Paper 2

true

Abstract

Researchers in the social sciences are increasingly drawing upon citizen science approaches for the collection of data. These approaches seek to involve non-professional scientists in the research process. However, there are concerns surrounding the quality of data produced by citizen scientists, which hinder further adoption of citizen science approaches. Protocols have been developed to detect and remove low quality contributions and ensure that data produced by citizen scientists are of sufficient quality; however, these protocols are difficult to apply for approaches such as ecological momentary assessment, which focus on generating data on subjective phenomena. Researchers using these approaches have tended to use engagement, typically operationalised as the quantity of contributions made by a particpant, as a proxy measurement for the quality of contributions. However, it remains unknown the extent to which data quality is a good proxy for data quality. If it is not, this approach can potentially create issues for the reproducibility of results by undermining the statistical power of the studies, and by creating a large amount of arbitrary researcher degrees of freedom. Using data from the Britain Breathing project, we show that .... Two or three sentences explaining what the **main result** reveals in direct comparison to what was thought to be the case previously, or how the main result adds to previous knowledge. One or two sentences to put the results into a more **general context**. Two or three sentences to provide a **broader perspective**, readily comprehensible to a scientist in any discipline.

Introduction:

Citizen science is an increasingly popular approach to research in the social sciences. However, concerns surrounding the quality of data produced by citizen scientists haved hindered further adoption of citizen science approaches (Aceves-Bueno, et al. 2017; Basiri, et al. 2019; Elliott & Rosenberg, 2019; Lukyanenko, et al. 2016; Riesch & Potter, 2014).

Whilst it has been shown that the quality of data provided by citizen scientists can rival that of professional scientists, there is more heterogeneity in the quality of data (Aceves Bueno, et al. 2017).

Identifying and removing low quality and careless responses remains essential for the both the credibility of the data, and for valid inferences to be made (Huang, 2018; Johnson & Sieber, 2013; Ternovski, 2022; McGonagle, et al. 2016). Huang (2018) demonstrates that insufficient effort responses can have a confounding effect on variables of interest, and can inflate observed correlations.

Ternovski argues that inattentive respondents can introduce "substantial measurement error and attenuation bias" (2022: 1). In a power analysis they found that almost four times more participants were needed to achieve 80% power for their statistical tests, than when no screening for attention was implemented.

In response to these concerns, approaches have been developed to mitigate concerns around data quality in citizen science studies, such as the implementation of quality assurance standards and protocols (Minghini, et al. 2017; Fonte, et al. 2017; Samulowska, et al, 2021).

This includes approaches that leverage large amounts of redundancy in the data collection or classification approach (Balázs, et al. 2021; Lintott et al. 2008), or approaches where strong priors about the distribution of likely observations enable the flagging of unlikely reports for further investigation (Salganik, 2019; Kelling et al. 2012).

However, these approaches are not universally applicable and are often developed in contexts where there is a relatively large amount of information about both participants, their contribution history, and the object or phenomena about which they are collecting data.

For some approaches, such as studies using experience sampling methodology (ESM) which typically focus on generating data about the subjective experiences of participants, these approaches are difficultly applicable, in particular when these studies are opt-in, and joining the study is as simple as downloading a smartphone application.

Given the difficulties in applying standard data quality protocols to ESM data, the most common approach used in these situations has been to use engagement, typically operationalised as the quantity of contributions submitted by a participant, as a proxy for the quality of the data, on the assumption that quality and quantity are associated, since both the attentiveness with which reports were completed and the number of reports completed are assumed to be at least partially determined by a participant's underlying motivation and level of engagement with the study (Doherty, et al. 2020; Geerharts, 2021).

Jaso, et al (2021) note that, for ESM studies, the "closest"best-practice" standard for cleaning data is the tendency to remove participants who do not meet an a-priori compliance cut off defined by the percent of surveys completed " (3).

However, excluding participants has several potential downsides, including "rejecting legitimate responses, reducing power, and reducing sample's representativness" (Ternovski, 2022). Furthermore the arbitrariness and large amount of researcher degrees of freedom for exclusion criteria can lead to multiple comparison problems, which can undermine the reproducibility of results (Steegen, et al. 2016; Gelman & Loken, 2013; Simonsohn, Simmons, & Nelson, 2020; Ioannidis, 2008). This is especially the case for exclusion criteria that are chosen post-hoc (Wicherts, et al. 2016: 5).

Furthermore, it remains to be established whether low engagement users do systematically produce lower-quality data. Our understanding of the motivations, levels of carelessness, and quality of the data produced by low-engagement participants remains limited (Eveleigh, et al. 2014., Jaso, et al. 2021., Welling, et al. 2021).

This paper uses data from Britain Breathing, a smartphone application-based ESM study on the symptoms of allergic rhinitis, and anti-hystamine prescription data from OpenPrescribe to explore whether low engagement users do provide less reliable data. Furthermore, by exploring several ways of operationalising engagement identified in the litterature, this paper tests how sensitive the impact of removing low-engagement participants to the operationalisation of engagement.

Finally this paper explores whether demographic differences between high and low engagement users, the key factor in determining how much rejecting low-engagement users will affect the sample's representativeness, are sensitive to how engagement is operationalised.

Methods:

Case study

This paper uses data from the Britain Breathing project and from OpenPrescribing.net .

Britain breathing data:

The Britain Breathing project is a citizen science study which used an smartphone-based experience sampling approach to collecting geolocated time series data on seasonal pollen allergy symptoms, also referred to as allergic rhinitis or hay-fever (Vigo, et al. 2018).

