Patient trust in the use of machine learning-based clinical decision support systems in psychiatric services: A randomized survey experiment

Erik Perfalk[1,2] ⓘ, Martin Bernstorff[1,2] ⓘ, Andreas Aalkjær Danielsen[1,2] ⓘ and Søren Dinesen Østergaard[1,2] ⓘ

[1]Department of Affective Disorders, Aarhus University Hospital – Psychiatry, Aarhus, Denmark and [2]Department of Clinical Medicine, Aarhus University, Aarhus, Denmark

Abstract

Background. Clinical decision support systems (CDSS) based on machine-learning (ML) models are emerging within psychiatry. If patients do not trust this technology, its implementation may disrupt the patient-clinician relationship. Therefore, the aim was to examine whether receiving basic information about ML-based CDSS increased trust in them.

Methods. We conducted an online randomized survey experiment in the Psychiatric Services of the Central Denmark Region. The participating patients were randomized into one of three arms: Intervention = information on clinical decision-making supported by an ML model; Active control = information on a standard clinical decision process, and Blank control = no information. The participants were unaware of the experiment. Subsequently, participants were asked about different aspects of trust and distrust regarding ML-based CDSS. The effect of the intervention was assessed by comparing scores of trust and distrust between the allocation arms.

Results. Out of 5800 invitees, 992 completed the survey experiment. The intervention increased trust in ML-based CDSS when compared to the active control (mean increase in trust: 5% [95% CI: 1%; 9%], $p = 0.0096$) and the blank control arm (mean increase in trust: 4% [1%; 8%], $p = 0.015$). Similarly, the intervention reduced distrust in ML-based CDSS when compared to the active control (mean decrease in distrust: $-3\%[-1\%; -5\%]$, $p = 0.021$) and the blank control arm (mean decrease in distrust: -4% $[-1\%; -8\%]$, $p = 0.022$). No statistically significant differences were observed between the active and the blank control arms.

Conclusions. Receiving basic information on ML-based CDSS in hospital psychiatry may increase patient trust in such systems.

EPA EUROPEAN PSYCHIATRIC ASSOCIATION

Introduction

Machine learning (ML) is based on the idea that machines (computers) can learn from historical data and be trained for pattern recognition (e.g., prediction). The perspectives on using ML to aid decision-making in the medical field are promising [1]. Indeed, prediction models based on ML have shown to be accurate in many clinical contexts with performance levels comparable to- or above those of clinicians [2].

Since patients are major stakeholders in the medical field, their acceptance of ML is of paramount importance with regard to successful implementation of tools based on ML [3,4]. Public and patient trust in ML in the healthcare setting has been surveyed before [5–7]. According to these studies, stakeholder trust in medical application of ML models relies on the knowledge that the final decision lies in the hands of the health professionals, that is that ML models are merely used for decision *support.* Moreover, increased information about ML models, including explainability (transparency of what drives prediction), is associated with increased trust in them [6,8,9].

To our knowledge, however, no prior surveys have focused on opinions on ML-based clinical decision support systems among patients receiving treatment in psychiatric services. This is an unfortunate gap in the literature, as such systems are gaining traction in the psychiatric field [10,11] and because the level of general and institutional trust is relatively low among some patient groups in psychiatry [12,13]. Therefore, the aim of this study was to investigate trust in ML-based clinical decision support systems among patients with mental disorders. Furthermore, we will test whether receiving information about ML-based clinical decision support systems will increase patients' trust in this technology.

Methods

Design

We conducted a randomized online survey experiment focusing on ML-based clinical decision support tools in psychiatric services. The study design and analysis plan were pre-registered and are available at https://doi.org/10.17605/OSF.IO/Z9385. The design of the study is shown in Figure 1.

Setting

The survey experiment was performed within the Psychiatric Services of the Central Denmark Region, which has a catchment area of approximately 1.3 million people. It comprises five public psychiatric hospitals, which provide free (tax-financed) inpatient, outpatient, and emergency psychiatric treatment.

Participants

Patients were eligible for participation if they were 18 years old and received treatment in the Psychiatric Services of the Central Denmark Region. Patients were ineligible if they had a forensic sanction, received coercive treatment, or had an organic mental disorder (ICD-10 code: F0X.X), or mental retardation (ICD-10 code: F7X.X). Based on power calculations (see below), we invited 6000 randomly drawn eligible patients using the online survey service SurveyXact [14] via "eBoks" – the secure digital mailing system used by Danish authorities to communicate with citizens [15]. The survey was distributed from May 26–31, 2023. A reminder was sent on June 12, 2023, to those not having responded yet. The participants provided informed consent for study participation by ticking a box and entering their unique social security number and name. The participants did not receive any monetary incentive for participation.

