# Package: public.ctn0094extra (via r-universe)

September 18, 2024

**Title** Helper Files for the CTN-0094 Relational Database

**Version** 1.0.4

**Date** 2023-11-21

**Description** Engineered features and ``helper'' functions ancillary to the 'public.ctn0094data' package, extending this package for ease of use (see <https://CRAN.R-project.org/package=public.ctn0094data>). This 'public.ctn0094data' package contains harmonized datasets from some of the National Institute of Drug Abuse's Clinical Trials Network (NIDA's CTN) projects. Specifically, the CTN-0094 project is to harmonize and de-identify clinical trials data from the CTN-0027, CTN-0030, and CTN-51 studies for opioid use disorder. This current version is built from 'public.ctn0094data' v. 1.0.6.

**License** MIT + file LICENSE

**Encoding** UTF-8

**Language** en-US

**LazyData** true

**Depends** R (>= 2.10), public.ctn0094data

**Imports** dplyr, magrittr, purrr, tibble, tidyr, utils

**Suggests** knitr, kableExtra, rmarkdown, scales, stringr

**Roxygen** list(markdown = TRUE)

**RoxygenNote** 7.2.3

**VignetteBuilder** knitr

**URL** https://ctn-0094.github.io/public.ctn0094extra/

**BugReports** https://github.com/CTN-0094/public.ctn0094extra/issues

**Repository** https://ctn-0094.r-universe.dev

**RemoteUrl** https://github.com/ctn-0094/public.ctn0094extra

**RemoteRef** HEAD

**RemoteSha** 236092d90141de283d655212c20d1c9359cb8ec5

# Contents



---

`CreateCTN30ProtocolHistory`
*Create a Subject History Table by Protocol*

---

## Description

Create a Subject History Table by Protocol

## Usage

```
CreateCTN30ProtocolHistory(
  randCTN30_df,
  start_int = -30,
  phase1Len_int = 98,
  phase2Len_int = 168
)
```

## Arguments

| | |
|---|---|
| `randCTN30_df` | A data frame of subject randomization days for CTN-0030. This data will have columns for `who`, `which`, and `when`. |
| `start_int` | When should the protocol timeline start? Defaults to 30 days before consent (allowing for self-reported drug use via Timeline Follow- Back data). |
| `phase1Len_int` | When should Phase I end per protocol? Defaults to 98 days (14 weeks). |
| `phase2Len_int` | When should Phase II end per protocol? Defaults to 168 days (24 weeks). |

## Details

We may want to perform SQL-like operations on a set of tables. This data table will form the "backbone" for future join operations. It creates one record per person in a study for each day in that study, and it is designed to work with subject-specific two-arm protocols (where Phase I and Phase II treatments could start on different days for each subject). For subjects who consented to treatment but were never randomized, we use an "intent to treat" philosophy and assign them an empty protocol timeline starting on the day of consent (day 0) until `phase1Len_int`.

NOTE: We expect this function to be used specifically to create a potential visit backbone for CTN-0030. For studies with fixed protocol lengths, please use `CreateSubjectHistory()` instead.

## Value

A tibble with columns `who` and `when`. Each subject will have one row for each day in the study range.

## Examples

```
# Subject A started Phase I on day 2 and was switched from treatment
#   Phase I to Phase II on day 45. Subject B started Phase I on day 6 and
#   was never switched to Phase II (either because they dropped out of the
#   study or because the treatment given to them in Phase I worked).
rand_df <- data.frame(
  who   = c("A", "A", "B", "B"),
  which = c(  1,   2,   1,   2),
  when  = c(  2,  45,   6,  NA)
)

# Based on this example data above, we expect to see potential contact
#   days per the protocol for subject A to range from -30 to 45 + 168.
#   For subject B, we expect this range to be from -30 to 6 + 98
CreateCTN30ProtocolHistory(rand_df)
```

---

`CreateHistory` *Create a Subject History Table*

---

## Description

Create a Subject History Table

## Usage

```
CreateHistory(
  rawData_ls,
  personsTable = "everybody",
  firstDay_int = NULL,
  lastDay_int = NULL
)
```