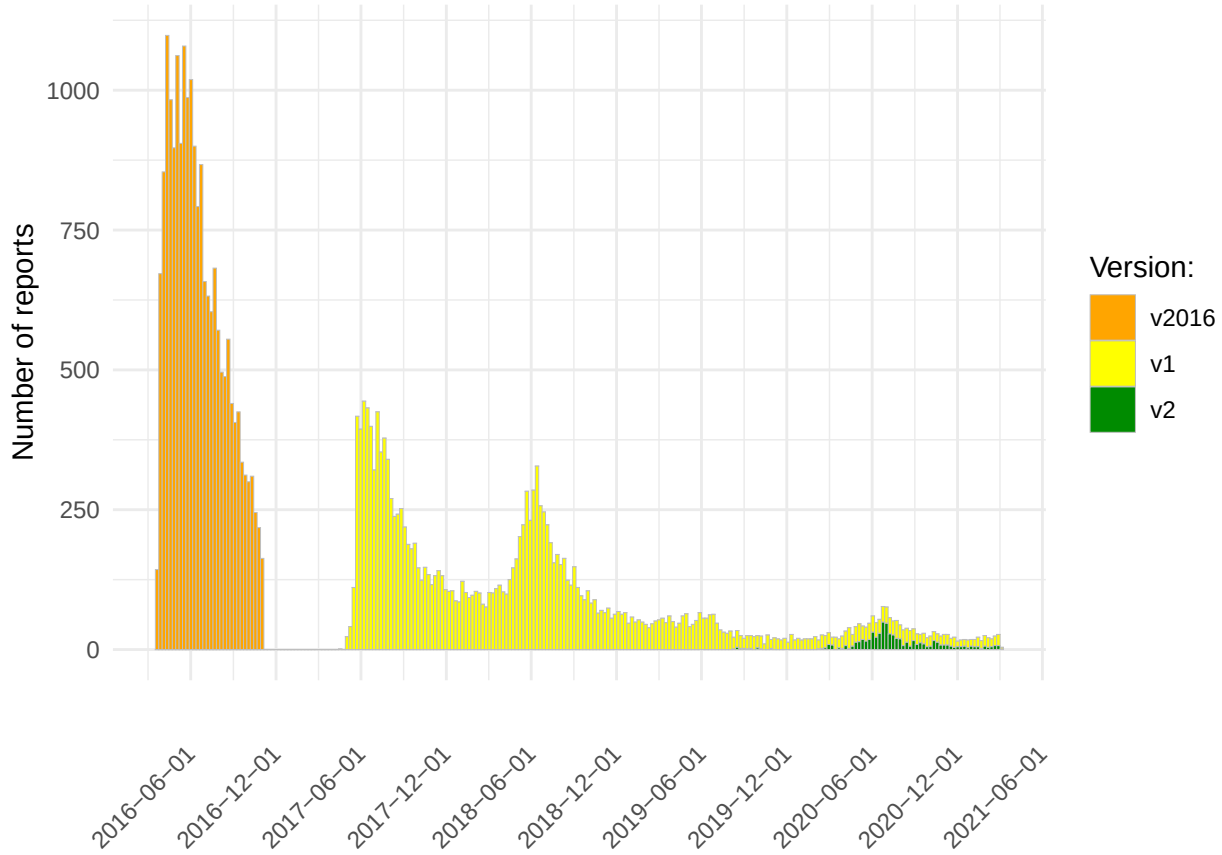
Participants are invited to download a mobile phone application and provide some basic information such as their gender, age, and allergy history. Users of the application can report symptoms at any time, and also at daily scheduled intervals. When they make a report, they are first asked how they are feeling. If they respond that they are feeling well, no further questions are asked, if they respond otherwise, they are

asked further questions about the severity of various symptoms ("nose", "eyes, and"breathing"), which they report on a four point scale ranging from 0, which signifies an absence of symptoms, to 3, which indicates severe symptoms.

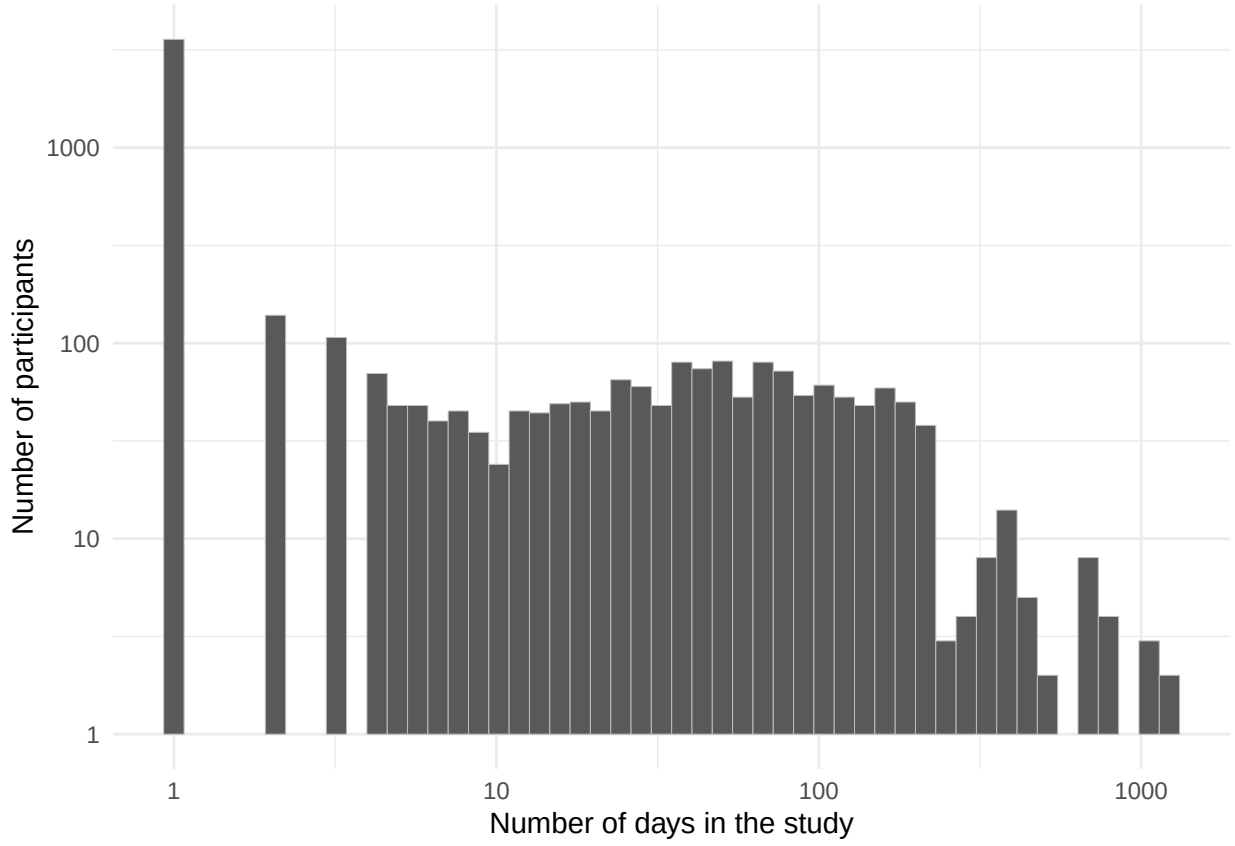
Available further information includes the gender of participants, their date of birth, known allergies, whether or not they are taking antihistamine medication, and the date, time, and location of each report submitted.

The data has good geographical coverage, with reports submitted from 95% of all postcode areas in the UK, with the average number of reports per postcode being 167 (idem: 89) **Should I update this with the latest figures? rather than quoting from the original vigo paper which doesn't have data from 2019 and 2020.**

Three versions of the application were deployed, v2016, v1 and v2, the use of which is shown in Figure x.



A common feature of citizen science data is high levels of participation inequality (Hackley, 2016). This feature can also be seen in the Britain Britain data. Fig X shows how many participants have reported on a given number of days. The log-log scale makes clear that the overwhelmingly most common behavior was for participants to make a single contribution. This is in line with findings from (XXXX add Short list of projects which show single contributions (Jaso and geerharts?, or maybe not just esm XXXX).



NHS England prescription data:

As a means of validating the data collected using the Britain Breathing application, Vigo, et al (2018) look at the correlation between the median lack of wellness reported on the Britain Breathing application and the number of antihistamines prescribed by general practitioners, finding that they are very strongly (r = 0.93) correlated. This data is openly available from OpenPrescribing.net. The smallest spatial unit at which this data is provided is the Clinical Commissioning group level.