After the survey was fielded, it came to our attention that participants using an Android-based device to access the "e-Boks" app could not access the hyperlink provided in the invitation. To solve this issue, we informed the participants of a solution to this technical problem when distributing the reminder. While this issue is likely to have reduced the response rate, we consider it unlikely to have introduced bias.

Power calculation

A power-analysis (alpha = 0.05, power = 0.80, two-sided) for a pairwise comparison, assuming a small intervention effect of Cohen's D = 0.2 (as seen in similar studies [8]), estimated that 1200 participants (400 per randomization arm) were required. A recent study in the same setting and using the same invitation procedure had a response rate of approximately 20% [16]. Based on this effect size and response rate, 6000 patients were invited to participate in the study.

Randomization

Participants were randomly allocated to one of three arms: intervention, active control, and blank control (for a description, see the "Survey experiment" section, below). As SurveyXact did not allow for standard 1/3, 1/3, 1/3 random allocation at the time of the study, the system was devised to assign the participants to one of the three arms based on the time they accessed the survey link in the invitation letter. Specifically, participants accessing the link in the intervals from $>= 0$ to 0.333 second; from >0.333 to 0.666 second, and from >0.666 to <1 second were assigned to the blank control arm, the active control arm, and the intervention arm, respectively. The participants were not aware that a randomization and intervention took place.

Baseline questionnaire

When entering the survey, irrespective of allocation arm, the participants initially filled in a baseline questionnaire regarding education level, current work status, household composition (adults and children), general trust, trust in technology, and perceived understanding of standard clinical decision-making as well as a perceived understanding of ML-based clinical decision support systems. Answers on trust and perceived understanding were provided on Likert scales from 0–10 (an English translation of the questionnaire in Danish is available in Supplementary Table 1: "Baseline questionnaire").

Survey experiment

The baseline questionnaire was followed by the experiment in which the participants received three different types of information based on the randomized allocation:

1. Intervention: Visual and text-based information pamphlet (slides within the electronic survey – see Supplementary Table 2: "Intervention") explaining how an ML-based clinical decision support system works and may aid clinical practice in psychiatric services.
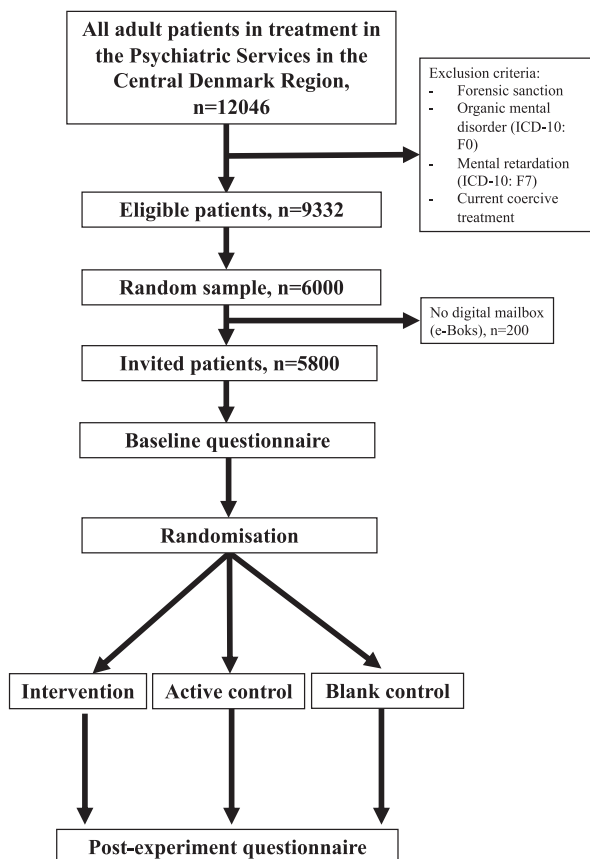

Figure 1. "Flowchart of study design and population".
"e-Boks: The secure digital mailing system used by Danish authorities to communicate with citizens".

2. Active control: Visual and text-based information pamphlet (slides within the electronic survey – see Supplementary Table 3: "Active control") explaining a standard clinical decision process in psychiatric services without the use of an ML-based clinical decision support system.
3. Blank Control: No information pamphlet.

Both information pamphlets in the online survey consisted of four slides each.

Post-experiment questionnaire (outcome measure)

After the survey experiment, the participants filled in a questionnaire aimed at measuring trust and distrust in ML-based clinical decision support systems in psychiatric services. Specifically, the respondents answered the following questions: 1. "I feel safe that mental health professionals can make decisions with the support of machine learning models." 2. "I trust that the Psychiatric Services can use machine learning models in a safe and appropriate way." 3. "I am concerned that use of machine learning models for decision support in psychiatry will increase the risk of error." 4. "I would like to have the opportunity to opt out of machine learning models being used for decision support in relation to my treatment in the psychiatric services." 5. "I am concerned that healthcare services, including the psychiatric services, are becoming too dependent on machine learning models." 6. "I am concerned that the use of machine learning models may lead to increased inequality in healthcare, including psychiatry." 7. "The advantages of using machine learning models for decision support in psychiatry outweigh the disadvantages." 8. "It is important to me that I can get an explanation of the basis on which a machine learning model recommends a given treatment". 9. "I am concerned that a machine learning model may make incorrect recommendations due to inaccuracies in my medical record." The questions were adapted from prior studies covering the same topic [6,17]. All questions were answered using an 11-level Likert scale ranging from 0 ("Totally disagree") to 10. ("Totally agree").

