Watching the Watchmen: Assessment-Biases in Waiting List Prioritization for the Delivery of Mental Health Services

Kreiseder, Fabian; Mosenhauer, Moritz

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Empfohlene Zitierung / Suggested Citation:
Kreiseder, F., & Mosenhauer, M. (2022). Watching the Watchmen: Assessment-Biases in Waiting List Prioritization for the Delivery of Mental Health Services. *European Journal of Management Issues*, *30*(1), 3-16. https://doi.org/10.15421/192201

European Journal of Management Issues

Volume 30(1), 2022, pp. 3-16

DOI: 10.15421/192201

Received: 31 December 2021
Revised: 26 January 2022
Accepted: 21 February 2022
Published: 03 March 2022

JEL Classification: I38, I31, H41

Watching the Watchmen: Assessment-Biases in Waiting List Prioritization for the Delivery of Mental Health Services

F. Kreiseder[‡], M. Mosenhauer[#]

Purpose: While the demand for mental health services increases, supply often stagnates. Providing treatment to those most in need is an important factor in its efficient distribution. We propose and conduct a statistical procedure for detecting rater-biases in patient prioritization tools.

Design / Method / Approach: We gather real-life data from 266 illness severity assessments in an Austrian publicly funded mental health service provider, including a rich set of covariates. To ensure robustness, we merge this data with determinants of mental health and assessment identified by previous research, such as weather or seasonal indicators.

Findings: We find statistically significant effects of rater-biases. These effects are robust to a large array of controls.

Practical Implications: A back-of-the-envelope calculation reveals that the identified rater effects can translate to large changes in the waiting times for patients. Misspecified treatment allocations may lead to worsened symptoms and potentially fatal outcomes.

Originality / Value: Although a growing literature focuses on patient prioritization tools, many articles study these in synthetic contexts using "vignettes". In comparison, our study adds external validity by considering real-life treatments in the field.

Research Limitations / Future Research: This study can be used as a starting point for deeper, causally focused studies.

Disclaimer: In accordance with publisher policies and our ethical obligations as researchers, we report that one of the authors is employed at a company that may be affected by the research reported in the enclosed paper. We have disclosed those interests fully.

[‡] *Fabian Kreiseder,*
Managing Shareholder,
PriorizR – famado GmbH, Austria,
e-mail: fabian.kreiseder@priorizr.com

[#] *Moritz Mosenhauer,*
Assistant Professor,
MCI Management Center Innsbruck, Austria,
e-mail: moritz.mosenhauer@mci.edu,
https://orcid.org/0000-0001-5275-0258

Paper type: Empirical

Keywords: patient prioritization tools, illness severity assessment, rater-based effects, mental health.

Reference to this paper should be made as follows:
Kreiseder, F., Mosenhauer, M. (2022). Watching the Watchmen: Assessment-Biases in Waiting List Prioritization for the Delivery of Mental Health Services. *European Journal of Management Issues*, 30(1), 3-16. doi:*10.15421/192201*.

ISSN 2519-8564 (print), ISSN 2523-451X (online). *European Journal of Management Issues*. 2022. Vol. 30(1)

Спостереження за спостерігачем: викривлена оцінка при пріоритизації черги надання послуг стосовно психічного здоров'я

Фабіан Крейседер [ǂ]
Моріц Мозенгауер [#]

[ǂ] *PriorizR – famado GmbH, Австрія*
[#] *MCI Центр менеджменту м. Інсбрук, Австрія*

Мета роботи: У той час як попит на послуги стосовно психічного здоров'я зростає, пропозиція часто стагнує. Надання лікування тим, хто найбільше потребує, є важливим фактором його ефективного розподілу. Ми пропонуємо та проводимо статистичну процедуру для виявлення викривлення оцінок у інструментах визначення пріоритетів пацієнтів.

Дизайн / Метод / Підхід дослідження: Ми зібрали реальні дані 266 оцінок тяжкості захворювання в австрійській державній установі психіатричної допомоги, включаючи багатий набір коваріацій. Для забезпечення надійності ми поєднали ці дані з детермінантами психічного здоров'я та оцінки, визначеними у попередніх дослідженнях, такими як погодні чи сезонні показники.

Результати дослідження: Ми виявили статистично значущий вплив оцінок-упереджень. Цей вплив є стійким до великої кількості контролів.

Практична цінність дослідження: Зворотний розрахунок показує, що виявлений вплив оцінок-упереджень може спричинити значні зміни у часі очікування пацієнтів. Неправильний розподіл черги на лікування може призвести до погіршення симптомів та потенційно смертельних наслідків.

Оригінальність / Цінність дослідження: Хоча все більше літератури присвячено інструментам визначення пріоритетів пацієнтів, багато статей вивчають їх у синтетичних контекстах, використовуючи «віньєтки». У порівнянні з цим наше дослідження додає зовнішню достовірність, розглядаючи реальні методи лікування в польових умовах.

Обмеження дослідження / Майбутні дослідження: Дане дослідження може бути використане як відправна точка для більш глибоких, причинно орієнтованих досліджень.

Заява про відмову від відповідальності: Відповідно до політики видавництва та наших етичних зобов'язань як дослідників, ми повідомляємо, що один з авторів працює в компанії, на яку може вплинути дослідження, представлене в цій статті. Ми розкрили ці інтереси.

Тип статті: Емпіричний

Ключові слова: інструменти визначення пріоритетів пацієнтів, оцінка тяжкості захворювання, вплив на оцінку, психічне здоров'я.

1. Introduction

Mental health services are confronted with an ever-widening mental health treatment gap. As a result, prolonged waiting times exacerbate symptoms *(Clark et al., 2018; Reichert & Jacobs, 2018)* and economic costs (*Rechnungshof [RH], 2019*). Recently mental health services started to adopt need-based waiting list strategies. In the course of such, patients are assessed and prioritized based on the resulting assessment score. With real-world data, we investigate to what extent those priority scores are independent of their raters. Because ideally who scores the patient should not affect the priority score and subsequently not a patient's access to mental health treatment.

Between 2005 and 2017 rates of major depressive episodes grew from 8.7% to 13.2% for adolescents between age 12 and 17; for young adults the rates from 2009 to 2017 inclined from 8.1% to 13.2% (*Twenge et al., 2019*). This may lead to a deteriorated performance in school, alcohol and drug consumption, bingeing, and suicidal ideation (*Glied & Pine, 2002*). The outbreak of the novel coronavirus has further deteriorated the mental health condition of many individuals (*Brooks et al., 2020; Talevi et al., 2020*). As a consequence, and despite the still prevailing stigma *(Corrigan et al., 2014)*, mental health service utilization has increased *(Lipson et al., 2019)*. However, while the demand has soared, the supply of mental health services stagnated, leaving many people untreated (*Mojtabai et al., 2016*).

Since privately financed treatment is for many not affordable (*Berufsverband Österreichischer PsychologInnen [BÖP] & Karmasin Research & Identity, 2020*), publicly funded treatment becomes often the only option available, but access to these services is connected with long waiting times (*Luigi et al., 2013*). Conversely, waiting time poses a major obstacle in accessing health care *(McIntyre & Chow, 2020)* and brings several other negative effects such as patient dissatisfaction *(Lizaur-Utrilla et al., 2016; Nottingham et al., 2018)*, patient anxiety (*Lizaur-Utrilla et al., 2016*), and significantly poorer health outcomes *(Clark et al., 2018; Reichert & Jacobs, 2018)*.

Therefore, the reduction of waiting time is a pressing issue for all stakeholders in mental health services. A possible solution to this problem might be the need-based allocation of scarce resources. The equivalent of this idea in healthcare might be found in patient prioritization tools. Such instruments assess and prioritize those patients with the highest level of need and allocate resources accordingly. But assessing patients on specific criteria is complex and sometimes subject to personal characteristics *(Raymond et al., 2017)*. This further resonates with the common understanding in literature that clinical judgment is not without its flaws *(Bell & Mellor, 2009)* and often prone to produce biased results *(Samuel & Bucher, 2017)*. These shortcomings are also reflected in the heterogeneous findings of earlier studies *(Déry et al., 2020; Harding & Taylor, 2013)* that investigated the tool's reliability and validity. Regardless, determining a patient's level of priority should not be influenced by the rater that conducts the scoring, especially not if a patient prioritization tool aims to provide a fair and transparent priority assessment (*Harding & Taylor, 2013*).

For our analysis, we obtained real-world data from the assessment center of an Austrian publicly funded mental health care provider. In this assessment center, incoming patients are scored by psychotherapists on several different dimensions, all of which aim to quantify a patient's level of need for treatment. With this scoring data, we attempt to measure the potential effect that each rater has on the resulting priority score of a patient. In an ideal scenario, no such effects should be measurable. Because a biased assessment would lead to unjustifiably prolonged waiting times for some individuals and thus to unwanted discrimination. Furthermore, being stuck on the waiting list, instead of receiving appropriate treatment may deteriorate existing symptoms and, in the worst case, it might produce fatal outcomes.

An important contribution of our paper is valuable real-world insights into the quality of patient prioritization tools. This is critical, given the fact that earlier works were mostly built on vignettes, neglecting the differing conditions in real-world settings, including stress, time pressure, and risk *(Patel et al., 2002)*; conditions that are, in fact, ubiquitous in mental health workers (*Rössler, 2012*). Furthermore, the concept of rater bias has been well addressed in other fields, such as in entrepreneurial settings (*Thomas, 2018*), in political beliefs (*Hibbing et al., 2014*), in forensic sciences (*Kassin et al., 2013*), and in grading (*Malouff, 2008*), to name a few. However, it has received notably less attention in patient priority assessments, and even more so for the ones employed in mental health settings. The findings of our study will hence add to the literature on rater bias. Ultimately, the results of this investigation are also highly relevant to practitioners as we highlight the concerns that are associated with the employment of such tools.