Procedure

Separating data into years

Britain Breathing data is available from April 2016 to April 2021. There were 200 982 contributions by 4 748 participants in 2016, 74 183 contributions by 335 participants in 2017, 65 534 contributions by 269 participants in 2018, 19 115 contributions by 78 participants in 2019, 16 446 contributions by 52 participants in 2020, and 182 contributions by 9 users in 2021.

Data was submitted throughout the year, with the exception of 2017, where data collection began in May, and in 2021, where data is available from January until the beginning of March

Data was analysed for each year, with the exception of 2021, due to the limited amount and temporal coverage of data for that year (fig x).

Clustering by level of engagement

The way in which engagement is understood and operationalised varies a lot between scholars and fields (Perski, et al. 2017, Yardley, et al. 2016). This paper seeks to establish how sensitive the impact of removing

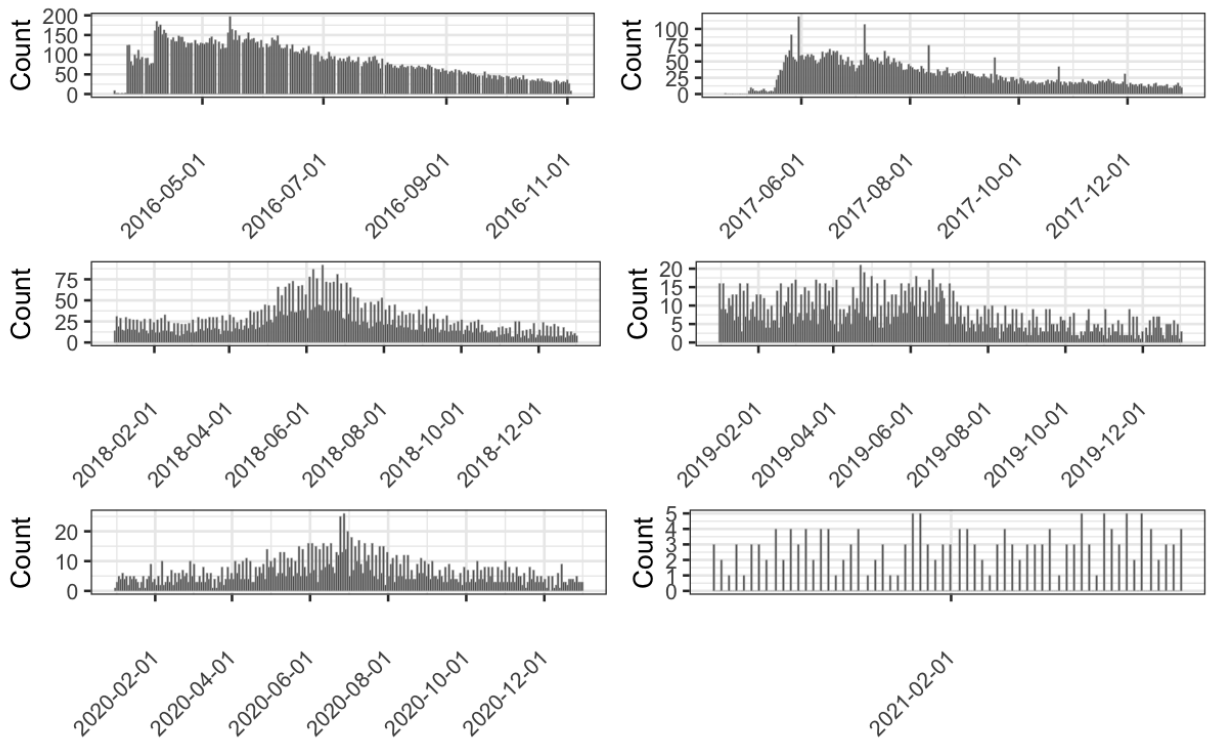


Figure 1: Fig x

disengaged users is to the way engagement is operationalised.

Threshold methods

Jaso, et al (2021) describe a general trend in the EMA/ESM literature to only include participants with a 70%–90% compliance rate, highlighting that this is often selected by convention and not empirically determined (2021:3). For example Sun, et al (2020) consider participants who completed less than 10 assessments to be unengaged, and excluded them for their analysis.

A more data driven applicationroach is found in a study by Kronkvist and Engstrom (2020), who split their participants into *abstainers*, who completed zero assessments, *dedicated participants* who completed a number of assessments one standard deviation or more above the average number of assessments completed by participants, and *occasional participants* who did not meet the criteria for the two previous groups.

Hidden Markov models

Druce, et al (2017) argue such approaches overlook the complexity of patterns of engagement. For example a participant may be strongly engaged with the application for a week, carefully filling in assessments every day, before having to cease contributing for various reasons. Under the approaches described above such contributions would not be labeled as highly engaged (and potentially excluded from analysis) for either missing a threshold such as 10 contributions, or for being below some deviation from the mean number of contributions.

Instead, they propose using first order hidden Markov models...

Participants were coded as engaged on a given day if they made at least one contribution to the application. This variable was assumed to be the result of participants being in one of three latent states of engagement

Comparing with prescription data

Data on antihistamine prescriptions is available from OpenPrescribing at the Clinical Commissioning Group (*CCG*) level. To enable comparisons with Britain Breathing data, reports were assigned to the CCG areas in which they were submitted using the sf package (Pebesma, 2018).

Vigo, et al (2018)... **XXXX I started to write up what had been done in vigo et al here before outlining what i planned to do, but it's not actually clear what was done in the original paper XXXX**

<—- Note: Amount vs Quantity

From the Open prescribe website:

"*Items* counts the number of times a medicine has been prescribed. It says nothing about how much of it has been prescribed (for that see quantity) as some presciptions will be for many weeks' worth of treatment while others will be much smaller.

***Quantity* is the total amount of a medicine that has been prescribed, but the units used depend on the particular form the medicine is in"**

I'm not sure exactly which variable to use here, items would seem to indicate how often people go to the doctor with Hayfever problems, quantity is the amount that has been prescribed, which I would assume is more to do with the gp's personal preferences than the patients? (my personal experience with hayfever has been to just buy over the counter loratadine, rather than go through the NHS, which can take weeks)

<—-

Group differences

To asses the impact of removing unengaged participants, this paper compares the characteristics of res

Demographics

Previous studies have explored associations between engagement and demographic characteristics.

Perski, et al (2017) report age, gender, education, employment and ethnicity as the most commonly found associations with engagement and attrition. Similarly, Druce, et al (2017) find that age differs notably between groups (with engaged participants being over 5 years older on average), and a substantially lower proportion of women were present in the "tourist" cluster. Kronkvist and Engstrom (2020) find that age and being female is associated with higher engagement in the study, as do Rintala, et al (2019). Turner, et al (2017) find that race/ethnicity, education, and age are are associated with higher levels of engagement.

(Would this ^ be better presented as a table?)