### Arguments

| | |
|---|---|
| rawData_ls | a list of tibbles returned by loadRawData |
| personsTable | What is the name of the data table that contains all the subject IDs? Defaults to "everybody". If no such table exists, then set this value to "none" and the unique subject IDs will be drawn from all tables in rawData_ls |
| firstDay_int | (OPTIONAL) What should be the first "day" counter for all subjects? This may be beneficial to set if you wish to have the History ignore all days before a certain point. |
| lastDay_int | (OPTIONAL) What should be the last "day" counter for all subjects? |

### Details

We may want to perform SQL-like operations on a set of tables. This data table will form the "backbone" for future join operations. It creates one record per person for each day in the study. The default behavior is to set the date range to the smallest observed value for "when" across all data tables to the largest observed value for "when" across all tables, inclusive. Necessarily, this is *very* sensitive to outliers or coding errors (for instance, if one subject has a "first day" 200 days before Study Day 0, then this History table will include days -200 to -1 for ALL subjects, regardless of any recorded contact in this period).

If this behavior is not desirable, you must specify a study range. For example, you may have a short recruitment period followed by a 6-month clinical trial. In this instance, you may want to ignore any data more than 30 days prior to Study Day 0 for each subject or more than a few weeks after the end of the trial. Thus, you would set `firstDay_int = -30L` and `lastDay_int = 6 * 30 + 14`.

### Value

A tibble with columns `who` and `when`. Each subject will have one row for each day in the study range.

### Examples

```
data_ls <- loadRawData(c("tlfb", "all_drugs", "everybody"))
CreateHistory(rawData_ls = data_ls, firstDay_int = -30)
```

---

## CreateProtocolHistory *Create a Subject History Table by Protocol*

---

### Description

Create a Subject History Table by Protocol

### Usage

```
CreateProtocolHistory(start_vec, end_vec, persons_df = "everybody")
```

### Arguments

| | |
|---|---|
| start_vec | a named integer vector with the number of days before subject consent when the subject history should start, per protocol |
| end_vec | a named integer vector with the length of the study phase of interest, per protocol |
| persons_df | Either the name of the data frame that contains all the subject IDs and their clinical trials (which defaults to "everybody"), or a data frame with this information. See "Details" for more information. |

### Details

We may want to perform SQL-like operations on a set of tables. This data table will form the "backbone" for future join operations. It creates one record per person in each study for each day in those studies (when persons_df is set to "everybody"), or it creates one record per person contained in the table persons_df for each day in those studies. The default behavior is to use the supplied "everybody" table for all consenting subjects in the CTN-0027, CTN-0030, and CTN-0051 clinical trials. However, users may only care about a smaller subset of these patients, so a subset of the "everybody" data frame can be supplied to the persons_df argument if desired.

NOTE: this function is only appropriate for trial with fixed start and end days (such as CTN-0027 or CTN-0051). For studies with variable-length (i.e., subject-specific) protocol lengths, please use CreateSubjectProtocolHistory() instead.

### Value

A tibble with columns who, project, and when. Each subject will have one row for each day in the study range.

### Examples

```
start_int <- c(`27` = -30L, `51` = -30L)
end_int   <- c(`27` = 168L, `51` = 168L)

CreateProtocolHistory(
  start_vec = start_int, end_vec = end_int
)
```

---

## derived_inductDelay *Derived Induction Delay Data*

---

### Description

This data set measures the number of days from a participant's randomization until they received their first dose of study drug.

### Usage

```
data(derived_inductDelay)
```

## Format

A tibble with 2,492 rows and 3 variables:

**who** Patient ID

**treatment** What treatment is prescribed: "Inpatient BUP", "Inpatient NR-NTX", "Methadone", "Outpatient BUP", "Outpatient BUP + EMM", "Outpatient BUP + SMM"

**inductDelay** How many days after being assigned to a treatment arm did the participant receive their first dose of study drug? Missing values indicate that the subject never received their first dose.

## Details

This data set is a derived data set. The inputs are the treatment and randomization data sets. The code to calculate the induction delay is given in `scripts/create_inductDelay_20210729.R`. The treatment arm is also included in this data set because the induction delay depends on the type of treatment. For example, inpatient treatment arms may have different protocols than outpatient treatment arms for determining how many days the subject must wait after randomization before treatment.

---

# `derived_raceEthnicity` *Derived Patient Race and Ethnicity Data*

---

## Description

Summarize the patients' self-reported race and ethnicity into four groups: "Non-Hispanic White", "Non-Hispanic Black", "Hispanic", and "Other".