The structure of this article is as follows: section 2 will discuss the role of clinical judgment and biases as well as extraneous factors in patient priority assessments, leading to our hypothesis and the utilized control variables. Section 3 introduces the research question. Section 4 covers the methodology, research model and design of this study. Section 5 presents the study results. Section 6 contains a discussion of the results and their limitations. Finally, section 7 contains concluding remarks, which are followed by the bibliography.

2. Theoretical Background

This section sets out the theoretical foundation of this study, by giving an overview of how errors in clinical judgment occur. It also deals with the associated concept of rater biases and concludes then with a set of extraneous factors that facilitate such errors and biases.

2.1. Clinical Judgment and Bias

Many patient prioritization tools rely on scoring processes to determine a patient's level of need. In general, scoring, as a form of measurement, requires "the assigning of numbers to observations (…) to quantify phenomena" *(Kimberlin & Winterstein, 2008, p. 2276)*. These observations are usually made by clinicians or any other trained rating personnel and are succeeded by a judgment that quantifies the respective phenomenon with the help of rating scales. However, many of the phenomena in health care are theoretical constructs that cannot be measured precisely (*Kimberlin & Winterstein, 2008*), as clinicians find themselves in a position where they have to infer psychological characteristics, which are internal in nature, from the external behavior of a person (*Reynolds & Suzuki, 2013*). For example, how do you properly assess the suffering of a patient due to his or her condition? Having depression might be seriously debilitating for one person but a minor negative effect for another. The quantifying of abstract concepts thus creates room for uncertainty, which in turn provides fertile ground for judgment errors (*Croskerry, 2002*).

And indeed, rater-based assessments are often found to be inaccurate, as raters tend to form categorical judgments about their examinees, but when these judgments are translated into ordinal or interval scales, conversion errors may arise *(Gingerich et al., 2011)*. *Bell & Mellor (2009)*, in their review on clinical judgments, also emphasized the lack of accuracy and reliability in many clinical judgments. In contrast, *Christensen-Szalanski et al. (1982)* advocated for the soundness of clinical judgment, provided that clinicians are experienced in their field. Whereas *López (1989)* found that the cognitive processes involved in clinical judgment are possibly influenced by irrelevant patient variables, which may consequently bias the judgment of clinicians, irrespective of the clinician's experience. In line with *Samuel & Bucher (2017)* who concluded in their review that naturalistic clinical assessments are just as biased as any other human assessment source, which questions the perception of clinicians' assessment abilities as the

gold standard for valid clinical judgment. Nevertheless, such biases lead to deviations from the objective truth and thus constitute a sometimes severe measurement error.

Judgment errors may also stem from systematic biases based on distortions in perceiving and/or processing information. They affect all kinds of human cognition and can thus not be attributed to one particular field nor the general cognitive ability of an individual (*West et al., 2008*), as they arise from both analytical and non-analytical thinking and can be the result of multiple causes (*Norman & Eva, 2010*). For example, their presence was observed in entrepreneurial settings (*Thomas, 2018*), in political beliefs *(Hibbing et al., 2014*), in forensic sciences (*Kassin et al., 2013*), and in grading (*Malouff, 2008*), to name a few. So, there is no reason to believe that priority assessments, conducted by therapists in mental health care settings, would be immune to such biases. Many scientific works have proven that bias is, in fact, also prevalent among clinicians *(Bowes et al., 2020; Hairston et al., 2019; Wolfson et al., 2000*). However, the extent of bias in patient priority assessments was not yet properly addressed in the literature. Still, given the multitude of findings in other settings that indicate the presence of bias in clinicians, it is assumed that bias is likely to play a role in patient priority assessments as well.

To summarize, the potential shortcomings of vignettes *(Patel et al., 2002*), the flaws in clinical judgment (*Bell & Mellor, 2009*), and the associated susceptibility to bias (*Samuel & Bucher, 2017*), as well as the mixed results on reliability and validity of patient prioritization tools that earlier studies generated *(Déry et al., 2020; Harding & Taylor, 2013*), emphasize the necessity of this study. Moreover, for a priority assessment to be fair and transparent, it must be independent of its rater. Our hypothesis, therefore, attempts to measure the independence of scoring results from their rater.

Hypothesis

H_1: The size of the initial score is associated with the therapist that conducted the scoring.

The following subsection will more deeply discuss the potential sources of biases that could influence a clinician's rating.

2.2. Extraneous Factors

Many researchers have supported the notion that ratings are influenced by variables that should, in fact, be irrelevant in the decision-making process *(FitzGerald & Hurst, 2017; McDermott et al., 2014; Murrie et al., 2013; van Ryn & Burke, 2000*), nevertheless, they regularly become apparent in the results. The following paragraphs will thus provide theoretical reasoning for the application of these variables as control variables in the analysis.

First, one of the most prominent biases is certainly discrimination towards gender or sex, as decades worth of research has shown. Often facilitated by explicit mechanisms such as stereotypes but also implicitly by an unconscious bias, such effects are widespread and were measured across many disciplines. For example, in hiring (*Chan & Wang, 2018*), in the workplace (*Wynn & Correll, 2018*), in science (*Roper, 2019*), in terms of disproportional media coverage (*Shor et al., 2019*), or access to healthcare (*Ulasi, 2008*). In other words, gender bias is omnipresent and creates vast social inequities.

Regardless, the gender or sex of a person should not play a role in assessing a patient or determining the priority of a referral. Yet, the literature indicates that underlying gender stereotypes may bias the assessment abilities of raters. *Earp et al. (2019)*, for example, uncovered in their study that adults rated boys to experience more pain than girls, although both were under equal clinical circumstances and showed the same pain behavior. The authors also mentioned that when they controlled for explicit gender stereotypes the effects were eradicated. In a comparable analysis, *Yourstone et al. (2008)* found that when the perpetrator in a hypothetical criminal case was female, psychiatric clinicians and psychology students tended to rather declare the person as legally insane than compared to cases with male offenders.

Given the evidence on gender bias in clinical judgments, the sex of the patient will constitute a control variable in this analysis.

Several patient prioritization tools include age as a determining factor of priority (*Hadorn & Steering Committee of the Western Canada Waiting List Project, 2000; MacCormick et al., 2003*). However, other studies also indicate that *unwanted* age bias becomes apparent as well in priority decisions (*Arslanian-Engoren, 2000; Arslanian-Engoren & Scott, 2016; Platts-Mills et al., 2010*). If not explicitly specified as a criterion of priority, patient age should not play a role in patient prioritization tools. In the underlying data of our study, age is indeed a priority criterion but thus also accounted for in the score. Beyond that, age should not considerably influence the scoring. Hence, we also control for potential effects of age on the priority score.

The prevalence of mental disorders is also likely to underly seasonal variations. Such variations, for example, were found by *Graaf et al. (2005)*, even though they were minor and they focused their study on nations with a warm maritime climate, they discovered that overall the occurrence of mental disorders in winter is higher than in summer. On a comparable note, *Slaunwhite et al. (2019)* investigated seasonal variations in psychiatric admissions to hospitals and found that for children and adolescents the highest rate of admissions was measured in February and for adults in May.

A different seasonal effect may become apparent due to Seasonal Affective Disorder (SAD). A fairly common condition *(Magnusson, 2000*) where affected individuals experience depression, along with other symptoms, in recurring seasonal intervals, mostly in winter (*Magnusson & Boivin, 2003*). Even though it can be treated quite successfully with light therapy, it affects approximately 1-3% of people living in moderate climate zones, with women being more likely to be affected by SAD than men *(Magnusson & Boivin, 2003*). Hence SAD appears to be fairly common in the population and its effects are thus likely to reveal seasonal differences in scoring results.

Furthermore, several studies reported on the deteriorated mental health due to the COVID-19 pandemic and lockdowns. A US study, conducted by *Adams-Prassl et al. (2020)*, revealed that mental health was reduced by .085 standard deviations, an effect that is exclusively driven by women, leading to a widening of the gender gap in mental health. Similar declines in mental health were found also in the Italian population (*Rossi et al., 2020*). An Austrian Study (*Pieh et al., 2020*) reported a surge in depressive and anxiety symptoms. Apart from females, this effect seemed to be most stressful for young adults and socially disadvantaged groups like low-income and jobless individuals. During the pandemic, we encountered a number of both strict and less strict lockdowns, which often lasted several weeks. The effects of such restrictions, as discussed above, could lead to spikes in the prevalence and severity of mental disorders.

Given the frequency of SAD and the consequences of the pandemic on mental health, it is probable to see seasonal fluctuations influencing the scoring results. We attempt to control for this scenario, by using the months in which the scoring was conducted, as a measure to reveal seasonal differences.

On another note, many scientific articles have pointed to the strong influence of weather on our decision-making. In most cases this relationship is not a direct one, rather it is moderated by mood. This means that certain weather conditions influence our mood and the mood, in turn, affects our decision-making.

In general, people in good moods tend to be more optimistic in their choices (*Hirshleifer & Shumway, 2003*). Thereby they often rely on their System 1 thinking (i.e. intuitive thinking) and are thus more prone to the application of heuristics (*Bless et al., 1996*). For example, *Murray et al. (2010)* have found that consumer spending tends to increase when negative affect declines as a result of increased exposure to sunlight. Examples are also found in the

financial world. *Hirshleifer & Shumway (2003)*, for instance, found in their study a strong correlation between sunshine and daily stock returns. Similarly, *Goetzmann & Zhu (2005)* looked in their study at returns on cloudy versus sunny days, but different to *Hirshleifer & Shumway (2003)* they could not identify any significant connections. Yet, they pointed out that such an effect was found in market makers' spreads, where spreads would increase on cloudy days.