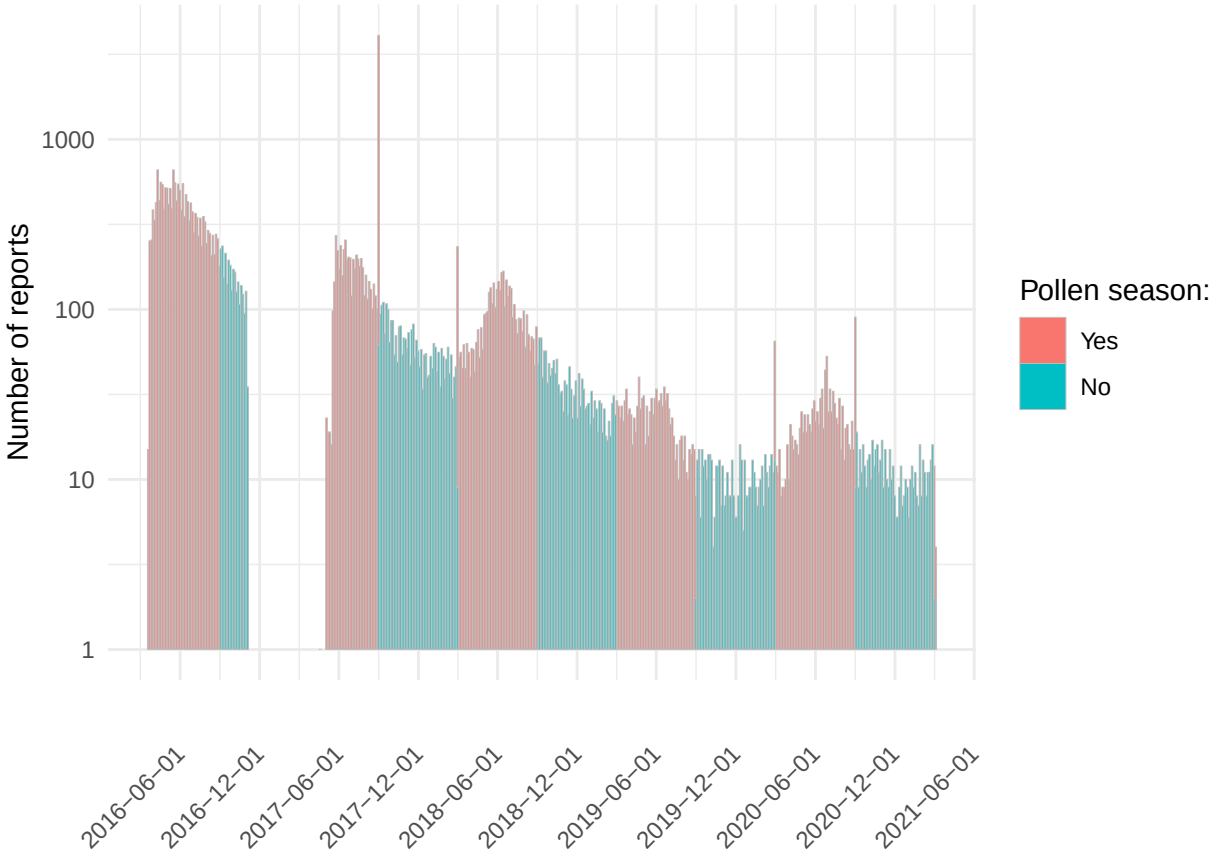
Britain breathing asks participants for their age and gender which will allow these characteristics to be compared. Furthermore the geotagged nature of the reports allows us to determine whether the reports are made in an urban or rural setting.

Symptom intensity

One possible factor in whether or not a participant chooses to contribute to the application on a given day may be the severity of the symptoms they are experiencing. Fig x shows that reports peak every year around pollen season. One possible mechanism is that as symptoms subside, participants stop reporting.

```
## Warning: Transformation introduced infinite values in continuous y-axis
```

```
## Warning: Removed 542 rows containing missing values (geom_bar).
```

