An n-Ary Word Sufficient Statistic for Sequential Univariate Categorical Values

Gabriel Odom, Laura Brandt, Ray Balise, and the CTN-0094 Team

2022-09-09

# Introduction and Motivation

## Background

Over 750,000 Americans have died from a drug overdose [since 1990](https://www.hhs.gov/opioids/about-the-epidemic/opioid-crisis-statistics/index.html) (HHS 2018). Even worse, after the onset of the COVID-19 pandemic, millions of Americans have self-reported [increases in their substance use](https://www.samhsa.gov/newsroom/press-announcements/202110260320) “a little more or much more” (SAMHSA 2021), and over 40 million Americans [needed treatment](https://zinniahealth.com/research/substance-use-disorder-treatment-by-state) for a substance use disorder (SUD) in 2020 (Liebhaber 2022). SUDs are treated with a [combination of pharmacological and psychological interventions](https://www.samhsa.gov/medication-assisted-treatment), known as medication assisted treatment (SAMHSA 2022). When evaluating the efficacy of new medication-assisted treatments for SUDs, clinicians require trial participants to provide frequent urine samples, known as urine drug screenings (UDS), which are tested for the presence of substances of misuse.

For a single substance or group of substances, outcomes of these tests are uni-dimensional and categorical. Common UDS categories include “substance positive”, “substance negative”, “improper urine sample temperature”, “missing urine”, and others. Further, because these tests are given sequentially over the course of a clinical trial, the tests are correlated within subject and across study week. Because of these properties, single-value summaries of the UDS pattern (such as the proportion of “substance positive” results, or the maximum number of consecutive “substance negative” results) are not *sufficient statistics* (they do not contain all of the information about each trial participants’ UDS pattern that was contained in the original UDS data) (Fisher and Russell 1922).

## Overview of this Paper

In this paper, we extend the concept of a binary word to multiple categories. We show that this construct sufficiently collapses longitudinal records into a single, compact “word”. Also, we give a few example algorithms to calculate some common single-value summaries from such a “word”. Finally, we introduce a standard and code-based library of algorithms (contained in the CTNote package (Odom, Brandt, and Balise 2022) for the R Computing Language (R Core Team 2022)) for treatment outcome definitions useful to evaluate medication-based treatments for SUDs.

# Methods

## “Words” as Sufficient Statistics for Longitudinal Categorical Data

We aim to create a sequence of letters and symbols that act as a sufficient statistic for a categorical random variable observed in a longitudinal manner. Our motivating example is a compact and sufficient representation of a subject’s weekly UDS results. For laboratory tests which detect a substance of interest in the urine, this statistic is a compact representation of the full pattern of substance use for an individual participant. To be of use, this summary statistic must have the following properties:

1. (*Machine-Readable*) It can be directly [parsed by a computer](https://opendatahandbook.org/glossary/en/terms/machine-readable/) (Handbook 2022).
2. (*Parsimonious*) It can be easily interpreted by a human.
3. (*Sufficient*) It represents all of the same information about the sequence of interest that would be present in the full data.

## Binary and -ary “Words”

[Binary words](https://www.oxfordreference.com/view/10.1093/acref/9780199235940.001.0001/acref-9780199235940-e-293) are a compact representation of a sequence of logical or binary variables (Clapham and Nicholson 2009). For example, let the random variable indicate whether or not a clinical trial participant visited their clinic in each week over a four-week period, and assume that we observe the pattern: “visit”, “visit”, “no visit”, “visit”. If “1” represents a visit and “0” represents no recorded visit, then we can represent this clinic visit pattern as the binary word “1101”. Similarly, we could represent these two categories symbolically as “V” for visit and “\_” for no visit, so our binary word becomes “VV\_V”. Notice that for a short sequence like our 4-week example here, the original data itself is both machine readable and human readable. However, parsimony (human readability) begins to suffer as sequences grow longer.

