IDENTIFYING SOURCES OF ZEROS: METHODOLOGICAL ADVANCES IN MODELLING ZERO-INFLATED COUNT DATA IN PUBLIC HEALTH RESEARCH

Lawal Olumuyiwa Mashood[1*], Abubakar Yahaya[2], Yahaya Zakari[2], Musa Tasi'u[2] and Abdulwasiu Oyelola Adegboye[3]

[1]Department of Statistics, Air Force Institute of Technology, Mando, Kaduna, Nigeria
[2]Department of Statistics, Ahmadu Bello University, Zaria, Kaduna, Nigeria
[3]Menzies School of Health Research, Charles Darwin University, Darwin, NT, Australia

*Corresponding Author email: lawal.mashood@afit.edu.ng

ABSTRACT

Count data exhibiting an abundance of zeros are prevalent in public health; nonetheless, differentiating between structural zeros (true absence) and sampling zeros (measurement constraints) continues to pose a significant methodological problem. Following PRISMA recommendations, we conducted a systematic literature review that distinguishes zero sources by searching PubMed, Web of Science, and Scopus (2000–2023) using Boolean operators and keywords associated with zero-inflated models, epidemiology, and health surveillance. Peer-reviewed, English-language research on structural and sampling zeros was one of the inclusion criteria. Fifteen papers were subjected to full-text review and quality evaluation following the screening of 2,695 records. In handling excess zeros, zero-inflated (ZI) and hurdle models fared better than conventional Poisson/negative binomial models. The research emphasised that epidemiological background should take precedence over fit statistics when selecting a model. ZI models, for instance, are suitable for data with both zero types. In contrast, hurdle models are suitable for situations with sampling zeros alone (such as hurdles to healthcare access). Among the main breakthroughs were Bayesian frameworks (like zinLDA) for microbiome data, which were able to separate structural zeros from technical noise with a sensitivity of >90%; spatially varying coefficient models that were able to distinguish between underreporting and cholera susceptibility zeros; and longitudinal algorithms (like LUMINATE) that were able to distinguish between biological zeros and technical zeros in microbial time-series. Model accuracy and inference validity are increased when zeros are explicitly classified. Based on sensitivity studies and the zero-generation process, researchers ought to use ZI/hurdle models.

Keywords: Zero-inflated models, structural zeros, sampling zeros, excess zeros, hurdle models, count data, epidemiology, systematic review.

1. INTRODUCTION

Count data with excess zeros are pervasive in epidemiological, environmental, and social science research. Such data frequently exhibit bimodal distributions combining true absence (structural zeros) and unobserved events due to sampling limitations (sampling zeros) (He *et al.*, 2014a; Tang *et al.*, 2017). This can arise in contexts like disease incidence reports (e.g., cholera outbreaks), microbiome analyses, drug consumption surveys, and healthcare utilisation studies, where zero counts represent either true absence (structural zeros) or unobserved events due to sampling limitations (sampling zeros). Structural zeros signify inherent non-susceptibility (e.g., populations immune to a disease), while sampling zeros stem from measurement errors, detection limits, or misreporting. Traditional count models (e.g., Poisson or negative binomial regression) fail to differentiate these sources, leading to biased estimates, inflated Type I errors, and flawed policy inferences.

The challenge intensifies in public health, where zero-inflated (ZI) data complicate disease surveillance, risk assessment, and evaluation of interventions. For instance, spatial models for cholera must distinguish regions with no cases due to immunity (structural zeros) versus underreporting (sampling zeros). Similarly, microbiome studies grapple with distinguishing bacterial absence from technical detection limits. Failure to distinguish these sources violates assumptions of standard count models (e.g., Poisson, negative binomial), leading to biased estimates, inflated Type I errors, and compromised inference (Rose *et al.*, 2006; Greene *et al.*, 2017). Despite methodological advances, such as zero-inflated Poisson (ZIP), zero-inflated negative binomial (ZINB), and hurdle models, guidance on selecting context-appropriate frameworks remains fragmented.

Hurdle models (Mullahy, 1986) offer an alternative, treating all zeros identically but modelling positive counts separately. While both frameworks address overdispersion, ZI models are theoretically preferred when structural zeros exist (Rose *et al.*, 2006; Zavras, 2019).

