Westat

Improving Lives Through Research®

Harnessing Paradata with Machine Learning to Inform Data Collection

Mengshi (Jack) Zhou, Gizem Korkmaz, Ting Yan, Hanyu Sun, Ryan Hubbard, Jill Carle, Rick Dulaney, Brad Edwards

WESTAT @ FCSM 2024

Introduction

- Cost of survey data collection is on the rise:
  - Survey response rates continue to decline despite more contact mode options (e.g., text, telephone).
  - The challenge is to identify cost and time-efficient ways to increase the likelihood of getting a response.
- Paradata contain a vast amount of information on when and how sampled persons are contacted and the outcome of each contact attempt.
- Can the analysis of paradata help inform the contact strategies and create efficiencies?

Related Work

- Paradata have been utilized to group households with similar contact history and outcomes (Durrant, G.B., Maslovskaya, O. and Smith, P.W., 2019).
- Propensity modeling informed by paradata predicts respondent likelihoods (Fang, Qixiang, et al., 2021).
- Studies used paradata to identify the optimal timing for reaching out in surveys (Shino, E., & McCarty, C. 2020).

Goals

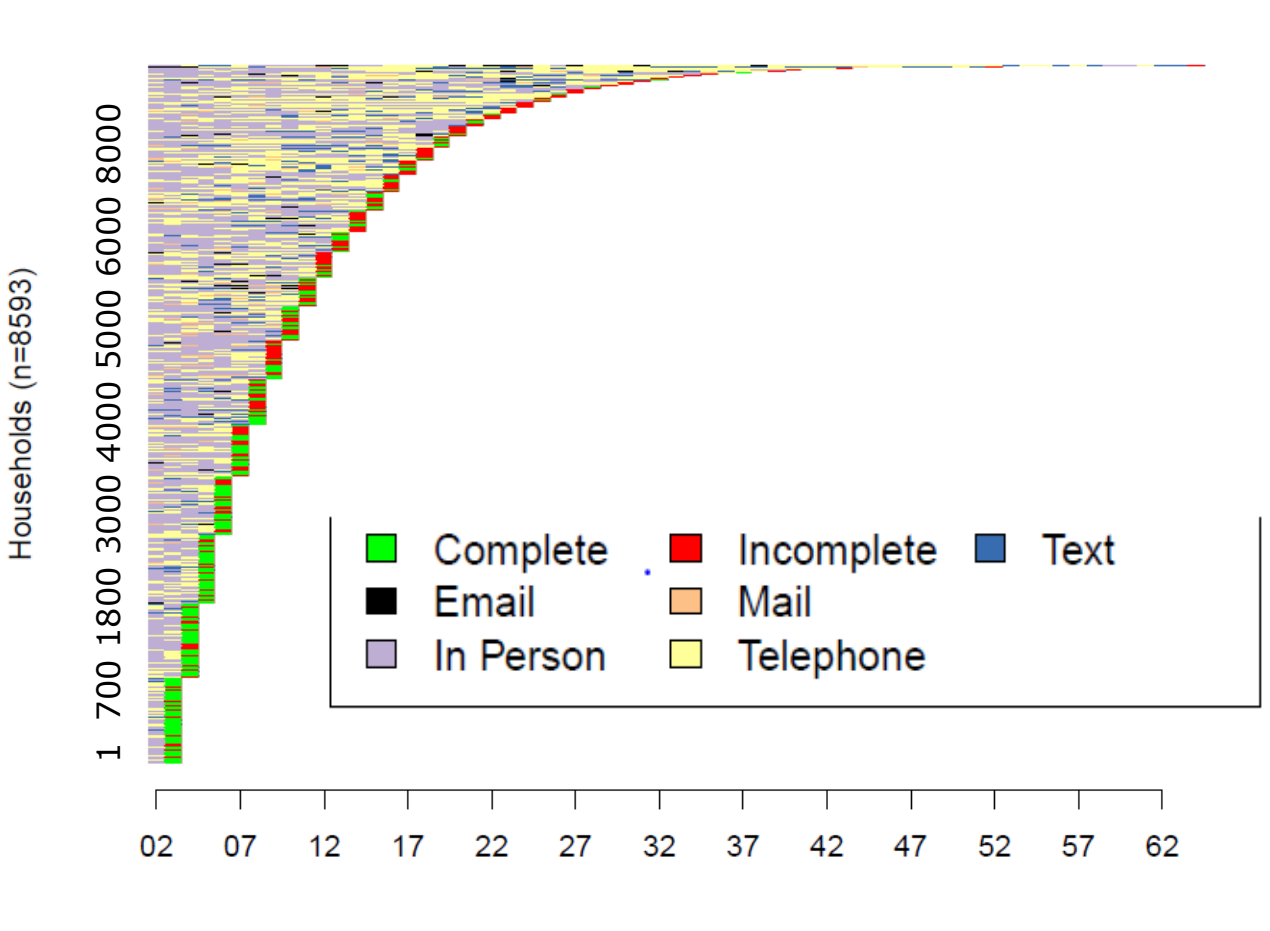
- This study aims to:
  - Understand how machine learning can be used to generate hypotheses from paradata to inform contact strategy.
  - Gather causal evidence for hypotheses.
  - Provide actionable insights for improving survey data collection.