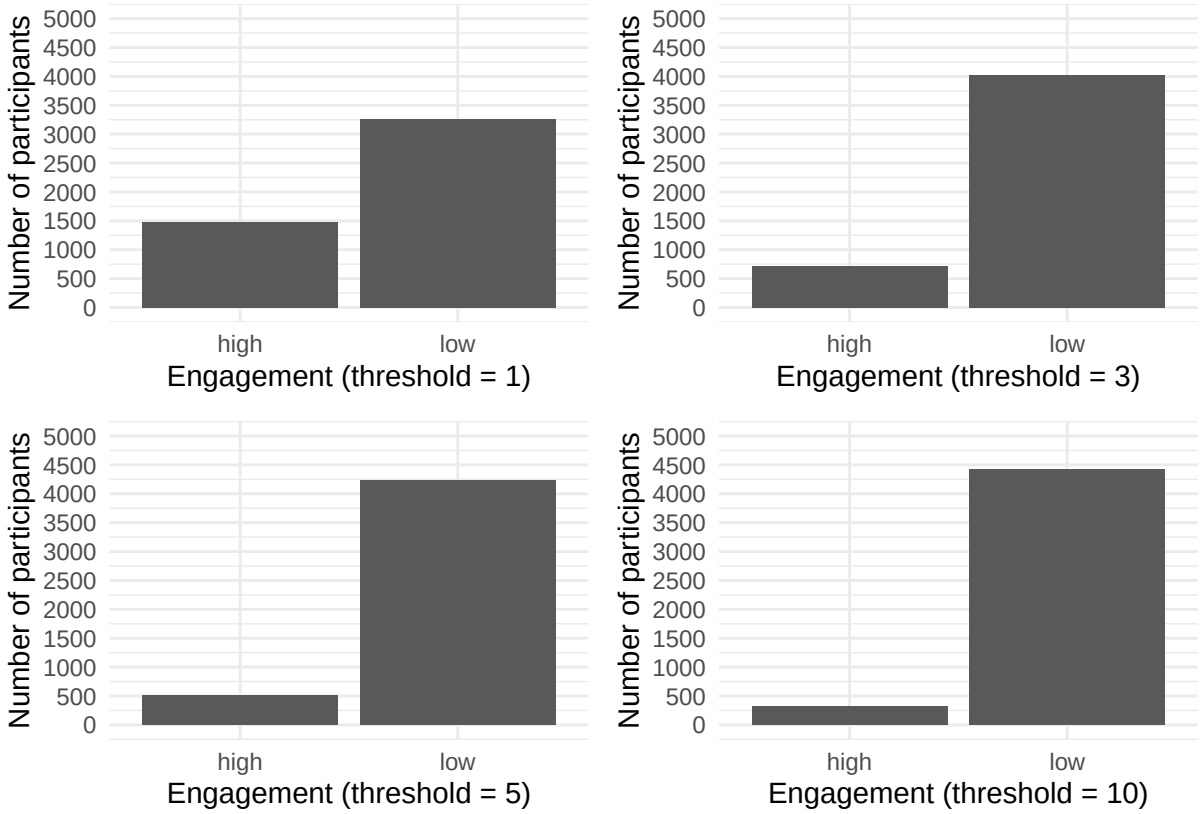


Data analysis

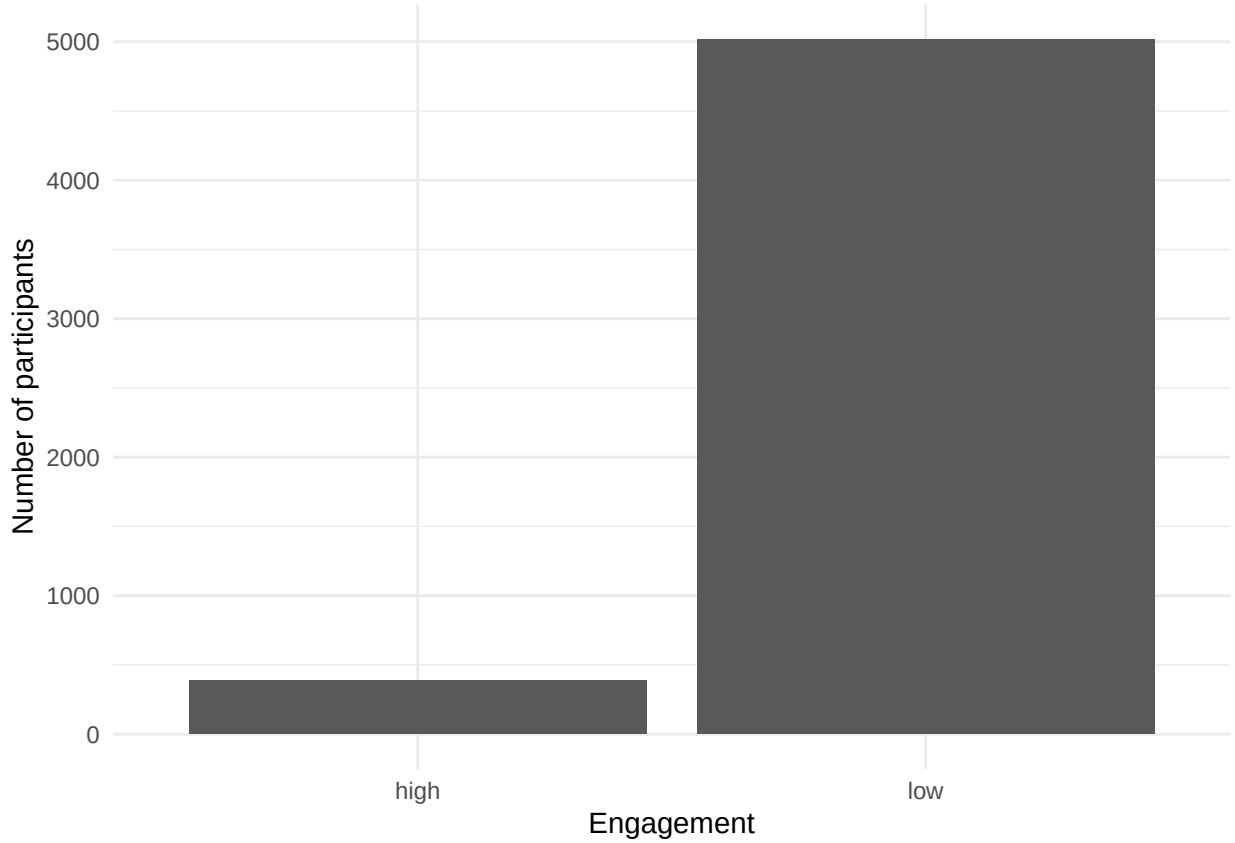
Clusterings

Simple clusterings

The most common way of operationalising engagement is using contribution thresholds (Jaso, et al. 2021). According to this approach, a participant is considered to not have engaged with the study if they contribute less than a certain number of reports. Figure X shows how participants in the first year would be classified according to this approach using 4 thresholds: 1, 3, 5, and 10.



Kronkvist and Engstrom (2020) split their participants into abstainers, who completed zero assessments, dedicated participants who completed a number of assessments one standard deviation or more above the average number of assessments completed by participants, and occasional participants who did not meet the criteria for the two previous groups. The mean number of contributions by participants in the first year of the study was ~4 and the standard deviation ~ 13. Figure X shows how participants would be classified according to this approach.



Hidden Markov Models

Druce, et al (2017) suggest using hidden Markov models as a better way of operationalising engagement. This paper applies the approach used in their paper to the Britain Breathing data.

First dummy variable was created for whether or a participant had made a contribution on a given day. One limitation of this approach is that some participants very occasionally made more than one report in a single day, potentially indicating higher engagement.

Hidden Markov models were then fit on this data using the depmixS4 R package (Visser, 2021) to estimate the latent level of each participant on a given day.

The model assumed that whether or not a participant would contribute on a given day was explained by them being in one of three latent states, high engagement (with a probability of contributing of 0.7) , low engagement (with a probability of contributing of 0.1), and disengaged (with a probability of contributing set at 1e-10, as the software does not allow for values of zero).

The disengaged state was assumed to be an absorbing state, meaning that once a participant was in the disengaged state, their probability of transitioning to another state was zero.

The model also assumed that all participants were in the highly engaged state on the day they began the study.

Finally participants were clustered according to the history of latent states inferred by the hidden Markov models using a Markov mixture model.

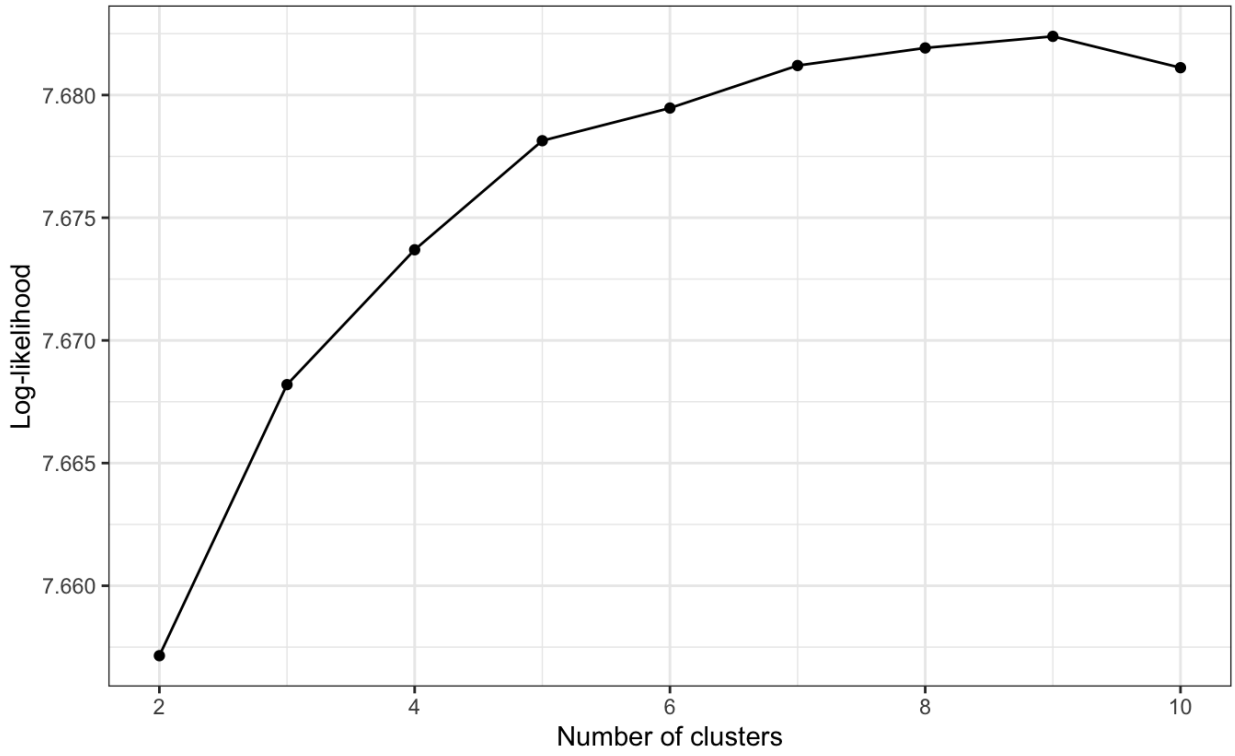
```
#Loading these manually as they take a long time to run

yr1_k4 <- readRDS(file = here::here("Clusterings", "yr1_k4"))
```

```
yr1_k4_clusters <- readRDS(file = here::here("Clusterings", "yr1_k4_clusters" ))
```