This systematic review addresses this gap by synthesising literature on modelling strategies for zero-inflated count data, focusing on methods that explicitly disentangle structural and sampling zeros. We evaluate applications across epidemiology, microbial ecology, and social sciences, highlighting innovations in Bayesian frameworks, spatial varying coefficient models, and longitudinal zero-detection algorithms. Our goal is to consolidate best practices for researchers navigating zero-inflated data complexities, ensuring robust statistical inference and informed public health decisions.

2. MATERIALS AND METHODS

The search strategy used is a content-based qualitative analysis to review the literature on different approaches for modelling count data with excess zeros, where researchers are unsure if it is sourced from structural or sampling zeros. The search strategy was meticulous in both organising and performing, using well-established frameworks to guarantee the comprehensiveness of the literature search. Databases, keywords, inclusion and exclusion criteria, and other precise and well-defined search parameters were employed as part of the research methodology. It is anticipated that this method will reduce prejudice and guarantee that a wide range of research covering different situations and viewpoints is included. Because the investigation covered a certain amount of time, the most recent and pertinent research findings could be obtained. The technique established the time parameter to offer transparency and contextual information for the assessment procedure.

An evidence-based minimum set of items for reporting in systematic reviews and meta-analyses, known as the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (Liberati *et al.*, 2009), was employed. Though PRISMA can serve as a foundation for publishing systematic reviews with goals other than evaluating interventions, its primary focus is on the reporting of reviews evaluating the techniques for modelling the coexistence of structural and sampling zeros. Because of the thorough approach, the results are more reliable and representative of the body of recent research.

The source of data used is a secondary type of online publications using the search terms listed in Table 1. A comprehensive search was conducted on all published papers and screened using three electronic databases: PubMed, Web of Science, and Scopus. The articles extracted are pertinent to proffering answers to the research questions. Every paper was sifted based on journal quality. Thematically retrieved keywords were funnelled through Boolean operators and truncations to improve the search strategy before being utilised in the chosen electronic reference databases. Relevant studies downloaded from the three databases were imported and screened using Rayyan (Ouzzani *et al.*, 2016).

Table 1: Searched String Terms

(count* AND data) AND ((true* OR impu* OR struct* OR inflat* OR excess*) AND (zero* OR null)) AND (model* OR regress* OR statistic* OR “machine learning” OR “deep learning” OR “artificial intelligence”) AND (“epidemiology” OR “disease” OR “public health” OR “global health” OR “one health” OR “health” OR “health surveillance” OR “surveillance” OR spread* OR mapping OR prevalence* OR incidence* OR mortalit* OR death* OR case* OR morbidit*)

The inclusion and exclusion criteria are listed in Table 2. Articles were independently screened using the information found in the article's title and abstract to determine whether or not to include them. To choose which articles would be included in the material evaluation, relevant articles based on the research questions were then evaluated for full-text review eligibility. Articles evaluated were reports, expert reviews, and/or commentary that lacked original research; only the relevant papers to the research questions were acknowledged.

Table 2: Inclusion and Exclusion Criteria

Criterion	Inclusion	Exclusion
Date	Studies done between **January 1, 2000** and **December 2023**	
Geographic location of study	Studies from **any country** in the world	
Language	Studies published in **English**	Studies published other than English
Participants	Studies focusing on **any age groups** as subject headings	
Peer Review	**Grey literature**, that is, papers without bibliographic information such as publication date/type, volume and issue numbers	**Non-peer-reviewed** literature
Reported outcomes	Studies whose outcomes of interest have been **reported**	Studies that are **self-reported** rather than using objective measures
Type of publication	**Original studies:** All published papers that have the potential to answer at least one research question	Papers such as publications, reviews, editorials, and letters that do not have any link with the research questions

3. RESULTS AND DISCUSSION

The preliminary stage of the search yielded 2,695 references, comprising 955 papers from PubMed, 312 from Web of Science, and 1,428 papers from Scopus. To get rid of duplicates and irrelevant studies, 1,771 possible duplicates were detected. Afterwards, a review and compilation of relevant papers was done using the titles and abstracts of these studies; 950 papers were deleted as a result of duplication. Out of the remaining 1,745 references examined, to were significant research that might have gone unnoticed in the initial search stage. Subsequently, 17 relevant research articles were chosen and subjected to further screening called full text screening. Lastly, these 17 references were subjected to the quality evaluation criteria. After the exercise, 15 papers were chosen and judged qualified to proffer answers to the research questions (See the PRISMA chart below in Figure 1). The selected papers were directly imported into the Mendeley reference manager.