Data

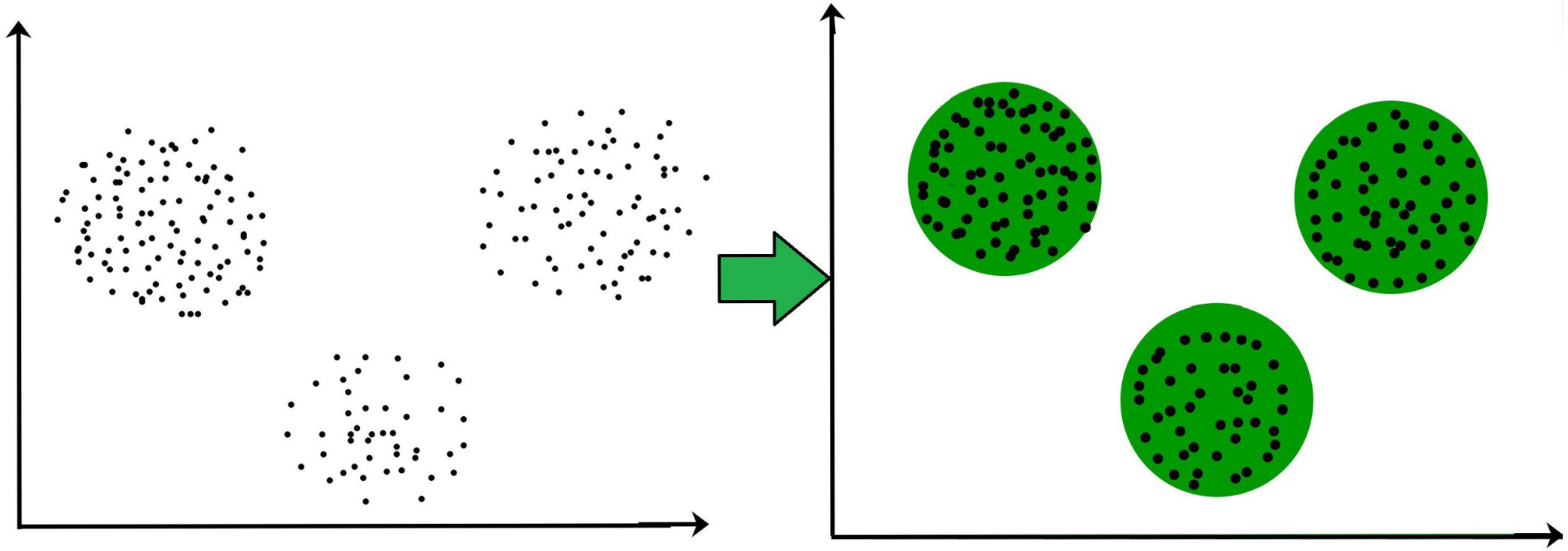
- Medical Expenditure Panel Survey (MEPS)
  - Nationally representative survey designed to explore healthcare use and expenditures in the U.S.
  - Conducted by Westat to support the Agency for Healthcare Research and Quality (AHRQ).
  - MEPS Household Component (MEPS-HC) collects data from a nationally representative sample of households, drawn from the National Health Interview Survey (NHIS).
  - High round 1 response rates are crucial to obtaining representative data.