Extensions of this concept have been [famously applied to the human genome](https://www.ncbi.nlm.nih.gov/Class/MLACourse/Original8Hour/Genetics/basepair.html) (Geer and Morrison 1999). The four nitrogen bases in DNA are abbreviated by their first letter: A for adenine, T for thymine, G for guanine, and C for cytosine. These four letters are collapsed by their position in the genome into a *quaternary* word (or a *quinary* word if U is included for uracil). For example, the “word” AAACCATTCACAATCAGACA expresses a sequence of 20 nucleic acid bases without loss of information about the bases (the “word” is *sufficient*). The “word” AAACCATTCACAATCAGACA is also much easier to read by a human than the condensed structural formula or a skeletal-structural formula of these 20 compounds (the “word” is *parsimonious*). Furthermore, such a quaternary word can be parsed by a computer with ease (the “word” is *machine readable*).

## Quinary UDS Pattern Legend

We now build from the examples above. In order to compactly summarize a participant’s pattern of substance use (for a particular class of substances) over time, we first define the following five-value legend:

* **+**: positive for the substance(s) in a specified window of time (a day, week, month, etc.) by urine screen (or participant self report, if such data are of interest)
* **–**: negative for the substance(s)
* **o**: subject failed to provide a urine sample
* \*: inconclusive results or mixed results (e.g. subject provided more than one urine sample in the time interval and they did not agree)
* \_: no specimens required (weekends, holidays, pre-randomization period, alternating visit days/weeks)

Obviously this legend can be modified and extended to better reflect the complexities of individual clinical trials, but we believe that the structure (representing visits / short time intervals as single symbols) will still be beneficial as a quick “snapshot” of the patient’s treatment results and adherence to trial protocol, and also as a computable summary of the subject’s behavior during the follow-up period and post-trial analysis phase.

## A Use Pattern Example

We will first observe the recorded opioid use data via urine screen for a subject after randomization.

who

what

source

when

2089

Opioid

UDSAB

0

2089

Opioid

UDSAB

7

2089

Opioid

UDSAB

15

2089

Opioid

UDSAB

21

2089

Opioid

UDSAB

28

2089

Opioid

UDSAB

56

While we observe the study days that this participant used opioids throughout the clinical trial, it’s difficult visualize the patient’s treatment trajectory using the data in this form. On the one hand, after some preliminary data cleaning, this data can be parsed by a computer (*machine readability*). Additionally, this table represents all of the opioid use for this subject (*sufficiency*). However, data in this form is **not** easy to directly interpret by a clinician or researcher. Can we immediately derive clinical intuition about this patient? Does the treatment appear to be effective?

In contrast, observe the opioid use pattern summary for this same participant: ++++---+--------------o-. Using the basic pattern definition legend above, we can clearly see that this subject had a challenging first month (++++), improved in the second month (---+), and then remained abstinent from opioids for the remainder of the clinical trial (--------------o-). Participant substance use pattern data in this form *is* easy to interpret by a clinician. Moreover, this symbolic representation of substance use can be parsed and summarized by a computer. Finally, notice that this pattern represents the entire course of treatment for this subject without loss of any clinically relevant information about their opioid misuse.

# Results

We mentioned above that these use pattern summaries are “machine readable” (that a computer can parse them). The software package CTNote:: is one such tool to enable computers to parse these use patterns (Odom, Brandt, and Balise 2022). This software package contains the following groups of routines (also known as functions):

* Handle missing UDS data with recode\_missing\_visits() and impute\_missing\_visits()
* Account for the study observation design or trial visit protocol with collapse\_lattice() and view\_by\_lattice()
* Detect a substance use pattern with detect\_subpattern() and detect\_in\_window()
* Measure the longest periods of consecutive behavior with measure\_retention() and measure\_abstinence\_period()
* Count the substance use events with count\_matches()

By executing various combinations of these routines, we were able write algorithms to define many common treatment outcome definitions used in SUD clinical trials. In our supplemental material, we include a “library” of algorithms to calculate each of these definitions. The input of each algorithm is a set of substance use pattern summaries (as a “word”) for all clinical trial participants; the output of each algorithm is a calculated treatment endpoint for each participant included in the input.