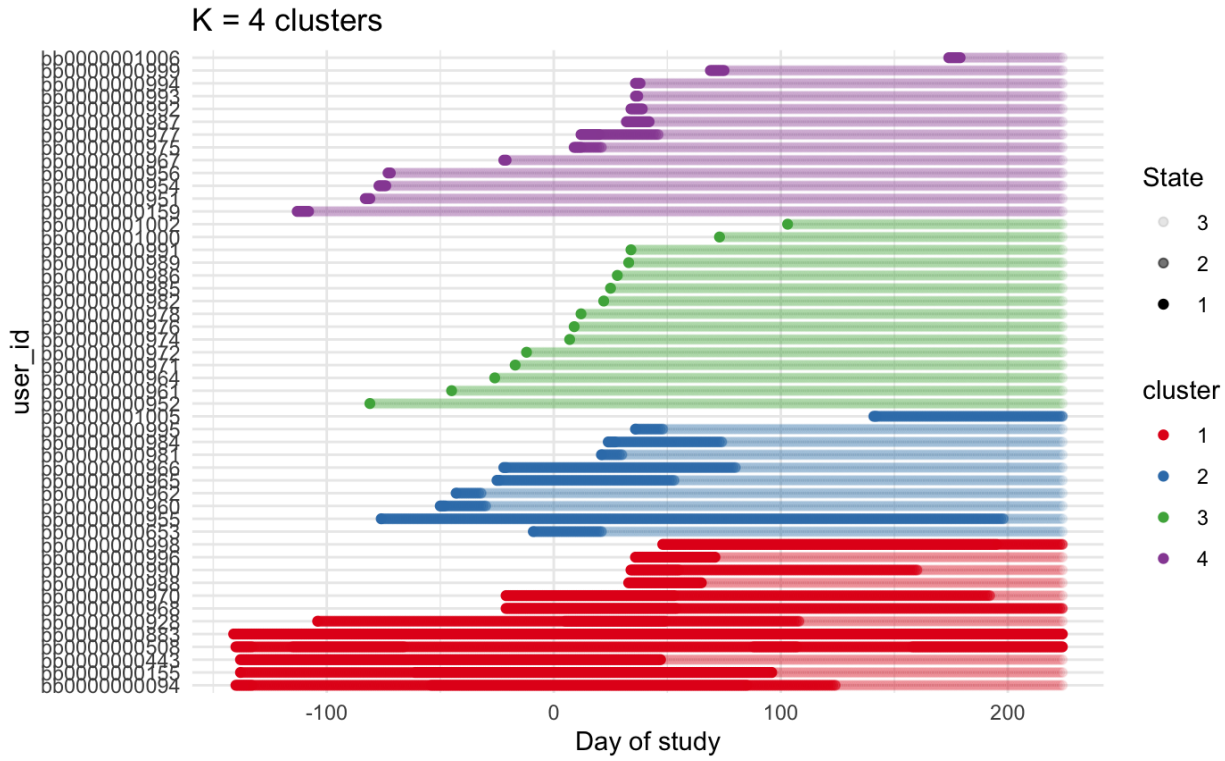
Using the elbow heuristic for detecting the optimal number of clusters we find that 4 clusters is optimal.

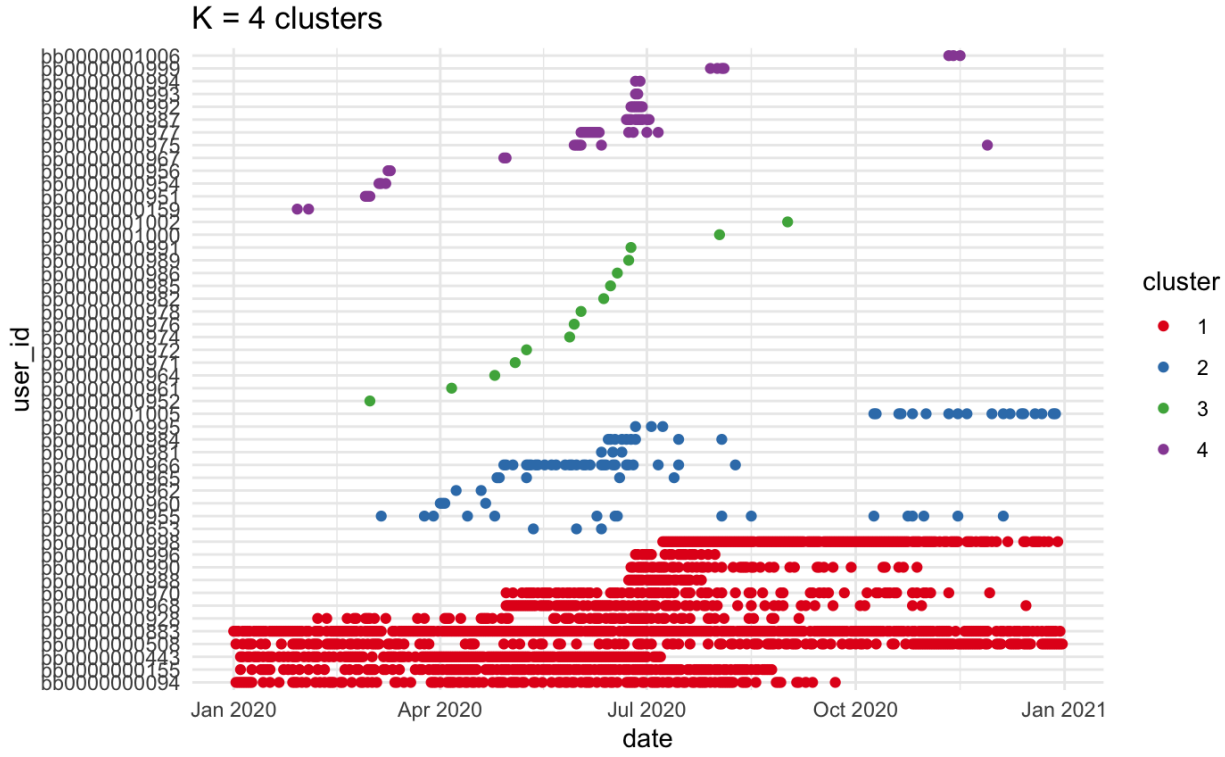
XXXX We don't actually find this at all this time, I found this the forst time I ran this and it is the most common result (there are also other reasons, mostly ease of interpretations, and consistency with the cloudy paper, why this is preferable). Also, things are much more consistant after the second year, likely a reflection of a smaller about of latent behaviours groupings given less participants XXXX



