Logistic Regression Analysis on the POLYPHARM Dataset

Kawthar Abdallah

1 Introduction

Polypharmacy, often defined as routinely taking a minimum of five medicines [1], presents a significant challenge in healthcare delivery, particularly in pediatric populations, where the long-term developmental and physiological effects of medication regimens remain poorly understood. The rising prevalence of chronic conditions, such as asthma, diabetes, and mental health disorders, in pediatric populations has led to an increased reliance on multiple medications, further exacerbating the risks associated with polypharmacy. While sometimes necessary for managing chronic diseases and medically complex conditions, it elevates the risk of adverse drug interactions [2], reduced medication adherence [3], and increased healthcare costs [4]. Understanding the predictors of polypharmacy is crucial for developing targeted interventions to mitigate these risks, particularly in vulnerable populations such as pediatric patients.

This project analyzes the POLYPHARM dataset to identify the demographic and clinical predictors of polypharmacy in pediatric patients aged 1-19. This dataset provides detailed demographic and clinical data, making it a valuable resource for achieving this objective. Given the unique developmental considerations and potential long-term effects of medication regimens in children and adolescents [5], understanding the factors contributing to polypharmacy in this age group is especially critical.

We chose logistic regression because it is well-suited for modeling binary outcomes and it aligns with established methodological guidelines [6]. Logistic regression is interpretable, which is important for clinical decision-making. It allows for the identification of key predictors and their effect sizes (odds ratios). Unlike predictive modeling efforts focused on generalizability, this approach allows for the identification of key predictors while providing clinically actionable insights. The larger goal is to understand *why* polypharmacy occurs in this population, rather than solely to predict its occurrence.

Our analysis identified age, gender, and comorbidities as significant predictors of polypharmacy, with older age and male gender associated with higher odds of polypharmacy, while the presence of comorbidities reduced the likelihood. These findings align with previous research on polypharmacy in various populations [2],[9], [10] and offer valuable insights for healthcare providers managing polypharmacy risks in younger individuals.

2 Background

Identifying patterns and predictors of polypharmacy is critical for designing evidence-based interventions that mitigate adverse risks and optimize patient care[3], particularly in pediatric populations where developmental considerations necessitate tailored approaches [7]. In pediatric populations, drug regimens can have long-term developmental effects, hence polypharmacy in this age group presents unique challenges due to ongoing developmental changes and the potential for long-term effects on growth and overall health [8]. Understanding the factors that contribute to polypharmacy in this age group is critical for improving medication management and patient outcomes.

Previous research has consistently identified demographic factors, such as age and gender, and clinical characteristics like comorbidities and healthcare utilization, as potential predictors of polypharmacy. For instance, older patients and those managing multiple chronic conditions are consistently found to have higher medication counts [2]. Gender differences also play a role, with females often experiencing higher rates of polypharmacy potentially due to differences in healthcare-seeking behaviors and prescribing practices [2],[9]. Additionally, disparities in healthcare access and prescribing patterns across racial and ethnic groups have been documented, further emphasizing the need to examine these factors in the context of polypharmacy [10]. Social determinants of health, such as socioeconomic status, access to healthcare, and health literacy, play a critical role in shaping medication use patterns and may contribute to disparities in polypharmacy prevalence across different populations. While these factors were not explored in this study, they represent important areas for future research to better understand the broader context of polypharmacy in pediatric populations.

This project draws from principles outlined in "Applied Logistic Regression" by Hosmer et al. (Chapter 3, Section 3.5), which emphasizes how logistic regression is useful for model that involve binary outcomes [6]. Logistic regression was chosen for this study because it is widely understood, interpretable, and well-suited for identifying relationships between predictors and a binary outcome like polypharmacy. Unlike predictive modeling efforts aimed at generalizability, this study prioritizes exploring relationships to provide actionable insights for healthcare providers.

Using data from the POLYPHARM dataset, which contains 3,500 observations and detailed demographic (e.g. age, sex) and clinical variables (e.g. comorbidities, healthcare utilization), this project uses logistic regression to bridge theoretical understanding (using statistical models) with practical applications (in this case, healthcare management). By identifying and quantifying key predictors, the study contributes to a growing body of literature aimed at addressing challenges related to polypharmacy.

3 Methods

This section explores the key features of the dataset, identifies relationships between variables, and visualizes trends to inform the modeling process.

