



JOHNS HOPKINS  
M E D I C I N E

## Use of R for implementing statistical methods in clinical research

Biostatisticians' efforts to promote open-source and useful R programs for producing efficient and reproducible results

Presented by: Ruizhe Chen, Hanfei Qi

October 8, 2024

# Who we are?

- Division of Quantitative Sciences, SKCCC, SOM ([Division of Quantitative Sciences - Johns Hopkins Sidney Kimmel Comprehensive Cancer Center \(jhmi.edu\)](#))



**Ruizhe Chen, Ph.D, M.S.**

Instructor of Oncology



**Hanfei Qi, M.S.**

Biostatistician

# Who we are?

- Division of Quantitative Sciences, SKCCC, SOM  
([Division of Quantitative Sciences - Johns Hopkins Sidney Kimmel Comprehensive Cancer Center \(jhmi.edu\)](https://www.jhmi.edu))
- **Biostatistics:** Clinical trial design and analysis, Statistical method and tools development (e.g., R packages: expectation-maximization (EM), minorization-maximization (MM), and proximal gradient algorithms)
- **Bioinformatics:** high-throughput genetic sequencing, transcriptional profiling, epigenetics, and single-cell and spatial multi-omics

# Objectives of the project (Why we do it?)

- Biostatisticians are also software developers (often not full-time; in statistical software languages: R)
- Objectives:
  - (1) New statistical method and corresponding R programs development: Tools to simulate multivariate zero-inflated count data
  - (2) Write statistical programs and demonstration examples for statisticians to better conduct statistical research/analysis (for efficiency, reproducibility, and education)
  - \*All R codes go on our division's GitHub page
    - [Biostatistics Posts | OncologyQS](#)

# Objective One

- Zero-inflated count data are commonly observed in toxicity/adverse events collected in oncology clinical trials
- There is a lack of methods and proper tools in simulating such data (ZIGP distributed)
- We proposed new methods in simulating univariate and multivariate ZIGP data (manuscript)
- We also wrote R functions to perform such tasks

# Random Number Generator for Zero-Inflated Generalized Poisson Data (RNGforZIGPD)



Qian Shi, PhD  
Mayo Clinic

Biometrical Journal XX (2024) XX, zzz-zzz / DOI: 10.1002/bimj.200100000

## Multivariate Zero-Inflated Generalized Poisson Data Generation Methods for Simulating Counts of Adverse Events

Ruizhe Chen\* <sup>1</sup>, Qian Shi <sup>2</sup>, and Hakan Demirtas <sup>3</sup>

<sup>1</sup> Division of Quantitative Sciences, Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University, 550 N Broadway, Suite 1103, Baltimore, MD, 21205, USA

<sup>2</sup> Department of Quantitative Health Sciences, Mayo Clinic, Harwick 8-28, 200 First Street SW, Rochester, MN 55905, USA

<sup>3</sup> Division of Epidemiology and Biostatistics, School of Public Health, University of Illinois Chicago, 1603 W Taylor St, Chicago, IL 60612, USA

Received zzz, revised zzz, accepted zzz

Counts of maximum grade adverse events collected in clinical trials are important measurements of treatments' toxicity and tolerability. Studying the frequencies and correlations of adverse event counts by types, treatment cycles, and grades can provide further insights into the toxicity profiles of the underlying treatments. A prerequisite to establish such statistical inferential methods is the ability to properly generate multivariate count data with designated event rates and correlation structures. In this article, we present three novel methods for simulating multivariate count data that follow zero-inflated generalized Poisson (ZIGP) distributions. The proposed methods can simulate arbitrarily specified pair-wise correlations within feasible ranges from the target distribution. We developed the methods under the NOrmal-To-Anything (NORTA) and Sample-Iterate (SI) data simulation frameworks. Our simulation study results show great performance of the proposed approaches in simulating ZIGP distributed count data with desired rate, scale, proportion of zeros, and correlation matrices. We apply the proposed methods in simulating AE counts based on the NCCTG Study N9741, a randomized multicenter phase III colorectal cancer study. The presented method can also have a broader applicability where we showcase a scenario in simulating counts of hospital visits based on a National Medical Expenditure Survey dataset.