On the other hand, when in a sad mood, people tend to evaluate their choices more critically (*Bless et al., 1996*) and follow the more logical rule (*Vries et al., 2012*). For example, *Bakhshi et al. (2014)* showed that restaurant recommendation ratings tended to be lower when submitted on rainy days (i.e. precipitation > 0). It also appears that weather influences the outcome of elections. A study conducted by *Meier et al. (2019)* revealed that rainy weather on election days favors parties with more conservative agendas. The authors explained this phenomenon with the reduced willingness to take risks on rainy days.

The effect of weather on our mood and ultimately on our decision-making is a well-known issue in the literature. Therefore, it is assumed that such an influence could also play a role in the scoring process of patient prioritization tools. With a selected set of weather variables, we attempt to control for such potential influences.

3. Research Question

This study intends to explore to what extent raters in patient prioritization tools influence the priority score that patients receive, which leads to the following research question:

How independently are priority scorings applied in the course of patient prioritization tools from the rater that conducted the scoring?

4. Data and Methods

This section starts with a description of the research setting and continues with an elaboration of the research model. It furthermore describes the sample, the data collection, and data analysis.

Our study takes place in an Austrian social facility that offers a set of services that are aimed to help and support people in psychological and social crises. One such service is the facility's psychotherapy department. There, more than 60 therapists, both employed and in partnership with the institution, are supporting roughly 3000 people annually, either in one of its five locations or in the private practices of its partners.

The state Vorarlberg is one of the main customers of this psychotherapy service. Therefore, it is mostly but indirectly financed by public resources. In turn, it can and must offer its patients affordable treatment plans, which is also why the demand for psychotherapy there is high and, as a result, its waiting lists fairly long.

In October 2020 the institution adopted a need-based waiting list strategy. This endeavor led to the introduction of an assessment center. A first contact point that determines a patient's degree of prioritization and, ultimately, how soon one sees a therapist. In the course of this assessment center, people go through an initial interview, which is usually scheduled to last 90 minutes. The interviewers are employed therapists, and each of them works at one of the five locations of this psychotherapy service. The number of raters per location ranges from 1 to a maximum of 3. Thus, the assignment of a therapist to a case is mostly random but somewhat depending on the location the patient visits. Still, a patient is usually not aware of which therapist works at which facility and can therefore not consciously decide on who will perform the scoring.

In the initial interview, patient presentations are assessed based on several different dimensions, including urgency, severity, and suffering. But other factors that were also found to be determinants of mental health, such as the social situation and age, are covered as well. These and other assessment criteria were gathered in a criteria catalog, which is depicted in Tab. 1. Additionally, to establish some sort of standardization, therapists were trained and given a detailed assessment guideline.

As presented in *Tab. 1*, therapists assign points within a specified range for each criterion. To control how much each component contributes to the overall score, apart from the range of points, each measure has furthermore a fixed weighting. After multiplying the points per measure with its corresponding weighting, the weighted scores are summed up and the result is the initial score that determines at what rank a patient enters the waiting list. The total clearing scores range from 0 to 122, where increasing values represent a higher level of need.

However, relevant for this study is only the sum of the raw scores (i.e. column 3 in *Tab. 1*; later referred to as "score sum"), which ranges between 0 and 46 and represents the dependent variable in our study. The unweighted score sum, instead of the weighted one, was chosen because it unadulteratedly reflects the judgment of a therapist on a given case. The sum, instead of one chosen criterion, because it expresses the accumulated judgment on a case. If any effects were to be found, the chances of measuring them would increase with the accumulation of all decisions made.

The score sum thus should quantify the patient's level of need, measured through the judgment of one therapist. Yet, as discussed in the literature review, clinical judgments are sometimes inaccurate and distorted by extraneous factors. However, if the goal is to establish a fair and equitable waiting list system that considers the individual level of need, the quantification of claimed criteria, which precedes the entering of a patient on the list, must be accurate and, more importantly, independent of who scores the patient. Because in an ideal scenario, the factor therapist should not play a role in determining the patient's need for treatment.

From the 29th of October 2020 until the 19th of March 2021, 306 incoming patients, which either directly applied or were referred by a practitioner, were assessed and added to the waiting list. 267 patients were included for the descriptive analyses since 37 cases were not attributable to a specific rater and 2 additional outliers were dropped as well. For the regression analysis, 266 observations were included, since one case offered no data on age. All variables ultimately used in this study are depicted in *Tab. 2*.

In the sample, 70.4% (n = 188) of patients were female and 29.6% (n = 79) male. The patient's age ranges from 8 to 71 with a mean of 37.6 and a median of 36.0. The majority of patients were in the age group between 30 and 49 (39.5%; n = 105), followed by 19-29 (28.6%; n = 76) and 50-64 (23.3%; n = 62). The cohorts 13 to 18 made up only 4.1% (n = 11), 65 to 71 2.6% (n=7), and 6 to 12 only 1.9% (n = 5). All patients from the sample resided in the state of Vorarlberg, Austria. The sample, therefore, reflects a cross-section of the adult population in Vorarlberg.

During the study period, nine therapists were actively assessing patients. All raters were female and between the ages 35 and 60. Assessments took place in five different locations: Bludenz (8.2%; n = 22), Dornbirn (24.7%; n = 66), Bregenz (36.0%; n = 96), Feldkirch (14.6%; n = 39), and Hohenems (16.5%; n = 44).

243 out of 267 patients received a preliminary diagnosis and some of them were diagnosed with more than one mental illness. Just under half of the diagnosed disorders were attributable to neurotic, stress-related, and somatoform disorders and roughly 40% to mood disorders. The distribution of disorder types of this sample is mostly in line with the distribution on a global level, where both anxiety disorders (F41) and depression (F32-F33) are also the most prevalent types of mental disorders (*James et al., 2018*).

Table 1: Overview of criteria used in assessing patients' level of need

Dimension	Assessment criteria	Range of points	Weighting	Maximum of weighted points per category
Urgency	Urgency	***0-3***	6	24
	Continuation of treatment after inpatient stay	***0-1***	6	
Severity	GAF score	***0-7***	2	26
	Global assessment of severity	***0-4***	3	
Suffering	Intensity of suffering	***0-10***	1	18
	Duration of suffering	***0-4***	1	
	Current significant increase in suffering	***0-1***	4	
Social situation	Financial situation	***0-3***	5	31
	Social support	***0-2***	3	
	Impact on others	***0-2***	5	
Motivation to change	Motivation to change	***0-4***	2	8
Age	Age	***0-3***	3	9
Capability to attend group therapy	Capability to attend group therapy	***0-1***	4	4
Mobility	Mobility	***0-1***	2	2
	Sum Score	**0-46**		**Total: 122**

***Source*:** developed by the authors

Table 2: Descriptive statistics of variables used in the analysis

	N	Minimum	Maximum	Mean	Std. Deviation
Score sum	267	13	45	28.91	5.854
Therapist B	267	0	1	.18	.382
Therapist C	267	0	1	.07	.258
Therapist D	267	0	1	.21	.411
Therapist E	267	0	1	.11	.316
Therapist F	267	0	1	.03	.181
Therapist G	267	0	1	.01	.086
Therapist H	267	0	1	.16	.372
Therapist I	267	0	1	.14	.346
Sex Patient	267	0	1	.30	.457
Age Patient	266	8	71	37.56	14.119
November	267	0	1	.22	.413
December	267	0	1	.25	.434
January	267	0	1	.24	.425
February	267	0	1	.19	.394
March	267	0	1	.08	.275
Avg. Temperature in C°	267	-6.7	17.0	3.599	4.1503
Precipitation in mm	267	0	49.4	3.121	6.8322
Atmospheric pressure in hPa	267	984.10	1035.60	1018.9079	10.24458
Rain	267	0	1	.49	.501
Snow	267	0	1	.17	.378
Valid N (listwise)	266				

Note: the table describes the number of observations (*N*), minimum, maximum, mean, and standard deviation of the main variables used in the analysis. Variables with a range from 0 to 1 are coded as dummy variables, their mean thus indicates their actual share of observations. In the case of Sex patient, this means that 30% of patients were male, as they were coded as 1. This logic does not apply to Rain and Snow, given that on a certain day there can be both rain and snow.

***Source*:** developed by the authors

The obtained datasets contained the score for each assessment criterion, as presented in *Tab. 1*. The addition of all criteria represents the score sum. The dimensions and their respective criteria are described as follows:

Urgency was measured on a scale from 0 to 3, where 0 stands for no symptomatic deterioration to be expected and 3 for severe deterioration of symptoms and potential for self-harm or harm to others. The adoption of this item was inspired by the priority criteria tool developed by *Coster et al. (2007)*, however, it was altered to fit the mental health context. *Continuation of treatment after inpatient stay* was answered with a simple yes or no, where yes equals 1 and 0 no. Further treatment is mainly about stabilization after an intense mental illness. People are treated as inpatients only until they are stable. Outpatient treatment is then used for relapse prevention and long-term stabilization. Thus, former inpatient patients are prioritized in this category.

The severity dimension contained two criteria. First, a slightly modified version of the *Global Assessment of Functioning (GAF) scale* as described in the *DSM-IV (American Psychiatric Association, 1994)*; in which the 100-91 interval receives no points and gradually down to the 30-21 interval, for which the therapist assigns the maximum of 7 points. The intervals 20-11 and 10-1 are not considered, as such severe and urgent cases are immediately referred to a crises team. Second, a greatly simplified version of the scale for the *global assessment of severity* (Endicott et al., 1976), which ranges from 0 to 4, where 0 represents little and 4 high severity.

A large body of research indicates that suffering adversely affects the overall psychological well-being of patients as well as the symptoms of anxiety and depression *(see e.g. Cowden et al., 2021; Samelius et al., 2010*). Thus, the third category measures suffering from three different perspectives. First, the *intensity of suffering* ranges from 0 for no psychological strain to 10 for extremely debilitating psychological strain. Second, *duration of suffering*, which spans from 0-6 weeks up to 1 year, with the respective points from 0 up to 4. Third, a binary item for a *current significant increase in suffering*, with 1 point for yes and 0 for no.