## Usage

```
data(derived_raceEthnicity)
```

## Format

A tibble with 3,560 rows and columns:

**who** Patient ID

**race** Self-reported race. Options are "White", "Black", "Other", and "Refused/missing".

**is_hispanic** Self-reported Hispanic/Latino ethnicity.

**race_ethnicity** Derived composite marker of race and ethnicity.

## Details

This data set contains a summary of self-reported race and ethnicity in four levels. Of note, the "Other" category includes 2 participants who marked "no" to the question of Hispanic/Latino ethnicity but refused to answer their race. Also, the "Other" category includes 33 participants for whom all information about race and ethnicity is missing. This data set is a derived data set; the script used to create it is `"scripts/create_raceEthnicity_20220816.R"`.

## derived_visitImputed *Imputed Patient Visit Data*

### Description

Given a series of weekly clinic visits described per protocol, this data marks subjects as present or missing.

### Usage

```
data(derived_visitImputed)
```

### Format

A tibble with 87,891 rows and columns:

**who** Patient ID

**when** Study day

**visitImputed** Marked as `"Present"` if the subject visited the clinic on that day, or `"Missing"` if the subject did not visit the clinic on a day we would have expected them to (based on regular weekly visits).

### Details

This contains planned visits. Not all appointments were kept. We indicate if an appointment was kept on a certain day by marking the subject as `"Present"` on that day. If the subject goes more than 7 days without a clinic visit, we mark the subject as `"Missing"` on days that are multiples of 7 from the randomization day. For subjects without a randomization day, weekly visits after day of consent are marked as `"Missing"` instead. This data set is a derived data set; the script used to create it is `"scripts/create_visitImputed_20210909.R"`.

NOTE: because our window is a strict weekly window, a subject who shows up for their clinic visit one or more days late will still be marked as missing on the day they were supposed to appear. This means that some subjects will be marked as having missed their weekly clinic visit on one day, but be present in the clinic the next.

## derived_weeklyOpioidPattern *Patient UDS Opioid Weekly Pattern Data*

### Description

Show the pattern of positive, negative, and missing urine drug screen (UDS) results for opioids by patient over the study protocol. Study "Week 1" starts the day after randomization (for patients who were randomized) or the day after signed consent (for patients who were not randomized).

## Usage

```
data(derived_weeklyOpioidPattern)
```

## Format

A tibble with 3,560 rows and columns:

**who** Patient ID

**startWeek** The start of the "word" is how many weeks before randomization? This should be -4 for most people, but can be as high as -8. Note that week 0 is included, so a value of -4 represents data in the 5th week before randomization; that is, 29-35 days prior to randomization. Most subjects have timeline follow-back data 30 days before consent, and delays between consent and randomization are common.

**randWeek1** Week of first randomization (1, if randomized; NA if not)

**randWeek2** Week of second randomization (only for CTN-0030)

**endWeek** The end of the "word" is how many weeks after randomization? This depends on the study protocol, but should be close to 16 or 24 weeks for most subjects.

**Baseline** A character string of symbols from `startWeek` to the last week before `randWeek1`. Symbols are as defined in `Phase_1`.

**Phase_1** A character string of symbols from `randWeek1` to `endWeek` (for subjects from CTN-0027 and CTN-0051) or the last week before `randWeek2` (for CTN-0030). These symbols are: `"+"` = the subject's UDS showed presence of an opioid in that week; `"-"` = the subject's UDS showed absence of an opioid in that week; `"*"` = two or more UDS in the same week, where >= 1 UDS was positive and >= 1 UDS was negative; `"o"` if the subject was supposed to visit a clinic per the study protocol but did not visit or did not complete a UDS screen; and `"_"` to represent weeks wherein the subject was not scheduled to provide UDS.

**Phase_2** A character string of symbols from `randWeek2` to `endWeek` for subjects from CTN-0030 only. Symbols are as defined in `Phase_1`.

