Detection and Remediation of Malicious Actors for Studies Involving Remote Data Collection

Bethany K. Bracken[1], John Wolcott[2], Isaac Potoczny-Jones[2], Brittany A. Mosser[3], Isabell R. Griffith-Fillipo[3] and Patricia A. Arean[3]
[1]*Charles River Analytics, 625 Mount Auburn St., Cambridge, MA, U.S.A.*
[2]*Tozny, LLC, 411 NW Park Ave. Ste 400, Portland, OR 97209, U.S.A.*
[3]*Department of Psychiatry & Behavioral Sciences, University of Washington, 1959 NE Pacific Street, Seattle, WA, 98195, U.S.A.*

Keywords: Remote Data Collection, Malicious Actors, Bots, Bad Actors.

Abstract: Although most human subjects research requires data collection by contacting local participants who visit a research site, some studies require increasingly large troves of data collected continuously during their typical daily lives using sensors (e.g., fitness trackers) and ecological momentary assessments. Long-term, continuous collection is becoming more feasible as smartphones become ubiquitous. To enable remote collection of these rich data sets while ensuring privacy, we built a system to allow secure and fully human-out-of-the-loop participant recruitment, screening, onboarding, data collection on smartphones, data transmission to the cloud, data security in the cloud, and data access by analysis and modeling teams. Study participants were paid for completion of daily ecological momentary assessments in keeping with standards of research equipoise, fairness, and retention strategies. However, our study attracted "malicious actors" who were pretending to be study participants, but were not, in order to receive payment. This opinion piece outlines how we initially detected malicious actors, and the steps we took in order to prevent future malicious actors from enrolling in the study. This opinion piece outlines several lessons learned that we think will be valuable for future studies that recruit, enroll, and maintain study participants remotely.

1 INTRODUCTION

Currently, most human-subjects data collection is done by recruiting participants through fliers or advertisements, and requiring that they visit the lab over the course of the study. However, this is costly, time-consuming, and results in decreasing subject retention with each required visit. Moreover, data collection in discrete time points only offers a small window into participants' lives and relies heavily on participant recall between visits to complete important behavioral and environmental data. Such data collection is flawed and rife with assumptions about data accuracy that may very well influence research into disease phenotyping, prediction analyses, and other important analyses (Areàn et al., 2016). With the advent of personal digital technology (e.g., fitness trackers, smartphone sensors), scientists are now in the position to collect such information as it happens in real time and with greater accuracy than ever before. For example, there are is an increasing number of studies to measure health outcomes over the longer term.

Our project, titled Health and Injury Prediction and Prevention Over Complex Reasoning and Analytic Techniques Integrated on a Cellphone App (HIPPOCRATIC App), requires just such a dataset. The goal of this study is to develop algorithms that enable continuous and real-time assessment of individuals' health by leveraging data that is passively and unobtrusively captured by smartphone sensors. While the potential medical outcomes are positive, the potential privacy outcomes are negative and invasive, so extraordinary care must be taken to protect both the security and privacy of user data throughout the data lifecycle.

To address this, we built a system to allow fully human-out-of-the-loop management of participants including participant recruitment, screening, onboarding, data collection on smartphones, data transmission to the cloud, data security in the cloud, and data access by analysis and modeling teams

Bracken, B., Wolcott, J., Potoczny-Jones, I., Mosser, B., Griffith-Fillipo, I. and Arean, P.
Detection and Remediation of Malicious Actors for Studies Involving Remote Data Collection.
DOI: 10.5220/0010805500003123
In *Proceedings of the 15th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2022) - Volume 5: HEALTHINF*, pages 377-383
ISBN: 978-989-758-552-4; ISSN: 2184-4305

