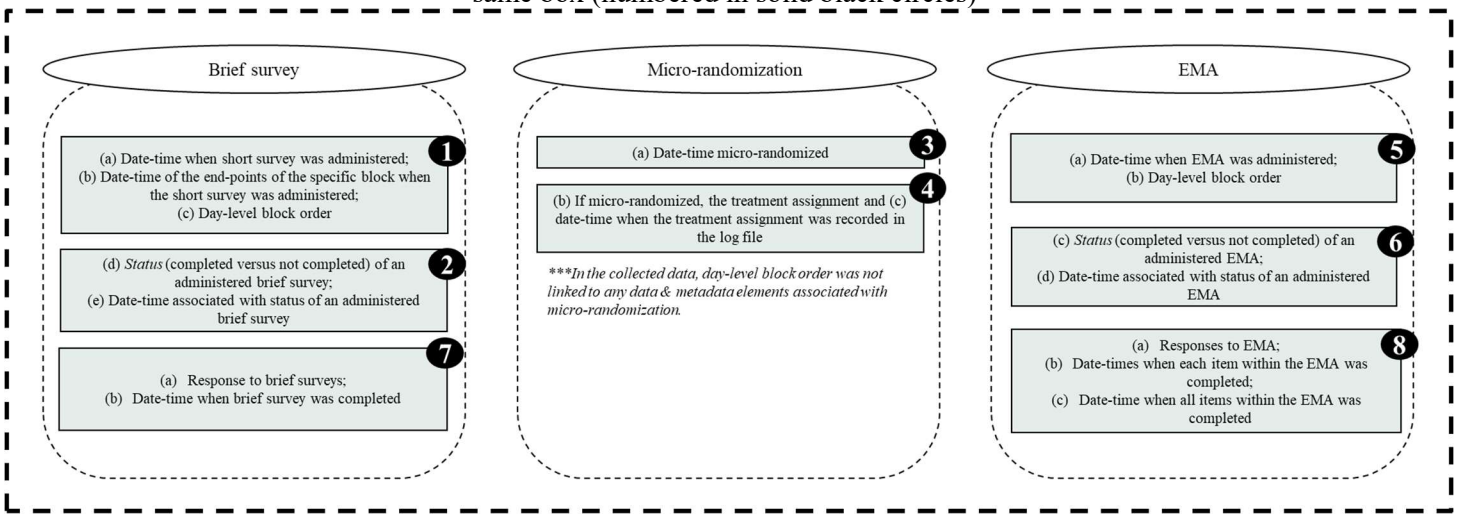
Supporting Information for
Mobile Assistance for Regulating Smoking Micro-Randomized Trial: Documentation on matching procedure

As a preliminary step to developing the matching procedure (record linkage algorithm), we mapped out which pieces of information had explicit links to each other in the collected data. The result of this step is displayed in Figure 1.

Figure 1: Distinct data & metadata elements in the collected data are depicted using small letters. Data & metadata elements that have an explicit link to each other in the collected data are depicted by grouping them together within the same box (numbered in solid black circles)



Takeaways from Figure 1 are summarized below:

- **Observation 1.** Data and metadata corresponding to brief surveys were distributed across three mutually exclusive groups: Box 1, 2, 7 in Figure 1. Explicit linkages only existed among data and metadata belonging to the same group.

- **Observation 2.** Data and metadata corresponding to micro-randomizations were distributed across two mutually exclusive groups: Box 3, 4 in Figure 1. Explicit linkages only existed among data and metadata belonging to the same group.

