The Por Nuestra Salud Study: Documentation on implemented study design and decisions made in data curation

Jamie Yap, Lindsey N. Potter, Brian Orleans, Stephanie M. Carpenter,
Inbal Nahum-Shani, David W. Wetter, Cho Y. Lam

August 17, 2021

1. About the Study

The Por Nuestra Salud (PNS; P60MD000503, PI: Wetter) study examined intrapersonal and contextual determinants of smoking cessation among Spanish-speaking Mexican American smokers attempting to quit smoking.

2. Data Curation Strategy

Our team has created open source documents and code provided in a GitHub repository https://github.com/jamieyap/PNS to (a) provide a detailed description of key assumptions and decisions made in data preprocessing and preparation implemented after the conduct of the study; (b) offer a workflow that can be reproduced by others; (c) provide end-users with curated datasets that can be traced directly back to the raw datasets used in constructing them, similar to the concept of *farm-to-table* traceability. Below, we highlight salient aspects of our approach to data curation, developed to improve transparency when working with mobile health (mHealth) data.

We viewed 'data curation' as a process that imposes interpretations upon the data and assigns relative value of one aspect of the raw data over another; an outcome of the process is a collection of 'curated datasets' – namely, datasets which capture the scientifically and practically informed sequence of transformations to the raw data most end-users should begin with in their work.

1. We created inter-linkable curated datasets that can be linked through a unique participant identifier and a unique EMA questionnaire identifier. This will allow end-users to trace a participant's story across all curated datasets. End-users will be able to determine the temporal relationship between observations within and across various data sources.

2. We used a snapshot-and-restore (S&R) strategy (i.e., see https://environments.rstudio.com/snapshot.html) to record software dependencies. Curation of the PNS study data is performed in R, a popular open-source software for data science, and its open-source packages. By employing a S&R strategy in tandem with open-sourced data curation code, end-users are more likely to reproduce the steps taken by project developers.

3. We focused on providing a conceptual description of decisions made during data curation as they impact the end-user's ability to do science with the PNS study data. Where appropriate, we emphasize principles end-users can apply in their own work.

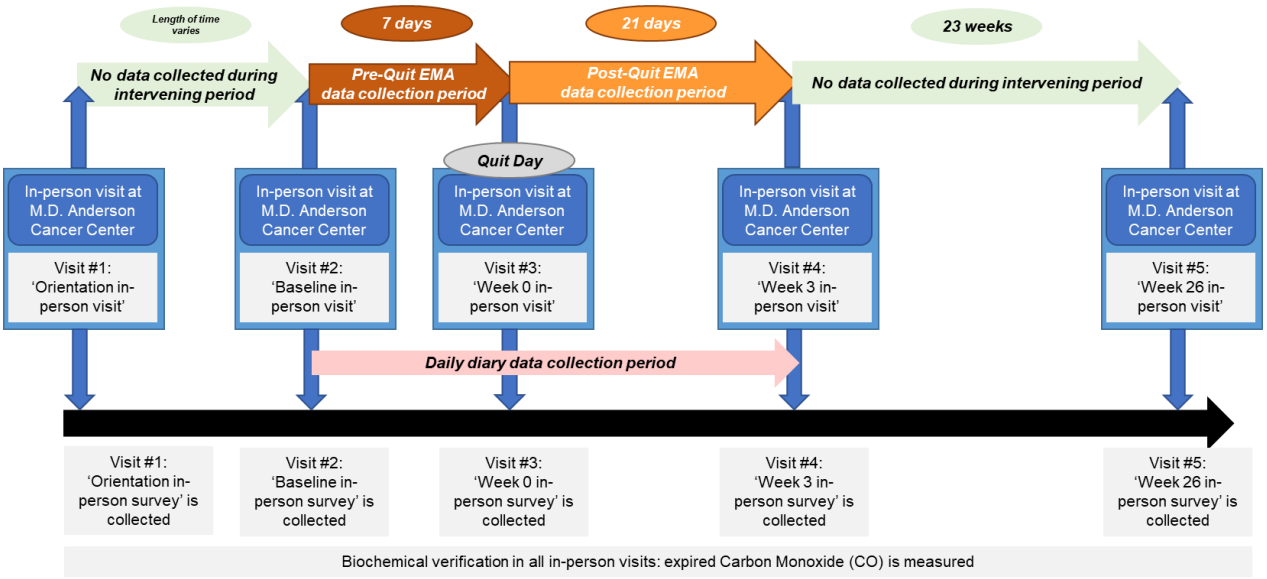
3. Study Design

Timeline of EMA data collection vis-a-vis in-person clinic visits

Five in-person visits at the M.D. Anderson Cancer Center occurred throughout the study, during which, a participant would complete an in-person survey as well as a battery of biochemical assessments (e.g., biochemical verification of expired Carbon Monoxide, among other biochemical assessments). The time frame of these in-person clinic visits with respect to various data collected during the conduct of the study is illustrated in Figure 1.

The study was designed so that participants completed up to four Ecological Momentary Assessment (EMA) questionnaires per day on their smartphone during a 28-day period: 7 days pre-quit (time frame between Visit#2 and Visit#3) and 21 days post-quit (time frame between Visit#3 and Visit#4). Additionally, participants also completed a daily diary survey on their smartphone at the end of each day during the same 28-day period.

Figure 1: Illustration of data collection time frame



During Visit#2 (Baseline in-person clinic visit)

A smart phone was provided to study participants during Visit #2. Together with study staff, each participant determines a date when they intended to quit smoking (in this documentation, we refer to this date as the participant's **'Projected Quit Date'**). Study staff then keyed-in the Projected Quit Date into the smart phone's EMA software, which is designed to switch from a **'Pre-Quit Mode'** to a **'Post-Quit Mode'** on 12AM on this date.

Study modes and assessment types

Both the Pre-Quit Mode and the Post-Quit Mode correspond to particular kinds of EMA questionnaires that were made available to participants.

Broadly, an EMA questionnaire could not only be launched at randomly selected times, but also after the participant indicated via a *button press* (i.e., the participant taps on their smartphone's

user interface) that they experienced a smoking-related event. When performing a button press (i.e., taps on the smartphone's user interface), the participant indicates whether the button press is associated with an intent to smoke, a smoking urge, or a recollection of cigarettes they had smoked in the past but had not yet reported in any of the EMA questionnaires they had completed thus far.

Participants were trained to recognize such events (e.g., a smoking urge) during Visit#2 (Baseline in-person clinic visit) and Visit #3 (Week 0 in-person visit). Table 1 lists the various kinds of EMA questionnaires available in each mode. The type of EMA questionnaire is captured in the **assessment_type** variable in the curated datasets.

Table 1 details the types of Self-Initiated EMA Questionnaires, as well as whether a button press is followed by an EMA Questionnaire. More specifically, following a Pre-Quit Smoking Part One and Post-Quit About to Slip Part One button press, the smartphone's EMA software is designed to launch a questionnaire following 20% - 100% of button presses (i.e., not all button presses are followed by a survey). The sampling rate varies participant-by-participant and depends upon individual's self-reported cigarettes smoked in previous EMA questionnaires leading up to the button press; unfortunately, more detailed information on the calculation of this sampling rate is not available. In contrast, the smartphone's EMA software is designed to always launch (i.e., 100% of the time) Pre-Quit Urge, Post-Quit Urge, and Post-Quit Already Slipped EMA questionnaires after the relevant button press is performed.

More specifically, the smartphone's EMA software is designed to selectively launch Pre-Quit Smoking Part One and Post-Quit About to Slip Part One EMA Questionnaires after a button press indicating an intent to smoke: the smartphone's EMA software is designed to launch these two types of EMA questionnaires only among 20% - 100% of total button presses associated with an intent to smoke. The specific sampling rate varies participant-by-participant and depends upon individual's self-reported cigarettes smoked in previous EMA questionnaires leading up to the button press. In contrast, the smartphone's EMA software is designed to always launch (i.e., 100% of the time) Pre-Quit Urge, Post-Quit Urge, and Post-Quit Already Slipped EMA questionnaires after the relevant button press is performed.

After Visit#2 (Baseline in-person clinic visit)

Of note, after Visit#2 (Baseline in-person clinic visit), participants had the ability to adjust their Projected Quit Date to be earlier or later than originally planned. If participants adjusted their Quit Date, study staff keyed-in the updated Quit Date into the smartphone's EMA software so that the switch from Pre-Quit Mode to Post-Quit Mode took effect on the **'Adjusted Quit Date'** rather than the original Projected Quit Date.

If there was an adjustment to the Projected Quit Date was made, then the smartphone's EMA software is designed to deliver Pre-Quit Mode EMAs prior to the Adjusted Quit Date and Post-Quit Mode EMAs after the Adjusted Quit Date. On the other hand, if there was no adjustment to the Projected Quit Date was made, then the smartphone's EMA software is designed to deliver Pre-Quit Mode EMAs prior to the Projected Quit Date and Post-Quit Mode EMAs after the Projected Quit Date.

Scope of documentation

This documentation focuses on curation of responses to EMA questionnaires from both pre- and post- quit periods; other data collected during the conduct of the study are beyond the scope of this documentation.

4. The Data Collected: An Overview

Content of EMAs: An Overview

EMA questionnaires contained items that asked a varied amount of information related to the quantity and timing of smoking events. Table 2 & Table 3 display specific information about the type of smoking questions asked in each type of questionnaire.

The first column displays the type of EMA questionnaire while the remaining columns display the type of information captured by each variable. The information is organized into groups based on whether the item(s) asked about quantity or timing of smoking. We direct readers of this documentation to the file **PNS EMA Codebook 07202010.docx** for a detailed description of items in each kind of EMA questionnaire and their possible responses.

The groupings of variables in both tables were not recorded as such in the raw data, but rather is a *scientifically-grounded preliminary step* performed during the data curation process to enable further data pre-processing.

Note: When pulling together information on one particular construct of interest (e.g., urge to smoke) from various kinds of EMA questionnaires, view Table 2 and Table 3 as a guiding template on how the *preliminary step* is performed.

EMA questionnaires also contained items that asked about mood, context, attention, urge to smoke, etc. Not all types of EMA questionnaires were identical, so end users should refer to the codebook to examine exactly what constructs were asked in each type of EMA questionnaire.

Additionally, we highlight that items within an EMA questionnaire can differed in terms of whether items were framed in the present moment, or in terms of a specific moment in the past, as displayed in Table 4.

Structure of raw data from EMA questionnaires: An Overview

Raw data from EMA questionnaires are stored in nine separate csv files, one for each type of EMA questionnaire described in Table 1. These nine files are structured in a similar manner.

Major features of the structure of the raw data as they relate to how they impact data curation are described next.

1. **Is there ground truth on each participant's Quit Date?** No. We noted that during the Visit#2 (Baseline in-person clinic visit), participants determined a date when they intended to quit smoking (i.e., the participant's 'Projected Quit Date'), but that participants had the ability to adjust their 'Projected Quit Date' to be *earlier* or *later* than originally planned (i.e., the participant's 'Adjusted Quit Date'). Thus, participant's *true Quit Date* is either of the following:

- The participant's 'Projected Quit Date' if no adjustment after the first in-person clinic visit was made
- The participant's 'Adjusted Quit Date' if an adjustment (to a date either earlier or later than the 'Projected Quit Date') after the first in-person clinic visit was made

However, we do not have ground truth on these true Quit Dates. Instead, we have multiple plausible but potentially conflicting date records from which these true Quit Dates may be inferred. One such potential source of information are timestamps recorded in the EMAs themselves, among other sources of information.

Hence, we developed a process to infer each participant's true Quit Date; this process is described in Section 8. From here onward, we will refer to these inferred Quit Dates as **'Working Quit Dates'** to emphasize the presence of uncertainty; these will be viewed as a participant's Quit Date for the purpose of data analysis. Working Quit Date will be captured by the variable **quit_hrts** in the curated datasets. In Section 10, we provide more detail about these timestamps.