(Bracken et al., 2020). Our approach improves privacy by allowing all stages of recruitment and participation without human access to private or Personally Identifying Information (PII). This requires a human-in-the-loop process for payment processing and support. Study participants were paid in keeping with standards of research equipoise and fairness. Providing incentive payments for completing study activities is a common strategy utilized by researchers to increase retention and engagement with study procedures (Wurst et al., 2020). Our Administration Dashboard allows for anonymized information review of all information required to address participants' concerns including: random unique user IDs (UUID's), surveys completed, incoming messages, and the ability to respond, all while preserving PII anonymity. This includes information such as information on date and amount of gift card delivery. However, since this is a remote study where no human has direct contact with any study participants, the study attracted "malicious actors" who faked upload of data in order to access payments. This is a common problem in research of this nature; methods for identifying bots and malicious actors are needed (Pozzar et al., 2020).

2 STUDY METHOD

Our study involved recruiting participants through social media (e.g., Facebook and Google advertisements). Participants visited a landing page that described the study and what participation entailed. Participants then completed an enrollment questionnaire. If participants were eligible to participate, they proceeded to read the consent form, take a short quiz to ensure they understood consent content, and then electronically sign the form. They were then sent a link to download our smartphone app. The smartphone app collected data from the smartphone sensors (e.g., accelerometer/gyroscope), but did not access any other app data (e.g., visits to social media sites or texts). The app also delivered a baseline survey asking general questions such as demographics, habits of smartphone use, and daily routine information, as well as shorter, twice-daily surveys asking questions about health (e.g., diagnosis with cold or flu), activity (e.g., sleeping patterns), and mood. Participation lasted up to 12 weeks, and participants were paid based on how many surveys they completed. Total potential payment for participants was $90 (in US dollars), with payment amount for the baseline and final (the longer surveys) being the largest, and the remaining payments split across the remaining surveys (twice daily for 12 weeks), increasing gradually throughout the 12 weeks.

Data collection successfully kicked off with recruiting starting March 15th, 2020 and subject onboarding beginning immediately after that. By the end of April, we saw over 3,000 subjects onboarded (driven in part by positive press coverage) and over 60,000 surveys uploaded as well as the smartphone sensor data for the participants. In May, 2020 we continued to monitor the platform usage as the study progressed through the first sixteen weeks of data collection including the completion of the 12 week study by some participants. In July, 2020, we started to observe a significant increase in study participant enrollment that was inconsistent with recruitment activity. We were excited about the numbers, but we also noticed some red flags that led to more investigation. The analysis eventually led to the conclusion that fraudulent participants were attempting to game the study to illegitimately obtain Amazon gift card incentives from the program.

Note that no systems were breached and no data was exposed. Malicious actor activity was limited to automation of fake users in order to receive payments. In addition, throughout the process, we remained in close communication with the University of Washington's (UW's) Institutional Review Board (IRB) about the fraudulent actors and the team's work to respond to that situation. Once observed, we analyzed the traffic and behavior, put models in place to help identify the malicious actors with increasing confidence, and deployed initial mitigation strategies. Over the course of the remainder of the study, we refined the rules used to detect and block these fraudulent users as well as to refine the enrollment and payment processes.

3 MALICIOUS ACTOR DETECTION AND REMEDIATION

3.1 Malicious Actor Detection

Our first indication that we had attracted malicious actors was that although we were not running new ads to drive recruitment and there was no additional press coverage, the daily enrollment numbers were rising rapidly from 50's per day into the 100's per day without an explanation. Figure 1 shows registration events from May 1st through the end of July 2020. The large spike around May 15th was expected due to the

app team having to redeploy an update to the iOS app. The unexpected ramp in registration activity started to become apparent in late June.

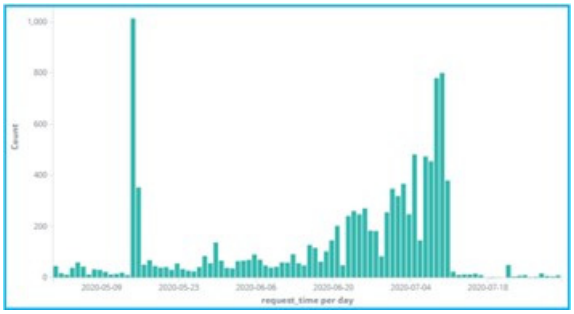

Figure 1: Registration events from May 1st to June 30th 2020.