Choice of primary outcome measure

As it is suboptimal to sum positively and negatively worded items (after inversion) [18], the three positively-worded "trust" items (1, 2, and 7) and the 5 negatively-worded "distrust" items (3, 4, 5, 6, and 9) were grouped a priori (see the pre-registered analysis plan). Item 8 was considered to be neutral and was therefore kept separate. Subsequently, principal component analyses were performed to test whether the trust and distrust items, respectively, were loaded onto latent components. The number of components was determined by analyzing the scree plot and choosing the number of components before the distinct break ("elbow") in the plot [19]. Subsequently, an item was considered to load onto a component if it had a loading of >0.40 or < −0.40 [20]. Based on the number of items loading onto the component, a trust total score (the sum of the positively-worded items) and a distrust total score (the sum of the negatively worded items) were constructed to be used as the outcome measures.

Handling of survey responses

After clicking the generic link to the survey, the patients identified themselves by manually inserting their social security number and their names. If the participants completed the questionnaire more than once, the first response was used. If the participants first made a partial response, including the randomization element, and subsequently completed the survey, while randomized to a different arm, then the full response was not included in the analyses (as these participants were unblinded to the randomization). As the outcome measures were placed at the end of the survey, only completers were included in the analyses. Questions could not be left blank. Thus, there were no missing values.

Supplementary data from electronic health records

The participants gave consent to extraction of sociodemographic (age, sex, civil status) and clinical data (the number of contacts to the psychiatric services including contact type (inpatient/outpatient) and the associated ICD-10 diagnoses, as well as the time since their first contact to the psychiatric services) from the electronic health records for the purpose of the study. Linkage was performed using the respondent's unique personal identification number [21], To define diagnostic subgroups, we considered the participant's most severe main diagnosis (registered from 2011 until the time of the survey) using the following ICD-10 hierarchy: F2x (psychotic disorders) > F3x (mood disorders) > F4x (anxiety- and stress-related disorders) > F5x (eating, sleeping, and other behavioral syndromes associated with physiological disturbances) > F6x (personality disorders) > F8x (developmental disorders including autism) > F9x (child and adolescent mental disorders) > F1x (substance use disorders).

Statistics

The sociodemographic and clinical characteristics were summarized using descriptive statistics. Potential differences in time used for survey completion between the randomization groups were tested using Mann–Whitney U test. As primary analyses, the level of trust/distrust in ML-based clinical decision support systems quantified by the trust/distrust scores were compared pairwise between the three randomization arms via two-sample t-tests. Equivalent, secondary analyses were conducted at the individual trust/distrust item level. The Pearson correlation coefficient was used to assess correlation between the latent trust and distrust mean scores. As robustness analyses, two-sample t-tests of trust/distrust in ML-based clinical decision support systems was performed across the three randomization arms, while stratifying by sex, age, diagnostic groups (affective/anxiety disorders versus psychotic/other disorders), socioeconomic factors (e.g., educational level, current work status), baseline knowledge of machine learning as decision support, and the level of general trust. The significance threshold was set at 0.05. Correction for multiple comparisons was not performed as the analyses were pre-registered and highly interdependent [22]. All data management and statistical analyses were performed using Rstudio version 2023.06.0 Build 421.

Ethics

Research studies based on surveys are exempt from ethical review board approval in Denmark (waiver no. 1-10-72-138-22 from the Central Denmark Region Committee on Health Research Ethics). The study was approved by the Legal Office in the Central Denmark Region (reg. no. 1-45-70-21-23) and registered on the internal list of research projects having the Central Denmark Region as data steward (reg. no. 1-16-02-170-23). Before the survey, to ensure that it was appropriate for the study population, we received feedback on

the questionnaire and the "intervention" and "active control" information pamphlets from two patients having received treatment in the Psychiatric Services in the Central Denmark Region. The respondents provided informed consent to participate in the study.

Role of the funding source

Results

A total of 992 invitees completed the survey (106 partial respondents were excluded) and Table 1 lists their sociodemographic and clinical characteristics as well as baseline information regarding trust in institutions and technologies. See Supplementary Table 4: "Clinical characteristics of the 1,098 randomized participants" for clinical characteristics of 1098 randomized participants (Full respondents + partial respondents).

