Smoking Data Report and Summary

Jason Ma

December 2018

1 Introduction

Before we dive into modeling and doing inference on the smoking data, it is critical that we understand the internal validity of the datasets and how they are related to each other. Doing so helps us design and adapt our model and gauge if assumptions we make are reasonable.

The main questions we are trying to answer are: (1) if each dataset makes sense in their own regard, and (2) if the datasets are consistent with one another. Both criteria must be met reasonably for any inferences to be made. This report summarizes our analysis of this dataset in addressing these questions. In particular, we aim to first introduce and provide high level explanation of what each dataset contains and what kind of patterns within each dataset and between them we should observe, then explain our analysis approach, and finally end with our findings.

2 The Smoking Data Set

We first provide an overview of the 3 data sources and the 4 datasets considered. There are 37 participants in the study, labeled as 201,202,..., 237. Due to some technical issues, the data are collected differently for participants 223-227, resulting in a different source. For the other participants, there are also two data sources: the original from the cloud, and the backup from the phone. In summary, the three data sources arelabeled as:

- Original. This source contains data from participants 201-222, 228-237.
- Backup. This source contains the backup data from participants 201-222, 228-237.
- Alternative. This source contains data from participants 223-227.

Within each data source, for each participant, there are 5 different datasets we are considering: puff-episode,puff-probability, random-ema, contingent-ema, and end-of-day-ema(eod-ema). Each of them is a time-series data collected over a span of 10 days, but they differ in what they measure and how they measure. Here is a brief description of each of them:

1. Puff-episode
    - Each event record(line of the data set) includes the participant ID, the timestamp of the smoking event that was detected.
2. Puff-probability.

- Each event record includes the participant ID, the timestamp of the Hands-to-Mouth(HTM) motion detected, and the estimated probability that this HTM event corresponds to a smoking event. Note that

3. Contingent-EMA

   - These data are **self-reported** by participants when they smoked.
   - Each event record includes the participant ID, the timestamp of the self-report, and the approximate time since the smoking event when the self-report was filled out, and some other qualitative quetions that assess the participant's emotional states, which have been associated with smoking.
   - Key question(s): "Approximately how long ago did you take the first puff from that cigarette?"
     (a) less than 5 minutes
     (b) 5-15 minutes
     (c) 15-30 minutes
     (d) more than 30 minutes

4. Random-EMA

   - These data are **self-reported** by participants at a random time in three separate time windows each day. If the participant misses the first one in a time window, a second one will be sent out at random time in the remaining duration of that window. A third one will be sent, if the second one is missed again. Then, no further random-ema will be sent out in that time window, regardless if the third one was filled out or not.
   - This ema asks similar questions to the contingent-ema
   - Key question(s):
     – "Did you puff since the last report?" (self-report of any type)
     – "Approximately how long ago did you take the first puff from that cigarette?"
       (a) 1-19 minutes
       (b) 20-39 minutes
       (c) 40-59 minutes
       (d) 60-79 minutes
       (e) 80-100 minutes
       (f) more than 100 minutes

5. End-of-Day-EMA

   - These data are **self-reported** by participants at the end of each day.
   - Participants are asked to check boxes representing the time windows when they smoked. They are also asked to answer some qualitative questions.
   - Key question(s):
     – "Please check the hour blocks when you took 1 or more puffs today."
       * 8am-9am, 9am-10am, ..., 7pm-8pm.

3 Are Original and Backup Consistent?

The obvious first question we want to ask is if the original and the backup data for participants 201-222,228-237 are consistent with each other? If not, then we would not know which data source to trust and how much bias our results are incurring as a result of using one over the other.

We ran multiple Python scripts to analyze the difference among the original and the alternative versions of our datasets. The following paragraphs explain our findings.

We first wanted to check if there are any participants that do not have backup data. We found that the following list of participants is not contained in the back-up data folder: 201, 203, 206, 210, 221, 229. Therefore, our only source of data for these participants comes from the original cloud data.