Second, the number of registered devices (users) that was being reported in our data tracking portal for experimenters (which reflected data directly from the smartphone apps) was much higher than what we were seeing in our data storage platform, TozStore, (which reflected true data upload metrics) – by as much as 3x. This indicated that many of these surveys were faked. Malicious actors were calling endpoints on the data tracking portal to indicate that a survey was uploaded, however no survey was actually uploaded.

Third, we started to see a large increase in the number of recent Android users, which was far too high in comparison to iOS users (initially a ratio of 2:1). Throughout the course of the study these numbers should track relatively closely, with the ratio of iOS to Android devices sold within the country in which participants are recruited. Figure 2 shows iOS vs. Android registration count divergence from the start of data collection. Figure 3 shows iOS vs. Android registration count divergence focusing on June and July.

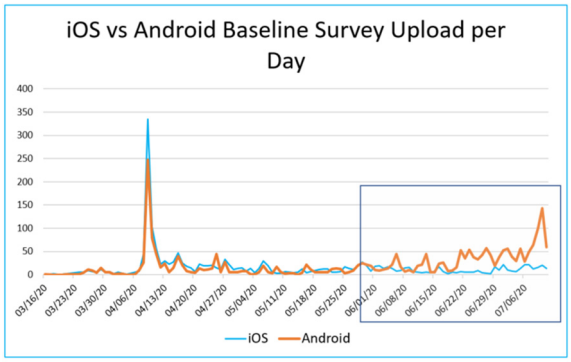

Figure 2: Number of Android and iOS devices registered from start of data collection.

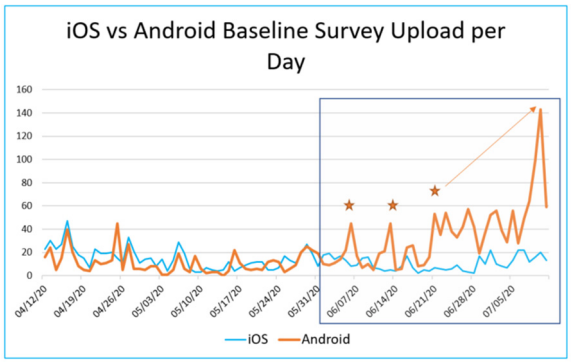

Figure 3: Number of Android and iOS devices registered in June and July of 2020.

Fourth, our baseline survey collected city information, which was a freeform field in the demographics section of the questionnaire. We analyzed this data and saw an inordinate number of participants reporting "Los Angeles" and “Brooklyn” as the city. This led to additional analysis of baseline survey metadata showing scripted responses that were repetitive and not representative of the expected demographics. Further analysis again found that in most cases, malicious actors did NOT upload daily surveys, which gave us confidence that most of the data collected was from legitimate study participants.

Fifth, we saw unusual IP traffic. Because it is very typical for services on the Internet to see traffic from a variety of IP addresses all over the world, and for some of that traffic to be large-scale automated bot traffic, the traffic itself did not raise any red flags. However, once we started the deeper analysis we determined much of the initial malicious actor traffic that was gaming the platform were concentrated in a small number of IP ranges in non-US countries that the study was not advertised in. These IP ranges were displaying automation-like behavior like repeated and fast endpoint access.

Our conclusion was that due to the nature of bad-actor activity, **the large majority of actual survey and sensor data was by legitimate study participants**. Furthermore, we were confident that we would be able to identify and remove the bad data.