Patients with a lower socio-economic status tend to wait longer for treatment (*McIntyre & Chow, 2020*), yet simultaneously they suffer disproportionately more under this burden *(Pathirana & Jackson, 2018*), which again puts a strain on already limited health care resources. Hence the institution included the social situation in its assessment criteria, which comprises the following three aspects. *Financial situation*, for which 0 represents full ability to self-finance psychotherapy and 4 for no or not enough resources to pay for treatment. *Social support* ranges from adequate support (0) to no support (2). *Impact on others*, where 1 represents an impact on adults, 2 an impact on children, and 0 no impact at all.

Decades' worth of research has shown that a patient's motivation for treatment and change is a strong predictor for good treatment outcomes (see e.g. *Keithly et al., 1980; Sifneos, 1978*). Thus, *motivation to change* constitutes another priority item during the assessment. It is evaluated by the scoring therapist and ranges from 0 for no motivation to 4 for highly motivated.

Given the steeply increasing prevalence rates of mental disorders among children and adolescents (*Twenge et al., 2019*) and the fact that early intervention and prevention have a higher probability for not only positive treatment outcomes but also improved long-term health as well as socio-economic gains (*Kieling et al., 2011*), the institution included the *age* dimension to prioritize younger patients. Patients from ages 0 to 12 receive 3 points, the age group 13 to 18 2 points, and 18 to 25 1 point. People older than 25 receive no points.

In this category, patients' *capability to attend group therapy* is evaluated. If a patient is incapable, the rater assigns 1 point and 0 if a capability is given. This criterion was implemented as the mental health service offers quick and uncomplicated access to group therapy sessions for patients on the waiting list as an early intervention measure. If, however, patients are ill-suited to group therapies, they are prioritized to sooner enter single therapy sessions.

For *mobility*, the rater assigns a 1 if the patient is dependent on public transport to get to the therapy, otherwise a 0. This criterion encompasses two rationales. On the one hand, Vorarlberg is a small province and patients like to have the possibility to remain anonymous and visit psychotherapists further away. On the other hand, it is also about equitable access to psychotherapy treatment. Patients that live in remote areas and/or with limited access to public transport receive additional points to compensate for their handicap in terms of mobility.

For therapists, a database was obtained that included an alias for the therapist's name, age, and sex, where 1 represents male and 0 female.

Another set of variables describes the metadata of the initial interview, including patient ID, evaluating therapist, and the location in which the scoring was conducted. The latter two were coded as dummy variables, i.e. every therapist and every location were assigned either 1 or 0, where 1 indicates the presence of the therapist/location and 0 otherwise. This data file was mainly used to connect the datasets for therapists and weather with the patient dataset.

To test the robustness of the results, we added several control variables including weather data. The daily weather information for each location was retrieved from the online weather database "meteostat". However, since not every location had its corresponding weather station, meteostat calculates its data with an interpolation method, which is an approximation of the actual value. The variables for weather included in the analysis are as follows: the *average temperature in degrees Celsius*, *precipitation in millimeter*, *atmospheric pressure in hectopascal*, and the binary variables for *rain* and *snow*, for which a value of 1 indicates snow- or rainfall on a given day, and 0 none.

To test the hypothesis, hierarchical linear regression was used, as it enables us (1) to measure if there is a statistically significant relationship (*p*-value) between a dependent and multiple independent variables, and if so, (2) how strongly this relationship applies (β-coefficient), by assessing and comparing the impacts of each regressor (*Alexopoulos, 2010*). Significance testing used α-level .05, two-tailed tests.

The hierarchical entry of independent variables allows us to check for potential moderating effects of the controls since we can determine the order in which each block of variables is added to the regression equation (*Jeong & Jung, 2016*). Thereby we can analyze the changes in therapist effects with each subsequent addition of a control variable.

5. Results

The response variable "score sum" ranges from 13 to 45 points, with a mean of 28.9 and a median of 28.0. Standard deviation was 5.9. *Tab. 3* depicts the result of each rater. The number of observations was 267 for the descriptive statistics, but only 266 in the following regression analysis, due to the missing data on age in one observation.

Fig. 1 provides further intuition for how the therapists' scorings relate to each other by indicating the score sum distributions for each therapist via box plots. Eye inspection suggests that not all therapists assign equal scores on average. While some therapists, such as E and F, assign relatively low scores, other therapists, such as B, C, G, and H, assign relatively high scores. Therapist D and I provide interesting border cases.

Table 3: Measures of central tendency by therapists

Therapists	Score sum				
	N	Mean	Median	Min	Max
A	22	26.5	27.0	20	34
B	47	30.4	28.0	18	41
C	19	32.3	29.5	25	42
D	57	28.3	29.0	17	39
E	30	24.8	27.0	15	34
F	9	24.7	27.5	13	32
G	2	31.5	30.0	21	42
H	44	30.4	28.0	16	44
I	37	30.1	30.0	20	45
Total	**267**	**28.9**	**28.0**	**13**	**45**

Source: developed by the authors

In subsections 5.1 and 5.2 we check whether these apparent differences across therapists are caused by outliers or whether they constitute robust differences in rating tendencies. To this end, we report the results of the hierarchical linear regression. The results are divided into the general analysis of our hypothesis and the adjacent robustness checks.

5.1. Rater-based effects on the score sum

Our hypothesis was tested using a multiple linear regression model. In the first model, 11.8% (*adj.* R^2) of variance in score sum was explained by therapists. ANOVA suggests that the regression model contains significant explanatory variables ($p < .001$).

The results presented in *Tab. 4* confirm that there are effects on the size of the scores depending on which therapist conducts the scoring. Therapists B ($p = .007$), C ($p = .001$), and H ($p = .008$) were below the statistical significance threshold of $p < .01$. Therapist I was with a p-value of .012 similarly close to $p > .01$ and therefore still considered. The largest statistically significant effect sizes were recorded in therapist C ($\beta = 5.77$), followed by therapist H ($\beta = 3.84$), B ($\beta = 3.84$), and I ($\beta = 3.76$).

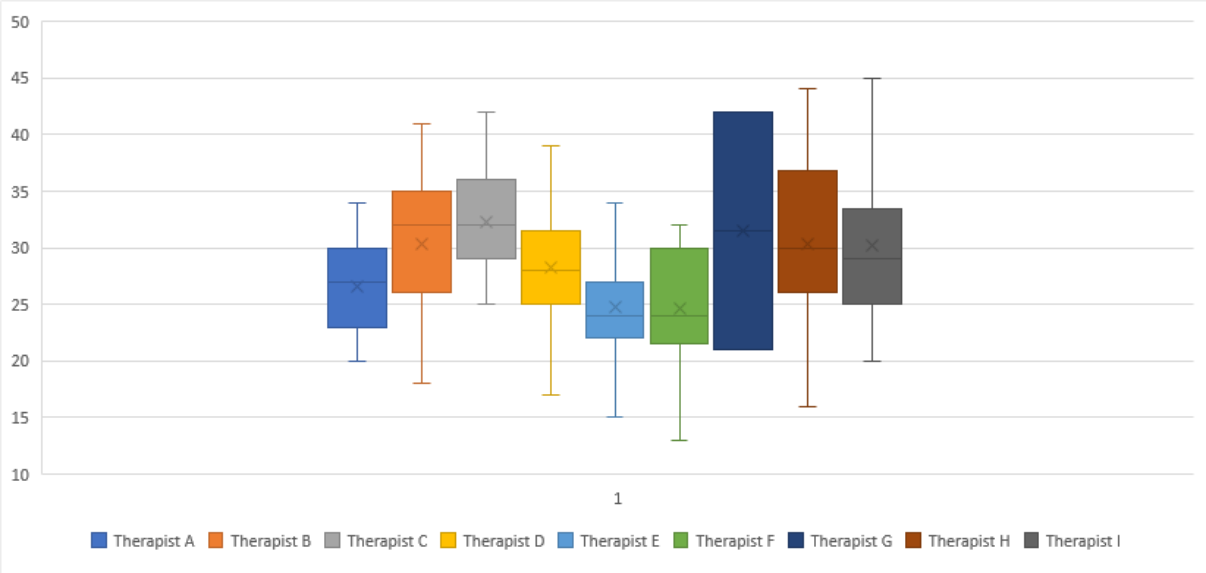

Figure 1. Box Plots of Sum Scores per Therapist

Source: developed by the authors

The presence of therapists B, C, H, and I was found to have predictive power over the size of the score patients receive. The null hypothesis can thus be rejected.

Table 4: Therapist-based effects on score sum (Model 1)

	Model 1	
	β	*p*
(Constant)	26.545	.000***
Therapist B	3.838	.007***
Therapist C	5.770	.001***
Therapist D	1.753	.205
Therapist E	-1.745	.259
Therapist F	-1.879	.389
Therapist G	4.955	.224
Therapist H	3.841	.008***
Therapist I	3.760	.012**
N	**266**	

Note: This table provides the coefficient estimates (β) of all rating therapists (excl. A) on the score sum (i.e. raw scores without weighting) including its respective p-value. Statistical significance is denoted as follows: *** for $p < .01$; ** for $p < .05$; * for $p < .10$.

Source: developed by the authors

5.2. Robustness checks

In this part, the robustness of the results received in subsection 5.1 is tested. In five consecutive models, control variables were added stepwise to monitor potential fluctuations in significance and effect size.

Across all five models, the percentage that explained the variance in score sum did not meaningfully change with each subsequent addition of controls. The *adjusted* R^2 ranged from 11.5% to a maximum of 13.6%.

A similar observation was made for the results of the ANOVA table. In all five instances, the variables included in the analyses were statistically significant with $p < .001$.

In the first step, we added patients' sex to control for any gender bias. As presented in *Tab. 5*, the results of the first robustness check (Model 2) indicate that the effects for therapists B, C, H, and I stay robust at their respective significance levels. Additionally, no major fluctuations in effect strength were detected. Therefore, we can conclude that in the presence of patients' sex, the effects of certain therapists remain robust.