3.1 Dataset

The POLYPHARM dataset comprises 3,500 observations across 14 variables, including demographic, clinical, and healthcare utilization data. Categorical variables, including `POLYPHARMACY` and `RACE`, were converted to factors to ensure proper handling during analysis. The types of variables that occurred were numeric (such as `AGE`) , binary nominal (such as `POLYPHARMACY`), multi-level nominal (such as `RACE`), and ordinal (such as `NUMPRIM`). Table 1 provides a summary of the key variables.

Variable	Description
`AGE`	Patient age (years)
`GENDER`	Gender (0 - Female, 1 - Male)
`RACE`	Race Categories (0 - White, 1 - Black, 2 - Other)
`NUMPRIM`	Number of primary diagnosis (0 - none, 1 - one, 2 - more than one)
`COMORBID`	Comorbidities (0 - None, 1 - Present)
`POLYPHARMACY`	Polypharmacy status (0 - None, 1 - Present)

Table 1: Key Variables in the Dataset

The dataset contained one missing value under the `URBAN` variable and also had 49 missing values under the `NUMPRIM` variable. Given the minimal impact (1.4% of observations) of the missing values, we opted for listwise deletion instead of imputation.

The dataset is publicly available in the `aplore3` R package, which includes datasets from Hosmer, Lemeshow, and Sturdivant's "Applied Logistic Regression" (3rd Ed., 2013) [6]. The dataset can be accessed at https://rdrr.io/cran/aplore3/man/polypharm.html.

3.2 Exploratory Data Analysis

The exploratory data analysis (EDA) phase aimed to uncover patterns and relationships within the POLYPHARM dataset, guiding the selection of predictors for the logistic regression model. The dataset revealed the following information:

Age distribution: Patients ranged from 1.17 to 18.92 years with an average age (mean) of 11.65 years. Polypharmacy cases had a higher mean age (12.19 years) compared to nonpolypharmacy cases (11.48 years), showing that older individuals are more likely to experience polypharmacy. Figure 1 suggests that patients with polypharmacy tend to be older than those without.

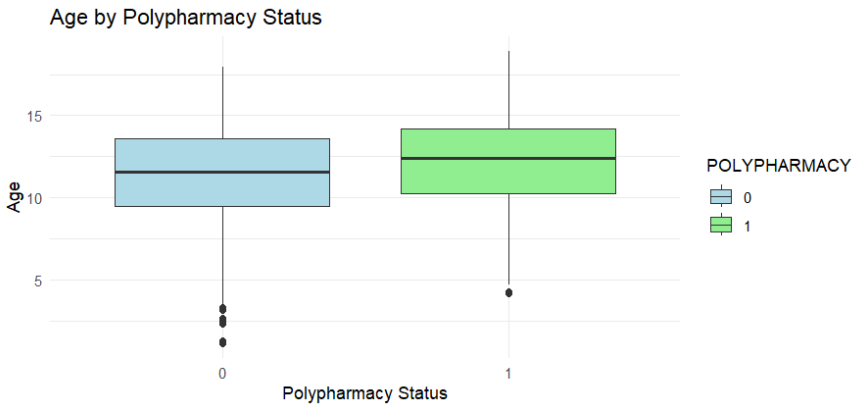

Figure 1: Age Distribution by Polypharmacy Status. All the five boxplot values are higher for the class where polypharmacy is present.

Gender Composition: Female patients accounted for 77.19% of the dataset, reflecting possible differences in healthcare-seeking behavior by gender. Figure 2 shows the polypharmacy outcomes for each gender:

