Exploring Bias(es) of Large Language Models in the Field of Mental Health – A Comparative Study Investigating the Effect of Gender and Sexual Orientation in Anorexia Nervosa and Bulimia Nervosa Case Vignettes

Rebekka Schnepper, Noa Roemmel, Rainer Schaefert, Lena Lambrecht-Walzinger, Gunther Meinlschmidt

Submitted to: JMIR Mental Health
on: March 01, 2024

Table of Contents

Exploring Bias(es) of Large Language Models in the Field of Mental Health – A Comparative Study Investigating the Effect of Gender and Sexual Orientation in Anorexia Nervosa and Bulimia Nervosa Case Vignettes

Rebekka Schnepper[1, 2] Dr rer nat; Noa Roemmel[1, 2]; Rainer Schaefert[1] Prof Dr Med; Lena Lambrecht-Walzinger[1] Dr med; Gunther Meinlschmidt[1, 2, 3, 4] Prof Dr

[1]Department of Psychosomatic Medicine University Hospital and University of Basel Basel CH
[2]Department of Digital and Blended Psychosomatics and Psychotherapy, Psychosomatic Medicine University Hospital and University of Basel Basel CH
[3]Division of Clinical Psychology and Cognitive Behavioural Therapy International Psychoanalytic University (IPU) Berlin Berlin DE
[4]Division of Clinical Psychology and Epidemiology Department of Psychology University of Basel Basel CH

Corresponding Author:
Rebekka Schnepper Dr rer nat
Department of Psychosomatic Medicine
University Hospital and University of Basel
Hebelstr. 2
Basel
CH

Abstract

Background: Large language models (LLMs) are increasingly used in the mental health field, with promising results in assessing mental disorders. However, correctness, dependability, and equity of LLM-generated information have been questioned. Amongst other, societal biases and research underrepresentation of certain population strata may affect LLMs. Because LLMs are already used for clinical practice, including decision support, it is important to investigate potential biases to ensure a responsible use of LLMs.

Objective: We aimed to estimate the presence and size of bias related to gender and sexual orientation produced by a common LLM, exemplified in the context of ED symptomatology and health-related quality of life (HRQoL) of patients with AN or BN.

Methods: We extracted 30 case vignettes (22 AN, 8 BN) from scientific articles. We adapted each vignette to create 4 versions, describing a female vs. male patient living with their female vs. male partner (2x2 design), yielding n=120 vignettes. We then fed each vignette into Chat Generative Pre-trained Transformer-4 (ChatGPT-4) thrice with the instruction to evaluate them by providing responses to two psychometric instruments, the RAND-36 questionnaire assessing HRQoL and the eating disorder examination questionnaire (EDE-Q). With the resulting LLM-generated scores, we calculated multilevel models (MLMs) with a random intercept for gender and sexual orientation (accounting for within-vignette variance), nested in vignettes (accounting for between-vignette variance).

Results: The MLM with N=360 observations indicated for the RAND-36 mental composite summary, a significant association with gender (conditional means: 12.8 for male and 15.1 for female cases; 95% CI of the effect=[-6.15, -0.35]; p=.037) but neither with sexual orientation nor an interaction effect (ps>.370). We found no indications for main or interaction effects of gender or sexual orientation for the EDE-Q overall score (conditional means: 5.59-5.65; ps>.611).

Conclusions: LLM-generated estimates of mental HRQoL in AN or BN case vignettes are at risk of being affected by cases' gender, with male cases scoring lower. Given the lack of real-world epidemiological evidence for such a pattern, our study highlights relevant risk of bias when applying generative AI in the context of mental health. Better understanding and mitigation of risk of bias related to gender and other factors, such as ethnicity or socioeconomic status, are highly warranted to ensure responsible use of LLMs when conducting diagnostic assessments or providing treatment recommendations.

(JMIR Preprints 01/03/2024:57986)
DOI: https://doi.org/10.2196/preprints.57986

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Exploring Bias(es) of Large Language Models in the Field of Mental Health – A Comparative Study Investigating the Effect of Gender and Sexual Orientation in Anorexia Nervosa and Bulimia Nervosa Case Vignettes

Rebekka Schnepper[1,2*], Noa Roemmel[1,2], Rainer Schaefert[1], Lena Lambrecht-Walzinger[1], & Gunther Meinlschmidt[1,2,3,4]

1) Department of Psychosomatic Medicine, University Hospital and University of Basel, Basel, Switzerland
2) Department of Digital and Blended Psychosomatics and Psychotherapy, Psychosomatic Medicine, University Hospital and University of Basel, Basel, Switzerland
3) Division of Clinical Psychology and Cognitive Behavioural Therapy, International Psychoanalytic University (IPU) Berlin, Berlin, Germany
4) Division of Clinical Psychology and Epidemiology, Department of Psychology, University of Basel, Basel, Switzerland

* Correspondence:
Rebekka Schnepper, University Hospital Basel, Hebelstr. 2, CH-4031 Basel, Switzerland

Phone: +41 61 328 46 33
E-Mail: rebekka.schnepper@usb.ch

Abstract

Background: Large language models (LLMs) are increasingly used in the mental health field, with promising results in assessing mental disorders. However, correctness, dependability, and equity of LLM-generated information have been questioned. Amongst other, societal biases and research underrepresentation of certain population strata may affect LLMs. Because LLMs are already used for clinical practice, including decision support, it is important to investigate potential biases to ensure a responsible use of LLMs.

Anorexia nervosa (AN) and bulimia nervosa (BN) show a lifetime prevalence of 1–2%, affecting more women than men. For men, sexual orientation was identified as a risk factor, with homosexual men having a higher risk of developing an eating disorder (ED) than heterosexual men. However, men are underrepresented in ED research and research on the association between gender and sexual orientation with prevalence, symptoms, and treatment outcomes of AN and BN is scarce.

Objective: We aimed to estimate the presence and size of bias related to gender and sexual orientation produced by a common LLM, exemplified in the context of ED symptomatology and health-related quality of life (HRQoL) of patients with AN or BN.

Methods: We extracted 30 case vignettes (22 AN, 8 BN) from scientific articles. We adapted each vignette to create 4 versions, describing a female vs. male patient living with their female vs. male partner (2x2 design), yielding n=120 vignettes. We then fed each vignette into Chat Generative Pre-trained Transformer-4 (ChatGPT-4) thrice with the instruction to evaluate them by providing responses to two psychometric instruments, the RAND-36 questionnaire assessing HRQoL and the eating disorder examination questionnaire (EDE-Q). With the resulting LLM-generated scores, we calculated multilevel models (MLMs) with a random intercept for gender and sexual orientation (accounting for within-vignette variance), nested in vignettes (accounting for between-vignette variance).

