

Peer Review

Review of: "Acceptability of chatbot versus General Practitioner consultations for healthcare conditions varying in terms of perceived stigma and severity (Preprint)"

Evgenii Arsentev¹

1. Independent researcher

1. *What the work claims*

I read this paper as asking one clear question. Does the stigma or the severity of a health problem change how willing people are to accept advice from a chatbot, from a General Practitioner (GP), or from a GP and chatbot together? The author also asks whether the characteristics of the participants make a difference (Section 2).

To answer it, the author ran an online experiment in which each participant rated health scenarios. Stigma, severity, and the consultation source were changed from one scenario to the next. The sample is 237 UK adults recruited through Facebook (Section 3.1). In my own words, the main claim is that for high-severity issues, people found the GP most acceptable, then the GP and chatbot together, then the chatbot alone. For low-severity issues, I understand the three sources were about equally acceptable. Stigma made the GP advantage smaller, and participant characteristics showed no significant effect (Section 4).

2. *Whether the conclusion holds on the data*

What I value most here is the design. The study was pre-registered on OSF and approved by the UCL ethics committee (Section 3). Severity, stigma, and source are crossed within subjects, so each participant is compared with themselves (Section 3). The health issues were chosen by a pilot of 37 people, not by the author's own judgement (Experimental conditions). The limitations section is candid about self-selection, missing data, and the gap between willingness and behaviour (Section 5.5). The severity

conclusion also holds well on the data. In Table 2 ("Sev X Info"), high severity gives GP 4.42 (0.06), GP-chatbot 3.44 (0.12), and chatbot 2.68 (0.12). At low severity, the three sources are close: GP 3.51, GP-chatbot 3.43, chatbot 3.48.

The stigma conclusion needs more care. The abstract says that more stigmatised issues increased chatbot acceptability, but in Table 2 ("Stig X Info") the chatbot mean is 3.14 (0.12) for high stigma and 3.02 (0.11) for low stigma, a difference of 0.12. Section 5.2 says stigma increases chatbot acceptance "if the severity of the health issue is low". Table 2 ("Sev X Stig X Info") shows the opposite direction at low severity, with chatbot 3.44 (0.13) for high stigma against 3.52 (0.13) for low stigma. The stigma difference for the chatbot appears at high severity instead, 2.85 (0.13) with high stigma against 2.51 (0.13) with low stigma. I would bring the abstract and Section 5.2 in line with these cell means, or add a sentence explaining why the text and the table differ.

I could not tell which sample was analysed. Section 4 gives 237 completed responses, and Section 5.5 says the excluded cases were almost one third of participants, but the number analysed is not reported. The degrees of freedom $F(2, 372)$ in the Abstract would correspond to 187 participants if the error term is $2 \times (n - 1)$, though that is my own inference. Is 187 the analysed n ? It would help to state the final n in Section 4, with the exclusions counted. On the statistics, I did not find a sphericity correction or post hoc pairwise comparisons in Section 3.5 or Section 4. Yet Sections 4 and 5 state orderings between specific sources. I would add the sphericity result (or the correction used) and pairwise tests with a stated adjustment for multiple comparisons, so those orderings rest on a test.

Last, the outcome is self-reported willingness for imagined conditions, not use, as the author says in Section 5.5. I would carry this sentence into the abstract and the conclusions, so that "acceptability" is read in that sense. With the stigma claim aligned to the tables, the analysed n stated, and the pairwise tests added, I think the conclusions would follow clearly from the data.

3. Reproducibility

I would like to thank the author for how much was made checkable here. The protocol and analysis plan are pre-registered at <https://osf.io/szgma/> (Section 3), the 12 conditions are listed in full ("3. Experimental conditions"), the wording of the willingness item and of the confidence item is quoted (Section 3.2), and the chatbot explainer video is linked (Section 3.4). That is a good base for anyone who wants to repeat the study.

Several things are still missing for a full replication. I would add the names of the 12 health issues used and their pilot stigma and severity scores, because the paper gives only the cut-offs (mean under 2 low, over 3.5 high, "3. Experimental conditions"). It would help to add a data availability statement and to name the statistical software. The exclusion rule and the number excluded should be stated exactly, since the paper says only "almost 1/3" (Section 5.5). I would also state whether the analysis matched the pre-registered plan or deviated from it, and list any deviations.

One small coding point. Prior chatbot knowledge is coded "Unsure (coded 0), yes (i), no (ii)" (Section 3.2), while Table 1 reports No / Yes / Missing, so the "Unsure" group is not shown. I would add it to the table, or explain where those participants went.

4. *What specifically to fix*

1. In the abstract, the severity-by-source interaction is given as $F(2, 372) = 118.14$, partial eta squared 0.388, while Table 2 ("Sev X Info") gives $F = 188.14$. With df 2 and 372, partial eta squared = $(2 \times F) / (2 \times F + 372)$. For $F = 118.14$ this is $236.28 / 608.28 = 0.388$, which matches the abstract, so 188.14 in Table 2 looks like a typo. I would correct it.
2. The abstract gives the stigma-by-source interaction as $F(2, 372) = 12.99$ with partial eta squared 0.65. By the same formula $(25.98 / 397.98)$ the value is about 0.065, so 0.65 looks like a misplaced digit.
3. It would help to report the analysed N and the exclusion rule (Section 4, Section 5.5), with the number excluded at each step. A reader can then follow how the sample was reached.
4. Section 5.2 ("if the severity of the health issue is low") does not fully match Table 2, "Sev X Stig X Info". At low severity, chatbot acceptability is 3.44 for high stigma against 3.52 for low stigma. At high severity it is 2.85 against 2.51. I would reword 5.2 and the abstract's stigma sentence to match these numbers.
5. Section 5.3 opens with "a GP-chatbot combination did increase the chatbot acceptability" and then says "participants were not more accepting of a GP-chatbot combination". One of the two sentences needs rewording, so the reader knows which result you mean.
6. Some table values need a second look. In Table 1, confidence in chatbot knowledge sums to $68 + 168 = 236$ with 0 missing, while the other rows sum to 237. In Table 3, the SEM of 0.74 for high confidence stands out against 0.08-0.13 elsewhere in that table. Table 3 also gives means without

test statistics, and adding F and p for each covariate would support the statement "none significant" (Section 4).

7. I would add how sphericity was handled, and pairwise comparisons with a correction. It would also help to list the 12 health issues with their pilot scores, and to state data availability. Two small things. Section 3.4 says "Patients were presented with the information page", and I think "participants" is meant. The numbering also jumps from 3.2 to "3. Experimental conditions" to 3.4, so it would help to renumber the sections.

5. Recommendation

Major revisions, though I want to say first that the design and the pre-registration are sound, and the severity result is clear in Table 2. The fixes I would ask for are about reporting, and none of them needs new data collection. It would help to check the two effect-size numbers in the abstract and Table 2, and to state the analysed N plainly. I would also align the stigma wording in Section 5.2 and the abstract, and add the missing post hoc tests.

Declarations

Potential competing interests: No potential competing interests to declare.