**Key words:** Random Number Generator; Zero-Inflation; Generalized Poisson; Pearson Correlation; Multivariate Count Data

Supporting Information for this article is available from the author or on the WWW under <http://dx.doi.org/10.1022/bimj.XXXXXXX>



Hakan  
Demirtas, PhD  
University of  
Illinois Chicago

# Invited Seminar Talk on the RNGforZIGPD Paper & the R Functions



**mathweb**  
Institutional Group • 58 people


Home Posts Events Files



## Stat Colloquium: Dr. Ruizhe Chen

Johns Hopkins University

Friday, October 11, 2024 · 11 AM - 12 PM

Mathematics/Psychology : 401 

**TITLE:** Multivariate Zero-Inflated Generalized Poisson Data Generation Methods for Simulating Counts of Adverse Events

**ABSTRACT:** Counts of maximum grade adverse events collected in clinical trials are important measurements of treatments' toxicity and tolerability. Studying the frequencies and correlations of adverse event counts by types, treatment cycles, and grades can provide further insights into the toxicity profiles of the underlying treatments. A prerequisite to establish such statistical inferential methods is the ability to properly generate multivariate count data with designated event rates and correlation structures. In this talk, we present methods for simulating multivariate count data that follow zero-inflated generalized Poisson (ZIGP) distributions. The proposed methods are developed based on the Normal-To-Anything (NORTA) and Sample-Iterate (SI) data simulation frameworks. In particular, we have adapted the NORTA with correlation adjustment by polynomial regression approach to the case of ZIGP distributed marginals. Our simulation study results show great performance of the proposed approaches in simulating ZIGP distributed count data with desired rate, scale, proportion of zeros, and correlation matrices. We apply the proposed methods in simulating AE counts based on the NCCTG Study N9741, a randomized multicenter phase III colorectal cancer study. The presented method can also enjoy a broader applicability where we showcase a scenario in simulating counts of hospital visits based on a National Medical Expenditure Survey dataset.

### Event Info

posted September 7, 2024

sponsor mathweb

tags [stat-colloq](#) [stat-colloq-f24](#)

share     

[add to calendar](#)

### Recent Events

-  Joint Math-Stat Colloquium: Dr. Daniel Reynolds  
Nov 22 at 11 AM
-  Stat Colloquium: Dr. Qing Mai  
Nov 8 at 11 AM
-  Stat Colloquium: Dr. Jing Li  
Oct 25 at 11 AM
-  Stat Colloquium: Dr. Akim Adekpedjou  
Oct 18 at 11 AM
-  Applied Mathematics Colloquium: Lili Du (UF)  
Nov 15 at 11 AM

# Random Number Generator for Zero-Inflated Generalized Poisson Data (RNGforZIGPD)



- Real Data Application (Manuscript):
  - The NCCTG Study N9741 was a randomized multicenter phase III study conducted to compare three treatment regimens: irinotecan and bolus fluorouracil plus leucovorin (IFL, control combination), oxaliplatin and infused fluorouracil plus leucovorin (FOLFOX), or irinotecan and oxaliplatin (IROX) in patients with previously untreated metastatic colorectal cancer.
- Codes and Examples available @
  - [Biostatistics Posts | OncologyQS](#)



# Objective Two

- To build a platform to summarize and standardize useful statistical software programs that are often used in producing statistical analysis reports and plans for oncology clinical trials.
  - GitHub page displays publicly available data and code
  - The Quantitative Sciences Division's GitHub Organization to store repositories (Internal)

# Oncology QS GitHub Page

- Instruction for generating Data and Safety Monitoring Boards (DSMBs) report.
- Code for efficacy and toxicity monitoring using Bayesian predictive probabilities.
- Tables: Table1; Cox PH models summary table;
- Plots: Kaplan-Meier curve; Forest plot; Swimmer plot.

# Future Plans

- \*An R package for RNGforZIGPD
- Existing documentations will improve as more team members contribute.
- More educational resources for fellows and coordinators from School of Medicine.
- General ideas on available statistical methods for oncology studies, along with their corresponding recommended R packages

***Thank you!***

**Sloan Foundation & Open-  
Source Office**