2. **Are partially completed/ignored EMA questionnaires also recorded in the raw data?** Yes. Raw data from EMA questionnaires are in *long format*, where a given participant's responses are stored in multiple rows, one for each EMA questionnaire successfully launched to a participant. The raw data contains records not only of completed EMA questionnaires, but also of EMA questionnaires that were partially completed or ignored. In the latter cases, items within an EMA questionnaire having no recorded responses are represented as blanks/no-spaces (i.e., as "") in the raw data.

3. **When a participant performs a button press immediately before an intention to smoke (see Table 1), is this event represented in the raw data, even if no EMA questionnaire was launched?** With *high likelihood*, when a participant indicates an *intention to smoke* via a button press, this event is recorded in the same raw data file containing completed/partially completed/ignored EMA questionnaires, regardless of whether any EMA questionnaire was launched (specifically, Pre-Quit Smoking Part One EMA or Post-Quit About to Slip Part One EMA).

   However, interpretation of records in the raw data is not without any ambiguity. In order to correctly attribute each row in the raw data to the actual events that led to such a record, we developed a process to tease apart not only records due to events described in #2 and #3 above, but also records due to the smartphone's EMA software not performing as designed (i.e., a software bug).

   This process is described in Section 6, and would, for example, allow us to tease apart which rows *should* be attributed to a participant performing a button press immediately before an intention to smoke but no EMA was launched from those rows which *should* be attributed to other events.

5. Participants to Exclude from all Analyses of Curated Data

A total of 200 participants were enrolled into the PNS study, of which 37 (18.5%) participants will be excluded from any analysis of the curated datasets, leaving at most 163 participants.

The rationale for excluding the 37 participants is as follows:

- No EMA questionnaires (any type) were launched in Post-Quit Mode (32 participants)
- Information on-hand is considered insufficient to infer the individual's pattern of smoking (5 participants). This determination is based on a visual inspection of how often and when EMA questionnaires were launched for each participant; Section 8 provides details of the method of constructing such plots that guided this decision.

Hence, any information from these 37 participants will be excluded from the curated datasets.

In the event that an end-user might see the need to inspect raw data concerning these 37 participants, they can do so using a curated dataset named **quit_dates_final.csv** which is created to enable such an investigation on these 37 participants. This dataset contains unique identifiers for all 200 participants, as well as a binary variable '**exclude'** that indicates whether a participant should be excluded from the curated datasets ('exclude' is equal to 1 if a participant is to be excluded from all analysis of curated datasets [N=37], and 0 otherwise [N=163]).

We note that, since none of the other curated datasets contains any information on the 37 participants, the variable 'exclude' was omitted from all other curated datasets (since all 37 participants would have a value of 1 for the variable 'exclude').

From here onward, this documentation will only apply to the 163 participants unless otherwise stated.

Table 1

Types of Self-Initiated EMA Questionnaires		When does participant initiate?
Pre-Quit Mode	**Post-Quit Mode**	
Pre-Quit Smoking Part One EMA	Post-Quit About to Slip Part One EMA	Immediately before an intention to smoke, participants will indicate (via a 'button press') that they intend to smoke using their smartphone. Subsequently, the smartphone's EMA software is designed to randomly sample which of these button presses will be followed by an EMA questionnaire. The sampling rate ranging between 20% and 100% is based on an individual's self-reported cigarettes smoked in previous EMA questionnaires. If one such occasion is selected, the smartphone's EMA software will deliver these types of EMA questionnaires.
Pre-Quit Smoking Part Two EMA	Post-Quit About to Slip Part Two EMA	Approximately 10 minutes after a Part One EMA is launched, the smartphone's EMA software automatically launches a Part Two EMA. Participants do not need to initiate this switch. *Although the smartphone EMA software launches Pre-Quit Smoking EMA Part Two & Post-Quit About to Slip Part Two EMA without needing any input from the participant, these two types of EMAs will still be viewed in this documentation as self-initiated EMAs.*
Pre-Quit Urge EMA	Post-Quit Urge EMA	Immediately after having felt the urge to smoke, participants will indicate (via a 'button press') that that they felt an urge using their smartphone. We note that the smartphone's EMA software was designed to launch these types of EMA questionnaires in 100% of these occasions (unlike the Pre-Quit Smoking Part One and Post-Quit About to Slip Part One EMA types).
—	Post-Quit Already Slipped <u>Note:</u> Post-Quit Already Slipped EMAs do not have a counterpart in "Pre-Quit Mode"	Immediately after the participant had a recollection of any cigarette they had smoked <u>in the past</u> but had not yet reported in any of the EMA questionnaires they had completed thus far. We note that the smartphone's EMA software was designed to launch this EMA type in 100% of these occasions (unlike the Pre-Quit Smoking Part One and Post-Quit About to Slip Part One EMA types).

Other Types of EMA Questionnaires		When is participant prompted?
Pre-Quit Mode	**Post-Quit Mode**	
Pre-Quit Random EMA	Post-Quit Random EMA	At randomly selected times within pre-specified time windows, the smartphone's EMA software will launch a Pre-Quit Random EMA or Post-Quit Random EMA.

Table 2

	Group 1	Group 2	Group 3
	Smoking quantity variables		**Timing of smoking events variables**
Types of Pre-Quit Mode EMA Questionnaires			
Random	**PreQRSmoking1** Since the last computer recording, have you smoked any cigarettes that you did not record in the computer? (*Possible Responses:* Yes, No)	**Smoking2_PreQ_Random** How many cigarettes did you smoke that you did not record? (*Possible Responses:* 0, less than 1, 1-2, 3-4, 5-6, 7-8, 9-10, more than 10)	**Smoking3** How long ago did you smoke the most recent cigarette that you did not record? (*Possible Responses:* 0-15 minutes, 16-30 minutes, 31-45 minutes, 46 minutes-1 hour, 1 hour and 1 minute – 1 hour and 15 minutes, 1 hour and 16 minutes – 1 hour and 30 minutes, 1 hour and 31 minutes – 1 hour and 45 minutes, 1 hour and 46 minutes – 2 hours, more than 2 hours)
Urge	**SmPQU1** Since the last computer recording, have you smoked any cigarettes that you did not record in the computer? (*Possible Responses:* Yes, No)	**Smoking2_PreQ_Urge** How many cigarettes did you smoke that you did not record? (*Possible Responses:* 0, less than 1, 1-2, 3-4, 5-6, 7-8, 9-10, more than 10)	**Smoking3** How long did you smoke the most recent cigarette that you did not record? (*Possible Responses:* 0-15 minutes, 16-30 minutes, 31-45 minutes, 46 minutes-1 hour, 1 hour and 1 minute – 1 hour and 15 minutes, 1 hour and 16 minutes – 1 hour and 30 minutes, 1 hour and 31 minutes – 1 hour and 45 minutes, 1 hour and 46 minutes – 2 hours, more than 2 hours)
Smoking Part One	*No item in the EMA questionnaire that can be mapped to this variable*	*No item in the EMA questionnaire that can be mapped to this variable*	*No item in the EMA questionnaire that can be mapped to this variable*
Smoking Part Two	*No item in the EMA questionnaire that can be mapped to this variable*	**CigJustNow** How many cigarettes did you just smoke? (*Possible Responses:* 0, less than 1, 1-2, 3-4, 5-6, 7-8, 9-10, more than 10)	*No item in the EMA questionnaire that can be mapped to this variable*

Table 3

	Group 1	Group 2	Group 3
	Smoking quantity variables		**Timing of smoking events variables**
Types of Post-Quit Mode EMA Questionnaires			
Random	**PostQRSmoking1** Since the last computer recording, have you smoked any cigarettes that you did not record in the computer? (*Possible Responses:* Yes, No)	**Smoking2_PostQ_Random** How many cigarettes did you smoke that you did not record? (*Possible Responses:* 0, less than 1, 1-2, 3-4, 5-6, 7-8, 9-10, more than 10)	**Smoking3** How long ago did you smoke the most recent cigarette that you did not record? (*Possible Responses:* 0-15 minutes, 16-30 minutes, 31-45 minutes, 46 minutes-1 hour, 1 hour and 1 minute – 1 hour and 15 minutes, 1 hour and 16 minutes – 1 hour and 30 minutes, 1 hour and 31 minutes – 1 hour and 45 minutes, 1 hour and 46 minutes – 2 hours, more than 2 hours)
Urge	**SmPostQU1** Since the last computer recording, have you smoked any cigarettes that you did not record in the computer? (*Possible Responses:* Yes, No)	**Smoking2_PostQ_Urge** How many cigarettes did you smoke that you did not record? (*Possible Responses:* 0, less than 1, 1-2, 3-4, 5-6, 7-8, 9-10, more than 10)	**Smoking3** How long ago did you smoke the most recent cigarette that you did not record? (*Possible Responses:* 0-15 minutes, 16-30 minutes, 31-45 minutes, 46 minutes-1 hour, 1 hour and 1 minute – 1 hour and 15 minutes, 1 hour and 16 minutes – 1 hour and 30 minutes, 1 hour and 31 minutes – 1 hour and 45 minutes, 1 hour and 46 minutes – 2 hours, more than 2 hours)
About to Slip Part One	*No item in the EMA questionnaire that can be mapped to this variable*	*No item in the EMA questionnaire that can be mapped to this variable*	*No item in the EMA questionnaire that can be mapped to this variable*
About to Slip Part Two	*No item in the EMA questionnaire that can be mapped to this variable*	**CigJustNow_PostQ_Slip2** How many cigarettes did you just smoke? (*Possible Responses:* 0, less than 1, 1-2, 3-4, 5-6, 7-8, 9-10, more than 10)	*No item in the EMA questionnaire that can be mapped to this variable*
Already Slipped	*No item in the EMA questionnaire that can be mapped to this variable*	**HowManyCig** How many cigarettes did you smoke during this slip? (*Possible Responses:* 0, less than 1, 1-2, 3-4, 5-6, 7-8, 9-10, more than 10)	**LastCig** How long ago did you smoke the last cigarette? (*Possible Responses:* 0-15 minutes, 16-30 minutes, 31-45 minutes, 46 minutes-1 hour, 1 hour and 1 minute – 1 hour and 15 minutes, 1 hour and 16 minutes – 1 hour and 30 minutes, 1 hour and 31 minutes – 1 hour and 45 minutes, 1 hour and 46 minutes – 2 hours, more than 2 hours)

Table 4

Type of EMA	Framing of Items	Example Items
• Pre- and Post-Quit Random EMA • Pre- and Post-Quit Urge EMA • Pre-Quit Smoking Part One; Post-Quit About to Slip Part One	Participants self-report on how they felt, their susceptibility to smoking, and their context either (1) in reference to the present moment or (2) in reference to the entire duration of time between the present EMA (i.e., the k^{th} EMA) and the prior EMA (i.e., the $(k-1)^{th}$ EMA)	Example items with framing (1): • "RIGHT NOW, I *feel* happy" • "RIGHT NOW, I have an urge to smoke" Example items with framing (2): • Since the last computer recording, how often have you found your attention drawn to cigarettes? • How long ago did you smoke the most recent cigarette that you did not record? • Since the last computer recording, I experienced or thought about a NEW OR ONGOING stressful issue or problem.
• Post-Quit Already Slipped EMA	Participants self-report on how they felt, their susceptibility to smoking, and their context in *terms of the past*, specifically in reference to the time they smoked the cigarette(s) reported in the present EMA.	• "PRIOR TO SMOKING, I *felt* happy" • "PRIOR TO SMOKING, I had an urge to smoke"
• Pre-Quit Smoking Part Two EMA • Post-Quit About to Slip Part Two EMA	Participants self-report on how they felt, their susceptibility to smoking, and their context either (1) in reference to the present moment or (2) in reference to the time they smoked the cigarette(s) reported in the present EMA	Example items with framing (1): • "RIGHT NOW, I *feel* happy" • "RIGHT NOW, I have an urge to smoke" Example items with framing (2): • "Did you change location in order to smoke?" • "Smoking was pleasurable?"

6. Teasing Apart Various Events Represented in the Raw Data

Unique identifiers in the raw data

Before proceeding, we note that all curated datasets can be linked through a unique participant identifier, captured by two equivalent variables named **'id'** (a 4-digit number) and **'callnumr'** (a 7-character alpha-numeric code).

The variable 'id' is a participant identifier retained as-is (unmodified) from the raw datasets in Table 5 (derived from the raw data variable, **'Participant_ID'**).