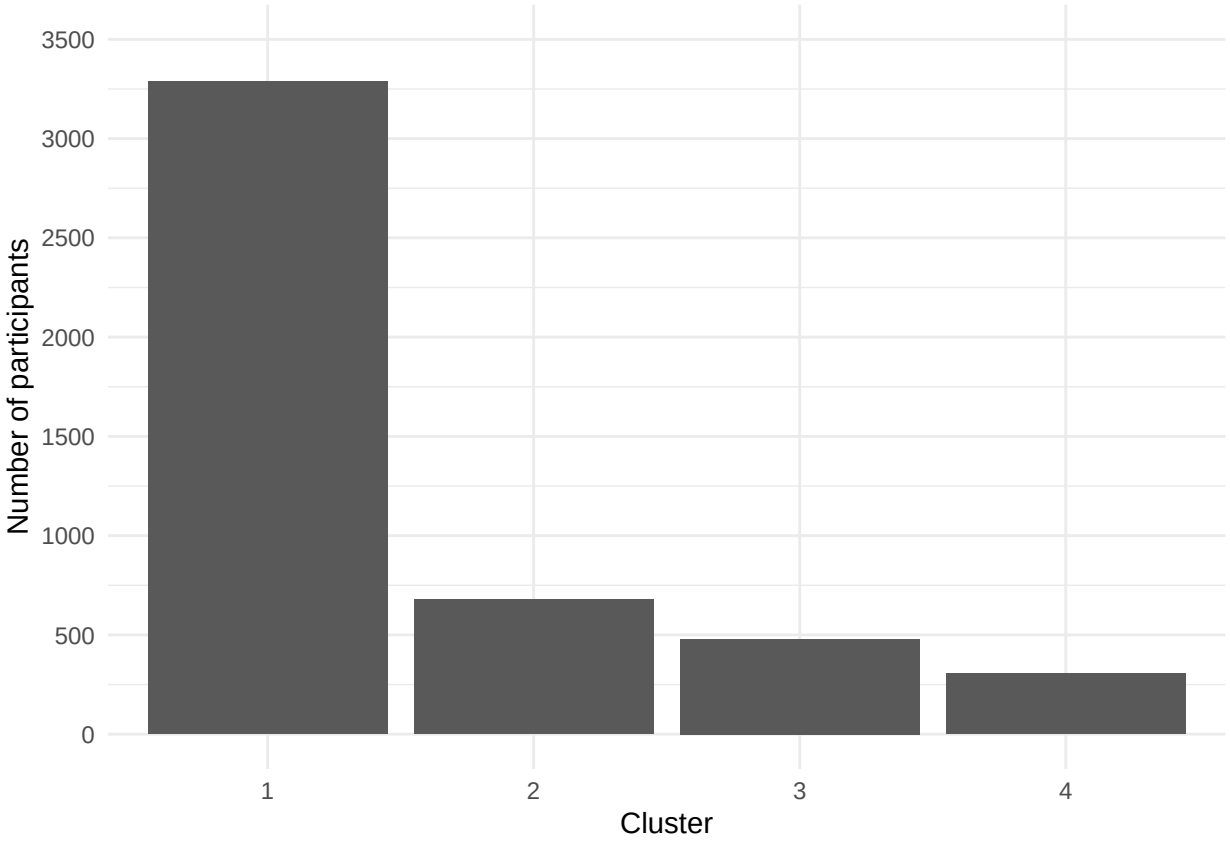
Cluster	Number of participants
1	526
2	475
3	3460
4	287

50 participants were randomly sampled from the study and visualized in figures x and x. Figure x shows the daily latent state assigned to the participants by the hidden Markov model, and the cluster assigned to the participant by the Markov mixture model. Figure x shows days on which the participant make a contribution, and the final cluster assigned to the participant by the Markov mixture model. **(I need to fix the x axis labels on these)**





Visually we can distinguish four reporting behaviors high (cluster 4, 526/4748), medium (cluster 2, 475/4784), low (cluster 4, 278/4784), and tourist (cluster 1, 3460/4748).



Correlation analysis

NHS prescription data is available from OpenPrescribing.net. However this data is only available for prescriptions made in England. Reports made outside of England (1697, or ~8.4% of reports) were therefore removed from the data, as they could not be linked to prescription data.

```
#Joining the cluster data with the prescription data

dfy1_all_clusters <- dfy1 %>%
  group_by(user_id) %>%
  count() %>%
  ungroup() %>%
  mutate(engagement_threshold_1 = if_else(n > 1, "high", "low"),
         engagement_threshold_3 = if_else(n > 3, "high", "low"),
         engagement_threshold_5 = if_else(n > 5, "high", "low"),
         engagement_threshold_10 = if_else(n > 10, "high", "low"),
         engagement_kron = if_else(n > 17, "high", "low")) %>%
  left_join(yr1_k4_clusters,by = "user_id") %>%
  dplyr::select(-prob.1,
          -prob.2,
          -prob.3,
          -prob.4,
          -n) %>%
  left_join(dfy1,by = "user_id")%>%
  st_as_sf(coords = c("longitude", "latitude"),
```

```
           agr = "constant",
           crs = "WGS84") %>%
  st_transform(crs = 27700)%>%
  st_intersection(ccgs) %>%
  st_join(ccgs, left = FALSE) %>%
  mutate(ccg = CCG21NM.y)
```

```
## Warning: attribute variables are assumed to be spatially constant throughout all
## geometries
```

```
dfy1_all_clusters %>%
  filter(ccg != "NHS VALE ROYAL CCG") %>%
  group_by(ccg, lubridate::month(date)) %>%
  summarise(reports = n(),
            mean_wellbeing = mean(how_im_doing))
```

```
## `summarise()` has grouped output by 'ccg'. You can override using the `.groups`
## argument.
```

```
## Simple feature collection with 668 features and 4 fields
## Geometry type: GEOMETRY
## Dimension:     XY
## Bounding box:  xmin: 148025.9 ymin: 28532.81 xmax: 652002.8 ymax: 653029.2
## Projected CRS: OSGB 1936 / British National Grid
## # A tibble: 668 x 5
## # Groups:   ccg [105]
##    ccg          `lubridate::mo~` reports mean_wellbeing                geometry
##    <chr>                   <dbl>   <int>          <dbl>            <GEOMETRY [m]>
##  1 NHS BARNSL~                 3       4          0.25  MULTIPOINT ((435254.1 40~
##  2 NHS BARNSL~                 4      13          0.308 MULTIPOINT ((417163.9 40~
##  3 NHS BARNSL~                 5      10          0.1   MULTIPOINT ((424306.2 40~
##  4 NHS BARNSL~                 7       2          0.5     POINT (444850 405796.8)
##  5 NHS BARNSL~                 8       3          0.667 MULTIPOINT ((433841.5 39~
##  6 NHS BASILD~                 3       5          0.6   MULTIPOINT ((559239.4 20~
##  7 NHS BASILD~                 4      11          1.09  MULTIPOINT ((557577.3 19~
##  8 NHS BASILD~                 5       7          0.857 MULTIPOINT ((559246 2003~
##  9 NHS BASILD~                 6       2          1     MULTIPOINT ((566839.9 18~
## 10 NHS BASILD~                 7       1          0     POINT (567811.1 190147.7)
## # ... with 658 more rows
```