The randomization led to the following distribution of full respondents across arms: blank control = 319 (partial respondents = 37), active control = 343 (partial respondents = 33), and intervention = 330 (partial respondents = 36). The median response time was 280 seconds (IQR: 167) for those allocated to the blank control arm, 368 seconds (IQR: 264) for the active control arm, and 388 seconds (IQR: 244) for the intervention arm. The response time for those in the active control arm and the intervention arm was statistically significantly longer than for those in the blank control arm (active control vs blank control: W:37551, $p < 0.0001$; intervention vs blank control: 32735, $p < 0.0001$) but did not differ statistically significantly from each other (active control vs intervention: W: 53222, $p = 0.18$).

The results of the principal component analysis are shown in Supplementary Table 5: "PCA component loadings for trust items (positively worded outcome items)" and Supplementary Table 6: "PCA component loadings for distrust items (negatively worded outcome items)" and Supplementary Figures 1 and 2. As expected a priori, the survey items were grouped into a trust component consisting of the three positively worded items and a distrust component consisting of the five negatively worded items. The

Table 1. Characteristics of the 992 participants with complete responses

| Variable | Overall n = 992 | | Randomization group | | | | | |
| | | | Blank control n = 319 | | Active control n = 343 | | Intervention n = 330 | |
	Median	IQR	Median	IQR	Median	IQR	Median	IQR
Age	35	25	34	25	37	26	36	24
Sex	n	%	n	%	n	%	n	%
Female	690	70%	215	67%	244	71%	231	70%
Most severe diagnosis[a]	n	%	n	%	n	%	n	%
Substance use disorders (F1)	5	1%	<5	<2%	<5	<2%	<5	<2%
Psychotic disorders (F2)	186	19%	53	17%	80	23%	53	16%
Affective disorders (F3)	540	54%	184	58%	180	52%	176	53%
Neurotic disorders (F4)	132	13%	36	11%	41	12%	55	17%
Eating, sleeping and sexual disorders (F5)	18	2%	<5	<2%	7	2%	7	2%
Personality disorders (F6)	38	4%	13	4%	15	4%	10	3%
Disorders of psychological development (F8)	7	1%	<5	<2%	<5	<2%	<5	<2%
Child and adolescent disorders (F9)	51	5%	21	7%	11	3%	19	6%
Other	15	2%	5	2%	<5	<2%	6	2%
	Median	IQR	Median	IQR	Median	IQR	Median	IQR
Total median contacts/year	11	12	10	11	11	13	11	11
Days since first contact to Psychiatric Services	1,293	2780	1,122	2521	1,654	2934	1,327	2835
Education	n	%	n	%	n	%	n	%
5-years higher education	113	11%	33	10%	40	12%	40	12%
4-years higher education	208	21%	66	21%	74	22%	68	21%
3-years higher education	95	10%	29	9%	36	10%	30	9%
High school	224	23%	83	26%	60	17%	81	25%
Primary school	196	20%	61	19%	73	21%	62	19%
Vocational school	68	7%	19	6%	28	8%	21	6%
Other	88	9%	28	9%	32	9%	28	9%
Current work status	n	%	n	%	n	%	n	%

Continued

Table 1. *Continued*

Variable	Overall n = 992		Randomization group					
			Blank control n = 319		Active control n = 343		Intervention n = 330	
Disability pension	182	18%	40	13%	78	23%	64	19%
Employed (full-time)	93	9%	35	11%	27	8%	31	9%
Employed (part-time)	43	4%	15	5%	15	4%	13	4%
Subsidized employment	60	6%	17	5%	21	6%	22	7%
Student (full time)	119	12%	48	15%	34	10%	37	11%
Student (part-time)	27	3%	10	3%	8	2%	9	3%
Retired	59	6%	17	5%	22	5%	20	6%
Sick leave	198	20%	69	22%	54	16%	75	23%
Unemployed	73	7%	24	8%	29	9%	20	6%
Vocational rehabilitation	138	138%	44	14%	55	16%	39	12%
Household adults	**n**	**%**	**n**	**%**	**n**	**%**	**n**	**%**
0	380	38%	121	38%	150	44%	109	33%
1	384	39%	128	40%	118	34%	138	42%
2	172	17%	56	18%	52	15%	64	19%
3 or more	56	6%	14	4%	23	7%	19	6%
Household children	**n**	**%**	**n**	**%**	**n**	**%**	**n**	**%**
0	772	78%	248	78%	280	82%	244	74%
1	104	10%	43	13%	24	7%	37	11%
2 or more	116	12%	28	9%	39	11%	49	15
	Median	**IQR**	**Median**	**IQR**	**Median**	**IQR**	**Median**	**IQR**
Baseline knowledge of treatment decisions in Psychiatry	7	3	7	3	6	4	7	4
Baseline knowledge of machine learning	4	7	4	7	4	7	3	6
General trust	6	5	6	4	6	5	5	5
Trust in technology	6	3	6	3	6	3	6	4
Trust in parliament	5	5	5	4	5	5	5	4
Trust in judicial system	7	4	7	3	6	4	7	3
Trust in the police	7	4	7	4	7	3	7	4
Trust in the European Union	5	3	5	3	5	3	5	4
Trust in healthcare services	7	3	7	3	7	3	7	3
Trust in psychiatric services	7	4	7	4	7	4	8	4

Note: Cell counts <5 are not specified due to risk of identification of individual patients.
[a]Most severe diagnosis in the period from 2011–2023. (F*) indicates the ICD-10 chapter.

trust and distrust sum scores were inversely correlated (Pearson correlation coefficient $t = -0.60$).