Data

- MEPS Household Component Spring 2022 Round 1
  - 8,593 households with at least two contact attempts from 10,071 households
  - 4,707 complete, 3,886 incomplete (complete rate: 54.8%)
  - Paradata: Contact mode (In-Person, Telephone, Text, E-mail, or Mail) of each contact.



Unsupervised Machine Learning: Clustering



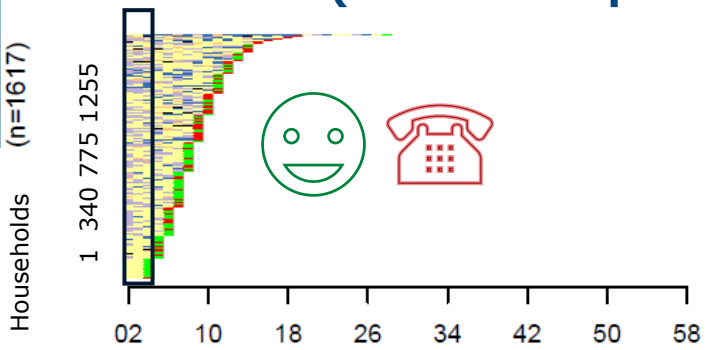
- Clustering algorithms automatically group similar observations.
- The measurement of similarities and number of groups are defined by humans.

Clustering Methods on Paradata

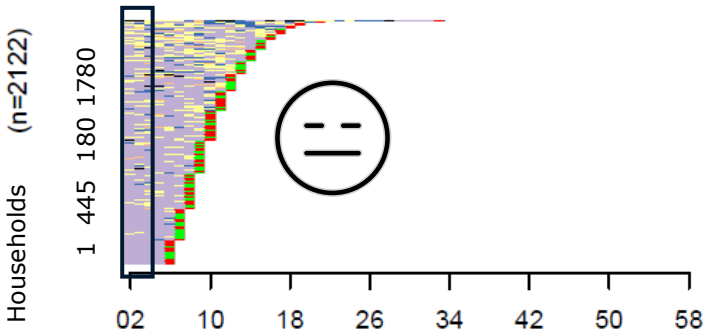
- **Observation**: Each household is represented by its contact mode sequence from contact 2 onward.
- **Similarity**: Calculated sequence distance using optimal matching (number of operation to transform sequences).
- **Clustering**: Partitioning Around Medoids algorithm

Clustering Results

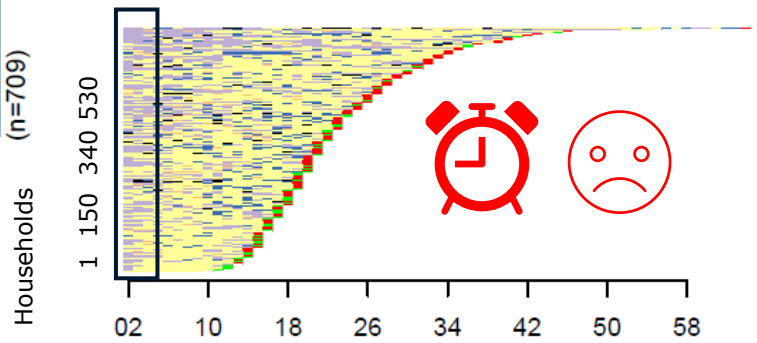
Cluster 1 (64.1% complete)