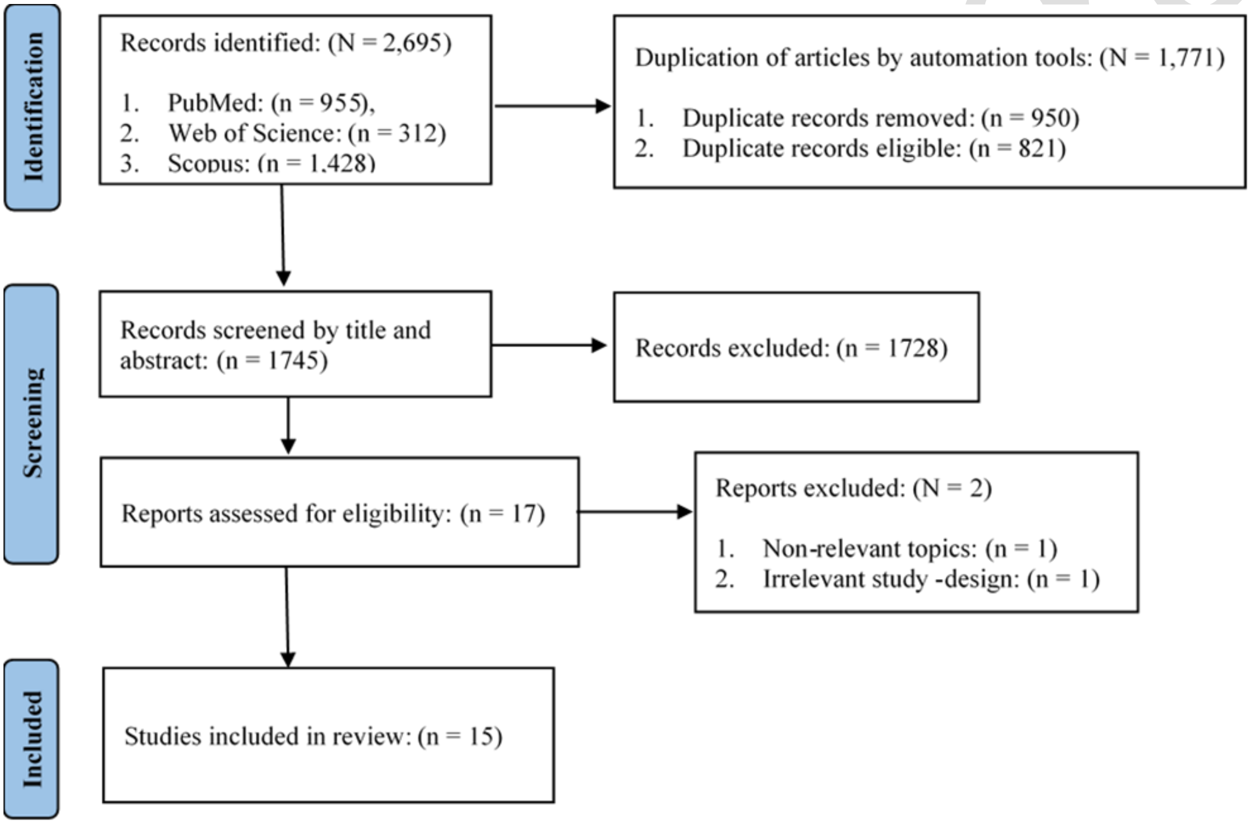
Identification
Records identified: (N = 2,695)
1. PubMed: (n = 955),
2. Web of Science: (n = 312)
3. Scopus: (n = 1,428)
Duplication of articles by automation tools: (N = 1,771)
1. Duplicate records removed: (n = 950)
2. Duplicate records eligible: (n = 821)
Screening
Records screened by title and abstract: (n = 1745)
Records excluded: (n = 1728)
Reports assessed for eligibility: (n = 17)
Reports excluded: (N = 2)
1. Non-relevant topics: (n = 1)
2. Irrelevant study -design: (n = 1)
Included
Studies included in review: (n = 15)

Fig. 1: PRISMA chart

Each article's data was extracted and subsequently presented as a table (see Table 3). Information on the article's origin, bibliographic details (such as author(s) names, year of publication, study country), study objective(s), data aspects (such as sources of data, surveillance data, sample, study design, data collection, data analysis techniques, and key findings were all included in the extracted data.