Results: The MLM with N=360 observations indicated for the RAND-36 mental composite summary, a significant association with gender (conditional means: 12.8 for male and 15.1 for female cases; 95% CI of the effect=[-6.15, -0.35]; p=.037) but neither with sexual orientation nor an interaction effect (ps>.370). We found no indications for main or interaction effects of gender or sexual orientation for the EDE-Q overall score (conditional means: 5.59-5.65; ps>.611).

Conclusions: LLM-generated estimates of mental HRQoL in AN or BN case vignettes are at risk of being affected by cases' gender, with male cases scoring lower. Given the lack of real-world epidemiological evidence for such a pattern, our study highlights relevant risk of bias when applying generative AI in the context of mental health. Better understanding and mitigation of risk of bias related to gender and other factors, such as ethnicity or socioeconomic status, are highly warranted to ensure responsible use of LLMs when conducting diagnostic assessments or providing treatment recommendations.

Keywords: Anorexia Nervosa; Artificial Intelligence (AI)**;** Bulimia Nervosa; ChatGPT; Eating Disorders; Generative AI; Large Language Models (LLMs); Responsible AI; Transformer;

Introduction

Large Language Models in the context of mental health

In recent years, there has been significant progress in the field of Artificial Intelligence (AI) [1]. In particular, the development of Large Language Models (LLMs), such as OpenAI's generative pre-trained transformer (GPT) models [2] or Google's LaMDA [3], has made the deployment of such algorithms accessible to researchers, clinicians, and the public alike [4]. With advancements in computational power and access to larger datasets, these models can now go beyond simple word counting [5] and actually account for the relationships between words [4, 6]. The technique of

modeling words in a large context has been referred to as transformer-based large language modeling [7]. This may not only facilitate the automatic analysis of large amounts of text data [8, 9], but, by modeling words in a large context, also allows the generation of meaningful text and the interactive use of this technology [4, 9]. Thus, the application of LLMs may improve efficiency and effectiveness of data processing in various fields – including health care [4].

Since psychology and psychotherapy research are primarily shaped by language, the potential of LLMs in this field is significant [1, 10]. This potential becomes even more meaningful when considering the contribution of mental disorders to the global disease burden [11] and acknowledging the persistent treatment gap in mental health care [12]. Especially in the field of psychological assessment, research on the use of LLMs is advanced [13]. For example, the employment of transformer language models on language patterns has resulted in remarkably high predictive accuracy on standardized well-being rating scales [14]. This procedure of employing LLMs to automatically generate psychological construct scores based on free text has been formally referred to as language-based assessment [13, 15]. Findings indicate comparable levels of validity and reliability of language-based assessments compared to standardized rating scales [14, 16]. Moreover, language-based assessments have the capacity to incorporate additional information beyond free text entries [13], such as user age [17].

LLMs have also been applied in the evaluation of clinical case vignettes and ChatGPT-4 has been shown to assess suicidality as reliable as mental health professionals [18]. Further, Chat-GPT 3.5's performance in the diagnostic assessment and advice on disease management in a study using 100 clinical vignettes has been rated as excellent by mental health professionals [19].

Biases and responsible AI

Despite the promising findings of using LLMs in the context of (mental) health, the issue of potential biases in information generated by LLMs has been raised. Because LLMs are increasingly being introduced into clinical practice, it is important to investigate potential biases to ensure a responsible use of Artificial Intelligence [20]and LLMs [21]. Since LLMs rely on training data which is directly or indirectly generated by humans, these models are likely to contain the same biases as the society in which they are created in [20-23]. This is especially critical in (mental) health care [24], where biases in LLMs may lead to discrimination of different social groups [21]. For example, ChatGPT 3.5 performed poorly in diagnosing an infectious disease known to be widely underdiagnosed [25]. Further, ChatGPT 3.5 made different treatment recommendations based on insurance status, which might introduce health disparities [26]. In the generation of clinical cases, ChatGPT-4 failed to create cases that depicted demographic diversity and relied on stereotypes when choosing gender or ethnicity [27]. Thus, the need for "fair AI" has been pointed out with the goal to develop prediction models that provide equivalent outputs for identical individuals who differ only in one sensitive attribute [28]. To avoid or at least reduce potential bias and move towards fair AI, this bias first needs to be conceptualized, measured, and understood [21].

The aim of this paper was to explore a potential bias in the evaluation of eating disorders (EDs), which have been subjected to stigma [29] and gender-biased assessment [30].

Eating Disorders AN / BN

Anorexia nervosa (AN) and bulimia nervosa (BN) are severe eating EDs with many medical complications, high mortality rates [31], and slow treatment progress with frequent relapses [32]. The lifetime prevalence to develop AN or BN is estimated to be 1-2% each [33]. Historically, AN and BN have been only described in women, and it was not until the 21st century that research started to systematically investigate EDs in men [34]. Today, men are estimated to account for approximately 10 – 25% of AN or BN cases [35, 36]. Research on gender difference in AN and BN presentation is scarce and inconclusive, with no clear findings with regards to genetic and

environmental factors that might explain differences in etiology or maintenance of EDs [37]. Likewise, findings on severity and treatment outcomes are ambiguous: While one study suggests that men diagnosed with AN might have faster and more frequent remission rates [38], another study found no difference [39]. Men might produce lower costs in outpatient treatment, however, this might be due to higher barriers to receive treatment [40]. Men have been found to be more stigmatizing than women towards people with EDs [41], and this internalized stigma might be one reason for the hesitancy to seek outpatient treatment.

In men, sexual orientation might increase the risk of developing an eating disorder, with more men with an ED or ED-related behavior identifying as homosexual compared to the general population [42, 43]. Further, independent of being diagnosed with an ED, homosexual men report more psychological distress than heterosexual men and in men with an ED, being homosexual was related to higher ED symptomatology [44]. In women, a review found no significant difference in overall disordered eating due to sexual orientation, but distinct patterns, with homosexual women reporting less restrictive eating behavior and more binge eating [45].

To conclude, only in the last two decades men were included in ED research and there are still many open questions related to the effect of gender on prevalence, symptoms, and treatment outcomes of AN and BN. With regards to sexual orientation, there is evidence for an association between identifying as homosexual and a higher risk of EDs in men but not in women.