Table 2 shows the response to the items focusing on trust and distrust in ML-based clinical decision support systems, across the three randomization groups. Notably, the median scores on the trust items were generally within the range of the trust general/institutional trust levels reported at baseline (available in Table 1).

Figure 2 shows the results of the primary analyses, which compare the three randomization groups with regard to the trust and the distrust sum score. The intervention increased trust in ML-based clinical decision support systems when compared to the active control (mean difference in trust: 5% [95% CI: 1%;9%], $t = -2.60$, df = 668.5, $p = 0.0096$) and the blank control arm (mean

difference in trust: 4% [95% CI: 1%;8%], $t = -2.43$, df = 645.0, $p = 0.015$). Similarly, the intervention reduced distrust in ML-based clinical decision support systems when compared to the active control (mean difference in distrust -3% [95% CI: -5%; -1%], $t = -2.60$, df = 670.4, $p = 0.021$) and the blank control arm (mean difference in distrust -4% [95% CI: -8%; -1%], $t = 2.30$, df = 646.53, $p = 0.022$). For both trust and distrust, there were no material or statistically significant differences between the active and the blank control arms (Trust: $t = 0.208$, df = 660.0, $p = 0.84$; Distrust: $t = -0.048$, df = 660.0). The equivalent results at the level of the Individual trust and distrust items are listed in Supplementary Table 7: "Single items from the post-experimental questionnaire with t-tests between groups"

Table 2. Individual item scores after the experiment

Outcome items (abbreviated[a])	Blank control		Active control		Intervention	
Trust items						
	Median	IQR	Median	IQR	Median	IQR
Item 1: I feel safe with ML models	6	4	6	4	6	5
Item 2: Trust in ML models	6	4	5	4	7	4
Item 7: The advantages outweigh disadvantages	5	3	5	3	5	3
Distrust items						
	Median	IQR	Median	IQR	Median	IQR
Item 3: Malpractice due to ML	5	4	6	5	5	5
Item 4: Possibility to opt out of ML	7	5	6	5	5	6
Item 5: Dependency on ML	6	5	7	5	6	4
Item 6: Inequality in healthcare due to ML	5	4	5	5	5	5
Item 9: Errors in EHR	8	5	8	5	8	5
Neutral item						
	Median	IQR	Median	IQR	Median	IQR
Item 8: Explainability of ML models	10	2	10	2	10	2

Abbreviations: ML: Machine learning. EHR: Electronic health record. Mean differences between groups for single items and results from T-tests are shown in Supplementary Table 7.
[a]The full phrasing of the items and their scoring range are available in the methods section.

and are in agreement with those from the analyses at the trust and distrust sum scores. Notably, the neutral item (importance of explainability of ML models) had an overall median of 10 (8–10) with a statistically significant difference between the blank control and the intervention arms (higher in the intervention arm), but not for the other comparisons.

The results of the analyses stratified by sex, age, diagnostic groups, educational level, current work status, baseline knowledge of machine learning as decision support, and the level of general trust are listed in Supplementary Table 8: "Results from intervention stratified by age, sex, diagnostic category, educational level, work status" and suggest that the intervention effect is generally consistent across the stratification groups with a few notable exceptions. Specifically, the intervention increased trust and decreased distrust in women, but not/less in men. Similarly, the intervention increased trust and decreased distrust for those with affective/anxiety disorders, but not/less for those with psychotic/other disorders.

Discussion

This randomized survey experiment among patients receiving treatment in psychiatric services showed that information on ML-based clinical decision support systems may increase patient trust in such systems. Notably, the results regarding the explainability of ML models suggest that this aspect was particularly important for the respondents.