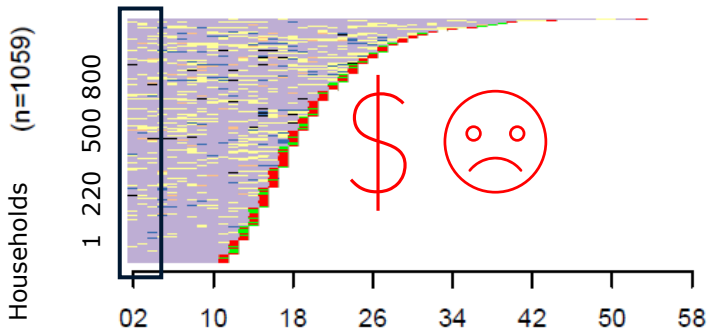
Cluster 2 (44.0% complete)



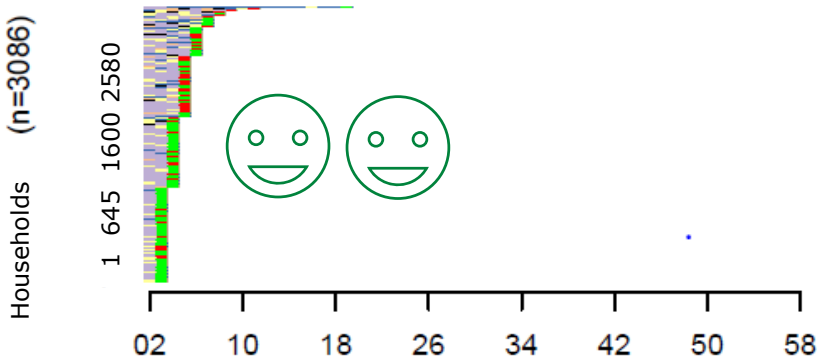
Cluster 3 (27.6% complete)



Cluster 4 (30.0% complete)



Cluster 5 (72.0% complete)



Clustering Results

		Cluster 1	Cluster 2	Cluster 3	Cluster 4	P
n		1617	2122	709	1059	
Contact 2-4 (%)	**Telephone**	**713 (44.1)**	**174 (8.2)**	**237 (33.4)**	**70 (6.6)**	<0.001
	In Person	**113 (7.0)**	**1255 (59.1)**	**181 (25.5)**	**705 (66.6)**	
	Mix	355 (22.0)	396 (18.7)	155 (21.9)	181 (17.1)	
	Other	309 (19.1)	297 (14.0)	136 (19.2)	103 (9.7)	
	Short	127 (7.9)	0 (0.0)	0 (0.0)	0 (0.0)	

Telephone: At least one telephone contact sequence during contact 2-4

In Person: Only in-person contact during contact 2-4

Mix: Mixed in-person and telephone contact during contact 2-4

Other: Other contact sequences during contact 2-4

Short: Contact sequence shorter than 4

Hypothesis

- **Hypothesis**:
  - Machine learning suggests that early telephone follow-ups after an initial in-person contact lead to a higher success rate.
- **Challenges in Testing Hypothesis:**
  - Households are not randomly assigned to contact modes.
  - Confounding factors such as household characteristics may make it difficult to determine the causal effects.

Target Trial Emulation

- **Question**: Is early telephone contact sequence associated with a higher success rate?
- **Cohort**: Households (Spring 2022, Round 1) with in-person first contact and telephone or in-person contacts during contacts 2-4
- **Exposure Group**:
  - 1. Treatment: Telephone sequence contact during contact 2-4.
  - 2. Control: Only in-person contact during contact 2-4.
- **Causal Inference**: Propensity score matching to mimic randomization.
- **Outcomes**: Response status during contacts 2-4 and days until completion.

Results

- **Propensity Score Matching**: Standardized differences are below 10%, indicating a well-balanced cohort.
- **Statistical Analysis**:
  - Telephone follow-ups were associated with increased likelihood of completion and positive response during contact 2-4.
  - The Kaplan-Meier analysis suggests that households with telephone follow-ups during contacts 2-4 complete the survey faster after contact 4.

Potential New Data Collection Strategies

- For households classified as likely to participate during the pre-round call period, prioritize telephone follow-ups after the initial in-person contact.
- For households without phone numbers, make efforts to match or obtain phone numbers before continuing with in-person contacts.
- Make appointment attempts during the pre-round call period.
- Ensure that face sheet information from NHIS data is available to guide the selection of the most effective interview mode for each household.