## Example: Abstinence Outcome

One of the [Schottenfeld et al. (2008)](https://doi.org/10.1016/s0140-6736(08)60954-x) treatment endpoint definitions is “the maximum consecutive days abstinent” from opioids, where missing clinic visits are imputed to represent a urine screen positive for the substance of interest (Schottenfeld, Chawarski, and Mazlan 2008). Therefore, the algorithm to calculate this definition for our example subject would be:

# The participant's use pattern summary; the %>% symbol is read "and then"  
"++++---+--------------o-" %>%   
 # change all missing visits to positive  
 recode\_missing\_visits(missing\_becomes = "+") %>%   
 # find the length of the longest abstinent period, in weeks  
 measure\_abstinence\_period()

## Example: Use Reduction Outcome

The [Haight et al. (2019)](https://doi.org/10.1016/S0140-6736(18)32259-1) treatment endpoint definition is the “percentage of negative UOS [urine opioid screen] from week 5 to week 24” (Haight et al. 2019). Therefore, the code to calculate this definition for our example subject would be:

count\_matches(  
 use\_pattern = "++++---+--------------o-",  
 match\_is = "-",  
 # Mixed results weeks count as half of a negative week  
 mixed\_results\_are = "\*", mixed\_weight = 0.5,  
 start = 5, end = 24,  
 # Return the proportion instead of the count  
 proportion = TRUE  
)

## Example: Relapse Outcome

The [Krupitsky et al. (2006)](https://doi.org/10.1016/j.jsat.2006.05.005) treatment endpoint has relapse defined as “3 consecutive positive UOS”, where missing clinic visits are imputed to represent a urine screen positive for the substance of interest (Krupitsky et al. 2006). Therefore, the algorithm to calculate this definition for our example subject would be:

"++++---+--------------o-" %>%   
 recode\_missing\_visits(missing\_becomes = "+") %>%   
 # change all mixed-results visits to positive  
 recode\_missing\_visits(missing\_is = "\*", missing\_becomes = "+") %>%   
 # detect if the subject had 3 weeks of positive or missing UDS in a row  
 detect\_subpattern(subpattern = "+++")

If we were interested in relapse from a “time-to-event” or reliability perspective (Kleinbaum and Klein 2012), the algorithm above changes only in the last line:

"++++---+--------------o-" %>%   
 recode\_missing\_visits(missing\_becomes = "+") %>%   
 recode\_missing\_visits(missing\_is = "\*", missing\_becomes = "+") %>%   
 # detect if the subject had 3 weeks of positive or missing UDS in a row, AND  
 # measure the number of weeks until this pattern \*starts\*  
 detect\_in\_window(  
 # Look at all possible 3-week windows, and  
 window\_width = 3L,  
 # check if any of them have all 3 positive UDS  
 threshold = 3L  
 )

# Discussion

## Use Pattern “Word” Limitations

While this substance use pattern summary satisfies the three conditions mentioned previously (machine readability, parsimony, and sufficiency), there are three main limitations:

### Accounting for Multivariate Categorical Variables

The first limitation surrounds the idea of poly-drug use. These use pattern summaries are substance or substance group specific (*univariate*, statistically speaking). Consider now the poly-drug use for our example trial participant:

who

what

source

when

2089

Cocaine

UDSAB

0

2089

Opioid

UDSAB

0

2089

Cocaine

UDSAB

7

2089

Opioid

UDSAB

7

2089

Opioid

UDSAB

15

2089

Cocaine

UDSAB

21

2089

Opioid

UDSAB

21

2089

Cocaine

UDSAB

28

2089

Opioid

UDSAB

28

2089

Cocaine

UDSAB

35

2089

Cocaine

UDSAB

43

2089

Cocaine

UDSAB

49

2089

Cocaine

UDSAB

56

2089

Opioid

UDSAB

56

2089

Cocaine

UDSAB

63

2089

Cocaine

UDSAB

70

2089

Cocaine

UDSAB

78

2089

Cocaine

UDSAB

84

2089

Cocaine

UDSAB

91

2089

Cocaine

UDSAB

98

2089

Cocaine

UDSAB

105

2089

Cocaine

UDSAB

112

2089

Cocaine

UDSAB

119

2089

Cocaine

UDSAB

126

2089

Cocaine

UDSAB

133

2089

Cocaine

UDSAB

140

2089

Cocaine

UDSAB

147

2089

Cocaine

UDSAB

154

2089

Cocaine

UDSAB

163

2089

Cocaine

UDSAB

168

2089

Cocaine

UDSAB

219

Recall from our examples above that this participant does appear to curb their weekly opioid and heroin use during the course of treatment. However, their cocaine use does *not* decrease over the same time interval. The **opioid** use pattern summary can not display information about concurrent **cocaine** use. While this is a limitation, one of our current areas of research involves expanding the support of substance use pattern “words” to include poly-substance use in a meaningful way (including the ability to preserve cross-substance correlations).