Table 3: Extracted article's data

S/N	Title	Bibliographic details			Study objectives	Data aspects		Methodology		Key findings
		Author's name	**Year of publication**	**Study country**		**Sources of data**	**Type of data (Surveillance, Cross-sectional**	**Method**	**Link function**	
1.	A GEE-type approach to untangle structural and random zeros in predictors	Peng Ye; Wan Tang; Jiang He; and Hua He	2018 Nov 26	USA	To address the issue of structural zeros in ZI count predictor data to distinguish between structural and random zeros	A randomised clinical study for teaching awareness and self-monitoring skills to indwelling urinary catheter users	Cross-sectional data	A semi-parametric GEE-type mixture model to simultaneously model the outcome and the ZI count predictor	Logit and Log	The GEE-type method performs well in practical settings and is more robust to model misspecification than the likelihood-based counterpart. It is interesting to apply the GEE-type method to analyse survival data with ZI predictors
2.	A zero-inflated latent dirichlet allocation model for microbiome studies	Rebecca A. Deek and Hongzhe Li	2021 Jan 22	USA, UK, and Australia	To develop a model that differentiates between structural and sampling zeros in microbial data using a Bayesian framework, and accurately identifies subcommunity structures and representative taxa	Dataset from the citizen scientists of the American Gut Project (AGP) and identify microbial communities characterised by different bacterial genera	Microbiome data	A zero-inflated Latent Dirichlet Allocation model (zinLDA) for sparse count data observed in microbiome studies		zinLDA provides better fits to the data and is able to separate structural zeros from sampling zeros with reasonable sensitivity and specificity

3.	A zero-inflated mixture spatially varying coefficient (SVC) modelling of cholera incidences	Frank Badu Osei, Alfred Stein, and Veronica Andreo	2022 Feb 9	Ghana	To investigate the spatial patterns of cholera counts with excess zeros. To estimate the zeros arising from the non-susceptible and susceptible populations, and To examine the performance of ZINB SVC in the determination of the spatially varying effects of precipitation and LST on cholera spatially	2014 cholera outbreak dataset acquired from the Centre for Health Information and Management (CHIM) of the Ghana Health Services (GHS); and precipitation data were obtained from the Climate Hazards Group Infrared Precipitation with Stations (CHIRPS) version 2.0 database	Cholera surveillance and precipitation data	ZI mixture SVC modeling framework	Logit and Log	The proportion of zeros in cholera cases varies spatially. If the objective is to distinguish between the sources of the zero counts, then it may not be important to use the Poison SVC as a benchmark for model comparison. On this, the authors conclude that the choice of a ZI mixture model over a Poisson SVC model should be based on the epidemiological significance for cholera monitoring rather than the model fit
4.	Analysing differences between microbiome communities using mixture distributions	Konstantin Shestopaloff; Michael D. Escobar; and Wei Xu	2018 Oct 28	Canada	To develop a method for analysing microbiome data that accurately estimates the proportions of structural and non-structural zeros; To model the	A cohort of first-degree relatives of Crohn's disease patients gathered as part of the Genetics, Environment, and	A cohort study and Microbiome data	A mixture model focused on a single operational taxonomic unit (OTU) and a Poisson distribution to model subject-specific counts. The Mixture components were defined based on the posterior		The proposed method accurately estimates the proportion of zeros for true presence-absence; models the underlying rate distribution for low counts effectively; and has good power for effectively detecting differences between structural and non-structural zeros, improving bias

					underlying rate distribution for low counts, accounting for the inherent sparsity of these observations, and To differentiate between structural and nonstructural zeros	Microbiome (GEM) study (Turpin *et al.*, 2016); and a study of environmental microbiomes collected in office environments limited to dust collection at university campuses in Toronto, Ontario, Canada, and Flagstaff, Arizona, USA (Chase *et al.*, 2016).		distribution of the Poisson rate, allowing for efficient computation of quantities required for permutation tests		identification and model fitting
5.	Efficient and accurate inference of mixed microbial population trajectories from longitudinal count data	Tyler Joseph; Amey P. Pasarkar; and Itsik Pe'er	2020 Jun 24	USA	To derive an inference algorithm for the model and variational parameters. To evaluate model performance through simulations, and To apply the model to real data for	The *C. diff* dataset contains both sequencing counts using 16S sequencing and biomass estimation using quantitative PCR	Time-series microbiome data	LUMINATE (longitudinal microbiome inference and zero detection)		LUMINATE can accurately distinguish biological zeros, when a taxon is absent from the community, from technical zeros, when a taxon is below the detection threshold; it outperforms other models: TGP-CODA (Äijö *et al.*, 2018) and MALLARD (Silverman, 2019) in terms of accuracy and efficiency; and the method runs faster than existing approaches without loss