Objective

We aimed to estimate the presence and size of bias related to gender and sexual orientation produced by ChatGPT-4, a common LLM, exemplified by their application in the context of eating disorder symptomatology and health-related quality of life (HRQoL) of patients with AN or BN. By providing clinical case vignettes to ChatGPT-4 and instructing it to take up the role of a clinical psychologist rating the vignettes, we sought to mimic the diagnostic process of an LLM-based ED assessment.

Methods

Vignette Selection and Modification

We searched PubMed and Google scholar up until October 2023 for vignettes in scientific articles published since 2000 that describe either AN or BN patients. A total of *n*=30 case vignettes were extracted from 12 different articles (published between 2001 and 2022). Of these vignettes, *n*=22 described AN patients and *n*=8 described BN patients. Most vignettes originally describe a female patient (*n*=28). We then adapted gender and sexual orientation each vignette to create 4 versions (2x2 design), describing a female vs. male patient living with their female vs. male partner (if either a marriage or an age ≥30 was mentioned, the term husband/wife was chosen, otherwise boyfriend/girlfriend). This resulted in *n*=120 adopted vignettes. Some information was removed due to content policy violations, i.e., drug abuse, self-mutilation, suicidal ideation or suicide attempts, sexual abuse, and traumatizing experiences. Further, details on the menstrual cycle were removed since they don't apply to male patients, as well as indications of height since they were unrealistically short for male patients. Last, some specific details not needed in this context were removed, e.g., study enrollment procedures and study-specific measures, medication plan, and the name of the hospital. See Table 1 for further details about the vignettes.

Table 1. Vignettes included in the study, search term, and information on parts that were removed, added, or changed.

Vignette	Search term	Removed	Changed	Added
1 [46]	Google scholar; October 13, 2023: (""case report"" OR ""case series"") AND (anorexia OR bulimia) AND psychotherapy", since 2000	GAF[a] score		AN[b] patient (implied in title of paper), sex, sexual orientation, living with boyfriend/girlfriend
2 [46]	See above	GAF score self-mutilation, suicide attempt		AN patient (implied in title of paper), sex, sexual orientation, living with boyfriend/girlfriend
3 [46]	See above	GAF score, amenorrhea		AN patient (implied in title of paper), sex, sexual orientation, living with boyfriend/girlfriend
4 [46]	See above	GAF score		AN patient (implied in title of paper), sex, sexual orientation, living with boyfriend/girlfriend
5 [46]	See above	GAF score	"School" changed to "university"	AN patient (implied in title of paper), sex, sexual orientation, living with boyfriend/girlfriend
6 [46]	See above	GAF score	"School" changed to "university"	AN patient (implied in title of paper), sex, sexual orientation,

				living with boyfriend/girlfriend
7 [46]	See above	GAF score		AN patient (implied in title of paper), sex, sexual orientation, living with boyfriend/girlfriend
8 [46]	See above	GAF score	“School” changed to “university”	AN patient (implied in title of paper), sex, sexual orientation, living with boyfriend/girlfriend
9 [46]	See above	GAF score	“Living with parents” changed to “living with boyfriend/girlfriend ”	AN patient (implied in title of paper), sex, sexual orientation
10 [46]	See above	GAF score, amenorrhea		AN patient (implied in title of paper), sex, sexual orientation, living with boyfriend/girlfriend
11 [46]	See above	GAF score, amenorrhea		AN patient (implied in title of paper), sex, sexual orientation, living with boyfriend/girlfriend
12 [46]	See above	GAF score, amenorrhea, suicide attempts		AN patient (implied in title of paper), sex, sexual orientation, living with boyfriend/girlfriend
13 [46]	See above	GAF score, amenorrhea, suicide attempts		AN patient (implied in title of paper), sex, sexual orientation, living with boyfriend/girlfriend
14 [46]	See above	GAF score, amenorrhea, suicide ideation, self-mutilation		AN patient (implied in title of paper), sex, sexual orientation, living with boyfriend/girlfriend
15 [47]	Pubmed, August 11, 2023: eating disorder filter for "case report", since 2000	menses, not sexually active		living with boyfriend/girlfriend
16 [47]	See above	medication details, menstrual cycle		living with boyfriend/girlfriend
17 [47]	See above	menstrual cycle		living with husband/wife (> 30 years)
18 [48]	See above	sexual abuse, drugs/alcohol, suicide		living with husband/wife
19 [48]	See above			living with boyfriend/girlfriend
20 [49]	See above	suicidal ideation		living with husband/wife (> 30 years)
21 [50]	Google scholar; October 13, 2023: (""case report"" OR ""case series"") AND (anorexia OR bulimia) AND psychotherapy", since 2000	substance abuse		living with boyfriend/girlfriend
22 [51]	See above	diagnostic manual		living with boyfriend/girlfriend

		& citation, name of measure, scientific consent, treated by author, height (unrealistic if changed to male sex)		
23 [52]	See above	city, education, menstrual irregularities, weight (unrealistic if changed to male sex)		living with boyfriend/girlfriend
24 [53]	See above	PTSD[c], sexual abuse, mens, study		living with boyfriend/girlfriend
25 [53]	See above	sexual abuse, PTSD, mens/menopause		living with husband/wife (> 30 years)
26 [53]	See above	Enrollment in study		living with boyfriend/girlfriend
27 [54]	See above			living with husband/wife (> 30 years)
28 [55]	See above			living with boyfriend/girlfriend
29 [56]	See above			living with husband/wife (> 30 years)
30 [57]	See above	height (unrealistic if changed to male sex)	“single” changed to “living with boyfriend/girlfriend ”	

[a] Global assessment of functioning
[b] Anorexia nervosa
[c] Post-traumatic stress disorder

Data Generation in ChatGPT-4

In three rounds, each vignette was fed into ChatGPT-4 with the instruction to evaluate them by providing responses to one of the two psychometric instruments. This resulted in a total of 720 vignette evaluations (120 vignettes * 3 rounds * 2 measures). ChatGPT-4 was opened in an internet browser (Google Chrome) with the chat history turned off to avoid a learning effect from the repeated evaluation of case vignettes. In the “custom instructions” settings, the instruction “Set the temperature of your replies to 0” was included. This instruction minimizes randomness in the text generation process and ensures maximum replicability, high precision, and factual accuracy. Data were generated between October and December 2023. See Textbox 1 for an example of a prompt.

Textbox 1. An example prompt for one of the 120 vignettes.