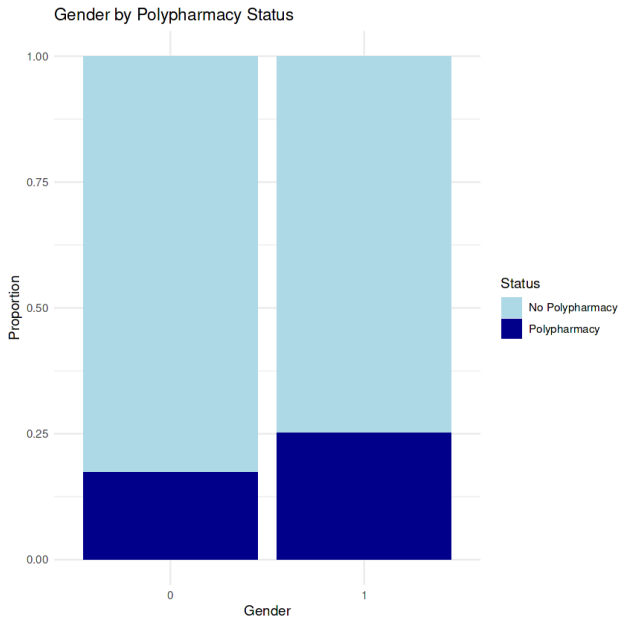

Figure 2: Gender Distribution by Polypharmacy Status

Polypharmacy prevalence: Approximately 23.4% of patients were classified as polypharmacy cases, highlighting its relevance in the dataset.

3.3 Statistical Tests

Pearson's Chi-squared test identified a significant association between `RACE` and `POLYPHARMACY` (X-squared = 11.417, df = NA, p = 0.002499). Since this test is the standard method for

assessing associations between categorical variables, it follows that we have support for the inclusion of race in the model. A two-sided Fisher's Exact test for count data that gives 0.002386 as p-value also indicates that polypharmacy status varies significantly across racial groups.

3.4 Correlation Analysis

To explore relationships between variables, we computed appropriate measures of correlation for each pair of variable types. The correlation measures for each pair is enumerated in the following table.

Variable Type Pair	**Correlation Measure**
Nominal vs. Nominal, Nominal vs. Ordinal, Nominal vs. Binary	Cramér's V
Numeric vs. Nominal	Eta-Squared
Numeric vs. Numeric	Pearson's Correlation
Numeric vs. Binary Nominal	Point-Biserial
Numeric vs. Ordinal, Ordinal vs. Ordinal, Binary vs. Ordinal	Spearman's Rank
Binary Nominal vs. Binary Nominal	Phi Coefficient

Table 2: Correlation measures for different variable type pairs.

We saw some several notable relationships:

- `AGE` **and** `YEAR`: A strong positive correlation ($r = 0.69$) suggesting these variables have a substantial overlap in the directionality of the information they represent.
- `NUMPRIM` **and** `ANYPRIM`: A near perfect correlation ($r = 0.99$) indicating these variables essentially capture the same information.
- `RACE` **and** `ETHNIC`: A moderate correlation ($r = 0.47$), suggesting some demographic clustering.

Figure 3 shows the heatmap of the matrix whose entries are the computed correlation measure between the corresponding pair of variables, with darker shades indicating stronger correlations. The heatmap provides a visual summary of the relationships between variables, guiding the selection of predictors for the logistic regression model.