Table 5

Raw Data Files Containing Responses to Pre-Quit Mode EMA Types	Raw Data Files Containing Responses to Post-Quit Mode EMA Types
Pre_Quit_Random.csv	Post_Quit_Random.csv
Pre_Quit_Urge.csv	Post_Quit_Urge.csv
Pre_Quit_Smoking.csv	Post_Quit_About_to_Slip.csv
Pre_Quit_Smoking_Part2.csv	Post_Quit_About_to_Slip_Part2.csv
	Post_Quit_Already_Slipped.csv

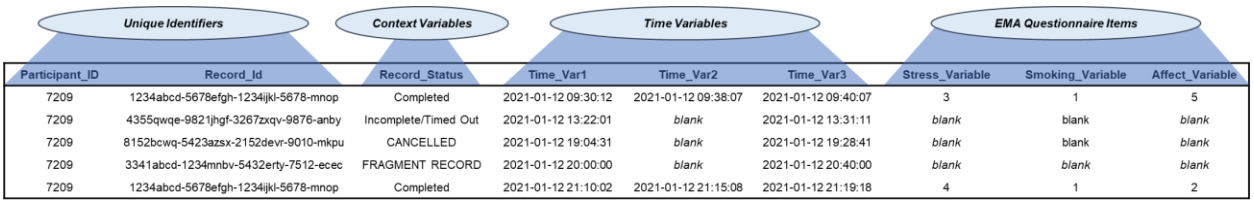
On the other hand, we noted that baseline survey raw data collected during Visit #2 also contained another variable, named 'callnumr' which study staff utilized as unique participant identifiers. We used records from study staff to match both participant identifiers together and added this participant identifier in the curated datasets as well. The original name of the identifier utilized during Visit #2 (i.e., 'callnumr') was retained in the curated datasets.

Finally, raw data files also provide a unique record (i.e., row) identifier: **'Record_ID'** is a 36-character alpha-numeric code associated with each record (i.e., row) in the raw data files. In the curated datasets, this variable is retained as-is (unmodified) and renamed **'record_id'**.

Structure of the raw data

The nine raw data files in Table 5 are in a tabular format and have a similar data structure. For each raw data file, all records (i.e., rows) belong to one and only one kind of EMA questionnaire. Figure 2 provides a grossly simplified schematic of columns and rows contained in these raw data files for one participant.

Figure 2



Participant_ID	Record_Id	Record_Status	Time_Var1	Time_Var2	Time_Var3	Stress_Variable	Smoking_Variable	Affect_Variable
7209	1234abcd-5678efgh-1234ijkl-5678-mnop	Completed	2021-01-12 09:30:12	2021-01-12 09:38:07	2021-01-12 09:40:07	3	1	5
7209	4355qwqe-9821jhgf-3267zxqv-9876-anby	Incomplete/Timed Out	2021-01-12 13:22:01	*blank*	2021-01-12 13:31:11	*blank*	blank	*blank*
7209	8152bcwq-5423azsx-2152devr-9010-mkpu	CANCELLED	2021-01-12 19:04:31	*blank*	2021-01-12 19:28:41	*blank*	blank	*blank*
7209	3341abcd-1234mnbv-5432erty-7512-ecec	FRAGMENT RECORD	2021-01-12 20:00:00	*blank*	2021-01-12 20:40:00	*blank*	*blank*	*blank*
7209	1234abcd-5678efgh-1234ijkl-5678-mnop	Completed	2021-01-12 21:10:02	2021-01-12 21:15:08	2021-01-12 21:19:18	4	1	2

All raw data files contain columns for **unique identifiers**, columns for **context variables**, columns for **time variables**, and columns for **EMA questionnaire items.** The raw data files differ only with respect to the specific columns pertaining to EMA questionnaire items.

Events represented in the raw data

A preliminary step in interpreting raw data records (i.e., rows) was to inspect the raw data variable, **'Record_Status'**. We observed that a value for 'Record_Status' is present in all records (i.e., rows) in the raw data, and further, has four possible values: 'Completed', 'Incomplete/Timed Out', 'CANCELLED', and 'FRAGMENT RECORD'. We note that the 'Record_Status' variable is retained as-is (i.e., unmodified) in the curated datasets and re-named **'record_status'**.

Unfortunately, we do not have documentation about what these values in the 'Record_Status' variable was meant to capture. At first sight, this variable appeared to capture whether a participant responded to an EMA questionnaire and the cause of non-completion. For example, 'Completed' could refer to records corresponding to when a participant completed the entirety of an EMA questionnaire; 'Incomplete/Timed Out' could refer to records corresponding to a time when a participant ignored an EMA questionnaire that was launched; 'CANCELLED' could refer to records corresponding to a time when a participant exercised the option of discontinuing an EMA questionnaire prior to responding to all items; 'FRAGMENT RECORD' could refer to records corresponding to when the smartphone's EMA software failed to record a portion of a participant's responses to items within an EMA questionnaire (e.g., possibly due to a software bug impacting completion of EMA questionnaire).

However, as we will show later, this interpretation is likely to be an insufficient characterization of the information captured in this variable. In particular, if end-users simply consider the variable 'Record_Status' alone and attempt to imbue a common-sense interpretation to its four possible values when assessing rates of EMA questionnaire (of any type) completion, then such a calculation will be inaccurate.

Towards a correct attribution of raw data records to events that caused them: our process

Initially, each row in the raw data file was inspected to determine whether there was a response recorded to any item within the EMA questionnaire; we construct the binary variable '**with_any_response'** for each record (i.e., row) as follows:

$$\textbf{with\_any\_response} = \begin{cases} 1, & \text{if a response was recorded for at least 1 item} \\ 0, & \text{if no response was recorded for all items} \end{cases}$$

Then, we jointly consider the variables 'with_any_response' and 'Record_Status' in light of the intended study design (i.e., see Figure 1 and Table 1). For each of the raw data files in Table 5, we tabulate the number of records (i.e., rows) having each possible combination of values of the pair of variables, 'with_any_response' and 'Record_Status'; the counts of each combination of values are displayed in Table 6.

Table 6

Type of EMA Questionnaire a Record (i.e., a Row) is Associated with	Value of 'with_any_response'	Value of 'Record_Status'				Row Totals
		'Completed'	'Incomplete/ Timed Out'	'CANCELLED'	'FRAGMENT RECORD'	
Post-Quit About to Slip Part One	0	0	3	702	20	725
Post-Quit About to Slip Part Two	0	0	31	0	34	65
Post-Quit Already Slipped	0	0	14	1130	24	1168
Post-Quit Random	0	29	2609	0	55	2693
Post-Quit Urge	0	8	9	595	21	633
Pre-Quit Random	0	8	806	0	57	871
Pre-Quit Smoking Part One	0	12	20	661	54	747
Pre-Quit Smoking Part Two	0	12	93	0	66	171
Pre-Quit Urge	0	8	18	823	76	925
Post-Quit About to Slip Part One	1	492	27	1	13	533
Post-Quit About to Slip Part Two	1	414	20	0	2	436
Post-Quit Already Slipped	1	471	28	1	7	507
Post-Quit Random	1	5455	197	0	13	5665
Post-Quit Urge	1	1273	51	0	6	1330
Pre-Quit Random	1	1662	133	0	2	1797
Pre-Quit Smoking Part One	1	991	87	0	11	1089
Pre-Quit Smoking Part Two	1	809	50	0	6	865
Pre-Quit Urge	1	997	83	0	13	1093
	Column Totals	12641	4279	3913	480	21313

From Table 6, we initially observe the existence of records that appear to violate *data integrity constraints* – or the logical rules of interpretation that information in a database must satisfy to facilitate consistent interpretation of records.

For example, inspection of records (i.e., rows) revealed the following:

- There are records (i.e., rows) for which no response was recorded for all items but were marked as 'Completed' (i.e., 'with_any_response=0 and 'Record_Status' = 'Completed').

- There are records (i.e., rows) for which no response was recorded for all items but were marked as 'FRAGMENT RECORD' (i.e., 'with_any_response=0 and 'Record_Status' = 'FRAGMENT RECORD').

These examples highlight how inspection of the raw data suggests that 'Record_Status' cannot be considered ground truth regarding questionnaire completion. As such, examining 'Record_Status' simultaneously with other complementary information (e.g., intended study design) may provide us with a more accurate view of what events could have led to the records (i.e., rows) in the raw data.

Now, we turn our attention to the existence of a high number of records (i.e., rows) for which no response was recorded for all items but the variable

'Record_Status' was marked as 'CANCELLED'.
Specifically, we observe that such cases occur only for the following types of EMA questionnaires:

- Pre-Quit Smoking Part One EMA & Post-Quit About to Slip Part One EMA
- Pre-Quit Urge EMA & Post-Quit Urge EMA
- Post-Quit Already Slipped EMA

Note: At times, 'data integrity constraints' are not enforced or only partially enforced by the data collection software. For example, when a software regards an EMA as 'Completed', we should expect to see no responses to any of the items. When violations to data integrity constraints occur, they must be identified and enforced during the process of data curation.

Pre-Quit Smoking Part One EMA & Post-Quit About to Slip Part One EMA (Button presses indicating an intent to smoke): Recall that the smartphone's EMA software is designed so that a portion of these button presses (i.e., a 20%-100% sampling rate) are followed by an EMA questionnaire. Hence, for Pre-Quit Smoking Part One EMAs and Post-Quit About to Slip Part One EMAs, we will attribute the presence of the 'with_any_response' = 0 & 'record_status' = 'CANCELLED' combination to the occurrence of button press which was not followed by a EMA questionnaire.

Pre-Quit Urge EMA & Post-Quit Urge EMA (i.e., button presses indicating an urge to smoke): Recall that the smartphone's EMA software is designed to so all of these button presses (i.e., 100% sampling rate) are followed by an EMA questionnaire. Hence, for Pre-Quit Urge EMAs and Post-Quit Urge EMAs, we will attribute the presence of the 'with_any_response' = 0 & 'record_status' = 'CANCELLED' combination to occurrence of an unknown technical issue. In other words, this combination suggests that Pre- and Post-Quit Urge button presses were not followed by an EMA questionnaire, which is inconsistent with the intended study design.**Post-Quit Already Slipped EMA (Button presses indicating a recollection of a past smoking event):** Recall that the smartphone's EMA software is designed so that all of these button presses (i.e., 100% sampling rate) are followed by an EMA questionnaire. Hence, for Post-Quit Already Slipped EMAs, we will attribute the presence of the 'with_any_response' = 0 & 'record_status' = 'CANCELLED' combination to an unknown technical issue. In other words, this combination suggests that a Post-Quit Already Slipped button press was not followed by an EMA questionnaire, which is inconsistent with the intended study design.

Finally, we turn our attention to those records associated with Self-Initiated EMA Questionnaires for which a response was recorded to at least one item. In contrast to the above cases discussed, we will instead attribute all such cases (regardless of value of 'Record_Status') to the occurrence of a button press that was indeed followed by an EMA questionnaire.

Therefore, as the attribution of records to the actual events that caused them cannot be accomplished with full certainty, it was necessary to develop an approach to *infer* what kind of information the variables in the raw data truly captured vis-à-vis the intended study design.

Attribution of Raw Data Records (i.e., Rows) to Events: Decision Rule Utilized

Our investigation which considered the variable 'Record_Status' simultaneously with other complementary information suggests that there are three types of mutually-exclusive events represented in the raw data:

Event A: EMA questionnaire (of any type) was launched *but* there is an indication of issues with smartphone's EMA software.

Event B: Participant performs a button press immediately before an intention to smoke, but no EMA questionnaire was launched (i.e., no Pre-Quit Smoking Part One EMA or Post-Quit About to Slip Part One EMA was launched).

Event C: EMA questionnaire was launched *and* there is no indication of issues with smartphone's EMA software.

Below, we display the decision rule we utilized to determine whether a record (i.e., row) in a raw data file is to be attributed to Event A, B, or C.

Note: Attribution of raw data records to the actual events that produced them cannot always be feasibly accomplished with full certainty.