### The Complexity-Parsimony Tradeoff

The second limitation is one of parsimony. Technically speaking, in order to ensure that use pattern summaries work regardless of a computer’s locale or operating system, we limit the symbols in the “word” to proper symbols from the [American Standard Code for Information Interchange](https://www.ascii-code.com/) (ASCII) list (AB 2022), which has been a computing standard [for decades](http://edition.cnn.com/TECH/computing/9907/06/1963.idg/) (Brandel 1999). Technically speaking, there are 128 [7-bit ASCII](https://www.ibm.com/docs/en/aix/7.1?topic=support-ascii-characters) symbols, but only 96 are visible (printable) characters (IBM 2022). Put simply, these 96 characters are all the symbols that a traditional North American computer keyboard can make. Therefore, we could define a substance use pattern “word” with any combination of these 96 printable symbols.

However, such a visual summary with 96 possible values would be incredibly challenging to interpret, consequently failing to meet the parsimony requirement (that our summary be easily read by a clinician). In the interest of simplicity, we chose the five symbols mentioned above (+, -, o, \*, and \_). We recognize that different clinical trial designs may necessitate the introduction of additional symbols to this use pattern summary, but we strongly recommend (due to human [memory constraints](https://doi.org/10.1037/h0043158)) that such lists be kept to seven symbols or fewer (Miller 1956). As an example, we discussed the benefit of adding special symbols for a “missing but excused” clinic visit (or for “urine sample of improper temperature”), but we ultimately did not because such instances were rare in our data.

### The Inflexibility of Computers

The third limitation is due to the cold and unfeeling logic of a computer. The computer cannot understand the many extenuating circumstances that pervade our human experience. Worth repeating is the old maxim of computing: “computers give you what you ask for, not what you want.” Consider a concrete example: assume a participant was supposed to visit the clinic for a weekly urine test on Thursday, but they came on Wednesday instead due to a conflict at their work; their urine sample from Wednesday came back “clean”. What would a nurse at the substance use clinic do in this case? Most people are reasonable; they would count the urine screen for that week as negative for illicit substances. What would a computer do? Without the direct input of a human to override the clinical trial protocol, a computer would count the negative urine on Wednesday but still mark the participant as a “failure to appear” / “missing urine screen” on Thursday. In many clinical trial protocols, missing urine screens are imputed as “positive” for the substance of interest, so now this participant’s use pattern summary is marked \* (mixed results) instead of - (negative results) for that week.

## Use Pattern “Word” Strengths

We believe this CTNote package and outcomes library will be useful to all future substance use disorder clinical trials in the National Institute of Drug Abuse (NIDA)’s [Clinical Trials Network](https://doi.org/10.1186/s13722-021-00238-6) (CTN) and in other substance use disorder research (Tai et al. 2021). While developing this construct, our work adheres to the [Open Knowledge](https://doi.org/10.1371/journal.pbio.1001195) philosophy of science (Molloy 2011), so the code for the CTNote package and all related algorithms are open-source.

**HELP!!** I need some help writing this part of the discussion. I think this quinary word structure is an awesome way to represent urine test results over time, but I don’t know how to sell a clinician or their statistician on this.

# References

AB, Injosoft. 2022. “The Extended ASCII Table.” *ASCII Code*. <https://www.ascii-code.com/>.

Brandel, Mary. 1999. “1963: The Debut of ASCII.” *CNN*. <http://edition.cnn.com/TECH/computing/9907/06/1963.idg/>.

Clapham, Christopher, and James Nicholson. 2009. “Binary Word.” In *The Concise Oxford Dictionary of Mathematics*. Oxford University Press. <https://www.oxfordreference.com/view/10.1093/acref/9780199235940.001.0001/acref-9780199235940-e-293>.

Fisher, R. A., and Edward John Russell. 1922. “On the Mathematical Foundations of Theoretical Statistics.” *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 222 (594-604): 309–68. <https://doi.org/10.1098/rsta.1922.0009>.