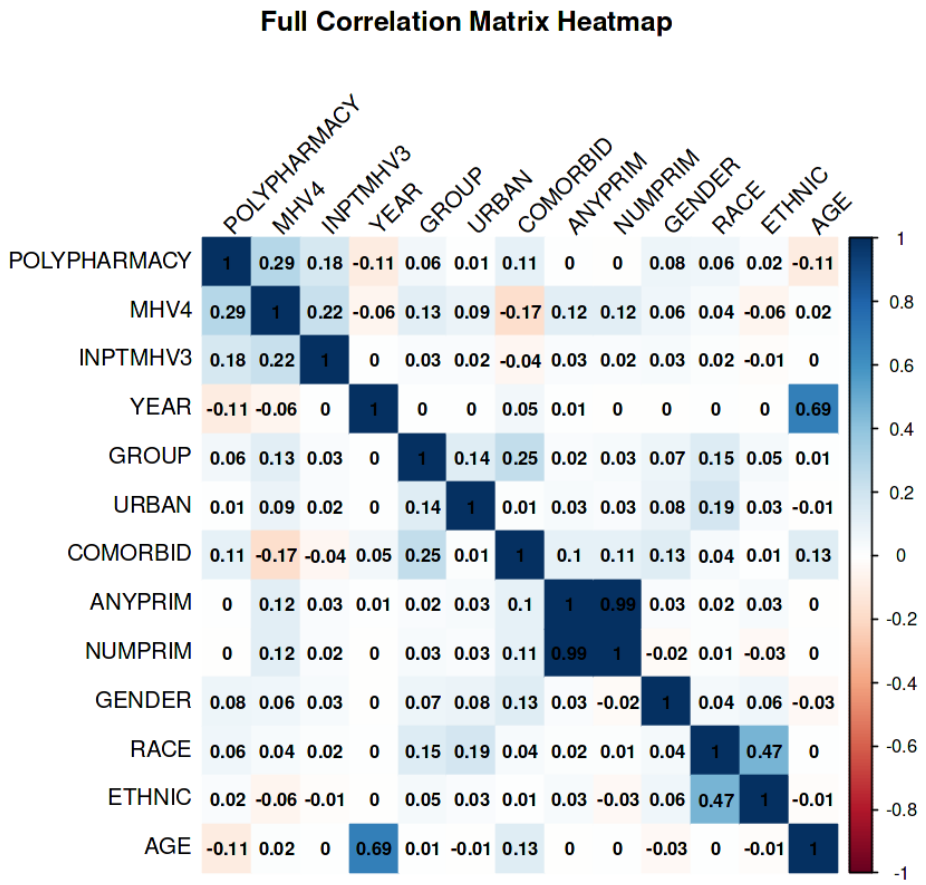

Figure 3: Correlation Matrix

3.5 Logistic Regression Model

The logistic regression model was fitted to the entire dataset to explore the relationships between the selected predictor variables and the likelihood of polypharmacy. The predictor variables included in the model were age (AGE), gender (GENDER), race (RACE, ETHNIC), number of primary diagnosis (NUMPRIM), and presence of comorbidities (COMORBID). These variables were selected based on their clinical relevance, insights from the exploratory data analysis, and prior research [2],[9], [10].

A visual inspection of the boxplot in Figure 1 suggests a potential linear relationship between age and the logit of polypharmacy status. The assumption of a linear relationship between continuous predictors and the logit of the outcome gives us some justification to include `AGE` as a continuous predictor.

The logistic regression model was specified as

$$logit(\texttt{P}) = \beta_0 + \beta_1 \texttt{AGE} + \beta_2 \texttt{GENDER} + \beta_3 \texttt{NUMPRIM} + \beta_4 \texttt{COMORBID} + \beta_5 \texttt{RACE} + \beta_6 \texttt{ETHNIC}.$$

In the equation, `P` represents the probability of polypharmacy occurrence, and the β coefficients represent the log odds for each predictor variable.

The logistic regression model was implemented using the 'glm' function in R with a binomial family. After fitting the logistic regression model, we assessed multicollinearity among the predictors used in the model by computing variance inflation factors (VIFs). All VIFs were below 5, indicating no significant multicollinearity among the included predictors and supporting the stability of the logistic regression coefficients.

Predictor	**GVIF**	**Df**	**GVIF$^{1/(2*Df)}$**
AGE	1.011251	1	1.005610
GENDER	1.014441	1	1.007195
NUMPRIM	1.013536	2	1.003367
COMORBID	1.028276	1	1.014039
ETHNIC	1.206754	1	1.098524
RACE	1.206079	2	1.047958

Table 3: R output from Variance Inflation Factor (VIF) computation. GVIF values below 5 indicate no significant multicollinearity.

4 Evaluation

This section presents the results of the fitted logistic regression model and discusses its performance. It is important to note that the model was fitted to the entire dataset to prioritize the exploration of relationships, rather than assessing its generalizability to an independent dataset.

4.1 Logistic Regression Results

Our logistic regression analysis identified the following significant predictors of polypharmacy, with odds ratios and confidence intervals indicating the strength and direction of these associations:

- Age: For every one-year increase in age, the odds of polypharmacy increased by 8%. (Odds Ratio = 1.08, 95% CI: 1.05-1.11, $p < 0.001$).
- Gender: Males were 1.49 times more likely to experience polypharmacy than females (Odds Ratio = 1.49, 95% CI: 1.21-1.84, $p \approx 0.0015$).
- Comorbidity: Patients with comorbidities had significantly lower odds of polypharmacy (Odds Ratio = 0.518, 95% CI: 0.39-0.66, $p < 0.001$).
- Race (`RACE1`): Compared to the reference category for race, patients belonging to `RACE1` have around 33% lower odds of having polypharmacy (Odds Ratio = 0.675, 95% CI: 0.52-0.84, $p < 0.001$).