To the best of our knowledge, this study represents the first investigation into whether providing information about ML as a decision-support tool may enhance patient trust in such systems, suggesting that this is indeed the case. This result is in line with that from the study by Cadario et al., predominantly targeting individuals from the general population, which investigated the effect of receiving information on the variables (shape/size/color) driving a malignant melanoma risk prediction algorithm [8]. In the study by Cadario et al, the participants either received information on

human healthcare providers versus ML algorithm's decision-making processes. Consistent with our study, it was found that a small amount of information about ML algorithms reduced "algorithm aversion", i.e., reduced the reluctance to utilize algorithmic support compared to human providers [8]. The results from Cadario et al. also suggest that explainability is positively associated with the uptake of prediction algorithms at the general population level. This is consistent with the patient perspective observed in this study in which the respondents agreed the most with the item stating "It is important to me that I can get an explanation of the basis on which a machine learning model recommends a given treatment." Analog findings have been reported in studies of patients (within, e.g., primary care, radiology, and dermatology) [7], clinicians (within, e.g. internal medicine, anesthesia, and psychiatry) [23], and the general population [5,6]. Taken together, these findings suggest that when developing ML models for healthcare, there should be an emphasis on explainability to ensure trust among both clinicians and patients.

The observed sex difference in the intervention effect (more effective in females compared to males) may be attributed to the fact that men generally reported higher baseline knowledge of machine learning and AI, which could likely have led to a ceiling effect (less "room" for increasing trust/reduce distrust in these technologies) [24]. That the intervention effect was more pronounced in patients with affective/anxiety disorders compared to those with psychotic/other disorders is likely partly driven by the sex difference outlined above, as there are relatively more females in the first (72%) compared to the latter patient group (63%). Furthermore, it has been reported that patients with psychotic disorders are particularly prone to have reduced trust in mental healthcare professionals [13], which may extend to the information received in the intervention of this study. Thus, it may be that a different type/dose of intervention is required to reach patients with psychotic disorders.

While the results of this study are indicative of a causal effect of the information intervention, the effect was numerically quite
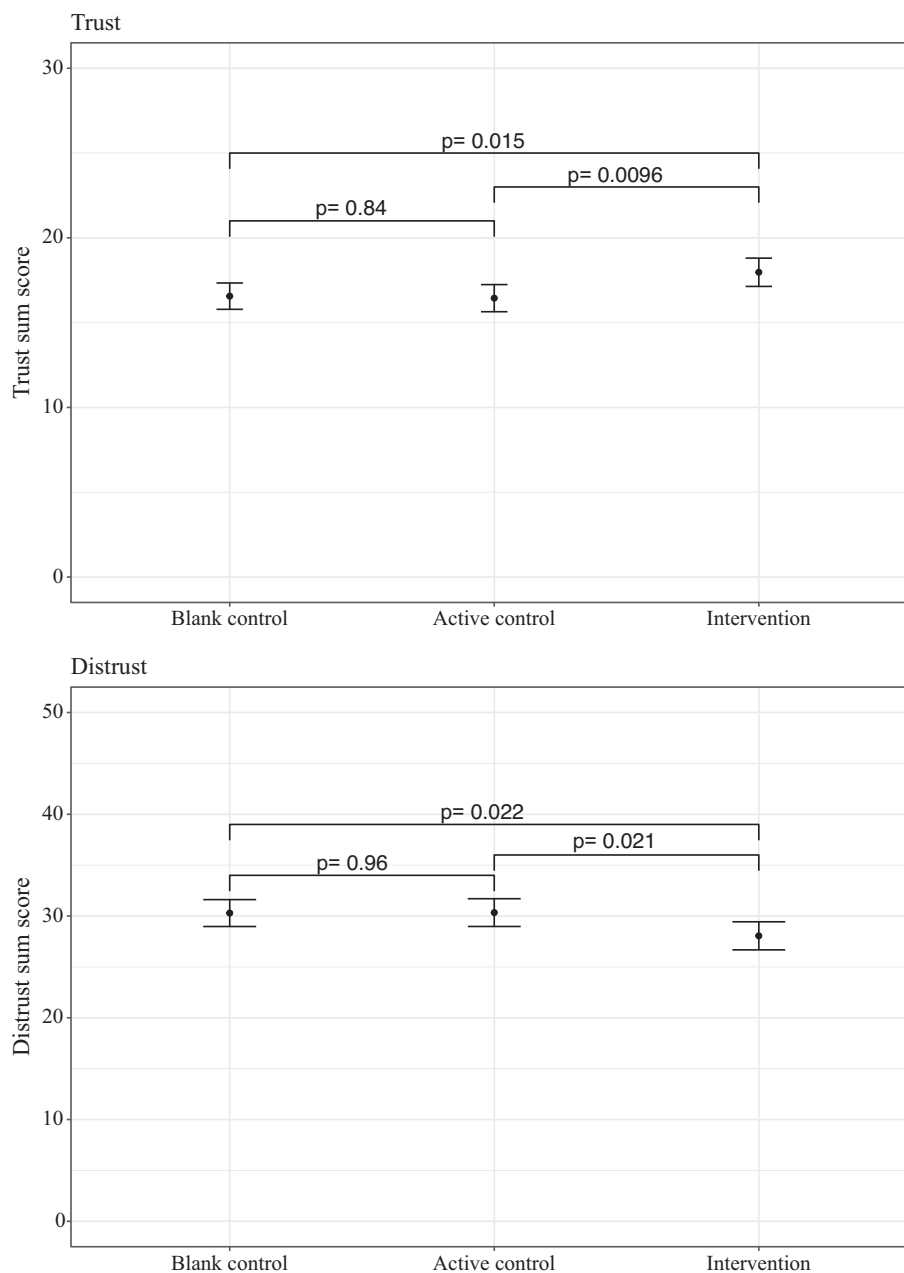
Trust