Take up the role of a clinical psychologist. Imagine that you see a patient described by the following case vignette.

“*A 21-year-old university student living with her boyfriend self-refers with concerns about her 7-year use of laxatives to control weight gain. She is eating daily without vomiting, but admits to binge-eating episodes three or four times weekly during the past 2 years. Compensatory vomiting stopped 6 months ago. She does not overexercise. Her BMI is low at 17.8, and her vital signs are normal. She admits to*

recent increased fatigue with occasional exertional dyspnea and daily diarrhea. She has been hospitalized twice in the past 3 years for dehydration not recognized as related to her laxative abuse.”

Based on the information given, what would be your best estimate regarding the following questions that refer to the case vignette:

So even though originally the questions are meant as self-report, apply them as questions to be replied as observer and provide the respective best estimate regarding the following questions that refer to the case vignette:

[*one of the two measures in their original format*]

Reply to each question with the reply categories:

[*original reply categories of the measure*]

If no estimate can be given for a question, code it as 999.

Provide the estimates as a simple table. In this table, provide each question as a new variable with the corresponding values in two columns, one column containing the question number in ascending order and one column containing ONLY the numerical values. Provide the entire table.

Measures

RAND 36-Item Short Form Health Survey Version 1.0 (SF-36)

The SF-36 [58] assesses HRQoL and consists of 8 subscales: physical functioning, bodily pain, role limitations due to physical health problems, role limitations due to personal or emotional problems, emotional well-being, social functioning, energy/fatigue, and general health perceptions. From these subscales, the mental composite summary (MCS; comprised of role limitations due to personal or emotional problems, emotional well-being, social functioning, and energy/fatigue), as well as a physical composite score (PCS) can be calculated. Evidence suggests that in EDs, MCS is more affected than PCS [59], thus this score was selected for this study. Further, the SF-36 includes a single item assessing perceived change in health which is not included in any of the subscales. Items are either answered with “yes/no” or on different Likert scales and then recoded to values ranging from 0 to 100, with higher scores indicating better HRQoL. To calculate the MCS, the authors have suggested an approach [60] in which first, the subscales are z-transformed using means and standard deviations from the general U.S. population; second, the subscales are aggregated by weighing them with coefficients from the general U.S. population; and third, a t-score transformation is performed (mean=50, *SD*=10). This approach has been criticized for distorting the raw scores, and it was found that simply calculating the MCS by forming the mean of the four subscales resulted in satisfactory validity [61]. In this study, the simple approach was chosen because on the one hand, only the MCS was investigated and therefore a potential correlation with the PCS would not pose a problem. On the other hand, the choice of population that the scores are z-standardized and weighed with makes assumptions on the origin of data that ChatGPT-4 were trained with, something that is not entirely known and therefore could distort our data.

Eating disorder examination-questionnaire (EDE-Q).

The EDE-Q [62] assesses ED symptomatology during the previous 28 days. It consists of four subscales: dietary restraint, weight concern, shape concern and eating concern. By calculating the mean of these subscales, a global score can be formed. Items are answered on a scale ranging from 0 to 6, with 6 reflecting the greatest severity or frequency of ED symptoms.

Statistical Analysis

Data from the ChatGPT-4 replies were copied to an Excel sheet, indicating the vignette number, gender, sexual orientation, and round number. Female gender and heterosexual orientation were coded as "0". We performed all analyses in RStudio [63]. For the main outcome analyses, we used the package "lme4" [64], which is suitable to calculate linear multilevel models (MLMs) with crossed random effects structure [65]. This approach was chosen to take the repeated evaluation (three rounds) of each vignette as well as the main and interaction effects of gender and sexual orientation into account. These MLMs included a random intercept for vignettes (accounting for between-vignette variance), as well as a random intercept for the gender * sexual orientation interaction nested in vignettes (accounting for within-vignette variance). This resulted in the formula:

$$Outcome \sim Gender*Orientation + (interaction(Gender, Orientation)|Vignette)$$

We plotted the results using ggplot2 [66].

Results

Descriptives

Table 2 shows the unconditional means of the MCS and EDE-Q. For the SF-36, there were 1.19% of missing values in items included in the MCS. For the EDE-Q, there were 0.76% of missing values in items included in the overall score (coded "999" by ChatGPT-4 and recoded to a missing value).

Table 2. Means and standard deviations of the two outcome measures for each of the four subgroups.

	MCS[a], mean score (SD)	EDE-Q[b], mean score (SD[c])
Female Gender		
Overall (*n*=180)	15.1 (15.6)	5.61 (0.52)
Heterosexual (*n*=90)	15.3 (16.3)	5.63 (0.49)
Homosexual (*n*=90)	14.8 (14.9)	5.60 (0.55)
Male Gender		
Overall (*n*=180)	12.8 (14.2)	5.65 (0.47)
Heterosexual (*n*=90)	12.1 (12.5)	5.64 (0.51)
Homosexual (*n*=90)	13.6 (15.7)	5.65 (0.42)

[a] Mental composite summary of the RAND 36-item short form survey

[b] Eating disorder examination questionnaire

[c] Standard deviation

Main Outcomes

For the MCS, the MLM with *N*=360 observations indicated a significant effect of gender with men having a lower MCS score (conditional means: 12.8 for male and 15.1 for female cases; see Figure 1), with no indications of an effect of sexual orientation or an interaction effect. For the EDE-Q overall score, there were no indications for significant main or interaction effects of gender or sexual orientation (conditional means: 5.59-5.65). See Table 3 for estimates of main and interaction effects, and respective p-values and confidence intervals of the estimates.

Table 3. Means and standard deviations of the two outcome measures for each of the four subgroups.