For records associated with *Pre-Quit Smoking Part One & Post-Quit About to Slip Part One* EMA questionnaires:

- Attribute a record (i.e., row) to Event A if
  - with_any_response = 0 and record_status = 'Completed', **or**
  - with_any_response = 0 and record_status = 'FRAGMENT RECORD', **or**

- Attribute a record (i.e., row) to Event B if
  - with_any_response = 0 and record_status = 'CANCELLED'

- Attribute a record (i.e., row) to Event C if
  - with_any_response = 1, **or**
    with_any_response = 0 and record_status = 'Incomplete/Timed Out'

For records associated with *all other types* of EMA questionnaires:

- Attribute a record (i.e., row) to Event A if
  - with_any_response = 0 and record_status = 'Completed', **or**
  - with_any_response = 0 and record_status = 'FRAGMENT RECORD', **or**
  - with_any_response = 0 and record_status = 'CANCELLED'

- Do not attribute any record (i.e., row) to Event B

- Attribute a record (i.e., row) to Event C if
  - with_any_response = 1, **or**
  - with_any_response = 0 and record_status = 'Incomplete/Timed Out'

Additional checks concerning Event B

We examined those records of Pre-Quit Smoking Part Two and Post-Quit About to Slip Part Two EMA questionnaires that were preceded by Pre-Quit Smoking Part One and Post-Quit About to Slip Part One EMA questionnaires, respectively.

Among records (i.e., rows) of Pre-Quit Smoking Part Two and Post-Quit About to Slip Part Two EMA questionnaires, we examined responses to the question, "How many cigarettes did you just smoke?". The number of records where the response was missing, zero, or more than zero was counted and displayed in columns named 'Missing', 'Zero', 'Greater than Zero', respectively in Table 7.

Table 7

Type of EMA Questionnaire Record (i.e., a Row) is Associated with	Missing	Zero	Greater than Zero	Row Totals
Post-Quit About to Slip Part Two	68	290	134	492
Pre-Quit Smoking Part Two	223	90	691	1004
Column Totals	291	380	825	1496

We observe that in majority of records associated with Post-Quit About to Slip Part Two EMA questionnaire (290 out of 492), participants reported zero cigarettes smoked.

We considered the possibility that participants reporting zero cigarettes in Part Two EMA questionnaires did not actually mean to perform a button press prior to the Part One EMA questionnaire. However, in light of the intended study design and the fact that participants were trained to perform a button press when they had an intent to smoke, we take the view that these records correspond to a time when a participant *did not* carry out their intent to smoke.

7. General Decisions in Data Curation

7.1 Constructing curated time variables

Rationale for constructing curated time variables:

Curated time variables are vital to the data curation process; their purpose is three-fold:

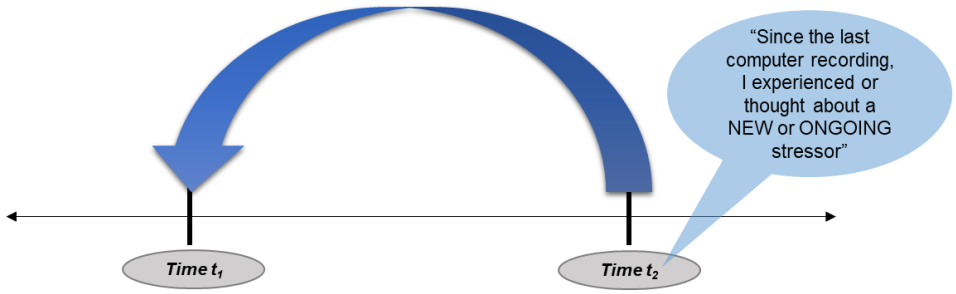
1. Subsequent data preparation steps will reference curated time variables when determining which records in the raw data will be retained or excluded in analyses of curated datasets.

2. There are multiple time variables recorded in the raw data. Curated time variables help to distill this information into a common set of time variables that will be utilized throughout all curated datasets. Having a common set of time variables facilitates a common language across all manuscripts utilizing the PNS study data, particularly when describing scientific hypotheses concerning timing of events.

3. Ordering of records in the raw data with respect to a time variable is a pre-requisite to approximating the time of occurrence of an event in between two EMA questionnaires. For example, if in an EMA questionnaire completed at t_2 a participant indicated they had experienced or thought about a NEW OR ONGOING stressful issue or problem since the

last recording, we are able to time-bound the participant's experience to have occurred between t_1 and t_2.

Figure 3



Constructing curated time variables at the EMA questionnaire-level:

For each record (i.e., row), the following four curated time variables were constructed:

1. Delivered Time (the variable '**delivered_hrts**' in the curated datasets):

    - Delivered Time is equal to the '**Initiated**' time variable in the raw data.
    - **For records attributed to Event B:** Delivered Time refers to the time when a participant performed the button press.
    - **For records attributed to Event C and associated with Random EMA questionnaires:** Delivered time refers to the time when the smartphone's EMA software launched a Random EMA questionnaire.
    - **For records attributed to Event C and associated with any type of Self-Initiated EMA questionnaire:** Delivered time refers to the time when a participant performed the button press preceding the launching of a Self-Initiated EMA questionnaire.

2. Begin Time (the variable '**begin_hrts**' in the curated datasets):

    - Begin Time is equal to the '**AssessmentBegin**' time variable in the raw data
    - **For records attributed to Event B:** Begin Time does not exist and is coded as a blank (i.e., '') in the curated datasets.
    - **For records attributed to Event C and associated with Random EMA questionnaires:** Begin Time refers to the time when a participant began completing the EMA questionnaire. Begin Time does not exist when an EMA questionnaire has no recorded response to any item (i.e., with_any_response=0); in these cases it is coded as a blank (i.e., ""). However, when a response to at least one item within an EMA questionnaire was recorded (i.e., with_any_response=1), then Begin Time is typically within a few minutes after Delivered Time.
    - **For records attributed to Event C and associated with any type of Self-Initiated EMA questionnaire:** *Identical interpretation as those records attributed to Event C and associated with Random EMA questionnaires.*

3. End Time (the variable '**end_hrts**' in the curated datasets):

- End Time is equal to either the '**AssessmentCompleted'** variable in the raw data or the '**AssessmentNotCompleted'** variable in the raw data, whichever of the two timestamps is available.
- **For records attributed to Event B:** End Time refers to the time when the software completes the sampling procedure (i.e., samples 20%-100% of button presses indicating intent to smoke); End Time is typically within a few *seconds* after Delivered Time.
- **For records attributed to Event C and associated with Random EMA questionnaires:** End Time refers either to the time when an EMA questionnaire was completed, or the point in time when the software determined that the EMA questionnaire was ignored or partially completed; regardless of whether an EMA questionnaire was ignored, partially completed, or fully completed, End Time is typically available and within a few *minutes* after Delivered Time. There are instances when there are no timestamps in the raw dataset that could be used for End Time (e.g., due to possible bug in software). In this case, this is left as a blank value (i.e., "") in the curated datasets.
- **For records attributed to Event C and associated with any type of Self-Initiated EMA questionnaire:** *Identical interpretation as those records attributed to Event C and associated with Random EMA questionnaires*.

Finally, we introduce a time variable that end-users may use to time-order records (i.e., rows) when simultaneously using records having with_any_response=1 or with_any_response=0 in analysis.

4. Aligned Time (the variable '**time_hrts'** in the curated datasets):

    - Aligned Time is equal to Begin Time if we observe a recorded response in at least one item in the raw data but equal to Delivered Time otherwise. In other words, for a particular record,

$$Aligned\ Time = \begin{cases} Begin\ Time, & if\ with\_any\_response = 1 \\ Delivered\ Time, & if\ with\_any\_response = 0 \end{cases}$$

    - **For records attributed to Event B:** Aligned Time is equal to Delivered Time (i.e., time when a participant performed the button press) since all records associated with Event B do not have any response recorded for all items.
    - **For records attributed to Event C and associated with Random EMA questionnaires:** Aligned Time refers to either when the smartphone's EMA software launched a Random EMA questionnaire or the time when a participant began completing the EMA questionnaire.
    - **For records attributed to Event C and associated with any type of Self-Initiated EMA questionnaire:** Aligned Time refers to either the time when a participant performed the button press or the time when a participant began completing the EMA questionnaire.

A note on time variables not captured in the raw data. We note that in the raw datasets displayed in Table 5, raw data variables pertaining to the time when each individual item within an EMA questionnaire was completed is not available. However, for records (any type)

associated with Event C, the four curated time variables above can be used to time bound responses within an EMA questionnaire.

Constructing curated time variables at the person-level

Three more curated time variables were constructed; these were based on an individual's Quit Date. We provide details on the determination of Quit Date in a separate section, Section 8.

- Quit Time (the variable '**quit_hrts**' in the curated datasets): Quit Time refers to 4AM on a participant's Quit Date.

- Start Study Time (the variable '**start_study_hrts**' in the curated datasets): Start Study Time refers to 12AM on the seventh day prior to 12AM on a participant's Quit Date.

- End Study Time (the variable '**end_study_hrts**' in the curated datasets): End Study Time refers to 12AM on the twenty-first day after 12AM on a participant's Quit Date.

We note that 4AM was used for Quit Time instead of 12AM as in Start Study Time and End Study Time to accommodate the fact that participants may sleep after midnight and hence may not have made an attempt at smoking cessation until after waking.

7.2 Time period of interest in analysis of curated datasets

Exclusion criteria based on Delivered Time

EMA questionnaires (i.e., records attributed to Event C) or button presses (i.e., records attributed to Event B) having Delivered Time either before Start Study Time or after End Study Time will be excluded from all data analysis. In other words, records outside of the study period are excluded from all curated datasets.

EMAs included in manuscripts will depend on the time period of interest:

a) A manuscript investigating both the Pre-Quit Period and Post-Quit Period will include only EMAs having Delivered Time that falls between Start Study Time and End Study Time
b) A manuscript investigating the Pre-Quit Period will include only EMAs having a Delivered Time that falls between Start Study Time and Quit Time
c) A manuscript investigating the Post-Quit Period will include only EMAs having Delivered Time that falls between Quit Time and End Study Time.

Further, a binary variable **'use_as_postquit'** was constructed for each EMA in the curated datasets. This variable can be used to identify EMAs for manuscripts focusing on the Pre-Quit Period only or Post-Quit Period only.

$$\textbf{use\_as\_postquit} = \begin{cases} 1, & \text{if Delivered Time falls between Quit Time and End Study Time} \\ 0, & \text{if Delivered Time falls between Start Study Time and Quit Time} \end{cases}$$

Important note. The variable 'assessment_type' in the curated datasets should not be used to determine when an EMA Questionnaire/button press was launched/performed in relation to a participant's Quit Date. The 'assessment_type' variable is simply used to determine which particular EMA questionnaire is associated with a record; the variable 'use_as_postquit' should

be used to make the determination as to whether an EMA Questionnaire/button press was launched/performed before or after Quit Date.

7.3 Main analysis and sensitivity analysis

In Section 4, we noted that there is no ground truth on each participant's Quit Date. Instead, multiple plausible but potentially conflicting records are available from which Quit Date needs to be inferred. For each participant, when all records under consideration match, we say that there is no ambiguity in the participant's Quit Date. On the other hand, for each participant, when all records under consideration do not match, we say that there is ambiguity in the participant's Quit Date.

When there is ambiguity in Quit Date, the participant's reported number of cigarettes smoked was compared against a set of candidate dates to infer a participant's likely true Quit Date. At least three domain experts on human behavior were engaged to visually inspect plots overlaying reported number of cigarettes smoked over the course of the study surrounding candidate dates (details and examples of the plots are discussed in Section 8); plots were constructed for each individual participant. If all domain experts were able to come to an agreement about which candidate date was most likely the participant's true Quit Date, then we say that there is low ambiguity in the participant's Quit Date. Otherwise, if not all domain experts are able to come to an agreement, then we say that there is high ambiguity in the participant's Quit Date.

Rationale: Participants who have an intent to quit smoking (i.e., participants in the intended study population) are expected to have a change in their smoking behavior in the days leading up to their intended Quit Date and in the days after their intended Quit Date. Hence, an approach in the vein of assessing inter-rater agreement was utilized.