Conclusion and Future Directions

- **Conclusion**: Machine learning can improve data collection strategies by generating actionable insights from paradata-focused hypotheses.
- **Future Work**:
  - Gather information on the impact of the new strategies.
  - Generate and evaluate additional hypotheses.
  - Develop machine learning models for contact recommendations.

References

- Durrant, Gabriele B., Olga Maslovskaya, and Peter WF Smith. "Investigating call record data using sequence analysis to inform adaptive survey designs." *International Journal of Social Research Methodology* 22.1 (2019): 37-54.
- Einarsson, Hafsteinn, Alexandru Cernat, and Natalie Shlomo. "Responsive and Adaptive Designs in Repeated Cross-National Surveys: A Simulation Study." Journal of Survey Statistics and Methodology 12.4 (2024): 906-931.
- Shino, Enrijeta, and Christopher McCarty. "Telephone survey calling patterns, productivity, survey responses, and their effect on measuring public opinion." Field Methods 32.3 (2020): 291-308.
- https://www.westat.com/importance-of-the-medical-expenditure-panel-survey-project/
- Studer, Matthias. "WeightedCluster library manual." A practical guide to creating typologies of trajectories in the social sciences with 2013.24 (2013): 33.
- https://www.geeksforgeeks.org/clustering-in-machine-learning/
- James, Gareth, et al. *An introduction to statistical learning*. Vol. 112. New York: springer, 2013.

Contacts:

Mengshi (Jack) Zhou, Ph.D.
Senior Data Scientist, Westat
jackzhou@westat.com

Gizem Korkmaz, Ph.D.
Associate Vice President, Westat
gizemkorkmaz@westat.com

Thank You!

Appendix

Clustering Methods on Paradata

- **Observation**: Each household is represented by its contact mode sequence from contact 2 onward.
- **Similarity**: Calculated sequence distance using optimal matching (numer of operation to transform sequences).
  - Substitution cost = 2
  - Insertion cost = 1
- **Clustering**: Partitioning Around Medoids algorithm

	Contact 2	Contact 3	Contact 4
Household 1	In-person	Telephone	In-person
Household 2	In-person	In-person	
Operation	None (0)	Substitution (2)	Insertion (1)