	MCS[a], estimate (p-value) [CI[c]]	EDE-Q[b], estimate (p-value) [CI[c]]
Gender	-3.25 (.037) [-6.15; -0.35]	-0.02 (.882) [-0.10; 0.14]
Sexual Orientation	-0.50 (.705)[-3.04; 2.05]	-0.03 (.674) [-0.14; 0.09]
Gender * Sexual Orientation	1.93 (.370) [-2.18; 6.04]	0.04 (.611) [-0.11; 0.19]

[a] Mental composite summary of the RAND 36-item short form survey
[b] Eating disorder examination questionnaire
[c] Confidence interval of estimate

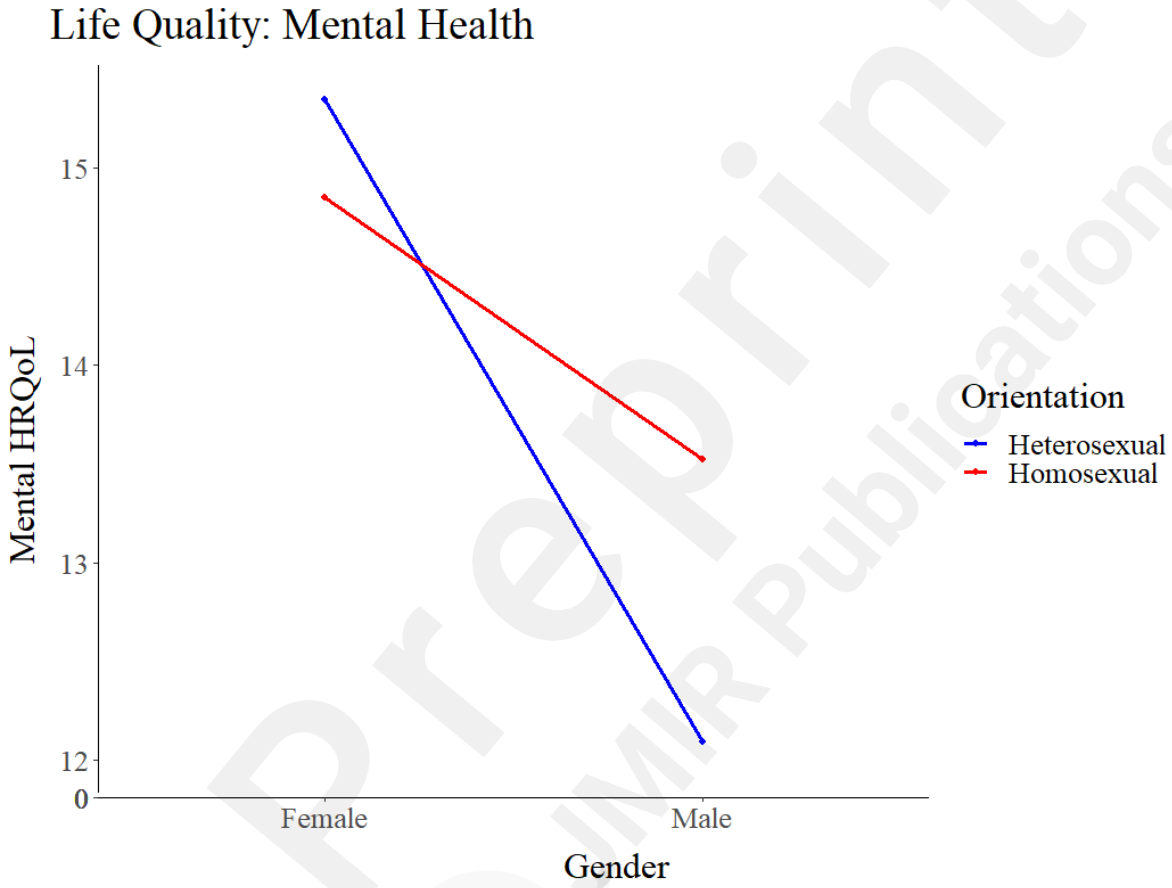

Figure 1. Lower HRQoL in men compared to women.

Discussion

Principal Results

We investigated if gender and sexual orientation in AN and BN case vignettes would influence mental HRQoL and ED severity estimates by ChatGPT-4, a commonly used LLM. Quadruples of 30 case vignettes from scientific articles were modified in a way that only information on gender and sexual orientation varied across vignettes of the same quadruple. Vignettes were then fed into ChatGPT-4 with the instruction to estimate scores of two widely used psychometric instruments for assessing HRQoL (MCS of the SF-36) and ED symptomatology (EDE-Q). Findings indicated no effect of gender or sexual orientation in ED severity. Of note, the EDE-Q scores were very high, which might have led to ceiling effects. For the MCS, there was an effect of gender but not for sexual orientation, with vignettes describing men resulting in lower MCS than vignettes describing

women. Thus, ChatGPT-4 assumed a greater impairment in mental HRQoL for men compared to women with similar ED severity. Since there is no evidence from previous studies that supports this finding, this can be considered a bias.

Interpretation

While the effect for gender was statistically significant, it is also important to look at the minimal clinically important difference (MCID), that is, to evaluate whether differences in scores would be clinically relevant [67]. For the MCS, the MCID has been estimated to be between 3 and 9 points [68, 69]. With a difference of 2.3, the gender effect found in this study was slightly below an MCID. However, a longitudinal study showed that MCS scores in ED patients only improved 1 to 6 points during two years of treatment even though ED symptoms improved markedly, which highlights the clinical relevance of below-MCID differences in MCS scores in subjects with ED[70].

Of note, the EDE-Q scores generated by ChatGPT-4 were around 1.6 points above the scores reported in ED samples [e.g., 71, 72, 73]. Likewise, the MCS scores generated by ChatGPT-4 were around 20 points below mean scores in other ED cohorts [74, 75]. This has implications on the evaluation of the above mentioned MCID, as potential floor effects need to be considered.

The gender bias delivered by ChatGPT-4 could be due to social roles assuming general lower mental problems in men than in women and consequently evoking more attention if mental problems are identified. Thus, ChatGPT-4 in this case could function as a mirror through highlighting possible prejudices and help us to correct them in real life. In the field of EDs, the role of gender, sexual orientation as well as the influence of stigmatization and biases in our society need to be understood better [45, 76].

Strengths and Limitations

Our study has several strengths: First, real vignettes from scientific publications were used and varied in a way that the distinct influence of gender and sexual orientation could be singled out. To our knowledge, this is the first study which tests a potential bias when instructing an LLM to evaluate clinical cases with the use of psychometric instruments. Second, while many studies mentioned in this paper have used ChatGPT-3.5, we used ChatGPT-4, which has been shown to perform better in the field of mental health (e.g., 18). Third, by applying repeated testing we reached a much larger sample size than other vignette studies, ensuring sufficient power for our analyses.
This study also has limitations. First, the gender ratio of the original vignettes was not balanced (only two male vignettes), which might have had an impact on the evaluation of these vignettes. However, this ratio approximately reflects the gender ratio of AN and BN in the general population. Second, even though we sought to set the temperature to 0 and followed available instructions to do so when using the applied interface, we could not verify whether the setting of the temperature via „custom instructions“ actually resulted in respective changes in the system setting of the temperature. Lastly, the deviations in EDE-Q and MCS scores raise the question whether scores generated by ChatGPT-4 can be transferred to scores reported in ED research and highlight that the use of LLMs for scoring patient vignettes is still in the fledging stages.