4.2 Model Fit and Residuals

The model's residual deviance decreased from 3776.9 (null model) to 3673.7, indicating a reasonable fit. To formally assess the overall improvement in model fit by including our predictor variables, we conducted a likelihood ratio test comparing our fitted model to the null model (an intercept-only model). Using the deviance reduction of 103.2 with a corresponding reduction of 8 degrees of freedom (corresponding to the number of predictor variables in our model), a likelihood ratio test yielded a highly significant p-value ($p < 0.001$). This provides further support that the included predictor variables contribute meaningfully to explaining the variation in polypharmacy status.

However, the residual deviance should be interpreted with caution, as the model was fitted to the entire dataset without cross-validation, which may limit its generalizability to other pediatric populations. Minimal extreme residuals suggest that the model adequately captures variability.

4.3 ROC Curve and AUC

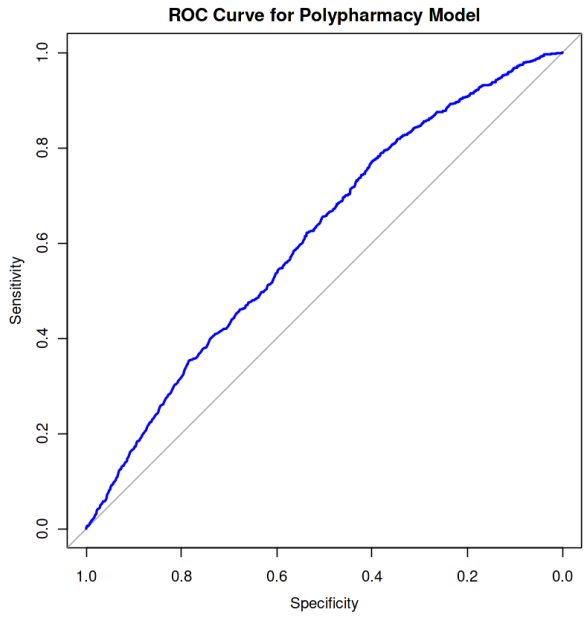

Figure 4: ROC Curve for Polypharmacy Model

The ROC curve (Figure 4) illustrates the trade-off between sensitivity and specificity. The Area Under the Curve (AUC) was 0.611, indicating moderate ability to discriminate between patients who experience polypharmacy and those who do not, within the dataset it was trained on. It is important to interpret this AUC in the context of the model being fitted to the entire dataset, without a separate validation set. While the AUC value suggests that the model can classify polypharmacy cases better than random guessing, there is room for improvement in its classification accuracy. Future studies should consider using advanced techniques such as cross-validation or machine learning algorithms to improve predictive ac-

curacy. These methods could help identify non-linear relationships and interactions between predictors, potentially leading to more accurate risk stratification and targeted interventions.

5 Conclusions and Future Work

This study identified significant demographic and clinical predictors of polypharmacy in a cohort of pediatric patients aged 1-19, providing actionable insights for healthcare providers to manage medication use more effectively in this population. Among the predictors examined, age emerged as the most statistically significant factor in determining polypharmacy risk. However, the presence of comorbidities showed the largest effect size, with these patients being nearly half as likely to experience polypharmacy compared to those without comorbidities. Gender also played a moderate role, with male patients showing a notably higher risk of polypharmacy compared to females.

Specifically, older age, male, absence of comorbidities, and not belonging to RACE1 were associated with a higher likelihood of polypharmacy. These findings highlight key subgroups within the pediatric population who may be at elevated risk of polypharmacy, necessitating closer monitoring and targeted medication management strategies to mitigate adverse outcomes. It is also notable that patients with comorbidities had lower odds of polypharmacy compared to those without comorbidities (Odds Ratio = 0.518, $p < 0.001$). This suggests that the presence of comorbidities in this dataset was associated with a reduced likelihood of polypharmacy, possibly due to differences in medication management, the nature of the conditions included in the comorbidity variable, or confounding factors such as age and healthcare utilization. The findings provide actionable insights for healthcare providers to manage polypharmacy risks effectively.

The analysis of the logistic regression model provides meaningful insights into polypharmacy risk factors while highlighting important areas for future investigation. While our model demonstrates better-than-random predictive ability (AUC = 0.611), the moderate discriminatory performance suggests that additional factors beyond our current predictors likely influence polypharmacy patterns in pediatric populations. These may include social determinants of health, specific medication classes, and healthcare system factors that were not captured in our dataset.

