213–20. doi:10.1037/h0026256.
23. Cohen J. A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement. 1960;20:37–46. doi:10.1177/001316446002000104.
24. Landis JR, Koch GG. The Measurement of Observer Agreement for Categorical Data. Biometrics. 1977;33:159. doi:10.2307/2529310.
25. Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika. 1951;16:297–334. doi:10.1007/BF02310555.
26. Taber KS. The Use of Cronbach's Alpha When Developing and Reporting Research Instruments in Science Education. Res Sci Educ. 2018;48:1273–96. doi:10.1007/s11165-016-9602-2.
27. Wright FC, Vito C de, Langer B, Hunter A. Multidisciplinary cancer conferences: a systematic review and development of practice standards. Eur J Cancer. 2007;43:1002–10. doi:10.1016/j.ejca.2007.01.025.
28. Huang RS, Mihalache A, Nafees A, Hasan A, Ye XY, Liu Z, et al. The impact of multidisciplinary cancer conferences on overall survival: a meta-analysis. Journal of the National Cancer Institute. 2024;116:356–69. doi:10.1093/jnci/djad268.
29. Berardi R, Morgese F, Rinaldi S, Torniai M, Mentrasti G, Scortichini L, Giampieri R. Benefits and Limitations of a Multidisciplinary Approach in Cancer Patient Management. Cancer Manag Res. 2020;12:9363–74. doi:10.2147/CMAR.S220976.
30. Sorin V, Klang E, Sklair-Levy M, Cohen I, Zippel DB, Balint Lahat N, et al. Large language model (ChatGPT) as a support tool for breast tumor board. NPJ Breast Cancer. 2023;9:44. doi:10.1038/s41523-023-00557-8.
31. Suresh K, Chandrashekara S. Sample size estimation and power analysis for clinical research studies. J Hum Reprod Sci. 2012;5:7–13. doi:10.4103/0974-1208.97779.