Implications and Future Directions

Our findings highlight the importance of examining biases in LLMs in the context of (mental) health care. Future studies should investigate the generalizability of these findings by exploring biases in other LLMs as well as in other fields of (mental) health. As ChatGPT-4 has been found to disregard conditions that are understudied [25], being aware of research and knowledge gaps as well as existing biases and stigma in society when using and training LLMs is of high importance. Further,

potential mitigation strategies for biases introduced by LLMs should be investigated. Even though AI is not widely used yet in the assessment of disorders, it is already used in assisting doctor's decision making [e.g., 66, 67]. Further, ChatGPT-3.5 has been used to generate more diverse and inclusive case vignettes to be used in medical education [77]. It has been proposed that in healthcare, specially trained LLMs are needed, as ChatGPT-4 was not intended to be used in a clinical context [78] and was deemed unreliable in offering personalized medical advice [26]. When training such LLMs, policy makers should make sure that measures are taken to minimize biases in the training material and that proposed frameworks for responsible AI [e.g., 79] are considered.

Conclusions

This study showed that ChatGPT-4 might exhibit a potential gender bias when evaluating ED symptomatology and mental HRQoL. Researchers as well as clinicians should be aware of potential biases when using LLMs to support clinical decision making. Better understanding and mitigation of risk of bias related to gender and other factors, such as ethnicity or socioeconomic status, are highly warranted to ensure responsible use of LLMs.

Acknowledgements

RS1 contributed to the conceptualization, methodology, and data collection, conducted the formal analysis, and wrote the original draft of the paper. NR contributed to the writing of the original draft. RS2 contributed to the conceptualization and manuscript review and editing. LL contributed to the conceptualization and manuscript review and editing. GM contributed to the conceptualization, methodology, formal analysis, writing the original draft, and manuscript review and editing. All authors read and approved the final submitted version of the paper.
RS1 is funded by the Swiss State Secretariat for Education, Research and Innovation (SERI, under funding number: 22.00094) in the context of a European Union (Horizon Europe) research consortium "Long Covid" (funding number: 101057553).

Conflicts of Interest

RS2 and GM received funding from the Stanley Thomas Johnson Stiftung & Gottfried und Julia Bangerter-Rhyner-Stiftung under projects no. PC 28/17 and PC 05/18, from Gesundheitsförderung Schweiz under project no. 18.191/K50001 and in the context of a Horizon Europe project from the Swiss State Secretariat for Education, Research and lnnovation (SERI) under contract number 22.00094 and from Wings Health in the context of a proof-of-concept study. GM received funding from the Swiss Heart Foundation under project no. FF21101, from the Research Foundation of the International Psychoanalytic University (IPU) Berlin under projects no. 5087 and 5217, from the German Federal Ministry of Education and Research under budget item 68606, from the Hasler Foundation under project No. 23004. GM is a co-founder, member of the board, and shareholder of Therayou AG, active in digital and blended mental healthcare. GM receives royalties from publishing companies as author, including a book published by Springer, and an honorarium from Lundbeck for speaking at a symposium. Furthermore, GM is compensated for providing psychotherapy to patients, acting as a supervisor, serving as a self-experience facilitator (‘*Selbsterfahrungsleiter*’), and for postgraduate training of psychotherapists, psychosomatic specialists, and supervisors. NR is a co-worker at Therayou AG, active in digital and blended mental healthcare. NR received funding from the Hasler Foundation under project No. 23004 and from Wings Health AG in the context of a proof-of-concept study.

Abbreviations

AI: artificial intelligence

AN: anorexia nervosa
BN: bulimia nervosa
ED: eating disorder
EDE-Q: eating disorder examination questionnaire
ChatGPT-4: Chat Generative Pre-trained Transformer-4
GAF: global assessment of functioning
HRQoL: health-related quality of life
LLM: large language model
MCID: minimal clinically important difference
MCS: mental composite summary
MLM: multilevel model
PCS: physical composite summary
PTSD: post-traumatic stress disorder

References

1. Demszky, D., et al., *Using large language models in psychology.* Nature Reviews Psychology, 2023.
2. OpenAi, et al., *GPT-4 Technical Report.* 2023.
3. *LaMDA: our breakthrough conversation technology.* Google, 2021.
4. Thirunavukarasu, A.J., et al., *Large language models in medicine.* Nature Medicine, 2023. **29**(8): p. 1930-1940.
5. Pennebaker, J.W., M.R. Mehl, and K.G. Niederhoffer, *Psychological Aspects of Natural Language Use: Our Words, Our Selves.* Annual Review of Psychology, 2003. **54**(1): p. 547-577.
6. De Choudhury, M., et al., *Predicting Depression via Social Media.* Proceedings of the International AAAI Conference on Web and Social Media, 2021. **7**(1): p. 128-137.
7. Vaswani, A., et al., *Attention is all you need.* Advances in neural information processing systems, 2017. **30**.
8. Liddy, E.D., *Natural Language Processing.* Encyclopedia of Library and Information Science, 2001.
9. Meyer, J.G., et al., *ChatGPT and large language models in academia: opportunities and challenges.* BioData Mining, 2023. **16**(1): p. 20,-s13040-023-00339-9.
10. Boyd, R.L. and H.A. Schwartz, *Natural Language Analysis and the Psychology of Verbal Behavior: The Past, Present, and Future States of the Field.* Journal of Language and Social Psychology, 2021. **40**(1): p. 21-41.
11. Rehm, J. and K.D. Shield, *Global Burden of Disease and the Impact of Mental and Addictive Disorders.* Current Psychiatry Reports, 2019. **21**(2): p. 10.
12. Chaulagain, A., et al., *WHO Mental Health Gap Action Programme Intervention Guide (mhGAP-IG): the first pre-service training study.* International Journal of Mental Health Systems, 2020. **14**(1): p. 47.
13. Kjell, O.N.E., K. Kjell, and H.A. Schwartz, *Beyond rating scales: With targeted evaluation, large language models are poised for psychological assessment.* Psychiatry Research, 2024. **333**: p. 115667.
14. Kjell, O.N.E., et al., *Natural language analyzed with AI-based transformers predict traditional subjective well-being measures approaching the theoretical upper limits in accuracy.* Scientific Reports, 2022. **12**(1): p. 3918.
15. Park, G., et al., *Automatic personality assessment through social media language.* Journal of Personality and Social Psychology, 2015. **108**(6): p. 934-952.
16. Kjell, O.N.E., et al., *Semantic measures: Using natural language processing to measure, differentiate, and describe psychological constructs.* Psychological Methods, 2019. **24**(1): p. 92-115.
17. Son, Y., et al., *World Trade Center responders in their own words: predicting PTSD symptom trajectories with AI-based language analyses of interviews.* Psychological Medicine, 2023. **53**(3): p. 918-926.
18. Levkovich, I. and Z. Elyoseph, *Suicide Risk Assessments Through the Eyes of ChatGPT-3.5 Versus ChatGPT-4: Vignette Study.* JMIR Ment Health, 2023. **10**: p. e51232.
19. Franco D'Souza, R., et al., *Appraising the performance of ChatGPT in psychiatry using 100 clinical case vignettes.* Asian J Psychiatr, 2023. **89**: p. 103770.
20. Ntoutsi, E., et al., *Bias in data-driven artificial intelligence systems—An introductory survey.* WIREs Data Mining and Knowledge Discovery, 2020. **10**(3).
21. Navigli, R., S. Conia, and B. Ross, *Biases in Large Language Models: Origins, Inventory, and Discussion.* Journal of Data and Information Quality, 2023. **15**(2): p. 1-21.