Once the malicious-actor activity was identified, our first concern was in halting payments to these actors while continuing to pay the legitimate study participants acting in good faith. We paused payments and implemented a new capability to allow applying an exclusion list when we ran our payment algorithm to pay participants. Once the exclusion list was available, we ran a catch-up payments cycle without paying the malicious actors in the exclusion list.

The first pass of the identification effort was focused on identifying unique payees in order to avoid sending incentives payments to malicious actors. This effort also provided an initial means to identify data that could be distinguished from the legitimate study participants' data that the data analytics teams could use for their analysis.

We reviewed and tested many approaches to detect malicious actors. We looked at baseline survey content, time-in-study and related behavior, existence of surveys and sensor data, registration email domains and formats, types of sensors uploaded, etc. The strongest indicators of malicious actors were in cross-referencing TozStore metadata and the smartphone app and payment data. As mentioned above, the data for this included all users of the system from the start of data collection; though note that this did not rely on the use of PII, i.e., no email addresses, location data, etc. were used. Future work could leverage a no-human-in-the-loop, secure compute approach to review raw GPS sensor data from the devices. This raw GPS data was encrypted and stored in TozStore, but due to its sensitive nature in terms of identifying information, this sensor type was not authorized for access by humans.

From the above efforts, an exclusion ruleset was developed per analysis and observed vs. expected study participation. See Exclusion Ruleset in Table 1. We also performed basic baseline survey analysis, though this was limited since it used information only available in the most recent surveys (e.g., city), but this turned out to be a good sanity check for future approaches to identify malicious actors based on survey responses as the small subset we identified were also flagged by the detection rules. False positives (not paying) are easy to correct whereas false negatives (paying malicious actors) are not. We paid participants using this exclusion list and then worked at refining our ruleset to rule-in some false positives (legitimate study participants). It should be noted we concluded that there were likely multiple malicious actors involved, or the same malicious actors using multiple approaches. We did see clear patterns of behavior from the majority of the identified malicious actors, but there were some behaviors unique to a smaller set of users appearing to try to game the study.

We developed a spreadsheet model to easily apply rules to flag users as "malicious actors." The model allowed for turning rules on and off to create a final exclusion list. The rules we applied to begin payments to participants again are shown in Table 1.

Table 1: Rules first applied.

Rule Description	Notes
Reused devices detected with repeated use of the same UUID.	Assume fraudulent behavior based on reusing devices; may include some valid users if include 2x times (difference = 140 if screen > 3 devices)
No baseline survey uploaded (smartphone app vs data storage database Mismatch): **Malicious actor if smartphone app received confirmation of baseline survey completion, but the baseline survey is not uploaded**	Indication that malicious actors are quickly "re-paving" devices to try again (the user's app indicates the survey was completed, but the malicious actor started over with a new install/registration before the survey was uploaded)
Long delay before registration: **Malicious actor if user signs up after a longer than normal period of time**	Assume they're caching codes
No registration date: **Malicious actor is assumed if user is not registered**	Likely caching registration codes

Table 2 shows rules that were considered, but for which we concluded that more analysis was needed to refine and qualify them further to increase confidence that they were accurate.

Table 2: Rules initially considered, but not applied.

Rule Description	Notes / Why They Were Not Applied
Baseline completed quickly (suspect data): **Malicious actor if completion time for SID1 is < 2 min (should take 5-10 minutes)**	The results are suspect due to very short or very long durations in the metadata – even for automation; more analysis is required
Participation duration check #1 - any uploads: **Malicious actor if user doesn't participate in the study for more than a N days (sans if recent registration)**	Applying this rule will include real people who just dropped after a short period; "N" is parameterized; default=7

Table 2: Rules initially considered, but not applied (cont.).