Group differences

Results

Discussion

Conclusion

References

Aceves-Bueno, E., Adeleye, A. S., Feraud, M., Huang, Y., Tao, M., Yang, Y., & Anderson, S. E. (2017). The accuracy of citizen science data: a quantitative review. *Bulletin of the Ecological Society of America*, *98*(4), 278-290.

Basiri, A., Haklay, M., Foody, G., & Mooney, P. (2019). Crowdsourced geospatial data quality: Challenges and future directions. *International Journal of Geographical Information Science*, *33*(8), 1588-1593.

Balázs, B., Mooney, P., Nováková, E., Bastin, L., & Arsanjani, J.J (2021). "Data Quality in Citizen Science" in Vohland, K., Land-Zandstra, A., Ceccaroni, L., Lemmens, R., Perelló, J., Ponti, M., ... & Wagenknecht, K. *The science of citizen science*. Springer Nature.

Brovelli, M. A., Minghini, M., Molinari, M., & Mooney, P. (2017). Towards an automated comparison of OpenStreetMap with authoritative road datasets. *Transactions in GIS*, *21*(2), 191-206.

Druce, K. L., McBeth, J., van der Veer, S. N., Selby, D. A., Vidgen, B., Georgatzis, K., ... & Dixon, W. G. (2017). Recruitment and ongoing engagement in a UK smartphone study examining the association between weather and pain: cohort study. JMIR mHealth and uHealth, *5*(11), e168.

Elliott, K. C., & Rosenberg, J. (2019). Philosophical foundations for citizen science. *Citizen Science: Theory and Practice*, *4*(1).

Fonte, C.C., Antoniou, V., Bastin, L., Estima, J., Arsanjani, J.J., Bayas, J.C.L., See, L. and Vatseva, R. (2017). Assessing VGI data quality. *Mapping and the citizen sensor*, 137-163.

Haklay, M. E. (2016). Why is participation inequality important?. Ubiquity Press.

Jaso, B. A., Kraus, N. I., & Heller, A. S. (2021). Identification of careless responding in ecological momentary assessment research: From posthoc analyses to real-time data monitoring. Psychological Methods.

Johnson, P. A., & Sieber, R. E. (2013). Situating the adoption of VGI by government. In *Crowdsourcing geographic knowledge* (pp. 65-81). Springer, Dordrecht.

Kronkvist, K., & Engström, A. (2020). Feasibility of gathering momentary and daily assessments of fear of crime using a smartphone application (STUNDA): Methodological considerations and findings from a study among Swedish university students. *Methodological Innovations*, *13*(3), 2059799120980306.

Lukyanenko, R., Parsons, J., & Wiersma, Y. F. (2016). Emerging problems of data quality in citizen science. *Conservation Biology*, *30*(3), 447-449.

McGonagle, A. K., Huang, J. L., & Walsh, B. M. (2016). Insufficient effort survey responding: An under-appreciated problem in work and organisational health psychology research. *Applied Psychology*, *65*(2), 287-321.

Minghini, M., Antoniou, V., Fonte, C.C., Estima, J., Olteanu-Raimond, A.M., See, L., Laakso, M., Skopeliti, A., Mooney, P., Jokar Arsanjani, J. and Lupia, F. (2017). "The relevance of protocols for VGI collection." in

OpenPrescribing.net (*2022). The DataLab, University of Oxford.* Available online: https://openprescribing.net Last accessed: XXXX

Pebesma, E. J. (2018). Simple features for R: standardized support for spatial vector data. *R J.*, *10*(1), 439.

Perski, O., Blandford, A., West, R., & Michie, S. (2017). Conceptualising engagement with digital behaviour change interventions: a systematic review using principles from critical interpretive synthesis. *Translational behavioural medicine*, *7*(2), 254-267.

Riesch, H., & Potter, C. (2014). Citizen science as seen by scientists: Methodological, epistemological and ethical dimensions. *Public understanding of science*, *23*(1), 107-120.

Samulowska, M., Chmielewski, S., Raczko, E., Lupa, M., Myszkowska, D., & Zagajewski, B. (2021). Crowdsourcing without Data Bias: Building a Quality Assurance System for Air Pollution Symptom Mapping. *ISPRS International Journal of Geo-Information*, *10*(2), 46.

Sun, J., Rhemtulla, M., & Vazire, S. (2020). Eavesdropping on Missing Data: What Are University Students Doing When They Miss Experience Sampling Reports?. Personality and Social Psychology Bulletin, 0146167220964639.

Ternovski, J., & Orr, L. (2022). A Note on Increases in Inattentive Online Survey-Takers Since 2020. *Journal of Quantitative Description: Digital Media*, *2*.

Vigo, M., Hassan, L., Vance, W., Jay, C., Brass, A., & Cruickshank, S. (2018). Britain Breathing: using the experience sampling method to collect the seasonal allergy symptoms of a country. *Journal of the American Medical Informatics Association*, *25*(1), 88-92.

Visser,I & Maarten Speekenbrink (2010). depmixS4: An R Package for Hidden Markov Models. *Journal of Statistical Software*, 36(7), 1-21. URL https://www.jstatsoft.org/v36/i07/.

Yardley, L., Spring, B. J., Riper, H., Morrison, L. G., Crane, D. H., Curtis, K., ... & Blandford, A. (2016). Understanding and promoting effective engagement with digital behaviour change interventions. *American journal of preventive medicine*, *51*(5), 833-842.

Supplementary materials:

The code for this project is available at https://github.com/NathanKhadaroo/BB_Paper.

The data for this project is not publicly available due to its potentially sensitive nature; however, researchers may request access from **XXXX who? XXXX**