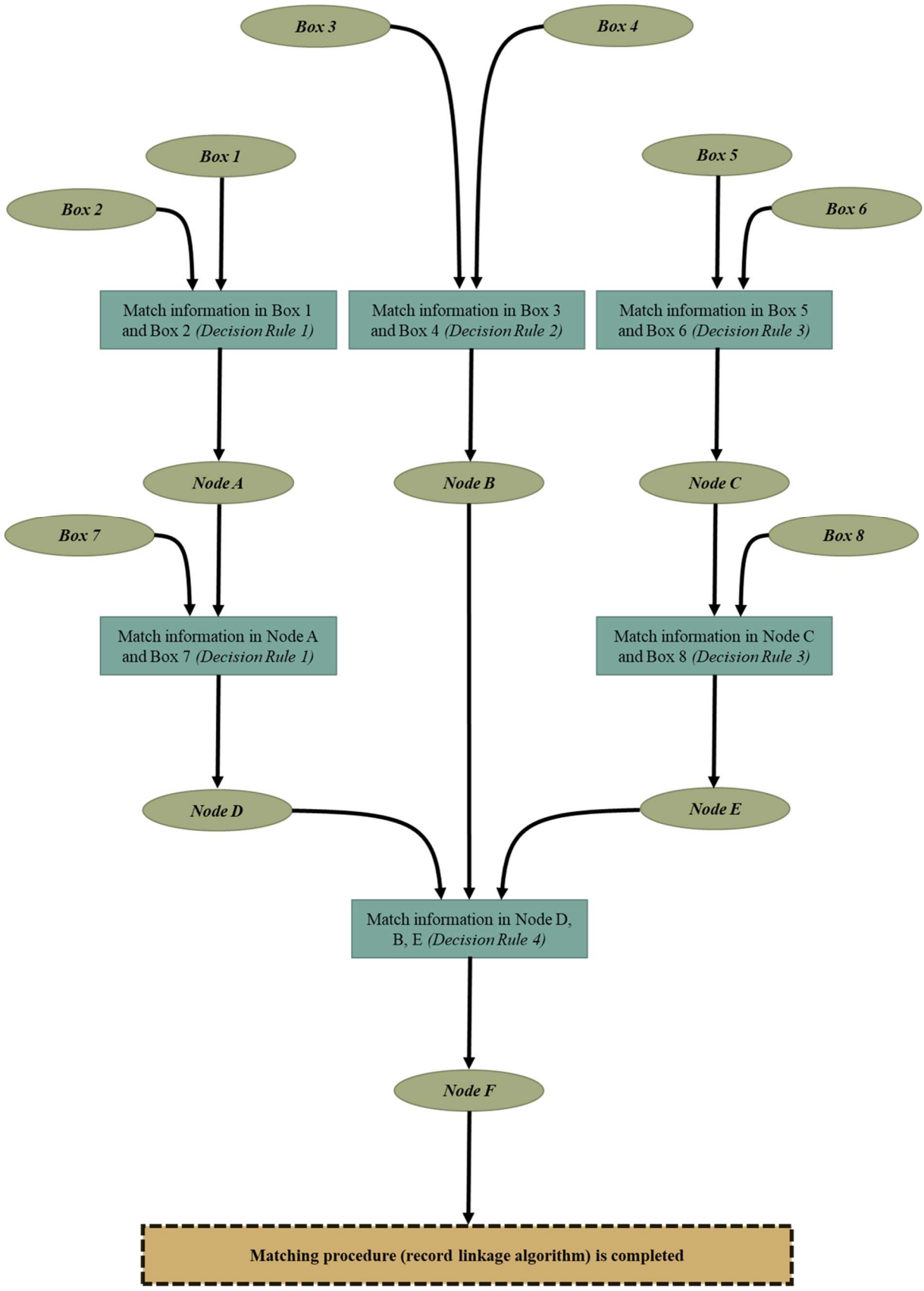
- **Observation 3.** Data and metadata corresponding to EMAs were distributed across three mutually exclusive groups: Box 5, 6, 8 in Figure 1. Explicit linkages only existed among data and metadata belonging to the same group.

- **Observation 4.** Day-level block order (i.e., whether the person-block was the 1st, 2nd, …, 6th block of the day) was present in the collected data and explicitly linked only to some of the metadata, specifically, date-times of brief survey administration and EMA administration (Figure 1, Box 1, 5). Data and metadata including treatment assignments, response to brief surveys, response to EMAs, and date-times of micro-randomization were not explicitly linked to day-level block order.

- **Observation 5.** Timestamps corresponding to when the ≈2-hour long person-blocks began and ended were present in the collected data and explicitly linked only to some of the metadata, specifically, date-times of brief survey administration and day-level block order (Figure 1, Box 1).

Using the mapping in Figure 1 as a starting point, we developed a matching procedure, displayed in Figure 2. The decision rules referred to in Figure 2 are provided below.