22. Walsh, C.G., et al., *Stigma, biomarkers, and algorithmic bias: recommendations for precision behavioral health with artificial intelligence.* JAMIA Open, 2020. **3**(1): p. 9-15.
23. Rahimi, P., C. Ecabert, and S. Marcel, *Toward responsible face datasets: modeling the distribution of a disentangled latent space for sampling face images from demographic groups.* arXiv preprint arXiv:2309.08442, 2023.
24. Chen, I.Y., P. Szolovits, and M. Ghassemi, *Can AI Help Reduce Disparities in General Medical and Mental Health Care?* AMA Journal of Ethics, 2019. **21**(2): p. E167-179.
25. Nacher, M., U. Françoise, and A. Adenis, *ChatGPT neglects a neglected disease.* The Lancet Infectious Diseases, 2024. **24**(2): p. e76.
26. Nastasi, A.J., et al., *A vignette-based evaluation of ChatGPT's ability to provide appropriate and equitable medical advice across care contexts.* Sci Rep, 2023. **13**(1): p. 17885.
27. Zack, T., et al., *Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study.* The Lancet Digital Health, 2024. **6**(1): p. e12-e22.
28. Mosteiro, P., et al., *Bias Discovery in Machine Learning Models for Mental Health.* Information, 2022. **13**(5): p. 237.
29. Roehrig, J.P. and C.P. McLean, *A comparison of stigma toward eating disorders versus depression.* International Journal of Eating Disorders, 2010. **43**(7): p. 671-674.
30. Gallagher, K.A., et al., *Evaluating gender bias in an eating disorder risk assessment questionnaire for athletes.* Eating Disorders, 2021. **29**(1): p. 29-41.
31. Westmoreland, P., M.J. Krantz, and P.S. Mehler, *Medical Complications of Anorexia Nervosa and Bulimia.* The American Journal of Medicine, 2016. **129**(1): p. 30-37.
32. Richard, M., *Effective treatment of eating disorders in Europe: treatment outcome and its predictors.* European Eating Disorders Review, 2005. **13**(3): p. 169-179.
33. Galmiche, M., et al., *Prevalence of eating disorders over the 2000–2018 period: a systematic literature review.* The American Journal of Clinical Nutrition, 2019. **109**(5): p. 1402-1413.
34. Gorrell, S. and S.B. Murray, *Eating Disorders in Males.* Child Adolesc Psychiatr Clin N Am, 2019. **28**(4): p. 641-651.
35. Hudson, J.I., et al., *The Prevalence and Correlates of Eating Disorders in the National Comorbidity Survey Replication.* Biological Psychiatry, 2007. **61**(3): p. 348-358.
36. Sweeting, H., et al., *Prevalence of eating disorders in males: a review of rates reported in academic research and UK mass media.* Int J Mens Health, 2015. **14**(2).
37. Timko, C.A., L. DeFilipp, and A. Dakanalis, *Sex Differences in Adolescent Anorexia and Bulimia Nervosa: Beyond the Signs and Symptoms.* Current Psychiatry Reports, 2019. **21**(1): p. 1.
38. Støving, R.K., et al., *Gender differences in outcome of eating disorders: A retrospective cohort study.* Psychiatry Research, 2011. **186**(2): p. 362-366.
39. Strobel, C., et al., *Long-term outcomes in treated males with anorexia nervosa and bulimia nervosa-A prospective, gender-matched study.* Int J Eat Disord, 2019. **52**(12): p. 1353-1364.
40. Bothe, T., J. Walker, and C. Kröger, *Gender-related differences in health-care and economic costs for eating disorders: A comparative cost-development analysis for anorexia and bulimia nervosa based on anonymized claims data.* International Journal of Eating Disorders, 2022. **55**(1): p. 61-75.
41. Brelet, L., et al., *Stigmatization toward People with Anorexia Nervosa, Bulimia Nervosa, and Binge Eating Disorder: A Scoping Review.* Nutrients, 2021. **13**(8): p. 2834.
42. Cao, Z., et al., *The association between sexual orientation and eating disorders-related eating behaviours in adolescents: A systematic review and meta-analysis.* European Eating Disorders Review, 2023. **31**(1): p. 46-64.
43. Boisvert, J.A. and W.A. Harrell, *Homosexuality as a Risk Factor for Eating Disorder Symptomatology in Men.* The Journal of Men's Studies, 2010. **17**(3): p. 210-225.
44. Strübel, J. and T.A. Petrie, *Sexual orientation, eating disorder classification, and men's*