## Details

This data set contains a "word" describing weekly non-study opioid use patterns as measured by UDS. Based on the substances screened in this data set, our list of substances classified as an opioid is: Oxymorphone, Opium, Fentanyl, Hydromorphone, Codeine, Suboxone, Tramadol, Morphine, Buprenorphine, Hydrocodone, Opioid, Methadone, Oxycodone, and Heroin. UDS results indicating the presence of one or more of these substances will be marked with `"+"` for that week. UDS results negative for these substances will be marked with `"-"`. If, by study protocol, the subject was supposed to visit a clinic to complete a UDS but they did not, then the visit for that week will be marked with `"o"`. If, by study protocol, the subject was NOT supposed to visit a clinic to complete a UDS for a given week, then visit for that week will be marked with `"_"`. This data set is a derived data set; the script used to create it is `"scripts/create_weeklyOpioidPattern_20211123.R"`.

NOTE: because our window is a strict weekly window, a subject could have both positive and negative UDS within the same 7-day period. In this case, the week is marked as `"*"`. Depending on the definition of treatment failure or treatment success desired, these dual-status indicators can be re-coded to `"+"` or `"-"` as appropriate. Also, some studies include a baseline UDS in the week of consent (before the subject was randomized to a treatment arm). Some subjects were randomized

in the same week as the week wherein consent was signed, while other subjects were randomized weeks later. We represent the weeks before consent and the variable number of weeks between consent and randomization with "_" if there were no baseline UDS visits in those weeks (this represents the pre-study period). For subjects who were never randomized, the weeks before consent are also marked as pre-study ("_"), and all subsequent protocol weeks are marked as missing.

---

`derived_weeklyTLFBPattern`

*Patient TLFB Opioid Weekly Pattern Data*

---

## Description

Show the pattern of positive and negative patient self-report (timeline follow-back, TLFB) results for opioids by patient over the study protocol. Study "Week 1" starts the day after randomization (for patients who were randomized) or the day after signed consent (for patients who were not randomized).

## Usage

```
data(derived_weeklyTLFBPattern)
```

## Format

A tibble with 3,560 rows and columns:

**who** Patient ID

**startWeek** The start of the "word" is how many weeks before randomization? This should be -4 for most people, but can be as high as -8. Note that week 0 is included, so a value of -4 represents data in the 5th week before randomization; that is, 29-35 days prior to randomization. Most subjects have timeline follow-back data 30 days before consent, and delays between consent and randomization are common.

**randWeek1** Week of first randomization (1, if randomized; NA if not)

**randWeek2** Week of second randomization (only for CTN-0030)

**endWeek** The end of the "word" is how many weeks after randomization? This depends on the study protocol, but should be close to 16 or 24 weeks for most subjects.

**Baseline** A character string of symbols from `startWeek` to the last week before `randWeek1`. Symbols are as defined in `Phase_1`.

**Phase_1** A character string of symbols from `randWeek1` to `endWeek` (for subjects from CTN-0027 and CTN-0051) or the last week before `randWeek2` (for CTN-0030). These symbols are: `"+"` = the subject reported the use of an opioid for two or more days in that week; `"-"` = the subject did not report use of an opioid in that week; `"*"` = the subject reported one day of opioid use in that week; `"o"` if the subject was supposed to report TLFB but did not; and `"_"` to represent weeks wherein the subject was not scheduled to provide TLFB (for example, more than 30 days before randomization).

**Phase_2** A character string of symbols from `randWeek2` to `endWeek` for subjects from CTN-0030 only. Symbols are as defined in `Phase_1`.

### Details

This data set contains a "word" describing weekly non-study opioid use patterns as reported by the subject in TLFB. Based on the substances reported by the subjects, our list of substances classified as an opioid is: non-study Buprenorphine, non-study Methadone, heroin, and "opioids" (which includes Oxymorphone, Opium, Fentanyl, Hydromorphone, Codeine, Suboxone, Tramadol, Morphine, Hydrocodone, and Oxycodone). TLFB reporting indicating the presence of two or more use days of these substances will be marked with `"+"` for that week. TLFB results positive for these substances for 0 days will be marked with `"-"`. TLFB results positive for these substances for 1 day in the week will be marked with `"*"` for that week. This data set is a derived data set; the script used to create it is `"scripts/create_weeklyTLFBPattern_20220511.R"`.

NOTE: some studies collected more TLFB data than others. Also, all times are marked starting with the week of randomization. We represent the weeks before randomization with `"_"` if no TLFB data was collected. For subjects who were never randomized, all subsequent protocol weeks are marked as missing (`"o"`).