Figure 2: Overall logic of matching procedure (record linkage algorithm)



Decision rules used by matching procedure

Decision Rule 1

Rationale:
This rule mimics the programming logic that responding to individual items in the brief survey must come after a push notification administering the brief survey but before the status[1] of the brief survey is designated by the smoking cessation app.

Notation:
In Part 1:

- Let tj denote date-time associated with record j in Box 1 when short survey was administered
- Let J denote the total number of records in Box 1
- Let M denote the total number of records in Box 2
- Let $s_1, \ldots, s_M$ denote date-time associated with record $1, \ldots, M$ of Box 2 when short survey status was recorded
- Let k^* denote the specific record in Box 2 matched to record j in Box 1

In Part 2:

- Let r_p denote the date-time associated with record p of Node A when short survey status was recorded
- Let P denote the total number of records in Node A
- Let Q denote the total number of records in Node D
- Let $u_1, \ldots, u_Q$ denote date-time associated with record $1, \ldots, Q$ of Node D when short survey was completed
- Let k^{**} denote the specific record in Node F matched to record p in Node A

Decision Rule 1:
Part 1
For each participant i:
For each record j in Box 1:
Starting with the differences $\{s_1 - tj, \ldots, s_M - tj\}$, take those which are positive and call that set Θ
k^* is the record in Box 2 that corresponds to the minimum of the differences in the set Θ
Part 2
For each participant i:
For each record p in Node A:
Starting with the differences $\{r_p - u_1, \ldots, r_p - u_Q\}$, take those which are positive and call that set Ω
l^{**} is the record in Box 7 that corresponds to the minimum of the differences in the set Ω
If $u_{l^*} \geq t_{k^*}$
Then set $k^{**} = l^{**}$
Else (if $u_{l^*} < t_{k^*}$)
No record in Box 7 is matched to record p in Node A

[1] ***Status*** pertains to a classification designated by the smoking cessation app to each administered short survey and EMA. Classifications are ***completed*** where date-time simply corresponds to when the participant completed the short survey or EMA, ***missed, timed out***, and ***canceled***, where date-time corresponds to roughly 10 min or less after the short survey or EMA was administered. The latter three classifications could be thought of as various ways non-completion was more specifically characterized by the smoking cessation app. 'Missed' corresponds to the least engagement with brief survey or EMA completion (since participant did not physically interact with the push notification); 'timed out' and 'cancelled' corresponds to more engagement with brief survey and EMA completion (since participant physically interacted with the push notification).

Decision Rule 2

Rationale:
This rule mimics the programming logic that treatment assignment is revealed and saved in a database after micro-randomization.

Notation:

- Let t_j denote date-time associated with record j in Box 3 when micro-randomization occurred
- Let J denote the total number of records in Box 4
- Let $s_1, \ldots, s_J$ denote date-time associated with record $1, \ldots, J$ of Box 4 when treatment assignment was recorded
- Let k^* denote the specific record in Box 4 matched to record j in Box 3

Decision Rule 2:
For each participant i:
For each record j in Box 3:
Starting with the differences $\{s_1 - tj, \ldots, s_J - tj\}$ take those which are positive and call that set Θ
k^* is the record in Box 4 that corresponds to the minimum of the differences in the set Θ

Decision Rule 3

Rationale:
This rule mimics the programming logic that responding to individual items in an EMA must come after a push notification administering the EMA but before the status[2] of the EMA is designated by the smoking cessation app.

Notation:
In Part 1:

- Let t_j denote date-time associated with record j in Box 5 when EMA was administered
- Let J denote the total number of records in Box 5
- Let M denote the total number of records in Box 6
- Let $s_1, \ldots, s_M$ denote date-time associated with record $1, \ldots, M$ of Box 6 when EMA status was recorded
- Let k^* denote the specific record in Box 6 matched to record j in Box 5

In Part 2:

- Let r_p denote the date-time associated with record p of Node C when EMA status was recorded
- Let P denote the total number of records in Node C
- Let Q denote the total number of records in Box 8
- Let $u_1, \ldots, u_Q$ denote date-time associated with record $1, \ldots, Q$ of Box 8 when EMA was completed
- Let k^{**} denote the specific record in Node E matched to record p in Node C