Then, the next natural question to ask is how different are the original and the backup data for the participants that do have both? We found that the result varies across the datasets, but in general the inconsistencies are tolerable, and we should use the backup data for participants who do have them because the backup data are in fact a super set of the original data for most cases.

3.1 Puff-Probability

In this section, we compare "puff-probability.csv" and "puff-probability-backup.csv" file. The original csv file contains 10617 rows, while the latter contains 10979 rows. There are many inconsistencies between the data sources. In aggregate, there are 840 entries in the backup data but not the original data; conversely, there are 2112 entries in the original data that are not in the backup data. A more detailed version of difference across sources by participants can be found in the corresponding .ipynb file in the github directory.

3.2 End of Day EMA

In this section, we compare "eod-ema.csv" and "eod-ema-backup.csv". The original csv file contains 202 rows, while the latter contains 195 rows. First, we note that participants **202,212** do not have this data in either source. In aggregate, there are **16** entries that are in the backup data but not the original data; conversely, there are **28** entries in the original data that are not in the backup data. Again, we found that for participants that are present in both sources, the backup data is a **strict superset** of the original data. In other word, the 28 extra rows in the original data come entirely from participants who are not present in the backup data. This means that we can safely just take the backup data for the participants who have them and the original data for the participants who don't.

3.3 Random EMA

In this section, we compare "random-ema.csv" and "random-ema-backup.csv". The original csv file contains 691 rows, while the latter contains 681 rows. In aggregate, there are **52** entries in the backup data but not the original data; conversely, there are **63** entries in the original data that are not in the backup data. Again, we found that for participants that are present in both sources, the backup data is a **strict superset** of the original data. A more detailed version of difference across sources by participants can be found in the corresponding .ipynb file in the github directory.

3.4 Event Contingent EMA

In this section, we compare "eventcontingent-ema.csv" and "eventcontingent-ema-backup.csv" file. The original csv file contains 180 rows, while the latter contains 185 rows. Participants **203,206,210,232,236** are missing from both data sources. This could either be just that there weren't event contingent situations that occurred to them, they felt too embarassed to respond, or an actual technical issue. In aggregate, there are **25** entries in the backup data but not the original data; conversely, there are **20** entries in the original data that are not in the backup data. Again, we found that for participants that are present in both sources, the backup data is a **strict superset** of the original data. A more detailed version of difference across sources by participants can be found in the corresponding .ipynb file in the github directory.

3.5 Summary

In summary, I believe that the inconsistencies are noticable yet tolerable. For participants who do not have backup data, we would just use the original data. For those who do, we would use the backup data because they are super sets of the original data in most cases.

On the other hand, I would like to point that there are a few participants whose data are inconsistent across many and in some cases all data streams: **207,208, 213**. Maybe it's worth asking the researchers at Memphis about these participants' data specifically.

4 Are the Data Sets Consistent with Each Other?

Now that we know that the datasets are mostly self-consistent, we move on to assess whether they are consistent with each other. For example, one question we might want to know is if a participant indicated that he/she smoked in the EOD-ema, are there any other measurements(contingent-ema, random-ema, or puff-probability) that corroborate this indication?

4.1 How reliable are the EOD-emas?

The example question above is an interesting question that's of both scientific and statistical interest. For us, we are interested in knowing how reliable are people's memory of their smoking behaviors reported in EOD-emas and if the lack of reporting true signals in the EOD-emas are indicative of smoking relapses or other events?

To investigate, we ran a Python script to compute the percentage of signals(indications of smoking events) reported in other measurements that were missed by EOD-ema. We computed the percentages for each pair of measurements in two ways: one by percentage of events missed, one by percentage of days missed. The second metric is considered because the first one is potentially inflated if there are multiple signals in a day but no signal in the EOD-ema for the same day. The following tables summarize our results.