In the second step, we tested the potential effects of patients' age. Note that we could not add therapists' age due to the relatively small number of raters and thus too high multicollinearity between the two. Nevertheless, the results of Model 3 reveal similarly small

changes in effect sizes and their respective *p*-values. As depicted in Tab. 5, all four of the before-mentioned therapists remained within their statistical significance threshold. The age of the patient does not seem to affect the therapist's judgment in this case, given that therapists B, C, H, and I still have an impact on the overall score patients receive, even in the presence of patients' age.

To control for seasonal influences and potential effects of the pandemic we added in Model 4 the months in which the scoring was conducted. As these variables were coded as binary variables, one variable had to be rejected. Since "October" had the least observations, it was the variable being excluded.

The presence of months altered the *p*-values of therapists and their respective effect sizes, most notably through the presence of March. Looking at *Tab. 5*, we find that therapist B's effect became slightly less significant, as its *p*-value exceeds now the .01 level. Also, its beta coefficient decreased by .4 points (.07 *standard deviations*). The same applies to H, whose *p*-value rose .036 and was thus now above the 0.01 level. Accordingly, its effect size shrank by .8 points (0.14 standard deviations). In therapist I we find the opposite; its beta coefficient grew by .7 points (.12 standard deviations) and its *p*-value (.004) fell below the .01 significance threshold. The effect of therapist C is highly robust; neither significance nor effect size was considerably different than in earlier models.

Although the addition of months to the model caused some minor fluctuations, all raters' effects remain significant, at least below the .05 level. The robustness of therapist C's effect along with the statistically significant effects of the other three therapists (B, E, and I) indicates that in their case the size of the score is still associated with the presence of said therapists, even in the company of the control variables for months.

Finally, we controlled for the mood-altering effects of the weather, which are found to influence a person's judgment. Thus, we added the variables *average temperature*, *precipitation*, *air pressure* as well as the dummy variables for *snow* and *rain* (see Model 5 in *Tab. 5*). With weather controls being added all therapists underwent minor changes in terms of effect size and statistical significance but generally remained stable.

Table 5: Therapist-based effects on score sum including controls (Model 2-5)

	Model 2		**Model 3**		**Model 4**		**Model 5**	
	β	***p***	***β***	***p***	***β***	***p***	***β***	***p***
(Constant)	26.638	.000***	27.598	.000***	26.197	.000***	27.464	.538
Therapist B	3.821	.008***	3.963	.006***	3.423	.017**	3.465	.020**
Therapist C	5.758	.001***	5.699	.001***	5.554	.001***	5.760	.001***
Therapist D	1.723	.215	1.741	.210	1.310	.345	1.697	.281
Therapist E	-1.753	.258	-1.761	.256	-2.021	.189	-1.479	.377
Therapist F	-1.887	.388	-1.681	.443	-1.922	.384	-1.517	.523
Therapist G	4.862	.234	4.748	.245	5.546	.192	5.622	.196
Therapist H	3.829	.008***	3.986	.006***	3.180	.036**	3.616	.029**
Therapist I	3.731	.013**	3.744	.013**	4.472	.004***	4.462	.007***
Sex Patient	-.256	.732	-.340	.650	-.399	.592	-.330	.661
Age Patient			-.026	.285	-.022	.377	-.019	.455
November					.906	.723	.626	.820
December					.327	.901	-.302	.918
January					1.651	.518	1.169	.674
February					2.368	.348	1.878	.483
March					4.670	.083*	4.319	.139
Avg. Temp.							-.010	.920
Precipitation							-.052	.412
Air pressure							-.001	.976
Rain							.245	.782
Snow							1.121	.340
N	**266**		**266**		**266**		**266**	

Note: This table provides the coefficient estimates (*β*) of all rating therapists (excl. A) on the score sum including the respective p-values. Added controls are depicted below the second dotted line. Statistical significance is denoted as follows: *** for $p < .01$; ** for $p < .05$; * for $p < .10$.

Source: developed by the authors

Fig. 2 isolates the regression coefficients from therapists showing robust divergences in rating behavior across our model specifications. Comparing the entries, we do not find that the inclusion of the control variables qualitatively changes the robustness of the predictors, as all regression coefficients remain below the .05 significance threshold at all times. Our hierarchical approach, however, allows us to add nuance to this judgment. While we do not find that sex or age of the patient alters either effect size or significance of the respective raters, discernible changes occur when adding both the month and weather controls. Hence, the rating behavior of some therapists does appear to be connected to external or seasonal influences, although the detected overlaps are small.

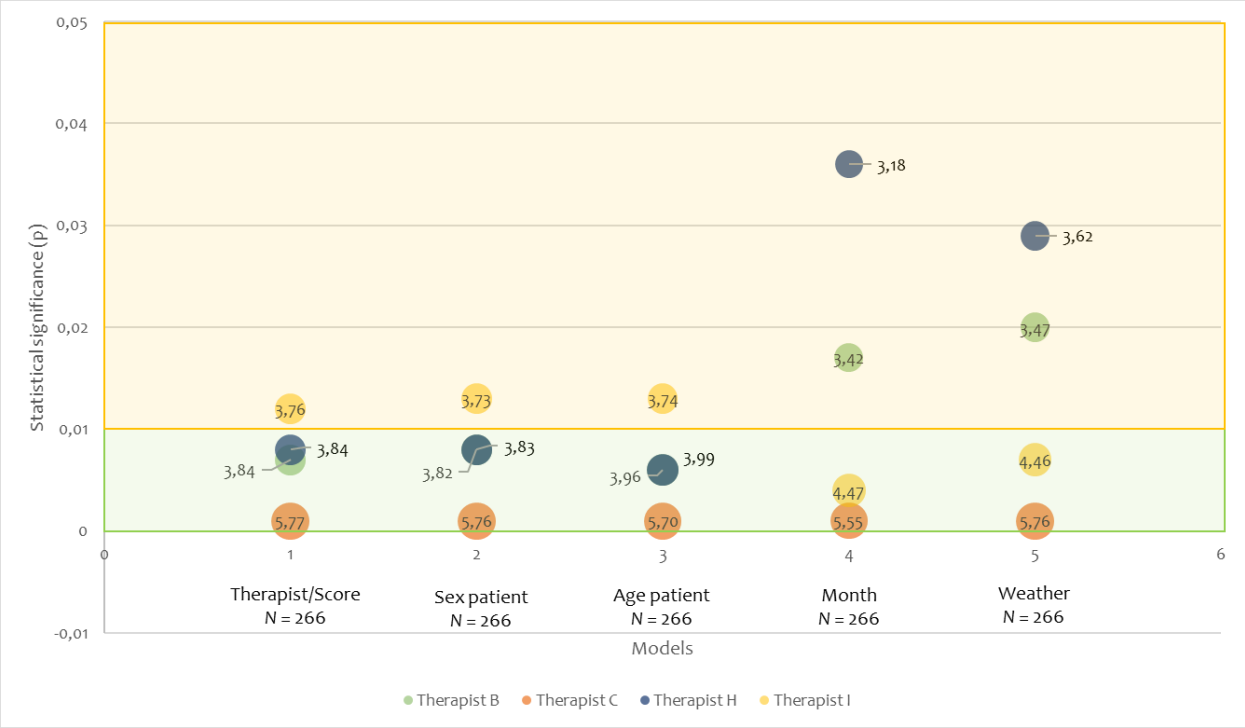

Figure 2. Significance of robust ($p < .05$) regression coefficients from therapists across regression models. Sex and age of the patient appear independent of assessment ratings. Month and weather, however, alter the significance of the observed effects

Source: developed by the authors

6. Discussion

This paper aimed to investigate how independently the scoring results, received in the course of patient priority assessments, are of their raters.

To test the effect of raters on the resulting score and thus the objectivity of scoring results, we used a hierarchical linear regression analysis loaded with data from the assessment center of a psychotherapy service. In six consecutive models, controls were added stepwise to the regression equation. This allowed us to observe potentially moderating effects of the control variables on therapists' effects. Overall, the percentage of variance in scores that was explained by the entered variables was low across all models (R^2 ranged from 11.5% to 13.6%). Although it was not the goal of this study to find all factors that fully explain the variance of the score, the low R^2 nonetheless indicates that at least other factors not included contain far more explanatory power over the size of the score. This notion was quickly tested in a separate regression analysis that used the results of all priority scoring items as independent variables. The results there showed that the variance of scores was, of course, fully explained by the priority criteria (R^2 = 1.0). Regardless, the significance levels received in the ANOVA table, suggest that the performed regression models provide indeed an explanatory contribution.

In an ideal scenario, the presence of a specific rater should not affect the size of the score that patients receive. The findings of this study, however, indicate that such effects are indeed present. In the main analysis (model 1), we exclusively observed the raters' effects, without adding any controls yet. The results confirmed the idea that scores are not independent of their rater. Four therapists were found to have statistically significant effects (3 with $p < .01$ and 1 with $p < .05$), with the biggest effect size at 5.77 (therapist C). This means that the mere presence of this specific therapist adds on average 5.77 points to the unweighted score sum of the respective patient. To put this in perspective 5.77 equals almost 1 standard deviation (.98 standard deviations) of the scores measured in this sample.

To further illustrate the magnitude of what 5.77 points mean in days waited on the waiting list, we conducted a quick back-of-the-envelope calculation. Note that our analysis used unweighted scores, thus we have to add the weighting, which varies across the criteria but on average adds 3.4 points for each point given. Multiplying the coefficient with the average weighting equals roughly 20 points. Then we looked at one of the patients that were scored by therapist C and already assigned to a treatment, so we can determine the number of days spent on the waiting list. One patient we found entered the waiting list on the 3rd of November, had an initial (weighted) score of 65, and waited 43 days on the list. Adjusting the patient's score by the therapist effect would result in only 45 instead of 65 points. For comparison, we picked a patient of another therapist, who did not show statistically significant effects. The other patient we found entered the waiting list on the 9th of November, had an initial (weighted) score of 43, but waited 77 days until that person received treatment. Thus, the patient waited more than a month longer for treatment, simply because that person was *not* rated by therapist C. We do not claim that this calculation is highly accurate nor sophisticated, also how soon patients leave the waiting list is slightly dependent on the availabilities of therapists, yet it is an approximation and illustrates our argument.