Decision Rule 3:
Part 1
For each participant i:
For each record j in Box 5:
Starting with the differences $\{s_1 - tj, \ldots, s_M - tj\}$, take those which are positive and call that set Θ
k^* is the record in Box 6 that corresponds to the minimum of the differences in the set Θ
Part 2
For each participant i:
For each record p in Node C:
Starting with the differences $\{r_p - u_1, \ldots, r_p - u_Q\}$, take those which are positive and call that set Ω
l^{**} is the record in Box 8 that corresponds to the minimum of the differences in the set Ω
If $u_{l^*} \geq t_{k^*}$
Then set $k^{**} = l^{**}$
Else (if $u_{l^*} < t_{k^*}$)
No record in Box 8 is matched to record p in Node C

[2] See footnote 1

Decision Rule 4

Rationale:
This rule mimics the programming logic applied at the block-level: (a) if a brief survey had already been administered, another brief survey will not be administered unless a new block commences; (b) administration of EMA is contingent on both administration of brief survey (but not completion of brief survey) and micro-randomization having already occurred; (c) micro-randomization is contingent on administration of brief survey (but not on completion of brief survey) having already occurred.

Notation:
In Part 1:

- Let t_j denote date-time short survey was administered associated with record j in Node D
- Let t_{j+1} denote date-time short survey was administered associated with record $j+1$ in Node D
- Let J denote the total number of records in Node D
- Let M denote the total number of records in Node B
- Let $s_1, \ldots, s_M$ denote date-time short survey was administered associated with record $1, \ldots, M$ of Node B
- Let k^* denote the specific record in Node B matched to record j in Node D

In Part 2:

- Let t_j denote date-time short survey was administered associated with record j in Node D
- Let t_{j+1} denote date-time short survey was administered associated with record $j+1$ in Node D
- Let J denote the total number of records in Node D
- Let k^* denote the specific record in Node B matched to record j in Node D
- Let s_{k^*} denote the date-time micro-randomized associated with k^* obtained from Node B in Part 1
- Let N denote the total number of records in Node E
- Let $u_1, \ldots, u_N$ denote date-time EMA was administered associated with record $1, \ldots, N$ of Node E
- Let k^{**} denote the specific record in Node E matched to record j in Node D

Decision Rule 4:
Part 1
For each participant i:
Order records in Node D according to increasing values of date-time short survey was administered
For each record j in Node D:
k^* is the record in Node B such that s_{k^*} lies within the interval $[t_j, t_{j+1}]$
Part 2
For each participant i:
For each record j in Node D:
k^{**} is the record in Node E such that $u_{k^{**}}$ lies within the interval $[s_{k^*}, t_{j+1}]$

More information (does not affect the validity of Decision Rule 4):
It is possible for blocks hinged on the current day to cross over[3] to the following day if the first block of the current day was scheduled to begin later in the day (specifically, after 10am at the participant's local time). Programming logic obeyed by the smoking cessation app dictated that when a particular block in the current day crosses over, two question surveys, micro-randomization, and EMAs are automatically forfeited (i.e., brief survey will never be administered, and subsequently, micro-randomization will never occur and EMA will never be administered), not only for the particular block that had crossed over, but also for any remaining blocks. When this happens, no data & metadata associated with these blocks that crossed over will appear in Boxes 1-8 of Figure 1 (e.g., since no brief survey and EMA will be administered, it follows that the corresponding date-times do not exist).

[3] The specific definition for 'cross over' we use in this appendix is the following: we say that a block hinged on the current day ***crosses over*** to the next day if (1) it is not the 1st block, i.e., it is the 2nd, …, 6th block; (2) it began after 12:00am on the next day; (3) the 1st block began on the current day. This definition encompasses exceptions to the general case described above. Specifically, blocks that began on the current day but ended on the next day do not count as having crossed over to the next day. In this special case, the smoking cessation app is programmed to permit administration of brief surveys, micro-randomization, and administration of EMA within the confines of the block even if they occur on the next day. As before, brief surveys, micro-randomizations, and EMAs of any remaining blocks will automatically be forfeited. However, administration of brief surveys on the next day will not be permitted to begin until all events associated with the block that crossed over have been completed.