	Original	Alternative	Back-up
% by entry	15.6	13.11	11.4
% by date	10	11.55	7.57

Table 1: Contingent vs. EOD EMA

	Original	Alternative	Back-up
% by entry	20.7	23.7	17.1
% by date	15.9	15.8	13.3

Table 2: Random vs. EOD EMA

	Original	Alternative	Back-up
% by entry	29.8	32.7	29.4
% by date	12.1	12.7	12.9

Table 3: Puff-Episode vs. EOD EMA

Now, we summarize our findings from the tables above. We see that the backup data are uniformly more consistent than the original data for all datasets. This is consistent with our result from the above section that the backup data are more complete for the participants included in them. We also notice that contingent-emas are more consistent than other measurements. This makes intuitive sense because the contingent-emas themselves are a more accurate measurement of smoking events as participants are supposed to fill it out as soon as they relapsed. Finally, we notice that the gap bewteen % by entry and % by date for puff-episode is much larger than the other two measurements. One logical explanation of this trend is that participants who relapsed into excessive smoking are also less likely to report their relapses in the EOD-emas. Nevertheless, the error rates are inflated for puff-episode because not all "smoking events" detected are actually smoking events.

In general, we see that the error percentage by date floats at around 15%. We claim that this is reasonable and tolerable for our purposes. After all, our models also try to account for the fact that people do not have perfect memory and existence of missing data, both at random and not at random.

4.2 Is Puff-probability useful at all?

Puff-probability datasets track all Hands-to-Mouth(HTM) motions for the participants. We understand that not all HTM motions correspond to actual smoking events, but at least some overlaps between events in puff-probability and signals in other measurements should be observed. In particular, we check for consistency between puff-probability with contingent-ema and EOD-ema. Checking for overlaps allows us to better understand two very related questions:

1. Does puff-probability do a decent job covering actual smoking events assuming that participants are honest and responsive.

2. How honest and responsive are the participants given that the puff-probability do cover actual smoking events?

A high average overlap between puff-probability and other measurements helps us answering the first question positively. On the other hand, a low variability in the overlaps between puff-probability and other measurements helps us answering the second question positively.

Our first attempt checks that for each signal(smoking event) recorded in self-reports if there is a HTM motion recorded within an hour plus-minus of the timestamp of the signal. The numbers reported are coverage percentage, which is defined as the total number of signals that have at least

one HTM event associated(occurred within one hour limit) divided by the total number of signals. Here is our result:

	Original	Alternative	Back-up
contingent-ema	89.4		
eod-ema	82.5	78.9	84.5

Table 4: HTM Coverage Table

Both coverage are above 80%, from which we assert that HTM event does a decent job covering true signals. In particular, the coverage rate for contingent-ema is as high as 89%, which is expected because the contingent-ema is a more accurate indication of true signals.

Our first attempt gave us results we were hoping, but the one hour plus-minus over the timestamp of the signals is too conservative in the case of contingent-ema. We would like to know how well HTM motions cover contingent-ema if we decrease the time limit to some smaller values such as 30 minutes or even just 5 minutes. This analysis also helps us on estimating the variance in people's ability to remember when they smoked. If people have perfect memory or always repond when they smoke and do so as soon as the relapse occurs, then the percentage should not go down at all. Here is our result:

Time limit	coverage %	average of HTMs per smoking event reported
60	89.4	4.06
30	83.8	3.39
15	75.0	2.87
5	50.0	1.80

Table 5: Contingent-ema vs. HTM sensitivity table

This result strengths our belief that puff-probability is in fact informative of smoking events. Even by decreasing the time limit to just 15 minutes, the coverage percentage is 75%. The decline in the coverage percentage also confirms our belief that people don't really know when they smoked precisely and that they were poor in filling out the contingent-ema when relapses occurred; only half of the time were participants able to do so within 5 minutes.

5 Summary

Through various numerical analysis of the datasets given, we conclude that that the datasets are self-consistent and there are associations among them. By using them together, we hope to construct an accurate predictive model that is able to infer smoking events in the future. Note that we do not conclude any statistical significance of our results as most of our analysis are purely numerical and did not utilize any statistical tests or rely on any statistical assumptions. Nevertheless, we believe that our procedures are enough to validate the datasets and then move on to the modeling phase of this project.