---

## loadRawData *Load Data Sets into a List*

---

### Description

Load Data Sets into a List

### Usage

```
loadRawData(dataNames_char)
```

### Arguments

| | |
|---|---|
| `dataNames_char` | Names of data sets to load |

### Details

We may want to perform SQL-like operations on a set of tables without loading each table into R's Global Environment separately. This function loads these data sets into a self-destructing environment and then returns a named list of these data sets.

### Value

Loads data sets specified into the current function environment for further evaluation (unused) and then returns these data sets as a named list

### Examples

```
loadRawData(c("tlfb", "all_drugs"))
```

MarkMissing *Code Empty Visit Values as "Missing" as Appropriate*

## Description

Given a complete timeline of potential subject visits per study protocol, mark certain visits as "Missing"

## Usage

```
MarkMissing(timeline_df, windowWidth = 7, daysGrace = 0)
```

## Arguments

| | |
|---|---|
| timeline_df | A data frame with columns who, when, visit and randomized. This data frame measures on which days the subjects visited the clinic (visit) and indicates when the subjects were randomized to Phase I of their respective studies (the randomized column). This data set will contain one (and only one) record per subject per day; and enough rows to cover all potential visits per the protocol length of the study. |
| windowWidth | How many days are expected between clinic visits? Defaults to 7, representing weekly clinic visits. |
| daysGrace | How many days late are subjects allowed to be for their weekly visit. Defaults to 0. Under this default behavior with weekly visits, a subject who visits the clinic on days 8 and 14 instead of days 7 and 14 will have a missing visit imputed for day 7. |

## Details

Most definitions of opioid use disorder treatment success or failure partially depend on a tally of the number of missed clinic visits. For example, a definition of early treatment failure could be "3 or more UDS positive for non-study opioids or missing visits within the first 28 days of randomization". Given a table of subject visits by day over the entire protocol timeline, this function will estimate when each subject missed a clinic visit (unfortunately, missed visits can often be improperly recorded in the patient logs; if such information is complete, using this function is unnecessary).

This estimation is conducted as follows: (1) first, for each subject, a regular grid of days is spread from the randomization day to the end of treatment by windowWidth; (2) next, we iterate over each day in this regular grid, and at each step we check the next windowWidth plus daysGrace days for a visit in that range, and we mark the day at the end of the window as "missing" if there are no visits in that range; (3) and finally, we combine these subject-specific data tables.

## Value

A copy of timeline_df with the column visitYM added. This column is a copy of the visit column with additional cells marking if a subject should have attended the clinic but did not.

## Examples

```
# TO DO
```

---

`MarkUse` *Mark Use Day by Subject*

---

## Description

Mark Use Day by Subject

## Usage

```
MarkUse(
  targetDrugs_char,
  drugs_df = NULL,
  reportSource = c("TFB", "UDSAB", "UDS"),
  retainEmptyRows = FALSE
)
```

## Arguments

| | |
|---|---|
| `targetDrugs_char` | A character vector including which drugs should be counted against the subject |
| `drugs_df` | A data frame with columns `who`, `when`, and `what`. This data frame measures which drugs were used by each subject over all days of treatment. This data set must also include a column `source`, which marks from which reporting source the drug use was recorded |
| `reportSource` | A character vector matching the source of the reported drug use. The options must be from Timeline Followback (`"TFB"`) questionnaires or daily urine drug screens (`"UDS"` or`"UDSAB"`). |
| `retainEmptyRows` | A logical flag to force rows for participants who did not have UDS positive for the substances listed in `targetDrugs_char` to be retained in the final results (with `NA` for `"when"` and `"source"`). Defaults to `FALSE` because the entire point of this function is to mark substance *USE*, not a lack thereof; however, this flag is needed for the vignette (because we forced the inclusion of a participant with no recorded UDS for pedagogical purposes). |

## Details

This function is basically just a fancy wrapper around some dplyr code. We just don't want the user to have to 1) know dplyr, or 2) write the code themselves.

## Value

A modification of the `drugs_df` data set: the columns are `"who"`, `"when"`, and `"source"`; each row corresponds to one use day per subject per use source (if, for instance, there is drug use for a particular day recorded in both TFB and UDS, then that day will have two rows in the resulting data set).

## Examples

```
MarkUse(c("Crack", "Pcp", "Opioid"))
```

# Index