Several key directions emerge for future research. First, longitudinal studies are needed to assess the long-term developmental impacts of polypharmacy in pediatric populations, particularly examining effects on growth and the risk of chronic conditions in adulthood. Second, investigating healthcare system factors, such as provider prescribing behaviors and access to pediatric specialists, could help explain variations in polypharmacy rates. Third, the role of social determinants of health and specific medication classes should be explored to develop more comprehensive predictive models. Finally, while our logistic regression approach provided valuable insights into key predictors, future studies might benefit from advanced modeling techniques, including machine learning algorithms, to uncover non-linear relationships and improve risk stratification accuracy.

This analysis contributes to the growing body of literature on polypharmacy in pediatric populations, providing valuable, actionable insights for healthcare providers. By understanding the factors associated with polypharmacy, clinicians can implement strategies to optimize

medication use, reduce the risk of adverse events, and improve health outcomes for young patients.

References

[1] World Health Organization. *Medication safety in polypharmacy*. WHO/UHC/SDS/2019.11; 2019.

[2] Matos A, Bankes DL, Bain KT, Ballinghoff T, Turgeon J. *Opioids, Polypharmacy, and Drug Interactions: A Technological Paradigm Shift Is Needed to Ameliorate the Ongoing Opioid Epidemic*. Pharmacy (Basel). 2020 Aug 25;8(3):154. doi: 10.3390/pharmacy8030154. PMID: 32854271; PMCID: PMC7559875.

[3] Cadogan CA, Ryan C, Hughes CM. *Appropriate Polypharmacy and Medicine Safety: When Many is not Too Many*. Drug Saf. 2016 Feb;39(2):109-16. doi: 10.1007/s40264-015-0378-5. PMID: 26692396; PMCID: PMC4735229.

[4] Maher RL, Hanlon J, Hajjar ER. *Clinical consequences of polypharmacy in elderly*. Expert Opin Drug Saf. 2014 Jan;13(1):57-65. doi: 10.1517/14740338.2013.827660. Epub 2013 Sep 27. PMID: 24073682; PMCID: PMC3864987.

[5] Baker C, Feinstein JA, Ma X, Bolen S, Dawson NV, Golchin N, Horace A, Kleinman LC, Meropol SB, Pestana Knight EM, Winterstein AG, Bakaki PM. *Variation of the prevalence of pediatric polypharmacy: A scoping review*. Pharmacoepidemiol Drug Saf. 2019 Mar;28(3):275-287. doi: 10.1002/pds.4719. Epub 2019 Feb 6. PMID: 30724414; PMCID: PMC6461742.

[6] Hosmer Jr, David W., Lemeshow S., Sturdivant R. *Applied logistic regression*. John Wiley & Sons, 2013.

[7] Taghy N, Ramel V, Rivadeneyra A, Carrouel F, Cambon L, Dussart C. *Exploring the Determinants of Polypharmacy Prescribing and Dispensing Behaviors in Primary Care for the Elderly-Qualitative Study*. Int J Environ Res Public Health. 2023 Jan 12;20(2):1389. doi: 10.3390/ijerph20021389. PMID: 36674148; PMCID: PMC9859068.

[8] Zanin A, Baratiri F, Roverato B, Mengato D, Pivato L, Avagnina I, Maghini I, Divisic A, Rusalen F, Agosto C, Venturini F, Benini F. Polypharmacy in Children with Medical Complexity: A Cross-Sectional Study in a Pediatric Palliative Care Center. Children (Basel). 2024 Jul 4;11(7):821. doi: 10.3390/children11070821. PMID: 39062270; PMCID: PMC11274911.

[9] Carmona-Torres JM, Rodríguez-Borrego MA, Laredo-Aguilera JA, Cobo-Cuenca AI, Santacruz-Salas E, López-Soto PJ. *Polypharmacy and associated factors: a gender perspective in the elderly*. Frontiers in Pharmacology. 2023 Apr 20;14:1189644. doi: 10.3389/fphar.2023.1189644.

[10] Briesacher BA, Limcangco R, Gaskin DJ. *Racial and Ethnic Disparities in Prescription Coverage and Medication Use*. Health Care Financing Review. 2003;25(1):63-76. doi: 10.1371/journal.pone.0159224.