Distrust

Figure 2. "Effect of the intervention on trust (top) and distrust (bottom) in machine learning model-based clinical decision support systems". The error bars represent confidence intervals.

small. This was expected as the intervention consisted of an electronic pamphlet with only four slides of text and pictures, which was administered only once. Furthermore, it can also be argued that the effect of the intervention may well be short-lived as the trust/distrust outcome was measured immediately after the intervention. For these reasons, future studies should ideally employ more comprehensive and repeated interventions, and longer time between the interventions and the outcome measurement.

Another aspect potentially affecting the effect size of the intervention is the timing of the survey. Specifically, as the survey was fielded at the end of May 2023, it came in the aftermath of the press coverage of the open letter signed by several high-profile tech leaders calling for a pause in the development of artificial intelligence until appropriate safeguards and legislation would be in place

[25]. This media coverage primarily focused on the security concerns and potential hazards associated with this emerging technology [26,27]. This somewhat negative press on artificial intelligence could have rendered the respondents resistant to the information in the intervention (reduced effect size). In contrast, however, it is also possible that the negative press had resulted in reduced baseline trust in the technology, which would then leave ample room for a positive effect of the information conveyed in the intervention (increased effect size). Based on the data at hand, we are, unfortunately, not able to determine the overall direction of this potential response bias.

There are limitations to this study, which should be taken into account by the readers when interpreting the results. First and foremost, the survey had a relatively low response rate, which

means that selection bias may be in play. Due to the lack of clinical and sociodemographic data on those not participating, we were unable to adjust the analyses for attrition. However, in a recent 2-wave survey in the same population during the COVID-19 pandemic (focusing on psychological distress/well-being among patients with mental disorders during the pandemic), we were granted force majeure access to clinical and sociodemographic data on non-respondents. This allowed us to employ inverse probability weighting to address the potential bias arising from non-response. Notably, this adjustment had no material impact on the results [16,28]. While this does not preclude selection bias in this study (a different survey topic and timing), it does suggest that this is unlikely to be a substantial problem. Second, we did not use a validated questionnaire for measuring trust and distrust in ML as a clinical decision support tool as such questionnaires, to our knowledge, have not yet been developed. With this inherent limitation in mind, we believe that the latent components of trust and distrust derived from principal component analysis of items with apparent face validity, represent reasonable outcome measures. Third, we did not have attention-check questions, which means that some participants may have responded inconsistently/arbitrarily to the questions. This would have introduced noise in the data and complicated signal detection. Yet, a signal (the intervention increasing trust and reducing distrust) was indeed detected. Fourth, with regard to generalizability, it should be borne in mind that Denmark is among the most digitalized countries in the world [29], and its inhabitants may, therefore, be more positive towards new technology. Thus, replication of the reported findings in other countries is warranted.

In conclusion, this survey experiment suggests that receiving information on ML-based clinical decision-support systems in hospital psychiatry likely increases patient trust in such systems. This is compatible with results from studies of other patient populations, clinicians as well as the general population. Thus, when taken together, the literature suggests that providing appropriate information to patients will be of importance when implementing ML-based clinical decision support systems.

References

[1] Beam AL, Drazen JM, Kohane IS, Leong T-Y, Manrai AK, Rubin EJ. Artificial Intelligence in medicine. N Engl J Med 2023;388:1220–1.

[2] Ben-Israel D, Jacobs WB, Casha S, Lang S, Ryu WHA, de Lotbiniere-Bassett M, et al. The impact of machine learning on patient care: A systematic review. Artif Intell Med 2020;103:101785.

[3] Salazar de Pablo G, Studerus E, Vaquerizo-Serrano J, Irving J, Catalan A, Oliver D, et al. Implementing precision psychiatry: A systematic review of individualized prediction models for clinical practice. Schizophr Bull 2021; 47:284–97.

[4] Frank D-A, Elbæk CT, Børsting CK, Mitkidis P, Otterbring T, Borau S. Drivers and social implications of Artificial Intelligence adoption in healthcare during the COVID-19 pandemic. Guidi B, editor. PLOS ONE. 2021;16:e0259928.

[5] Young AT, Amara D, Bhattacharya A, Wei ML. Patient and general public attitudes towards clinical artificial intelligence: A mixed methods systematic review. Lancet Digit Health 2021;3:e599–e611.