Rule Description	Notes / Why They Were Not Applied
Participation duration check #2 - surveys (similar to above): **Malicious actor if user submits surveys for less than N days (sans if recent registration)**	Applying this rule will include real people who just dropped after a short period; "N" is parameterized; default=7
Participation duration check #3 - sensor data: **Malicious actor if user loads sensor data for less than N days (sans if recent registration)**	Data is incomplete; also, users could initially turn off sensor collection while still submitting surveys & MFCC; "N" is parameterized; default=7

3.2 Malicious Actor Remediation

Once we identified malicious actors, we paused study recruitment for 45 days while we integrated several mitigation strategies. We then re-started the study, but continued to integrate additional strategies as the study progressed. Our mitigation strategies were as follows.

First, the initial mitigations we deployed were designed based on the initial red flags we saw that alerted us to the malicious actors. We (1) paused incentive payments, (2) modified the enrollment website to pause enrollments, (3) disabled the backend registration endpoints as we saw some malicious actors were bypassing the website to call the end-point directly, and (4) blocked access to all connections from the suspicious IPs outside the regions we advertised in.

Second, we made changes to the smartphone app to mitigate automation of the survey fulfillment and other gaming, including (1) updating the app to detect rooted devices, geo location, and device emulation; (2) detecting and blocking previously used Device IDs, and (3) invalidating unused registration codes.

Third, we made several changes to our payment process including modifying the secure payments processing software to receive an exclusion list of malicious actors to not pay. We performed dry runs to test payment totals with and without the list of rules initially applied (see Table 1).

Fourth, we made several modifications to our study methods (with an university IRB and government Human Research Protection Office (HRPO)-approved modification in place) to prevent future malicious actors from enrolling. (1) We first deployed a CAPTCHA mechanism within the landing and consent webpages to improve automated fraudulent activity deterrence. (2) Participants were required to provide certain information (e.g., zip code, state, height, weight) and allow collection of passive data from accelerometer and gyroscope sensors. These requirements limited the ability of malicious actors to create numerous accounts using a single mobile device and provided more data to inform other mitigation efforts. (3) We modified the payment cycles to run approximately on a monthly basis rather than weekly. This allowed us the time to run an analysis step prior to payments processing in order to refine the exclusion ruleset and update the exclusion list, and to give malicious actors less time to detect our methods and adapt. The exclusion list generation spreadsheet also provides a list of malicious actors that we have leveraged to separate out the good study data from malicious actor data so that we can share them as completely different data sets with the data analysis teams. (4) We also reverted to relying on the data in our primary database (TozStore) rather than the data tracking portal connected to the smartphone app as the source of truth for completed surveys when calculating the payment amounts.

As expected, attempts by fraudulent participants to game the study continued, but the mitigations slowed them down. We monitored the effectiveness of our combined remediation efforts as recruitment and registration efforts ramped up. We continually monitored registration logs as well as CAPTCHA challenge failures. Unexpected rates of either registration or CAPTCHA challenge failures are an indication of malicious actor activity. The CAPTCHA is a deterrence, but it is a statistics game, so we expected some malicious actors to adapt and use means to bypass the challenge (e.g., humans vs. bots). The alerts were deployed to prompt analysis and expansion of the IP/VPN blocklist. This is an effective means to slow down fraudulent registrations while the malicious actors spin up new VPNs.

We implemented an Access Control List (ACL) mechanism using a static list of CIDR blocks (IP address ranges) that are known to originate outside the country of interest (in our case, outside of the United States). This approach allows for adding access for specific countries if the program wants to expand outside the US (e.g., Canada, Australia, New Zealand, etc.). We noticed that within a few hours of enacting IP address blocks, the malicious actor traffic transitioned to VPNs in the US. This ACL mechanism was then also used to block all VPNs. We investigated methods and services for automated detection of VPNs, but as we find additional malicious actor IP addresses and VPNs, new ones could be easily added to the list.

4 MIS-IDENTIFIED MALICIOUS ACTORS

There were several cases in which our exclusion rules mis-characterized a legitimate study participant as a malicious actor. For example, one exclusion rule triggered when no smartphone data were uploaded to the database, although survey data was uploaded. However, there were instances in which the smartphone app malfunctioned and did not upload sensor data for legitimate study participants.