With differing amounts of ambiguity in Quit Date, two sets of analysis are thus recommended for end-users of the curated datasets:

1. Main Analysis: In main analysis, participants who do not have ambiguity in their Quit Date *and* participants who have ambiguity (low or high) in their Quit Date will be included in analysis.

2. Sensitivity Analysis: In sensitivity analysis, only participants who do not have ambiguity in their Quit Date and participants who have low ambiguity in their Quit Date will be included in analysis (i.e., participants who have high ambiguity in their Quit Date will be excluded).

A binary variable defined below was constructed for each record (i.e., row) in the curated datasets; this variable can be used to identify participants having ambiguity in their Quit Date:

$$\textbf{sensitivity} = \begin{cases} 0, & \text{if there is } high \text{ ambiguity in the participant's Quit Date} \\ 1, & \text{if there is } no \text{ ambiguity or } low \text{ ambiguity in the participant's Quit Date} \end{cases}$$

Table 8 summarizes joint criterion on the variables '**use_as_postquit**' and '**sensitivity**' by type of manuscript and type of analysis.

Table 8

Scope of manuscript	Main analysis	Sensitivity analysis
Only Pre-Quit Period	EMAs associated with either '**sensitivity**'=0 or '**sensitivity**'=1 will be included. Only EMAs associated with '**use_as_postquit**' = 0 will be included.	Only EMAs associated with '**sensitivity**' = 1 will be included. Only EMAs associated with '**use_as_postquit**' = 0 will be included.
Only Post-Quit Period	EMAs associated with either '**sensitivity**'=0 or '**sensitivity**'=1 will be included. Only EMAs associated with '**use_as_postquit**' = 1 will be included.	Only EMAs associated with '**sensitivity**' = 1 will be included. Only EMAs associated with '**use_as_postquit**' = 1 will be included.
Both Pre-Quit Period and Post-Quit Period	EMAs associated with either '**sensitivity**'=0 or '**sensitivity**'=1 will be included. EMAs associated with either '**use_as_postquit**'=0 or '**use_as_postquit**'=1 will be included.	Only EMAs associated with '**sensitivity**' = 1 will be included. EMAs associated with either '**use_as_postquit**'=0 or '**use_as_postquit**'=1 will be included.

7.4 Structure of curated datasets with respect to Events A, B, C

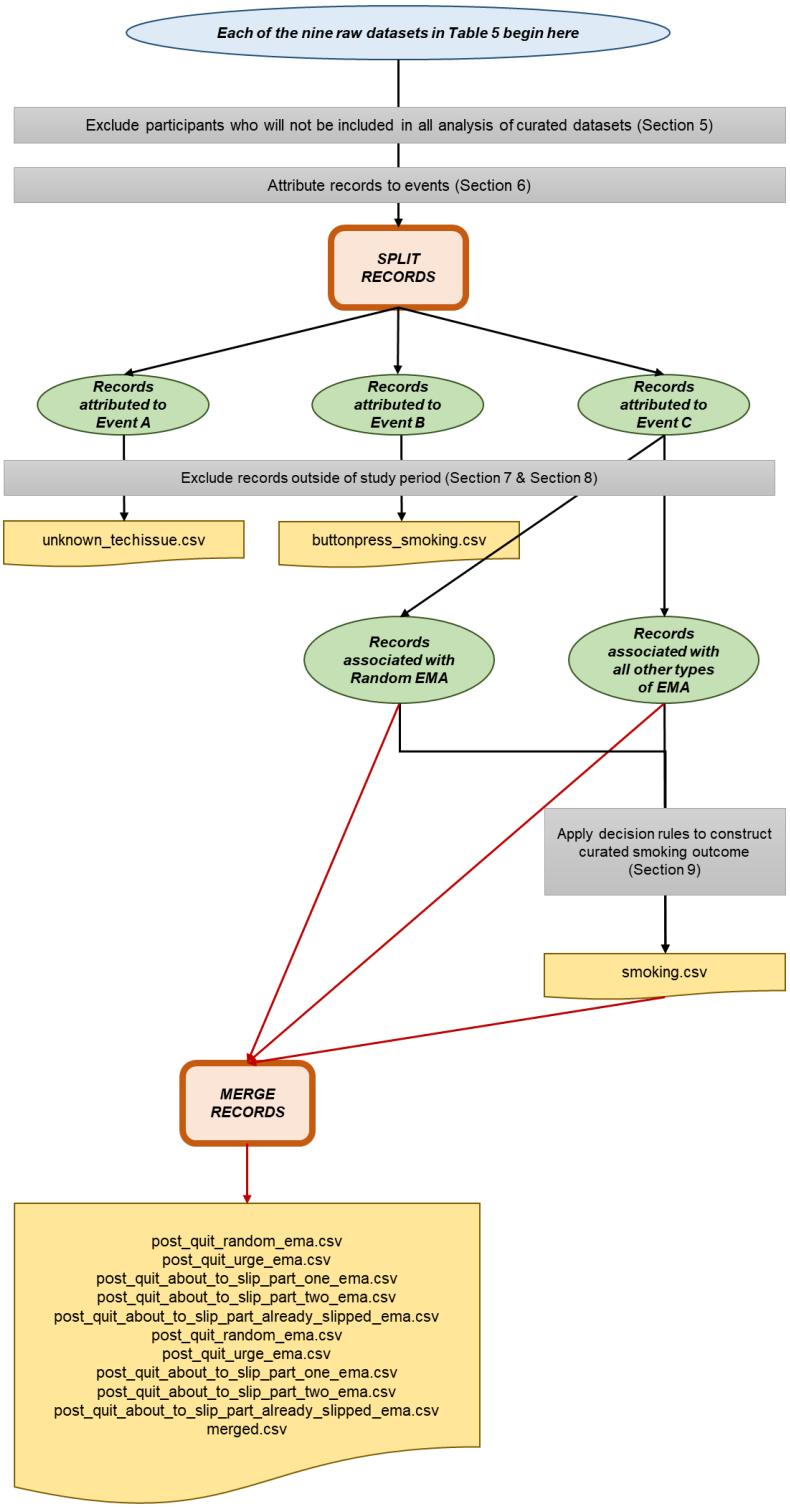
The curated datasets will be structured so that records attributed to Events A or B will be in a separate database from records attributed to Event C. However, participant data across all kinds of databases may still be linked via the participant's unique identifier (i.e., the variable, 'id'). Table 9 displays a listing of names of curated datasets and whether all records within a dataset were attributed to Events A, B, or C.

Table 9

Event A	Event B	Event C
Curated datasets whose records come from one and only one kind of EMA questionnaire		
None	*None*	Box 8c pre_quit_random_ema.csv pre_quit_urge_ema.csv pre_quit_smoking_part_one_ema.csv pre_quit_smoking_part_two_ema.csv post_quit_random_ema.csv post_quit_urge_ema.csv post_quit_about_to_slip_part_one_ema.csv post_quit_about_to_slip_part_two_ema.csv post_quit_already_slipped_ema.csv
Curated datasets whose records come from more than one kind of EMA questionnaire		
Box 8a unknown_techissue.csv	Box 8b buttonpress_smoking.csv	Box 8d smoking.csv merged.csv

Figure 4 provides a flowchart tracing each stage of the data curation process, beginning from the raw data in Table 5 to the curated data files in Table 9.

Figure 4



8. Quit Dates

8.1 Date records used towards inferring a participant's True Quit Date

Different sets of date records are present and could be used to provide evidence for the day most likely to be a participant's true Quit Date. We collected these dates into one file named **alldates_annotated.xlsx**.

1. **Date Record#1** is a date recorded in the baseline in-person visit (Visit #2).

    - Specifically, this date is obtained from the '**quitday'** column in the raw data file named **PNSBaseline.csv**. We note that this date is *likely* the participant's **'Projected Quit Date'**.

    - In **alldates_annotated.xlsx**, this date is provided the variable '**quitday'**.

2. **Date Record#2** is a date within records of study staff.

    - Study staff records (specifically, the file **PNS_FU_Dates_030512.xlsx**) contain a date in the column '**EMA_Qday'**. We note that this date is *likely* either a participant's **'Projected Quit Date'** or **'Adjusted Quit Date'**. However, we cannot ascertain which of these dates correspond to '**EMA_Qday**'.

    - In **alldates_annotated.xlsx**, this date is provided in the variable '**EMA_Qday'**.

3. **Date Record#3** is the date associated with the first record (i.e., row) after the smartphone's EMA software switched to Post-Quit Mode.

    - The date associated with this particular record was calculated using all available Post-Quit Mode EMAs in the raw data.

    - In **alldates_annotated.xlsx**, this date is provided in two variables: **postquit.earliest.longformatdate** (in year-month-day hour-minute-second format) and **postquit.earliest.shortformatdate** (in year-month-day format)

4. **Date Record#4** is the date associated with the last record (i.e., row) when the smartphone's EMA software was still in Pre-Quit Mode.

    - The date associated with this particular record was calculated using all available Pre-Quit Mode EMAs in the raw data.

    - In **alldates_annotated.xlsx**, this date is provided in two variables: **prequit.latest.longformatdate** (in year-month-day hour-minute-second format) and **prequit.latest.shortformatdate** (in year-month-day format)

5. **Date Record #5** are dates within EMA raw data files

- Each EMA raw data file in Table 5 contains the five variables, '**Quit_Date1'**, '**Quit_Date2'**, '**Quit_Date3'**, '**Quit_Date4'**, '**Quit_Date5'**. At first glance, we considered the possibility that these variables were meant to capture a move from Projected Quit Date to Adjusted Quit Date (e.g., **'Projected Quit Date**' would be in '**Quit_Date1'** while **'Adjusted Quit Date**' would be in '**Quit_Date2**'). However, examination of the raw data showed that throughout all EMA raw data files in Table 5, the variables '**Quit_Date2'**, '**Quit_Date3'**, '**Quit_Date4'**, and '**Quit_Date5'** have missing values. On the other hand, all participants have a date in '**Quit_Date1'**.

- The variable **'Quit_Date1'** will not be included in **alldates.csv**. Rationale: A comparison of the variable **'Quit_Date1'** against the variable '**EMA_Qday'** in records of study staff shows that for a given participant, the recorded year-month-day in these two variables match exactly. Moreover, for a given participant, if **'EMA_Qday'** is missing in records of study staff, this participant will not have data in any of the EMA raw data files (and hence, will have missing **'Quit_Date1'** records as well).

We additionally note that we observe instances when dates recorded by study staff (i.e., **'EMA_Qday'**) fall after the time when the first EMA in Post-Quit Mode was delivered (i.e., **'postquit.earliest.longformatdate'**). For example, **'postquit.earliest.longformatdate'** is on March 16, 2010 but **'Quit_Date1'**and **'EMA_Qday'** are both on March 17, 2010; further examples could be seen in the file **alldates_annotated.xlsx**. To the best of our knowledge, the software switched from Pre-Quit Mode to Post-Quit Mode only upon manual activation by study staff. Since participants did not have the ability to initiate the switch, interpreting **'EMA_Qday'** as 'Projected Quit Date' is called into question and hence we have ambiguity in these data sources.

8.2 Distinguishing participants who have ambiguity in their Quit Date from those who do not

Participants for whom all the following dates match are considered to have no ambiguity in their Quit Date, while participants for whom at least two of the following dates do not match are considered to have ambiguity in their Quit Date.

- Date Record#1: A date recorded in the baseline in-person visit (Visit #2)
- Date Record#2: A date within records of study staff.
- Date Record#3: A date associated with the first record (i.e., row) after the smartphone's EMA software switched to Post-Quit Mode.

In the file **alldates_annotated.xlsx**, ambiguity status is indicated by the variable '**is.equal'** (=1 if a participant has no ambiguity in their Quit Date, =0 otherwise).

- For participants who are considered to have no ambiguity in their Quit Date, the matched date in the records above will be used as their **'Working Quit Date'**.

- For participants who are considered to have ambiguity in their Quit Date, the process described next was used to determine their **'Working Quit Date'**.

8.3 General procedure for determining 'Working Quit Date' for participants who have ambiguity in their Quit Date

We earlier noted that participants who have an intent to quit smoking (i.e., participants in the intended study population) are expected to have a change in their smoking behavior in the days leading up to their intended Quit Date and in the days after their intended Quit Date.