In the following four models, we added stepwise the controls. Adding the patient's sex, caused no substantial fluctuation in effect

sizes, indicating the absence of gender bias in our sample. However, due to all therapists being female, we cannot say if a different picture would have emerged, when male raters had been present. We assume that if such gender-related effects were indeed existent, they would only be measurable at the level of the respective priority criterion. For example, the study of *Earp et al. (2019)* showed that boys were rated to experience more pain than girls, even though they showed the same symptoms. We quickly tested with an additional regression analysis if this applies to our sample too. However, the results show that the effects of therapists' ratings for intensity of suffering do not change with patients' sex being added to the regression. Still, we encourage future studies to conduct investigations of potential biases also at the sublevel, i.e. regression analyses with the priority items as dependent variables.

When controlling for patients' age, we again found no fluctuations in effect sizes that would be indicative of any bias towards age. Due to multicollinearity between therapists and their age we could not assess if age on the rater side would affect the scores.

However, when we added the months to the model, we noticed some changes in *p*-values and effect sizes, although all effects of therapists that were measured before to be statistically significant also remained significant at $p < .05$. To our surprise, the month of March showed to have a significant effect ($p < .1$) on the size of the score, even though not strong, it was still greater compared to other months. It must be mentioned that the number of ratings recorded in March, due to the cut-off point for sample collection on the 19th of March, was only a third of the number of ratings in other months (on average 59 ratings per month). However, as discussed in subsection 2.2, we cannot exclude the possibility that the effect recorded in March is attributable to the increased prevalence of SAD in winter months *(Magnusson & Boivin, 2003)* or the general increased prevalence rates of mental disorders in winter (*Graaf et al., 2005*) and spring (*Slaunwhite et al., 2019*). Similarly, it could also be caused by overall deteriorated mental health due to the effects of the Covid-19 pandemic, as expressed by some authors *(Adams-Prassl et al., 2020; Pieh et al., 2020; Rossi et al., 2020)*. Regardless, the observed impact of March vanishes in the subsequent models, suggesting little robustness of the effect.

As shown in 2.2, many authors have pointed out that weather influences our mood, and mood, in turn, influences our decision-making. However, in this research, we could not identify any effects due to certain weather phenomena, as no substantial effect changes were detected when weather controls were added. In other studies, authors investigated the influences of sunny (*Hirshleifer & Shumway, 2003; Murray et al., 2010*) and cloudy (*Goetzmann & Zhu, 2005*) weather on our decision-making. Unfortunately, we had no access to data such as sunlight or cloudiness.

Overall, we can answer our proposed research question by demonstrating that the effects of therapists on the size of the score are indeed measurable and that these stay relatively robust in the presence of the included controls. The example provided earlier has demonstrated how this effect can prefer or discriminate one patient over another, simply by being rated by two different therapists. However, if this effect is also measurable in other settings, and therefore truly reliable, can only be determined in future studies. Also, to what extent certain heuristics and biases, as well as other imperfections of the clinical judgment, led to these outcomes can only be revealed in upcoming experiments that limit their research focus on the detection of biases in patient priority assessments. Although the controls used in our analysis considered a few of the potential biases, they could not explain the observed deviations in our sample. In general, we find that the concept of rater bias was given much attention in other areas but not in patient prioritization tools and even less so for the ones employed in a mental health context. An issue that needs to be studied more thoroughly in the future, as our results demonstrate.

On a final note, our findings add to the literature that described the questionable reliability and validity of patient prioritization tools (*Déry et al., 2020; Harding & Taylor, 2013*), by indicating the influence that therapists have on the outcome of priority scorings. This research is also highly relevant for practitioners, as it provides valuable knowledge about the weakness of such ratings. Given the findings in our analysis, we recommend that further training of therapists could reduce the stated effects, as suggested by *Harries & Gilhooly (2011)*. Thereby, the training would be most effective if it educates about the decision-making processes and the inherent pitfalls (*Bell & Mellor, 2009*). Ultimately, this would facilitate better decision-making and thus provide increased reliability and validity of patient prioritization tools.

This study is not without limitations. Additional to the ones already mentioned in the discussion, further limitations that need to be considered are listed below. First and foremost, we cannot tell to what extent the actual level of need of patients has caused the results in the four therapists with statistically significant effects. It might be that these therapists stood out simply because they were incidentally assigned cases that had indeed more severe symptoms or fulfilled any other priority criteria relevant in the scoring, compared to the ones of their peers, which resulted in them assigning on average higher scores to their patients. This makes the internal validity of the study to some extent questionable. Future research would be well advised to find methods that control for the actual level of need.

Due to limitations in data availability for individual therapists, our analysis cannot claim to provide a comprehensive overview of rater discrepancies. To a lesser extent, this concerns therapists identified as divergent from the rest of the sample. While for Therapist B, H, and I, we have more than 30 observations each, we have only 19 observations for therapist C. Although the respective rater-fixed effect is highly robust ($p<.001$), further observations would help bolster the meaningfulness of the detected effects. To a larger extent, data limitations concern therapists for which no effect has been identified. Fig. 1 suggests that therapists F and G assign, on average, the lowest and highest ratings, respectively, and would thus be natural candidates for receiving stronger attention. Yet, likely due to low sample sizes of 9 and 2 observations for therapists F and G, our analysis does not flag the rating behaviors as exceptional. Hence, although our analysis can confirm our main hypothesis by reliably detecting some assessment biases, it can likely not detect all of them.

Another limitation is the novelty of the tool used at the institution. The involved therapists were confronted with an entirely new and standardized technique to determine a patient's level of need. Although they were trained prior to the introduction of the tool, they could perhaps require further familiarization with the way that a patients' priority is assessed now. The lack of experience might therefore have threatened the content validity of the priority criteria, i.e. therapists might have interpreted the items differently or falsely. We advise the institution's stakeholders to run the analysis again sometime in the future.

External validity (i.e. generalizability) might be somewhat limited too, in particular for two reasons. First, the sample data was collected using a consecutive sampling method. Although this method is less prone to sampling bias than simple convenience sampling (*Schuster & Powers, 2005*) it still falls into the category of non-probability sampling methods, which are more likely to produce biased samples (*Suresh, 2014*). Second, the sample includes only patients from semi-rural areas, which makes generalizing these findings to urban regions a bit problematic, given the differing demands and conditions.

A final limitation might be that the results are based on secondary data. As *Kimberlin & Winterstein (2008)* have pointed out, secondary data is usually collected for a different purpose. Although we are confident that the data collected was appropriate to answer the research question, we cannot exclude the possibility that some data was falsely recorded in the institution's client information system, from which the data used in this study originates.

7. Conclusion

This study aimed to investigate rater-based effects in the scoring results of patient prioritization tools utilized in mental health services. Based on the results of the hierarchical linear regression, we can conclude that some therapists indeed demonstrate to have a statistically significant influence over the size of the resulting priority score. The results further indicate that these effects can lead to unwanted discrimination and consequently to unjustifiably prolonged waiting times for some patients, which thwarts the idea of a fair and equitable prioritization of patients. Based on the conclusions, practitioners should consider further training of raters, with a particular focus on the pitfalls involved in decision-making processes. Since our study could not identify the causes for these effects, we encourage other researchers to more thoroughly investigate biases that may lead to such effects. Our study has contributed to the notion of previous research that the quality of patient prioritization tools is sometimes worrying. Furthermore, our insights add to the literature on inconsistent clinical judgment, and even more so they provide much-needed information about the quality of patient prioritization tools used in mental health settings.

8. Funding

This study received no specific financial support.

9. Competing interests

In accordance with publisher policies and our ethical obligations as researchers, we report that one of the authors is employed at a company that may be affected by the research reported in the enclosed paper. We have disclosed those interests fully.

References

Adams-Prassl, A., Boneva, T., Golin, M., & Rauh, C. (2020). *The impact of the coronavirus lockdown on mental health: Evidence from the US*. doi:10.17863/CAM.57997.

Alexopoulos, E. C. (2010). Introduction to multivariate regression analysis. *Hippokratia, 14*(Suppl 1), 23–28. PMCID.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders: DSM-IV (4th ed.)*. American Psychiatric Association.

Arslanian-Engoren, C. (2000). Gender and age bias in triage decisions. Journal of Emergency Nursing, 26(2), 117–124. doi:10.1016/S0099-1767(00)90053-9.

Arslanian-Engoren, C., & Scott, L. D. (2016). Women's perceptions of biases and barriers in their myocardial infarction triage experience. *Heart & Lung : The Journal of Critical Care, 45*(3), 166–172. doi:10.1016/j.hrtlng.2016.02.010.

Bakhshi, S., Kanuparthy, P., & Gilbert, E. (2014). Demographics, weather and online reviews. In C.-W. Chung, A. Broder, K. Shim, & T. Suel (Eds.), *Proceedings of the 23rd international conference on World wide web - WWW '14* (pp. 443–454). ACM Press. doi:10.1145/2566486.2568021.

Bell, I., & Mellor, D. (2009). Clinical judgements: Research and practice. *Australian Psychologist*, 44(2), 112–121. doi:10.1080/00050060802550023.

Berufsverband Österreichischer PsychologInnen, & Karmasin Research & Identity. (2020). *Psychische Gesundheit in Österreich.* Retrieved from https://www.boep.or.at/download/5ef991483c15c8588f00001a/BOEP-Studie_Psychische_Gesundheit_in_Oesterreich.pdf.