Geer, Renata, and Janet Morrison. 1999. “Genetics Review: Base Pairs.” <https://www.ncbi.nlm.nih.gov/Class/MLACourse/Original8Hour/Genetics/basepair.html>.

Haight, Barbara R., Susan M. Learned, Celine M. Laffont, Paul J. Fudala, Yue Zhao, Amanda S. Garofalo, Mark K. Greenwald, et al. 2019. “Efficacy and Safety of a Monthly Buprenorphine Depot Injection for Opioid Use Disorder: A Multicentre, Randomised, Double-Blind, Placebo-Controlled, Phase 3 Trial.” *The Lancet* 393 (10173): 778–90. <https://doi.org/10.1016/S0140-6736(18)32259-1>.

Handbook, Open Data. 2022. “Machine Readable.” *Open Data Handbook*. <https://opendatahandbook.org/glossary/en/terms/machine-readable/>.

HHS, US. 2018. “Opioid Crisis Statistics.” Text. *HHS.gov*. <https://www.hhs.gov/opioids/about-the-epidemic/opioid-crisis-statistics/index.html>.

IBM. 2022. “ASCII Characters.” <https://prod.ibmdocs-production-dal-6099123ce774e592a519d7c33db8265e-0000.us-south.containers.appdomain.cloud/docs/en/aix/7.1?topic=support-ascii-characters>.

Kleinbaum, David G., and Mitchel Klein. 2012. “Introduction to Survival Analysis.” In *Survival Analysis: A Self-Learning Text*, edited by David G. Kleinbaum and Mitchel Klein, 1–54. Statistics for Biology and Health. New York, NY: Springer. <https://doi.org/10.1007/978-1-4419-6646-9_1>.

Krupitsky, Evgeny M., Edwin E. Zvartau, Dimitry V. Masalov, Marina V. Tsoy, Andrey M. Burakov, Valentina Y. Egorova, Tatyana Y. Didenko, et al. 2006. “Naltrexone with or Without Fluoxetine for Preventing Relapse to Heroin Addiction in St. Petersburg, Russia.” *Journal of Substance Abuse Treatment* 31 (4): 319–28. <https://doi.org/10.1016/j.jsat.2006.05.005>.

Liebhaber, Lauren. 2022. “Over 1 Million Americans Are Seeking Treatment for Substance Use Disorder—Here’s How It Breaks down by State.” *Zinnia Health*. <https://zinniahealth.com/research/substance-use-disorder-treatment-by-state>.

Miller, George A. 1956. “The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information.” *Psychological Review* 63: 81–97. <https://doi.org/10.1037/h0043158>.

Molloy, Jennifer C. 2011. “The Open Knowledge Foundation: Open Data Means Better Science.” *PLoS Biology* 9 (12): e1001195. <https://doi.org/10.1371/journal.pbio.1001195>.

Odom, Gabriel, Laura Brandt, and Raymond Balise. 2022. “CTNote: Clinical Trial Network Outcomes, Treatments and Endpoints.” <https://github.com/CTN-0094/CTNote>.

R Core Team. 2022. “R: A Language and Environment for Statistical Computing.” Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>.

SAMHSA, US. 2021. “SAMHSA Releases 2020 National Survey on Drug Use and Health.” *SAMHSA.gov*. <https://www.samhsa.gov/newsroom/press-announcements/202110260320>.

———. 2022. “Medication-Assisted Treatment (MAT).” *SAMHSA.gov*. <https://www.samhsa.gov/medication-assisted-treatment>.

Schottenfeld, Richard S., Marek C. Chawarski, and Mahmud Mazlan. 2008. “Maintenance Treatment with Buprenorphine and Naltrexone for Heroin Dependence in Malaysia: A Randomised, Double-Blind, Placebo-Controlled Trial.” *The Lancet* 371 (9631): 2192–2200. <https://doi.org/10.1016/S0140-6736(08)60954-X>.

Tai, Betty, Ronald Dobbins, Quandra Blackeney, David Liu, and Landhing Moran. 2021. “The NIDA Clinical Trials Network: Evolving, Expanding, and Addressing the Opioid Epidemic.” *Addiction Science & Clinical Practice* 16 (1): 28. <https://doi.org/10.1186/s13722-021-00238-6>.