To infer participants true Quit date (i.e., arrive at the 'Working Quit Date'), we conducted a process which involved at least three domain experts in human behavior performing a visual inspection of reported number of cigarettes smoked in EMAs with respect to the three dates described above.

The example plots illustrate a participant with no ambiguity in their Quit Date (Panel A, Table 10) and another participant with low ambiguity in their Quit Date (Panel B, Table 10).

Interpreting the horizontal axis of the plots

The horizontal axis represents the specific days within the study period relative to time zero. Time zero is the time associated with the first record after the smartphone software switched to Post-Quit Mode (i.e., Date Record#3 described in Section 8). Time zero is represented with a vertical line that is shaded a transparent gray. As the other candidate dates are represented by vertical dashed lines, when they coincide with time zero, an overlap between the candidate dates can be visualized.

For example, Panel A in Table 10, shows that time zero is represented by a shaded gray vertical line. Further, there is a black vertical dashed line on day zero which suggests that both '**quitday**' '**EMA_Qday**' were on day zero (i.e., '**quitday**'=0 and '**EMA_Qday**'=0).

This suggests there is no ambiguity in quit date, such that the date recorded in the baseline raw data file, within records of study staff, and the date on which the first record (i.e., row) after the smartphone's EMA software switched to Post-Quit Mode coincided on exactly the same day.

On the other hand, Panel B in Table 10 shows that time zero is represented with a shaded gray vertical line and that **'EMA_Qday'** (vertical dashed black line) also falls on day zero. However, **'quitday'** (also a black dashed line) falls five days prior to day zero (i.e., **'EMA_Qday'**=0 and **'quitday'**=-5).

Hence, although date within records of study staff and the time when the smartphone's EMA software switched to Post-Quit Mode coincided on the same day, the quit date recorded in the baseline raw data file was five days prior to these two dates.

Records (i.e., rows) utilized to construct the plots

All records in the raw data, regardless of whether they were eventually attributed to Event A, B, or C are utilized to construct the plots.

- **Records (i.e., rows) which would eventually be attributed to Event A:** These records represent some indication of issues with the smartphone's EMA software; these records will not have a response recorded for all EMA items, including items asking participants to report number of cigarettes smoked.

  Records corresponding to Event A will be represented as either solid yellow (for records associated with Self-Initiated EMA questionnaires) or grey (for records associated with

Random EMA questionnaires) dots in the plots. These annotations convey that an interaction between the participant and their smartphone may have occurred within the vicinity of the candidate Quit Dates.

- **Records (i.e., rows) which would eventually be attributed to Event B:** These records represent the occurrence of a button press immediately before an intention to smoke, but no EMA questionnaire was launched; these records will not have a response recorded for all EMA items, including items asking participants to report number of cigarettes smoked.
  Records corresponding to Event B will be represented as solid yellow (since these records are associated with Self-Initiated EMA questionnaires, but never with Random EMA questionnaires) dots in the plots. Again, these annotations convey that an interaction between the participant and their smartphone may have occurred within the vicinity of the candidate Quit Dates.

- **Records which would eventually be attributed to Event C:** These records represent the event that an EMA questionnaire was launched with no indication of issues with the smartphone's EMA software. These records may or may not have a response recorded to any of the EMA items.

  When a participant reports their number of cigarettes smoked, this is annotated by either a solid blue dot (zero cigarettes smoked) or a solid red dot (more than zero cigarettes smoked). A vertical line above the solid red dot indicates the maximum number of cigarettes the participant meant to report (see Section 9.2 for details on calculating number of cigarettes smoked in the plots).

  These annotations convey that an interaction between the participant and their smartphone may have occurred within the vicinity of the candidate Quit Dates, and further, displays whether any cigarettes were smoked within the vicinity of the candidate Quit Dates.

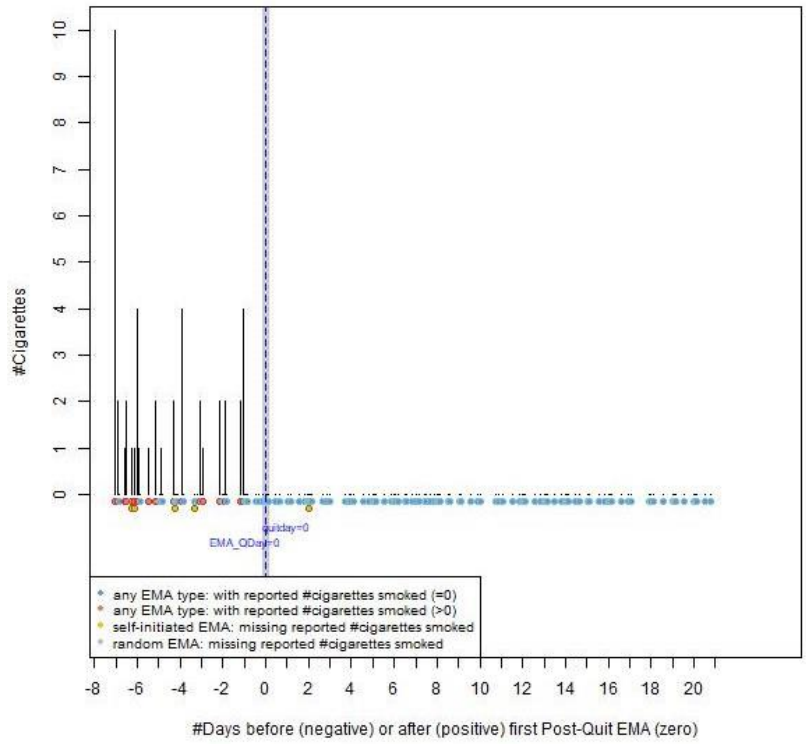
  Now, on the other hand, if a participant does not report their number of cigarettes smoked, this is represented as either solid yellow or grey dots in the plots.

The location of the solid dots relative to time zero was determined by calculating the number of days elapsed before (negative) or after (positive) time zero using the curated time variable, Aligned Time.

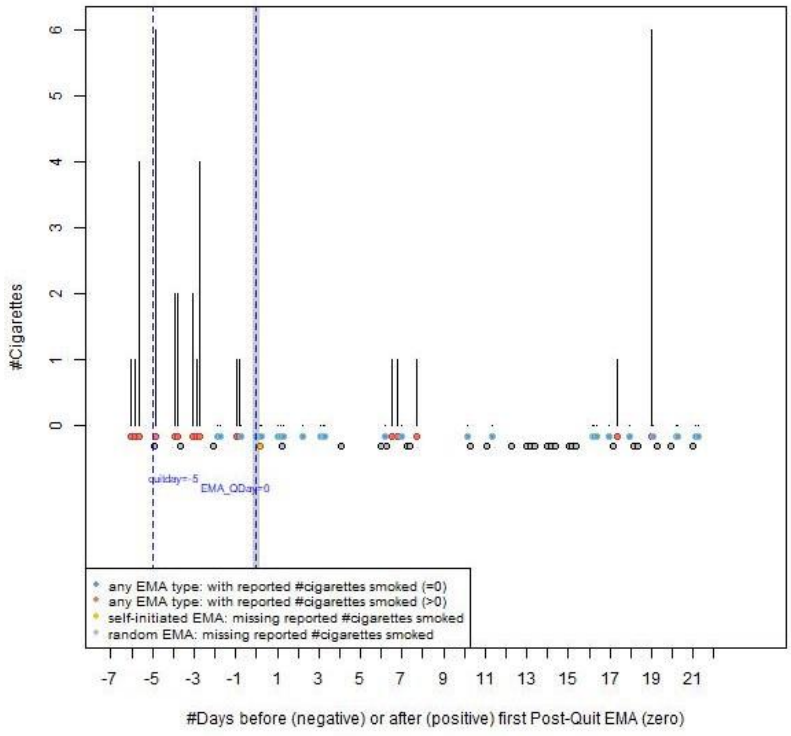
We note that days having no points (e.g., Day 9 on Panel B in Table 10) represent the fact that no records (i.e., rows) exist for those particular days. This scenario is possible if, for example, the participant switched off their smartphone for the entire day.

Table 10

Panel A: Data from a participant with no ambiguity in their Quit Date



Panel B: Data from a participant with low ambiguity in their Quit Date



9. Construction of curated smoking variables

9.1 Decisions common to all curated smoking variables

Records associated with any type of Post-Quit Mode EMA or Pre-Quit Mode EMA will be utilized in constructing the curated smoking variables. However, only records (i.e., rows) which have been attributed to Event C will be utilized to construct curated smoking variables. Additionally, records (i.e., rows) may be excluded depending on whether a response to any item was recorded. Specifically,

- Pre-Quit Random EMAs and Post-Quit Random EMAs having no recorded response to any EMA item (i.e., having with_any_response=0) will *not* be used to construct curated smoking variables.

- Pre-Quit Random EMAs and Post-Quit Random EMAs having a recorded response to any EMA item (i.e., having with_any_response=1) will be used to construct curated smoking variables.

- Self-initiated EMAs having no recorded response to any EMA item (i.e., having with_any_response=0) will be used to construct curated smoking variables

- Self-initiated EMAs having a recorded response to any EMA item (i.e., having with_any_response=1) will be used to construct curated smoking variables.

The remaining sections begin with identifying relevant variables within each type of EMA questionnaire which will be used during construction of the smoking variable

each type of EMA questionnaire. Recall that in Table 2 and Table 3, these items were grouped according to the following information:

- **Group 1:** Items focusing on whether a smoking event occurred (e.g., yes/no)
- **Group 2:** Items focusing on number of cigarettes smoked (e.g., 0,1,2,3, … cigarettes)
- **Group 3**: Items focusing on timing of smoking events (e.g., 0-15 minutes ago)

These tables also noted when a particular type of EMA questionnaire did not contain items relevant to the three groups.

9.2 Decisions specific to the construction of curated smoking quantity variables

Construction of curated smoking quantity variable included the following preliminary steps:

1. Resolve inconsistencies between codebook and raw data: For the Pre-Quit Random, Pre-Quit Urge, Post-Quit Random, Post-Quit Urge EMA types, coding of responses in codebook did not match coding of responses in the raw data. In particular, the category '2' appears to have been omitted in the codebook's coding of responses (see Panel A in Table 11). We note that for the Pre-Quit Smoking Part Two, Post-Quit Smoking Part Two, Post-Quit Already Slipped EMA types, coding of responses in codebook match. However, for the Pre-Quit Smoking Part Two and Post-Quit About to Slip Part Two EMA types, the category '2' appears to have been omitted in the raw data's coding of responses (see

Panel B in Table 11). Unlike other EMA types, the Post-Quit Already Slipped EMA type did not have the issues presented by the coding of responses (see Panel C in Table 11).

Table 11

Panel A

Description	Codebook	Raw Data
0	0	0
Less than 1	1	1
1-2	3	2
3-4	4	3
5-6	5	4
7-8	6	5
9-10	7	6
More than 10	8	7

Panel B

Description	Codebook	Raw Data
0	0	0
Less than 1	1	1
1-2	3	3
3-4	4	4
5-6	5	5
7-8	6	6
9-10	7	7
More than 10	8	8

Panel C

Description	Codebook	Raw Data
0	0	0
Less than 1	1	1
1-2	2	2
3-4	3	3
5-6	4	4
7-8	5	5
9-10	6	6
More than 10	7	7

Decision: For the Pre-Quit Random, Pre-Quit Urge, Post-Quit Random, Post-Quit Urge EMA types, use scale implied by the raw data. Hence, for all EMA types, use the coding of responses displayed in Panel C in Table 11.
Rationale: An encoding error is likely to have occurred in the coding of responses in codebook/raw data.

2. Align scales of variables in Group 2 so that all Group 2 variables are on the same scale: Only Pre-Quit Smoking Part Two and Post-Quit About to Slip Part Two EMA types require re-coding of values. The re-coding was performed as shown in Table 12.

Table 12

Description	Old Value	New Value
0	0	0
Less than 1	1	1
1-2	3	2
3-4	4	3
5-6	5	4
7-8	6	5
9-10	7	6
More than 10	8	7

3. Merge all Group 2 variables into one variable, **'rawdata.qty'** and merge all Group 1 variables into one variable, **'rawdata.indicator'**: We note that the variables **'rawdata.qty' and 'rawdata.indicator'** are simply *intermediate variables* and will not be included in the curated datasets.