Bless, H., Schwarz, N., & Kemmelmeier, M. (1996). Mood and Stereotyping: Affective States and the Use of General Knowledge Structures. *European Review of Social Psychology, 7*(1), 63–93. doi:10.1080/14792779443000102.

Bowes, S. M., Ammirati, R. J., Costello, T. H., Basterfield, C., & Lilienfeld, S. O. (2020). Cognitive biases, heuristics, and logical fallacies in clinical practice: A brief field guide for practicing clinicians and supervisors. *Professional Psychology: Research and Practice, 51*(5), 435–445. doi:10.1037/pro0000309.

Brooks, S. K., Webster, R. K., Smith, L. E., Woodland, L., Wessely, S., Greenberg, N., & Rubin, G. J. (2020). The psychological impact of quarantine and how to reduce it: rapid review of the evidence. *The Lancet, 395*(10227), 912–920. doi:10.1016/S0140-6736(20)30460-8.

Chan, J., & Wang, J. (2018). Hiring Preferences in Online Labor Markets: Evidence of a Female Hiring Bias. *Management Science, 64*(7), 2973–2994. doi:10.1287/mnsc.2017.2756.

Christensen-Szalanski, J. J., Diehr, P. H., Bushyhead, J. B., & Wood, R. W. (1982). Two studies of good clinical judgment. *Medical Decision Making, 2*(3), 275–283. doi:10.1177/0272989X8200200303.

Clark, D. M., Canvin, L., Green, J., Layard, R., Pilling, S., & Janecka, M. (2018). Transparency about the outcomes of mental health services (IAPT approach): an analysis of public data. *The Lancet, 391*(10121), 679–686. doi:10.1016/S0140-6736(17)32133-5.

Corrigan, P. W., Druss, B. G., & Perlick, D. A. (2014). The Impact of Mental Illness Stigma on Seeking and Participating in Mental Health Care. *Psychological Science in the Public Interest, 15*(2), 37–70. doi:10.1177/1529100614531398.

Coster, C. de, McMillan, S., Brant, R., McGurran, J., & Noseworthy, T. (2007). The Western Canada Waiting List Project: Development of a priority referral score for hip and knee arthroplasty. *Journal of Evaluation in Clinical Practice, 13*(2), 192-6; quiz 197. doi:10.1111/j.1365-2753.2006.00671.x.

Cowden, R. G., Davis, E. B., Counted, V., Chen, Y., Rueger, S. Y., VanderWeele, T. J., Lemke, A. W., Glowiak, K. J., & Worthington, E. L. (2021). Suffering, Mental Health, and Psychological Well-being During the COVID-19 Pandemic: A Longitudinal Study of U.S. Adults With Chronic Health Conditions. *Wellbeing, Space and Society*, 2, 100048. doi:10.1016/j.wss.2021.100048.

Croskerry, P. (2002). Achieving quality in clinical decision making: cognitive strategies and detection of bias. *Academic Emergency Medicine,* 9(11), 1184–1204. doi:10.1111/j.1553-2712.2002.tb01574.x.

Déry, J., Ruiz, A., Routhier, F., Bélanger, V., Côté, A., Ait-Kadi, D., Gagnon, M.-P., Deslauriers, S., Lopes Pecora, A. T., Redondo, E., Allaire, A.-S., & Lamontagne, M.-E. (2020). A systematic review of patient prioritization tools in non-emergency healthcare services. *Systematic Reviews,* 9(1), Article 227, 1–14. doi:10.1186/s13643-020-01482-8.

Earp, B. D., Monrad, J. T., LaFrance, M., Bargh, J. A., Cohen, L. L., & Richeson, J. A. (2019). Featured Article: Gender Bias in Pediatric Pain Assessment. *Journal of Pediatric Psychology, 44*(4), 403–414. doi:10.1093/jpepsy/jsy104.

Endicott, J., Spitzer, R. L., Fleiss, J. L., & Cohen, J. (1976). The global assessment scale. A procedure for measuring overall severity of psychiatric disturbance. *Archives of General Psychiatry, 33*(6), 766–771. doi:10.1001/archpsyc.1976.01770060086012.

FitzGerald, C., & Hurst, S. (2017). Implicit bias in healthcare professionals: A systematic review. *BMC Medical Ethics, 18*(1), 19. doi:10.1186/s12910-017-0179-8.

Gingerich, A., Regehr, G., & Eva, K. W. (2011). Rater-based assessments as social judgments: Rethinking the etiology of rater errors. *Academic Medicine, 86*(10), 1-7. doi:10.1097/ACM.0b013e31822a6cf8.

Glied, S., & Pine, D. S. (2002). Consequences and correlates of adolescent depression. *Archives of Pediatrics & Adolescent Medicine, 156*(10), 1009–1014. doi:10.1001/archpedi.156.10.1009.

Goetzmann, W. N., & Zhu, N. (2005). Rain or Shine: Where is the Weather Effect? *European Financial Management, 11*(5), 559–578. doi:10.1111/j.1354-7798.2005.00298.x.

Graaf, R. de, van Dorsselaer, S., Have, M. ten, Schoemaker, C., & Vollebergh, W. A. M. (2005). Seasonal variations in mental disorders in the general population of a country with a maritime climate: Findings from the Netherlands mental health survey and incidence study. *American Journal of Epidemiology, 162*(7), 654–661. doi:10.1093/aje/kwi264.

Hadorn, D. C., & Steering Committee of the Western Canada Waiting List Project (2000). Setting priorities for waiting lists: defining our terms. *Cmaj, 163*(7), 857–860. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/pmc80512/.

Hairston, D. R., Gibbs, T. A., Wong, S. S., & Jordan, A. (2019). *Clinician Bias in Diagnosis and Treatment*. In M. M. Medlock, D. Shtasel, N.-H. T. Trinh, & D. R. Williams (Eds.), Racism and Psychiatry (pp. 105–137). Springer International Publishing. doi:10.1007/978-3-319-90197-8_7.

Harding, K. E., & Taylor, N. (2013). *Triage in Nonemergency Services*. In R. Hall (Ed.), International Series in Operations Research & Management Science. Patient Flow (Vol. 206, pp. 229–250). Springer US. doi:10.1007/978-1-4614-9512-3_10.

Harries, P., & Gilhooly, K. (2011). Training Novices to Make Expert, Occupationally Focused, Community Mental Health Referral Decisions. *British Journal of Occupational Therapy, 74*(2), 58–65. doi:10.4276/030802211X12971689813963.

Hibbing, J. R., Smith, K. B., & Alford, J. R. (2014). Differences in negativity bias underlie variations in political ideology. *Behavioral and Brain Sciences, 37*(3), 297–307. doi:10.1017/S0140525X13001192.

Hirshleifer, D., & Shumway, T. (2003). Good Day Sunshine: Stock Returns and the Weather. *The Journal of Finance, 58*(3), 1009–1032. doi:10.1111/1540-6261.00556.

James, S. L., Abate, D., Abate, K. H., Abay, S. M., Abbafati, C., Abbasi, N., ... & Briggs, A. M. (2018). Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. *The Lancet, 392*(10159), 1789-1858.. doi:10.1016/S0140-6736(18)32279-7

Jeong, Y., & Jung, M. J. (2016). Application and Interpretation of Hierarchical Multiple Regression. *Orthopedic Nursing, 35*(5), 338–341. doi:10.1097/NOR.0000000000000279.

Kassin, S. M., Dror, I. E., & Kukucka, J. (2013). The forensic confirmation bias: Problems, perspectives, and proposed solutions. *Journal of Applied Research in Memory and Cognition, 2*(1), 42–52. doi:10.1016/j.jarmac.2013.01.001.

Keithly, L. J., Samples, S. J., & Strupp, H. H. (1980). Patient motivation as a predictor of process and outcome in psychotherapy. *Psychotherapy and Psychosomatics, 33*(1-2), 87–97. doi:10.1159/000287417.

Kieling, C., Baker-Henningham, H., Belfer, M., Conti, G., Ertem, I., Omigbodun, O., Rohde, L. A., Srinath, S., Ulkuer, N., & Rahman, A. (2011). Child and adolescent mental health worldwide: evidence for action. *The Lancet, 378*(9801), 1515–1525. doi:10.1016/S0140-6736(11)60827-1.

Kimberlin, C. L., & Winterstein, A. G. (2008). Validity and reliability of measurement instruments used in research. *American Journal of Health-System Pharmacy, 65*(23), 2276–2284. doi:10.2146/ajhp070364.

Lipson, S. K., Lattie, E. G., & Eisenberg, D. (2019). Increased Rates of Mental Health Service Utilization by U.S. College Students: 10-Year Population-Level Trends (2007-2017). *Psychiatric Services, 70*(1), 60–63. doi:10.1176/appi.ps.201800332.

Lizaur-Utrilla, A., Martinez-Mendez, D., Miralles-Muñoz, F. A., Marco-Gomez, L., & Lopez-Prats, F. A. (2016). Negative impact of waiting time for primary total knee arthroplasty on satisfaction and patient-reported outcome. *International Orthopaedics, 40*(11), 2303–2307. doi:10.1007/s00264-016-3209-0.

López, S. R. (1989). Patient variable biases in clinical judgment: Conceptual overview and methodological considerations. *Psychological Bulletin, 106*(2), 184–203. doi:10.1037/0033-2909.106.2.184.

Luigi, S., Michael, B., & Valerie, M. (2013). OECD health policy studies waiting time policies in the health sector: What works? *Oecd Publishing*. Retrieved from https://read.oecd.org/10.1787/9789264179080-en?format=pdf.

MacCormick, A. D., Collecutt, W. G., & Parry, B. R. (2003). Prioritizing patients for elective surgery: A systematic review. *ANZ Journal of Surgery, 73*(8), 633–642. doi:10.1046/j.1445-2197.2003.02605.x

Magnusson, A. (2000). An overview of epidemiological studies on seasonal affective disorder. *Acta Psychiatrica Scandinavica, 101*(3), 176–184. doi:10.1034/j.1600-0447.2000.101003176.x.