[6] Ploug T, Sundby A, Moeslund TB, Holm S. Population preferences for performance and explainability of Artificial Intelligence in health care: Choice-based conjoint survey. J Med Internet Res 2021;23:e26611.

[7] Reading Turchioe M, Harkins S, Desai P, Kumar S, Kim J, Hermann A, et al. Women's perspectives on the use of artificial intelligence (AI)-based technologies in mental healthcare. JAMIA Open 2023;6:ooad048.

[8] Cadario R, Longoni C, Morewedge CK. Understanding, explaining, and utilizing medical artificial intelligence. Nat Hum Behav 2021;5:1636–42.

[9] Yarborough BJH, Stumbo SP. Patient perspectives on acceptability of, and implementation preferences for, use of electronic health records and machine learning to identify suicide risk. Gen Hosp Psychiatry 2021;70: 31–7.

[10] Danielsen AA, Fenger MHJ, Østergaard SD, Nielbo KL, Mors O. Predicting mechanical restraint of psychiatric inpatients by applying machine learning on electronic health data. Acta Psychiatr Scand. 2019;140:147–57.

[11] Bernert RA, Hilberg AM, Melia R, Kim JP, Shah NH, Abnousi F. Artificial Intelligence and suicide prevention: A systematic review of machine learning investigations. Int J Environ Res Public Health 2020;17:5929.

[12] Kopacz MS, Ames D, Koenig HG. Association between trust and mental, social, and physical health outcomes in veterans and active duty service members with combat-related PTSD symptomatology. Front Psychiatry 2018;9:408.

[13] Verhaeghe M, Bracke P. Stigma and trust among mental health service users. Arch Psychiatr Nurs 2011;25:294–302.

[14] https://www.survey-xact.dk/. 2023-08-31 [Internet]. [cited 2023 Aug 31]. Available from: https://www.survey-xact.dk/

[15] Ebert JF, Huibers L, Christensen B, Christensen MB. Paper- or Web-Based Questionnaire Invitations as a Method for Data Collection: Cross-Sectional Comparative Study of Differences in Response Rate, Completeness of Data, and Financial Cost. J Med Internet Res 2018;20:e24.

[16] Kølbæk P, Jefsen OH, Speed M, Østergaard SD. Mental health of patients with mental illness during the COVID-19 pandemic lockdown: A questionnaire-based survey weighted for attrition. Nord J Psychiatry. 2021;1–10.

[17] Esmaeilzadeh P, Mirzaei T, Dharanikota S. Patients' Perceptions toward human–Artificial Intelligence interaction in health care: Experimental Study. J Med Internet Res 2021;23:e25856.

[18] Timmerby N, Nørholm V, Rasmussen N-A, Lindberg L, Andreasson Aamund K, Bech P. A major clinimetric dilemma in self-reported outcome scales: Mixing positively and negatively worded items. Psychother Psychosom 2017;86:124–5.

[19] Abdi H, Williams LJ. Principal component analysis. Princ Compon Anal.

[20] Guadagnoli E, Velicer W. Relation of sample size to the stability of component patterns. Psychol Bull 1988;103:265–75.

[21] Pedersen CB. The danish civil registration system. Scand J Public Health 2011;39:22–5.

[22] Rothman KJ. No adjustments are needed for multiple comparisons: Epidemiology. 1990;1:43–6.

[23] Diprose WK, Buist N, Hua N, Thurier Q, Shand G, Robinson R. Physician understanding, explainability, and trust in a hypothetical machine learning risk calculator. J Am Med Inform Assoc 2020;27: 592–600.

[24] Gillespie N, Lockey S, Curtis C. Trust in artificial Intelligence: A five country study. 2021.

[25] Pause Giant AI. Experiments: An Open Letter. [Internet]. Future of Life institute; 2023 [cited 2024 Mar 20]. Available from: https://futureoflife.org/open-letter/pause-giant-ai-experiments/

[26] Gregory A, Hern A. AI poses existential threat and risk to health of millions, experts warn [Internet]. The Guardian; 2023 [cited 2024 Mar 20]. Available from: https://www.theguardian.com/technology/2023/may/10/ai-poses-existential-threat-and-risk-to-health-of-millions-experts-warn

[27] Metz C, Schmidt G. Elon Musk and Others Call for Pause on A.I., Citing 'Profound Risks to Society' [Internet]. The New York Times; 2023 [cited 2024 Mar 20]. Available from: https://www.nytimes.com/2023/03/29/technology/ai-artificial-intelligence-musk-risks.html

[28] Kølbæk P, Gil Y, Schmidt FCL, Speed M, Østergaard SD. Symptom severity and well-being of patients with mental illness during the COVID-19 pandemic: A two-wave survey. Nord J Psychiatry 2023;77:293–303.

[29] Digital Economy and Society Index (DESI) – Denmark. European commision; 2022.