Table 13

	rawdata.indicator	**rawdata.qty**
Pre-Quit EMA		
Random	**PreQRSmoking1**	**Smoking2_PreQ_Random**
Urge	**SmPQU1**	**Smoking2_PreQ_Urge**
Smoking Part One	*Set to missing value*	*Set to missing value*
Smoking Part Two	*Set to missing value*	**CigJustNow**
Post-Quit EMA		
Random	*PostQRSmoking1*	**Smoking2_PostQ_Random**
Urge	*SmPostQU1*	**Smoking2_PostQ_Urge**
About to Slip Part One	*Set to missing value*	**Set to missing value**
About to Slip Part Two	*Set to missing value*	**CigJustNow_PostQ_Slip2**
Already Slipped	*Set to missing value*	**HowManyCig**

4. Assign a value to '**smoking_qty'** based on value of Group 1 & 2 variables
    - If **'rawdata.qty'** *is not missing*, then use the rule in Table 14.

Table 14

Description	Value of **rawdata.qty**	Value of **smoking_qty**
0	0	0
Less than 1	1	.5
1-2	2	1.5
3-4	3	3.5
5-6	4	5.5
7-8	5	7.5
9-10	6	9.5
More than 10	7	10

Note: It is possible for a participant to report zero cigarettes smoked in Pre-Quit Smoking Part Two and Post-Quit About to Slip Part Two EMA types. When this is the case, we leave their reported number of cigarettes smoked unmodified, i.e., **'smoking_qty'** will be equal to zero in these cases.

Rationale: The button press could possibly have effectively functioned as a smoking cessation intervention itself and reduced the participant's likelihood of lapse to smoking.

- If **'rawdata.qty'** *is missing*, then use the following rule:
    - If **'rawdata.indicator'** = 0, then set **'smoking_qty'**=0
    - If **'rawdata.indicator'** is missing, then **'smoking_qty'** remains a missing value

A note on constructing plots to infer Quit Date. The plots inspected by domain experts described in Section 8 include a visual indicator of number of cigarettes smoked. The calculation of the number of cigarettes smoked to produce those plots is identical to all steps described in Section 9.2, except that the correspondence in Table 15 was used to produce plots rather than that described in Table 14. In other words, if a participant reported 'Less than 1' cigarettes smoked, then this would be depicted as 1 cigarette smoked in the plots inspected by domain experts (but a value of 0.5 would be given to the **'smoking_qty'** in the curated datasets end-users will utilize in analysis).

Table 15

Description	Value of **rawdata.qty**	Value of **smoking_qty**
0	0	0
Less than 1	1	1
1-2	2	2
3-4	3	4
5-6	4	6
7-8	5	8
9-10	6	10
More than 10	7	11

9.3 Decisions specific to the construction of a curated smoking indicator variable

After constructing the **'smoking_qty'** variable, we now construct a binary variable, named **'smoking_indicator'** in the curated datasets to indicate whether there is any indication of smoking between the current EMA and previous EMA (i.e., the EMA immediately preceding the current EMA); below, t_1 and t_2 denote the time associated with the previous EMA and current EMA, respectively.

$$smoking\_indicator = \begin{cases} 1, \textit{there is an indication of} \text{ smoking between } t_1 \text{ and } t_2 \\ 0, \textit{there is} \text{ an } \textit{indication of} \text{ no smoking between } t_1 \text{ and } t_2 \\ \textit{missing, there is no indication of the presence or absence of smoking} \text{ between } t_1 \text{ and } t_2 \end{cases}$$

The **'smoking_indicator'** variable may be constructed using the following rule:

- If **'rawdata.indicator'** = 0, then set **'smoking_indicator'**=0
- If **'rawdata.indicator'** = 1 and **'smoking_qty'**=0, then set **'smoking_indicator'**=0
- If **'rawdata.indicator'** = 1 and **'smoking_qty'**>0, then set **'smoking_indicator**=1
- If **'rawdata.indicator'** is missing, determine the value of **'smoking_indicator'** based on EMA type:
  - If assessment type is Pre-Quit Smoking Part Two/Post-Quit About to Slip Part Two/Post-Quit Already Slipped and **'smoking_qty'**=0 then set **'smoking_indicator'**=0
  - If assessment type is Pre-Quit Smoking Part Two/Post-Quit About to Slip Part Two/Post-Quit Already Slipped and **'smoking_qty'**>0 then set **'smoking_indicator'**=1
  - If assessment type is Pre-Quit Smoking Part One/Post-Quit About to Slip Part One then **'smoking_indicator'** remains a missing value.

Rationale for constructing **'smoking_indicator'** based on **'smoking_qty'**: At first sight, it may appear more straightforward/consistent with the chronological order of questions posed to a participant within an EMA to initially construct **smoking_indicator** prior to **smoking_qty**. However, we note that questions in Group 2 (see Table 2 and Table 3) allow participants the option to report zero cigarettes smoked. Hence, the approach we have taken is to place more weight on information provided in Group 2 variables compared to information provided in Group 1 variables.

9.4 Decisions specific to the construction of a curated smoking time variable

We now describe the construction of a curated smoking time variable, named **'smoking_delta_minutes'** in the curated datasets. This variable captures the time when the most recent cigarette reported in **'smoking_qty'** was smoked *relative to the current EMA*. Hence, **'smoking_delta_minutes'** is displayed in terms of number of minutes prior to the current EMA. This involved several steps:

1. Align scales of variables in Group 3 so that all Group 3 variables are on the same scale
   Variables in Group 3 did not have issues concerning inconsistencies in coding of values in the codebook and raw data that variables in Group 2 had. However, we note that the Pre-Quit Random, Pre-Quit Urge, Post-Quit Random, Post-Quit Urge EMA types were

coded as in Panel A in Table 16 while the Post-Quit Already Slipped type EMA was coded as in Panel B in Table 16.

Table 16

Panel A

Description	Value
0-15 minutes	1
16-30 minutes	2
31-45 minutes	3
46 minutes – 1 hour	4
1 hour and 1 minute – 1 hour and 15 minutes	5
1 hour and 16 minutes – 1 hour and 30 minutes	6
1 hour and 31 minutes – 1 hour and 45 minutes	7
1 hour and 46 minutes – 2 hours	8

Panel B

Description	Value
0-15 minutes	0
16-30 minutes	1
31-45 minutes	2
46 minutes – 1 hour	3
1 hour and 1 minute – 1 hour and 15 minutes	4
1 hour and 16 minutes – 1 hour and 30 minutes	5
1 hour and 31 minutes – 1 hour and 45 minutes	6
1 hour and 46 minutes – 2 hours	7
More than 2 hours	8

Hence, re-coding was performed on Pre-Quit Random, Pre-Quit Urge, Post-Quit Random, Post-Quit Urge as in Table 17.

Table 17

Description	Old Value	New Value
0-15 minutes	1	0
16-30 minutes	2	1
31-45 minutes	3	2
46 minutes – 1 hour	4	3
1 hour & 1 minute – 1 hour & 15 minutes	5	4
1 hour & 16 minute – 1 hour & 30 minutes	6	5
1 hour & 31 minute – 1 hour & 45 minutes	7	6
1 hour & 46 minute – 2 hours	8	7
More than 2 hours	9	8

2. Merge all Group 3 variables into one variable, **'rawdata.timing'**
   We note that the variable **'rawdata.timing'** is simply an *intermediate variable* and will not be included in the curated datasets.

Table 18

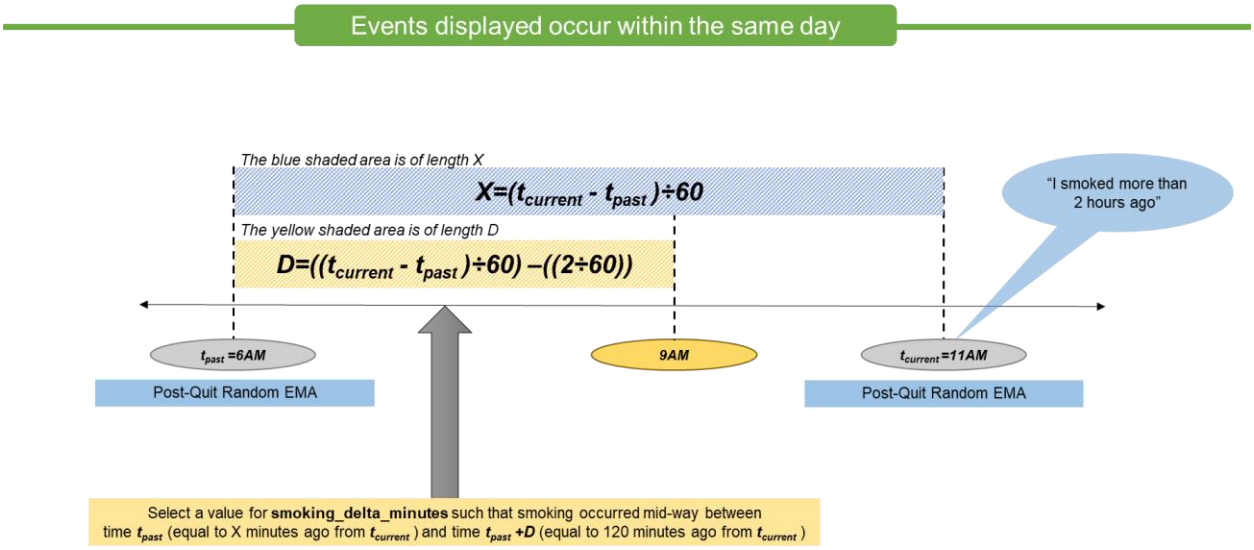
	rawdata.timing
Pre-Quit EMA	
Random	**Smoking3**
Urge	**Smoking3**
Smoking Part One	*Set to missing value*
Smoking Part Two	*Set to missing value*
Post-Quit EMA	
Random	**Smoking3**
Urge	**Smoking3**
About to Slip Part One	*Set to missing value*
About to Slip Part Two	*Set to missing value*
Already Slipped	*LastCig*

3. Use the rule below to assign a value to **'smoking_delta_minutes'** when **'rawdata.timing'** is not missing:
   - If the value of **'rawdata.timing'** is between 0-7, then **'smoking_delta_minutes'** is the midpoint of the corresponding time interval (e.g., the midpoint of 0-15 minutes is 7.5 minutes). In other words, the midpoint

of time intervals presented to participants as possible response options was taken as the **'smoking_delta_minutes'**.

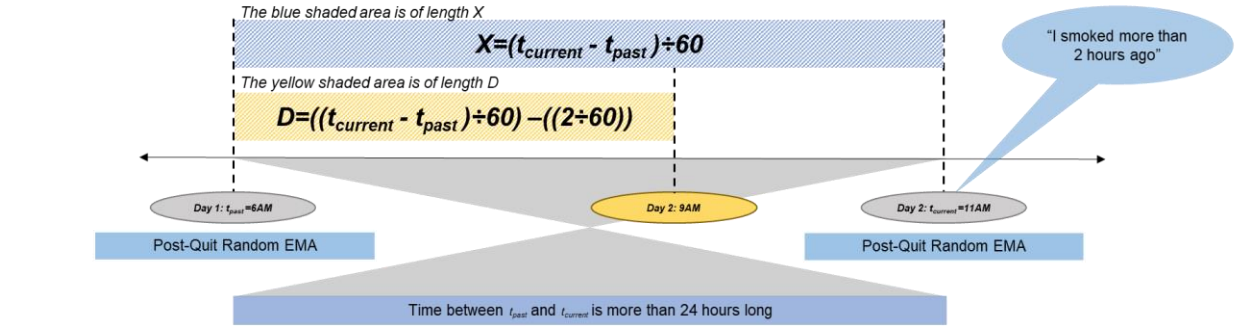
- If the value of **'rawdata.timing'** is 8 (i.e., more than 2 hours ago), then then **'smoking_delta_minutes'** is the midpoint of 120 minutes ago to X minutes ago, where X = $\frac{t_{current} - t_{past}}{60}$ and $t_{current}$ denotes the value of Aligned Time (**time_hrts'**) of the current EMA and t_{past} denotes the value of Aligned Time (**time_hrts**) of the EMA immediately preceding the current EMA. An example of this scenario is depicted in Figure 5.