Magnusson, A., & Boivin, D. (2003). Seasonal affective disorder: An overview. *Chronobiology International, 20*(2), 189–207. doi:10.1081/CBI-120019310.

Malouff, J. (2008). Bias in Grading. *College Teaching, 56*(3), 191–192. doi:10.3200/CTCH.56.3.191-192.

McDermott, P. A., Watkins, M. W., & Rhoad, A. M. (2014). Whose IQ is it? Assessor bias variance in high-stakes psychological assessment. *Psychological Assessment, 26*(1), 207–214. doi:10.1037/a0034832.

McIntyre, D., & Chow, C. K. (2020). Waiting Time as an Indicator for Health Services Under Strain: A Narrative Review. *Inquiry : A Journal of Medical Care Organization, Provision and Financing, 57*, 46958020910305. doi:10.1177/0046958020910305.

Meier, A. N., Schmid, L., & Stutzer, A. (2019). Rain, emotions and voting for the status quo. *European Economic Review*, 119, 434–451. doi:10.1016/j.euroecorev.2019.07.014.

Mojtabai, R., Olfson, M., & Han, B. (2016). National Trends in the Prevalence and Treatment of Depression in Adolescents and Young Adults. *Pediatrics, 138*(6). doi:10.1542/peds.2016-1878.

Murray, K. B., Di Muro, F., Finn, A., & Popkowski Leszczyc, P. (2010). The effect of weather on consumer spending. *Journal of Retailing and Consumer Services, 17*(6), 512–520. doi:10.1016/j.jretconser.2010.08.006.

Murrie, D. C., Boccaccini, M. T., Guarnera, L. A., & Rufino, K. A. (2013). Are forensic experts biased by the side that retained them? *Psychological Science, 24*(10), 1889–1897. doi:10.1177/0956797613481812.

Norman, G. R., & Eva, K. W. (2010). Diagnostic error and clinical reasoning. *Medical Education, 44*(1), 94–100. doi:10.1111/j.1365-2923.2009.03507.x.

Nottingham, Q. J., Johnson, D. M., & Russell, R. S. (2018). The Effect of Waiting Time on Patient Perceptions of Care Quality. *Quality Management Journal, 25*(1), 32–45. doi:10.1080/10686967.2018.1404368.

Patel, V. L., Kaufman, D. R., & Arocha, J. F. (2002). Emerging paradigms of cognition in medical decision-making. *Journal of Biomedical Informatics, 35*(1), 52–75. doi:10.1016/S1532-0464(02)00009-6.

Pathirana, T. I., & Jackson, C. A. (2018). Socioeconomic status and multimorbidity: A systematic review and meta-analysis. *Australian and New Zealand Journal of Public Health, 42*(2), 186–194. doi:10.1111/1753-6405.12762.

Pieh, C., Budimir, S., & Probst, T. (2020). The effect of age, gender, income, work, and physical activity on mental health during coronavirus disease (COVID-19) lockdown in Austria. *Journal of Psychosomatic Research*, 136, 110186. doi:10.1016/j.jpsychores.2020.110186.

Platts-Mills, T. F., Travers, D., Biese, K., McCall, B., Kizer, S., LaMantia, M., Busby-Whitehead, J., & Cairns, C. B. (2010). Accuracy of the Emergency Severity Index triage instrument for identifying elder emergency department patients receiving an immediate life-saving intervention. *Academic Emergency Medicine : Official Journal of the Society for Academic Emergency Medicine, 17*(3), 238–243. doi:10.1111/j.1553-2712.2010.00670.x.

Raymond, M.-H., Demers, L., & Feldman, D. E. (2017). Differences in Waiting List Prioritization Preferences of Occupational Therapists, Elderly People, and Persons With Disabilities: A Discrete Choice Experiment. *Archives of Physical Medicine and Rehabilitation, 99*(1), 35-42. doi:10.1016/j.apmr.2017.06.031.

Rechnungshof. (2019). Bericht des Rechnungshofes: Versorgung psychisch Erkrankter durch die Sozialversicherung (BUND 2019/8). Wien. Retrieved from https://www.rechnungshof.gv.at/rh/home/home/Versorgung_psychisch_Erkrankter_SV.pdf.

Reichert, A., & Jacobs, R. (2018). The impact of waiting time on patient outcomes: Evidence from early intervention in psychosis services in England. *Health Economics, 27*(11), 1772–1787. doi:10.1002/hec.3800.

Reynolds, C. R., & Suzuki, L. A. (2013). *Bias in psychological assessment: An empirical review and recommendations*. In Handbook of psychology: Assessment psychology, Vol. 10, 2nd ed (pp. 82–113). John Wiley & Sons, Inc.

Roper, R. L. (2019). Does Gender Bias Still Affect Women in Science? *Microbiology and Molecular Biology Reviews : MMBR, 83*(3). doi:10.1128/MMBR.00018-19.

Rossi, R., Socci, V., Talevi, D., Mensi, S., Niolu, C., Pacitti, F., Di Marco, A., Rossi, A., Siracusano, A., & Di Lorenzo, G. (2020). Covid-19 Pandemic and Lockdown Measures Impact on Mental Health Among the General Population in Italy. *Frontiers in Psychiatry*, 11, 790. doi:10.3389/fpsyt.2020.00790.

Rössler, W. (2012). Stress, burnout, and job dissatisfaction in mental health workers. *European Archives of Psychiatry and Clinical Neuroscience, 262*(S2), S65-9. doi:10.1007/s00406-012-0353-4.

Samelius, L., Wijma, B., Wingren, G., & Wijma, K. (2010). Lifetime history of abuse, suffering and psychological health. *Nordic Journal of Psychiatry*, 64(4), 227–232. doi:10.3109/08039480903478680.

Samuel, D. B., & Bucher, M. A. (2017). Assessing the assessors: The feasibility and validity of clinicians as a source for personality disorder research. *Personality Disorders, 8*(2), 104–112. doi:10.1037/per0000190.

Schuster, D. P., & Powers, W. J. (2005). Translational and experimental clinical research. Lippincott Williams & Wilkins.

Shor, E., van de Rijt, A., & Fotouhi, B. (2019). A Large-Scale Test of Gender Bias in the Media. *Sociological Science*, 6, 526–550. doi:10.15195/v6.a20

Sifneos, P. E. (1978). Motivation for Change A Prognostic Guide for Successful Psychotherapy. *Psychotherapy and Psychosomatics, 29*(1/4), 293–298. doi:10.1159/000287144.

Slaunwhite, A. K., Ronis, S. T., Peters, P. A., & Miller, D. (2019). Seasonal variations in psychiatric admissions to hospital. *Canadian Psychology/Psychologie Canadienne, 60*(3), 155–164. doi:10.1037/cap0000156.

Suresh, S. (2014). Nursing Research and Statistics (2nd ed.). Elsevier Health *Sciences APAC*. Retrieved from http://gbv.eblib.com/patron/FullRecord.aspx?p=2004159.

Talevi, D., Socci, V., Carai, M., Carnaghi, G., Faleri, S., Trebbi, E., Di Bernardo, A., Capelli, F., & Pacitti, F. (2020). Mental health outcomes of the CoViD-19 pandemic. *Rivista Di Psichiatria, 55*(3), 137–144. doi:10.1708/3382.33569.

Thomas, O. (2018). Two decades of cognitive bias research in entrepreneurship: What do we know and where do we go from here? *Management Review Quarterly, 68*(2), 107–143. doi:10.1007/s11301-018-0135-9.

Twenge, J. M., Cooper, A. B., Joiner, T. E., Duffy, M. E., & Binau, S. G. (2019). Age, period, and cohort trends in mood disorder indicators and suicide-related outcomes in a nationally representative dataset, 2005-2017. *Journal of Abnormal Psychology, 128*(3), 185–199. doi:10.1037/abn0000410.

Ulasi, I. (2008). Gender bias in access to healthcare in Nigeria: A study of end-stage renal disease. *Tropical Doctor, 38*(1), 50–52. doi:10.1258/td.2007.060160.

van Ryn, M., & Burke, J. (2000). The effect of patient race and socio-economic status on physicians' perceptions of patients. *Social Science & Medicine, 50*(6), 813–828. doi:10.1016/s0277-9536(99)00338-x.

Vries, M. de, Holland, R. W., Corneille, O., Rondeel, E., & Witteman, C. L. (2012). Mood effects on dominated choices: Positive mood induces departures from logical rules. *Journal of Behavioral Decision Making, 25*(1), 74–81. doi:10.1002/bdm.716.

West, R. F., Toplak, M. E., & Stanovich, K. E. (2008). Heuristics and biases as measures of critical thinking: Associations with cognitive ability and thinking dispositions. *Journal of Educational Psychology, 100*(4), 930–941. doi:10.1037/a0012842.

Wolfson, A. M., Doctor, J. N., & Burns, S. P. (2000). Clinician judgments of functional outcomes: How bias and perceived accuracy affect rating. *Archives of Physical Medicine and Rehabilitation, 81*(12), 1567–1574. doi:10.1053/apmr.2000.16345.

Wynn, A. T., & Correll, S. J. (2018). *Combating Gender Bias in Modern Workplaces*. In B. J. Risman, C. M. Froyum, & W. J. Scarborough (Eds.), Handbooks of Sociology and Social Research. Handbook of the Sociology of Gender (pp. 509–521). Springer International Publishing. doi:10.1007/978-3-319-76333-0_37.

Yourstone, J., Lindholm, T., Grann, M., & Svenson, O. (2008). Evidence of gender bias in legal insanity evaluations: A case vignette study of clinicians, judges and students. *Nordic Journal of Psychiatry*, 62(4), 273–278. doi:10.1080/08039480801963135.