One function of the human-out-of-the-loop participant handling approach that we developed, but that is outside the scope of this paper (Bracken et al., 2020) is a portal through which experimenter teams can communicate with participants. The experimenter sees only the random ID assigned to the participant, but the participant receives communication within the study application's chat feature and/or emails through the email address they signed up for the study with (mapping between the two occurs in the cloud out of reach of the human experimenters). Through this portal using anonymous communication and case-by-case analysis of user participant activity information, we identified many of the mis-labelled participants who we then re-characterized as good participants after email exchanges. Catch up runs of incentive payments were performed for these users and their data was reclassified as good for use by analysis teams.

5 CONCLUSIONS

We built a system to allow fully human-out-of-the-loop management of patients including patient recruitment, screening, onboarding, data collection on smartphones, data transmission to the cloud, data security in the cloud, and data access by analysis and modeling teams. However, since no human has direct contact with any study participants, the study attracted "malicious actors" who faked upload of data in order to access payments. We identified and put into place mechanisms to block malicious actors. As expected, attempts by fraudulent participants to game the study continued, but the mitigations slowed them down.

However, we believe that this work to identify and prevent malicious actors has had several positive results. First, the lessons learned here can result in improvement of future remotely conducted studies by integrating these malicious actor mitigation strategies from study initiation.

Second, it improved the study outlined here. It caused us to closely monitor study data, which has led to higher confidence results. It has improved dataset quality for the data analysis teams, and reduced the burden of dataset cleanup. The process has identified data integrity and upload issues that otherwise would have been missed until late in the data collection process. These would not have been found until data analysis teams were deeper into their analysis. In addition, malicious actor identification and early analysis of profiles has led to improved quality assurance of the smartphone app used in the study.

In future studies, we will also explore integration of additional strategies not used in this study. We can use data that was deemed too sensitive for humans to access (e.g., email addresses, IP addresses, GPS location) to identify potential malicious actors. This can be done without humans accessing the data as we have now developed a tool for humans to apply analysis techniques to data that may be identifiable that is stored in the cloud, then pull down the results of the analysis that are not identifiable. For example, a researcher can write code that will access and search for matching IP addresses, then only see the randomly assigned participant IDs that have matching IP addresses.

ACKNOWLEDGEMENTS

This material is based upon work supported by United States Air Force and DARPA under Contract No. FA8750-18-C-0056 entitled Health and Injury Prediction and Prevention Over Complex Reasoning and Analytic Techniques Integrated on a Cellphone App (HIPPOCRATIC App). The views, opinions and/or findings expressed are those of the author and should not be interpreted as representing the official views or policies of the DoD or the U.S. Government.

REFERENCES

Bracken, B.K., Potoczny-Jones, I., Wolcott, J., Raffaele, E., Woodward, L., Gogoel, C., Kiourtis, N., Schulte, B., Arean, P.A., and Farry, M. Development of Human-Out-of-the-Loop Participant Recruitment, Data Collection, Data Handling, and Participant Management System. *Proceedings of the Annual International Human Factors and Ergonomics Society, October 5-10, 2020.*

Pozzar, R., Hammer, M. J., Underhill-Blazey, M., Wright, A. A., Tulsky, J. A., Hong, F., Gundersen, D. A., & Berry, D. L. (2020). Threats of bots and other bad actors to data quality following research participant

recruitment through social media: Cross-sectional questionnaire. *Journal of Medical Internet Research*, *22*(10), e23021.

Wurst, R., Maliezefski, A., Ramsenthaler, C., Brame, J., & Fuchs, R. (2020). Effects of Incentives on Adherence to a Web-Based Intervention Promoting Physical Activity: Naturalistic Study. *Journal of Medical Internet Research*, *22*(7), e18338.