Figure 5



- If $t_{current} - t_{past} > 24\ hours$ then set **'smoking_delta_minutes'** to a missing value, i.e., the rule described above is not applied. An example of this scenario is depicted in Figure 6.

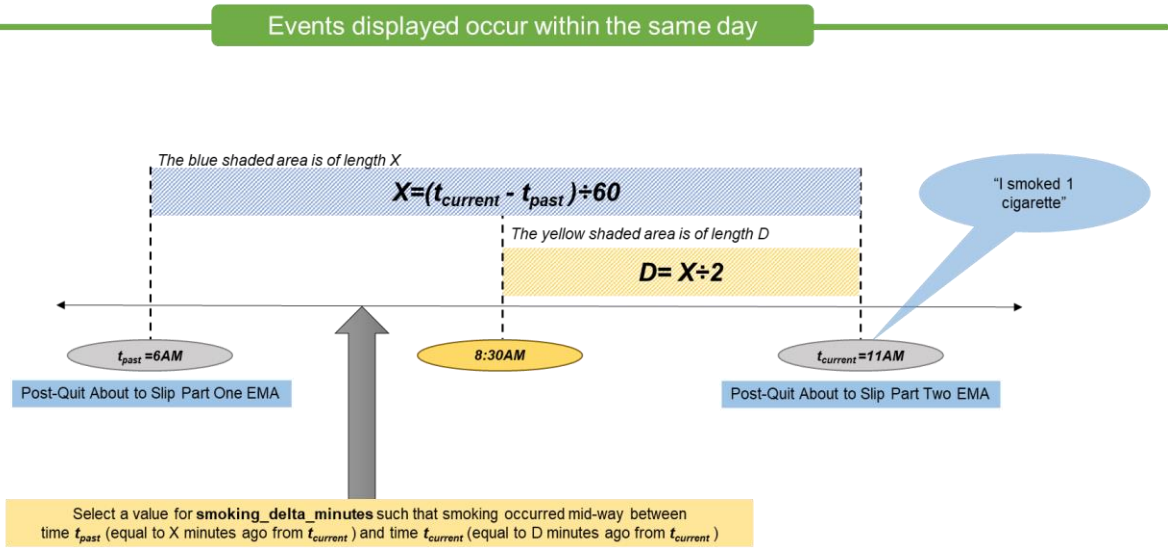
Figure 6



Note: Since Aligned Time is in seconds, X needs to be divided by 60.

4. Use the rule below to assign a value to **'smoking_delta_minutes'** when **'rawdata.timing'** is missing, determine the value of **'smoking_delta_minutes'** based on EMA type:
    - If EMA is a Pre-Quit Smoking Part Two or Post-Quit Already Slipped Part Two type EMA and their corresponding Part One assessments are available, then set **'smoking_delta_minutes'** to be halfway through the Part One and Part Two assessments. That is, $X = \frac{1}{2} \cdot \left(\frac{t_{current} - t_{past}}{60}\right)$. An example of this scenario is depicted in Figure 7.

Figure 7



- If EMA is a Pre-Quit Smoking Part Two type EMA and its corresponding Part One assessment is *not* available then set **'smoking_delta_minutes'** to be one-half the average value of $\frac{t_{current} - t_{past}}{60}$, calculated among all Pre-Quit Smoking Part Two EMAs which have their corresponding Part One EMAs.
- If EMA is a Post-Quit About to Slip Part Two type EMA and its corresponding Part One assessment is *not* available then set **'smoking_delta_minutes'** to be one-half the average value of $\frac{t_{current} - t_{past}}{60}$, calculated among all Post-Quit About to Slip Part Two EMAs which have their corresponding Part One EMAs.
- If EMA is a Pre-Quit Smoking Part One or Post-Quit Already Slipped Part One type EMA, then set **'smoking_delta_minutes'** to be a missing value.

When information between **'rawdata.timing'** & time elapsed between two consecutive EMA is inconsistent: If the midpoint obtained from the above calculation is greater than the time elapsed between the current EMA and the EMA immediately preceding the current EMA, then recalculate the midpoint to be half-way through the previous EMA and the current EMA. That is, $X = \frac{1}{2} \cdot \left(\frac{t_{current} - t_{past}}{60}\right)$.

When information between **'smoking_delta_minutes'** and **'smoking_qty'** is inconsistent: We use the rule below to determine **'smoking_delta_minutes'**

- If **'smoking_qty'**=0, **'smoking_delta_minutes'** will be set to a missing value

  Note: This scenario is applicable when participant reports 0 cigarettes smoked in a Group 2 variable and reports a value for a Group 3 variable.

- If **'smoking_qty'** is missing, **'smoking_delta_minutes'** will be set to a missing value

10. Timestamps provided in the curated datasets

Timestamps in the raw datasets: how they were used in the data curation process and assumptions regarding calculation of number of hours elapsed since a time-origin of interest
All timestamps in the raw datasets are only provided in human-readable format, and further, they were in the participant's local time (which was in Houston, Texas for all participants). To facilitate calculation of *number of hours elapsed since a time-origin of interest* (e.g., number of hours elapsed since Quit Time, number of hours elapsed since Start Study Time), numeric format timestamps were created.

We note that these numeric format timestamps will not be provided to end-users in the curated datasets but are described here for end-users who would like to probe deeper into the assumptions underlying the data curation process.

In the R code utilized to create the curated datasets, the creation of numeric format timestamps were a notable intermediate step in calculating the number of hours elapsed since a time-origin of interest (e.g., number of hours elapsed since the quit date).

- **Example 1.** If we found that a Random EMA was delivered more than 504 hours after 12AM on a participant's Quit Date (i.e., 21 days × 24 hours per day = 504 hours), then this particular EMA was excluded from all curated datasets (see Section 7.1)
- **Example 2.** If we found that a participant reported '0 cigarettes smoked' exactly at 12noon the day immediately after the first Post-Quit EMA, then we would represent this as a blue dot located (i.e., an EMA in which zero cigarettes were reported) at (location vis-à-vis horizontal-axis, location vis-à-vis vertical-axis)=(0.5, 0) in data visualizations used towards inferring a participant's Quit Date (see Table 10 in Section 8).

Process of constructing curated datasets: Method used to calculate number of hours elapsed since a time-origin of interest
In the code used to construct the curated datasets, numeric format timestamps are suffixed by "unixts". Specifically, these are the following time variables which are direct analogues of time variables introduced in previous sections suffixed by "hrts" (see Section 7.1).

- "start_study_unixts" (Start Study Time)
- "quit_unixts" (Quit Time)
- "end_study_unixts" (End Study Time)
- "delivered_unixts" (Delivered Time [in reference to a particular EMA])

- “begin_unixts” (Begin Time [in reference to a particular EMA])
- “end_unixts” (End Time [in reference to a particular EMA])
- “time_unixts” (Aligned Time [in reference to a particular EMA])

We display a sample R code snippet in the box below illustrating how the conversion between human-readable format and numeric format was implemented in the data curation process.

```
# Conversion from human-readable timestamp “2009-10-16 10:33:40” to numeric format timestamp
my_hrts <- as.POSIXct(strptime("2009-10-16 10:33:40",
                               format = "%Y-%m-%d %H:%M:%S",
                               tz = "UTC"))
my_unixts <- as.numeric(my_hrts)
print(my_unixts)
# The output will be 1255689220
# Now, let’s convert back to human-readable format
converted_time <- as.POSIXct(1255689220,
                             origin="1970-01-01",
                             tz="UTC")
print(converted_time)
# The output will be "2009-10-16 10:33:40 UTC"
```

A rationale for encoding human-readable format time as a number is that doing so may facilitate a smoother process of data-manipulation. For example, if t1 and t2 are two timestamps in numeric format, and t1 comes before t2 (e.g., t1 may be equal to 1610564920 and t2 may be equal to 1610591344), one can simply use the following formula to calculate number of hours elapsed: (t2 - t1) ÷ (60 × 60) – there are some nuances to this calculation that are discussed next.

An important takeaway of what we label as “unixts” variables is that all of them are calculated to represent the number of seconds elapsed since in January 1, 1970 in a participants’ local time zone, not in “UTC” time zone. Note that this is slightly different from a common computer science term of UNIX epoch time (which they are *not*), which is the equivalent but expressed in reference to “UTC”. With that said, in our R code that created this data (e.g., the code snippet displayed above), you might see that we used “UTC” as a reference in our calculated times, in hopes of better ensuring portability of code and reproducibility of curated datasets across computers (i.e., two computers can execute the same code and yield identical curated datasets as output).

A limitation of this method is that this method ignored changes to and from Daylight Savings Time (DST). In other words, calculation of number of hours elapsed since a time-origin of interest only accounted for the absolute passage of time, not hours elapsed as the participant may have perceived it to have been.

We continue the examples above to illustrate the potential influence of this limitation to the curated datasets:

- **Example 1 (continued).** If DST took effect (i.e., **from the participant's perspective,** their time zone shifted from UTC-6 to UTC-5) a few days after Quit Date (e.g., perhaps DST took effect 3 days after Quit Date), then from the participant's perspective, there will be 505 hours elapsed between Quit Date and 12AM on the 21st day after 12AM on a participant's Quit Date (instead of 504 hours).
- **Example 2 (continued).** If DST took effect (i.e., **from the participant's perspective,** their time zone shifted from UTC-6 to UTC-5) the day when the first Post-Quit EMA was delivered, then from the participant's perspective there will be 13 hours elapsed between 12AM and 12noon of that day. In this case, if we had accounted for DST, the blue dot should have been located at blue dot located at (location vis-à-vis horizontal-axis, location vis-à-vis vertical-axis)=(0.54, 0), instead of (0.5,0). Here, 0.54 = 12/24 + 1/24.

A short guide for end-users of curated datasets: use of time variables suffixed by "hrts"

If you plan to do any manipulation using the time variables, there are a few steps to take:

1) Read in (convert) the "hrts" variable to a variable that is stored as a datetime in your local programming language.
2) When doing any adding and subtracting of time in respect to the datetime, make sure to do so in a way that properly accounts for any changes to and from DST. In SAS, this can be accomplished by adding 'dtday' in the INTX function or in R, one can add days() via the R package *lubdridate* (Grolemund and Wickham, 2011) – this is demonstrated in the examples below.
3) Do visually or programmatically double-check that the date times were read in correctly and that math conducted on the variables works as intended.

Reading in and manipulating a datetime in R: Suppose a participant had a Random EMA launched on 2009-10-16 10:33:40 and we want to create a variable 3 weeks from that EMA. Converting "2009-10-16 10:33:40" to a datetime and then adding 3 weeks is accomplished as follows:

```
# Conversion from human-readable timestamp to numeric format timestamp

# You can also read in an "hrts" vector instead of a single date

my_dt <- as.POSIXct(strptime("2009-10-16 10:33:40",

                                    format = "%Y-%m-%d %H:%M:%S",

                                    tz = "America/Chicago"))

# Output will be "2009-10-16 10:33:40 CDT"

print(my_dt)

library(lubridate) #Load lubridate package

#Add a fixed number of days to a datetime

# Output will be "2009-11-06 10:33:40 CST"

my_dt+days(21)
```

Reading in and manipulating a datetime in SAS: let's repeat the same task in SAS

```
data temp;
      format my_dt my_dt_plus_3weeks datetime.; *initialize vars;
      my_dt = input("2009-10-16 10:33:40",anydtdtm.); *input datetime;
      put my_dt; *16OCT09:10:33:40;
      my_dt_plus_3weeks = INTNX('dtday',my_dt,21,'same'); *add 3 weeks;
      put my_dt_plus_3weeks; *06NOV09:10:33:40;
      *notice the time components are equal;
run;
```

11. References

1. Grolemund G, Wickham H (2011). "Dates and Times Made Easy with lubridate." Journal of Statistical Software, 40(3), 1–25. https://www.jstatsoft.org/v40/i03/