psychosocial well-being. Psychology of Men & Masculinities, 2020. **21**(2): p. 190-200.
45. Dotan, A., R. Bachner-Melman, and S.C. Dahlenburg, *Sexual orientation and disordered eating in women: a meta-analysis.* Eating and Weight Disorders - Studies on Anorexia, Bulimia and Obesity, 2021. **26**(1): p. 13-25.
46. García-Anaya, M., A. Caballero-Romo, and L. González-Macías, *Parent-Focused Psychotherapy for the Preventive Management of Chronicity in Anorexia Nervosa: A Case Series.* Int J Environ Res Public Health, 2022. **19**(15).
47. Olson, A.F., *Outpatient management of electrolyte imbalances associated with anorexia nervosa and bulimia nervosa.* J Infus Nurs, 2005. **28**(2): p. 118-22.
48. Gurevich, M.I., M.K. Chung, and P.J. LaRiccia, *Resolving bulimia nervosa using an innovative neural therapy approach: two case reports.* Clin Case Rep, 2018. **6**(2): p. 278-282.
49. Manuelli, M., et al., *Changes in eating behavior after deep brain stimulation for anorexia nervosa. A case study.* Eat Weight Disord, 2020. **25**(5): p. 1481-1486.
50. González-Macías, L., A. Caballero-Romo, and M. García-Anaya, *Group family psychotherapy during relapse. Case report of a novel intervention for severe and enduring anorexia nervosa.* Salud mental, 2021. **44**(1): p. 31-37.
51. Safer, D.L., C.F. Telch, and W.S. Agras, *Dialectical behavior therapy adapted for bulimia: a case report.* Int J Eat Disord, 2001. **30**(1): p. 101-6.
52. Srinivasa, P., et al., *Case report on anorexia nervosa.* Indian J Psychol Med, 2015. **37**(2): p. 236-8.
53. Berman, M.I., K.N. Boutelle, and S.J. Crow, *A case series investigating acceptance and commitment therapy as a treatment for previously treated, unremitted patients with anorexia nervosa.* Eur Eat Disord Rev, 2009. **17**(6): p. 426-34.
54. Barbosa Pinto, M., et al., *A Case Report of Anorexia Nervosa - the "perfect" woman.* European Psychiatry, 2022. **65**(S1): p. S583-S584.
55. Laser, E. and M. Sassack. *Treating bulimia with hypnosis and low-level light therapy: a case report*. in *Mechanisms for Low-Light Therapy VII*. 2012. SPIE.
56. Morgan, C.D. and C. Marsh, *Bulimia nervosa in an elderly male: a case report.* Int J Eat Disord, 2006. **39**(2): p. 170-1.
57. Sansone, R.A., A. Naqvi, and L.A. Sansone, *An unusual cause of dizziness in bulimia nervosa: a case report.* Int J Eat Disord, 2005. **37**(4): p. 364-6.
58. Hays, R.D., C.D. Sherbourne, and R.M. Mazel, *The rand 36-item health survey 1.0.* Health Economics, 1993. **2**(3): p. 217-227.
59. Jenkins, P.E., et al., *Health-related quality of life among adolescents with eating disorders.* Journal of Psychosomatic Research, 2014. **76**(1): p. 1-5.
60. Ware, J.E. and I. New England Medical Center Hospital Health, *SF-36 physical and mental health summary scales : a user's manual*. 1994, Boston: Health Institute, New England Medical Center Boston.
61. Andersen, J.R., et al., *Correlated physical and mental health composite scores for the RAND-36 and RAND-12 health surveys: can we keep them simple?* Health and Quality of Life Outcomes, 2022. **20**(1): p. 89.
62. Fairburn, C.G. and S.J. Beglin, *Assessment of eating disorders: interview or self-report questionnaire?* Int J Eat Disord, 1994. **16**(4): p. 363-70.
63. Allaire, J., *RStudio: integrated development environment for R.* Boston, MA, 2012. **770**(394): p. 165-171.
64. Bates, D., et al., *Fitting linear mixed-effects models using lme4.* arXiv preprint arXiv:1406.5823, 2014.
65. Bliese, P.D., *Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis*, in *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions.* 2000, Jossey-Bass/Wiley: Hoboken, NJ, US. p.

349-381.
66. Wickham, H. and H. Wickham, *Data analysis*. 2016: Springer.
67. Dettori, J.R., D.C. Norvell, and J.R. Chapman, *Clinically Important Difference: 4 Tips Toward a Better Understanding.* Global Spine J, 2022. **12**(6): p. 1297-1298.
68. Ferguson, R.J., A.B. Robinson, and M. Splaine, *Use of the reliable change index to evaluate clinical significance in SF-36 outcomes.* Qual Life Res, 2002. **11**(6): p. 509-16.
69. Samsa, G., et al., *Determining clinically important differences in health status measures: a general approach with illustration to the Health Utilities Index Mark II.* Pharmacoeconomics, 1999. **15**(2): p. 141-55.
70. Padierna, A., et al., *Changes in health related quality of life among patients treated for eating disorders.* Qual Life Res, 2002. **11**(6): p. 545-52.
71. Jennings, K.M. and K.E. Phillips, *Eating Disorder Examination-Questionnaire (EDE-Q): Norms for Clinical Sample of Female Adolescents with Anorexia Nervosa.* Arch Psychiatr Nurs, 2017. **31**(6): p. 578-581.
72. Aardoom, J.J., et al., *Norms and discriminative validity of the Eating Disorder Examination Questionnaire (EDE-Q).* Eat Behav, 2012. **13**(4): p. 305-9.
73. Jennings, K.M. and K.E. Phillips, *Eating Disorder Examination-Questionnaire (EDE-Q): Norms for a Clinical Sample of Males.* Arch Psychiatr Nurs, 2017. **31**(1): p. 73-76.
74. Padierna, A., et al., *The health-related quality of life in eating disorders.* Qual Life Res, 2000. **9**(6): p. 667-74.
75. Doll, H.A., S.E. Petersen, and S.L. Stewart-Brown, *Eating disorders and emotional and physical well-being: associations between student self-reports of eating disorders and quality of life as measured by the SF-36.* Qual Life Res, 2005. **14**(3): p. 705-17.
76. O'Connor, C., et al., *How do people with eating disorders experience the stigma associated with their condition? A mixed-methods systematic review.* J Ment Health, 2021. **30**(4): p. 454-469.
77. Bakkum, M.J., et al., *Using artificial intelligence to create diverse and inclusive medical case vignettes for education.* Br J Clin Pharmacol, 2024. **90**(3): p. 640-648.
78. Li, J., et al., *ChatGPT in healthcare: A taxonomy and systematic review.* Computer Methods and Programs in Biomedicine, 2024. **245**: p. 108013.
79. Ray, P.P., *Benchmarking, ethical alignment, and evaluation framework for conversational AI: Advancing responsible development of ChatGPT.* BenchCouncil Transactions on Benchmarks, Standards and Evaluations, 2023. **3**(3): p. 100136.