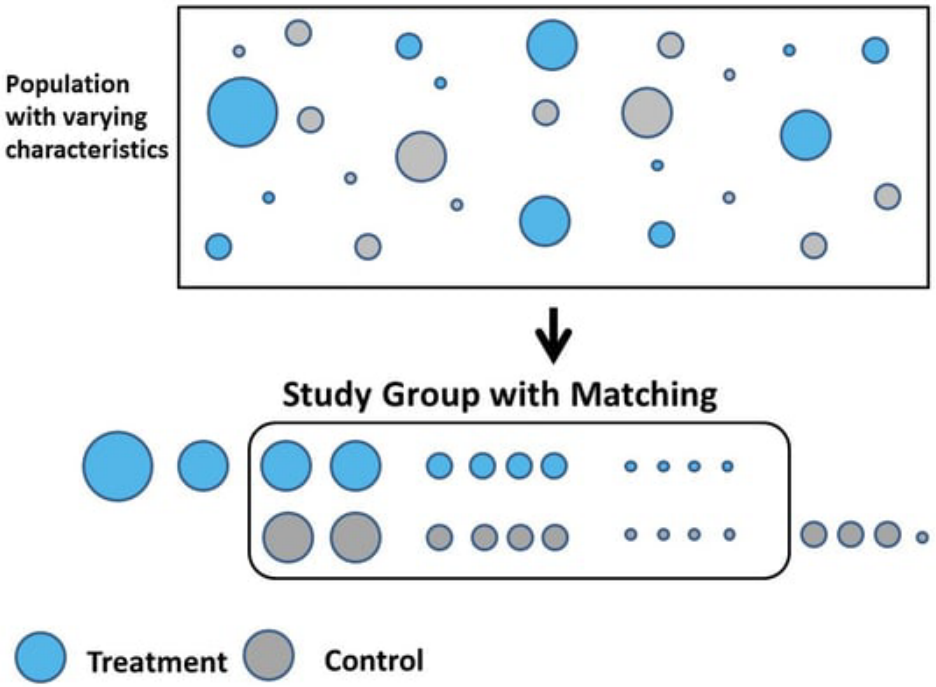
Propensity Score Matching

		Treatment (At least one telephone sequence contact during contact 2-4)	Control (Only in-person contact during contact 2-4)	Standard Difference (%)
n		764	764	
ACR	Likely	296 (38.74)	292 (38.22)	1.07
	Unlikely	77 (10.08)	75 (9.82)	0.87
	Mutual	58 (7.59)	61 (7.98)	-1.48
	Unknown	333 (43.59)	336 (43.98)	-0.79
Gender	Female	412 (53.93)	406 (53.14)	1.58
	Male	342 (44.76)	354 (45.16)	-0.79
	Unknown	10 (1.31)	13 (1.70)	-3.45
Race	White	532 (69.63)	537 (70.29)	-1.42
	Black	109 (14.27)	102 (13.35)	2.62
	Asian	61 (7.98)	57 (7.46)	1.93
	Multiple	11 (1.44)	10 (1.31)	1.10
	Other	1 (0.13)	0 (0.00)	3.62
	Unknown	65 (6.54)	64 (7.59)	-4.23
Age	=< 25	65 (8.51)	64 (8.38)	1.13
	26-54	532 (68.46)	519 (67.93)	0.47
	>=65	176 (23.04)	181 (23.69)	-1.55
House Size	1	211 (27.62)	223 (29.19)	-3.51
	2-4	476 (62.30)	472 (61.78)	1.08
	>= 5	77 (10.08)	69 (9.03)	3.48
Outcome01	Success	259 (33.90)	240 (31.41)	5.25
	Fail	524 (66.10)	505 (68.59)	-5.25

- All standard differences are below 10%.
- Means our analytic cohort is well-balanced across the potential confounders.
- Effectively mimicking randomization.

Causal Inference: Propensity Score Matching

- Propensity score matching (PSM) is a commonly used method to establish causality from observational studies.
- PSM pairs observations from different groups based on their likelihood of receiving treatment given a set of observed covariates.
- PSM emulates the randomization of treatment in observational studies by balancing the distribution of observed covariates between the treatment and control groups.

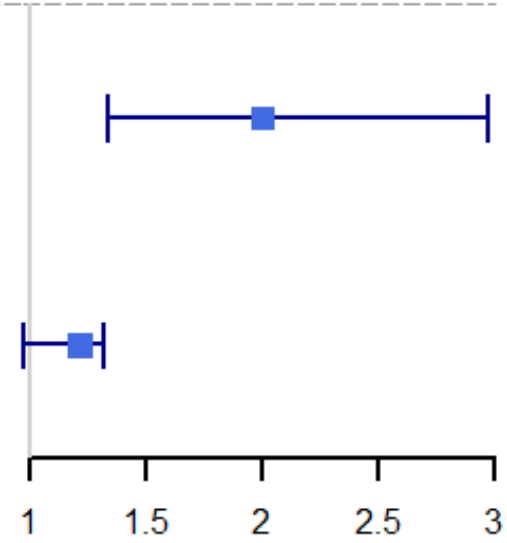


Zhou, Qiao, et al. "Exploratory Compatibility Regularity of Traditional Chinese Medicine on Osteoarthritis Treatment: A Data Mining and Random Walk-Based Identification." *Evidence-Based Complementary and Alternative Medicine* 2021.1 (2021): 2361512.

Outcome during contact 2-4

- A multinomial logistic regression model suggests:
  - Households with telephone follow-ups are 2 times more likely to complete during contact 2-4 as compared to those with in-person follow-ups.
  - Households with telephone follow-ups are 21% more likely to have a positive response during contact 2-4 as compared to those with in-person follow-ups.

Outcome	Treatment	Control	OR	P_value
Complete	79	46	2.00	<0.0001
Response	397	383	1.21	0.082



Days until the complete interview after contact 4

- The Kaplan-Meier curve suggests that households with telephone follow-ups during contacts 2-4 complete the survey faster after contact 4.