					relative abundance estimation					of accuracy.
6.	Fitting a distribution to microbial counts: Making sense of zeroes	Ana Sofia Ribeiro Duarte, Anders Stockmarr, and Maarten Nauta	2015 Mar 2	Denmark	To demonstrate why the LOQ should be excluded from the analysis of microbial data. To develop a model that estimates the parameters of a distribution of microbial concentrations without assuming any LOQ; and To develop a new method that differentiates between artificial and true zeroes in microbial counts	A contaminated food unit assumed as a product containing one or more colony forming units (CFU)	Microbial data	Developed a method that fits a discrete zero inflated Poisson-lognormal (PLN) distribution to raw plate count data to estimate true prevalence of contamination and within-lot distribution of concentrations without assuming a limit of quantification (LOQ).		The new method for analysing quantitative microbial data is effective in estimating true prevalence and concentration parameters, making it a valuable tool for analysing quantitative microbiological risk assessment enumeration data; The method has the ability to differentiate between artificial and true zeroes in microbial counts, providing more accurate estimates; The study suggests that the concept of a theoretically established LOQ may lead to false interpretations of zero counts and recommends avoiding the use of LOQ in the analysis of microbial data; The new method provided accurate estimates of mean, standard deviation, and prevalence, especially at low true prevalence levels and low expected standard deviation; The model performs better with lower standard deviations or higher means for μ, while underestimating σ across most scenarios.
7.	Misreporting and econometric modelling of zeros in	William H. Greene, Mark N. Harris, Preety	2017 Aug 04	Australia	To develop a modelling approach to analyse misreporting	Data from the Australian National Drug	Longitudinal data	The Double ZIOP (DZIOP) model was proposed to address misreporting in		The model suggests that misreporting significantly affects the incidence of cannabis use; It estimates that

	survey data on social bads: An application to cannabis consumption	Srivastava, and Xueyan Zhao			in drug consumption data, focusing on cannabis use	Strategy Household Survey is used for the analysis		sensitive responses like illegal drug consumption from survey data		17% of reported zeros for cannabis use are due to misreporting, highlighting the impact of inaccurate reporting on survey data; and the findings emphasise the importance of considering misreporting in modelling drug consumption and its implications for policy-making
8.	Modelling repeated count measures with excess zeros in an epidemiological study	Rahul Gupta; Rhonda D. Szczesniak; and Maurizio Macaluso	2015 Aug	Alabama	To compare the ZIP-mixed model with traditional Poisson and negative binomial models in analysing count data with excess zeros in an epidemiological study on female condom use problems	Data collected through interviews and questionnaires	Longitudinal data	Used a longitudinal ZIP mixed effects model to analyse repeated assessments of female condom use problems in women attending STD clinics	Logit and Log	The ZIP-mixed regression model provided better fit compared to Poisson and negative binomial models; employing ZI models in epidemiologic research provides improved data fit and a richer interpretation of condom use problem determinants, meanwhile ignoring excess zero counts in skewed data may lead to underestimation of variability and incorrect inferences; and the ZIP-mixed regression model was more informative in assessing female condom failure compared to traditional models, highlighting age and beliefs as significant predictors
9.	Modelling Time Series of Counts in Epidemiology	Alexandra M. Schmidt and Jonny Everson Scherwinski Pereira Hua	2011 Feb 01	Brazil	To review and propose generalised dynamic models for time series of count data, focusing on overdispersio	Artificial dataset and weekly observed counts of dengue fever cases in	Temporal data	Generalised dynamic Poisson models considering Poisson-gamma, and Poisson-log-normal (Poi-LN) mixture models and their		The study concludes that the proposed models, including zero-inflated versions, are effective in analysing count data with overdispersion and excess zeros, such as dengue fever cases, providing insights into

					n and excess zeros, and apply these models to analyse dengue fever cases	the district of Rio de Janeiro. The dataset was obtained from the Rio de Janeiro Municipal Health Secretariat (SMS-RJ)		ZI versions.		the presence of the disease even when no cases are observed. The ZIP model effectively captures the observed processes' presence or absence and provides valuable estimates for model comparison and forecasting. ZI models indicated a high probability of observed zeros coming from the Poisson part.
10.	On the Use of Zero-Inflated and Hurdle Models for Modelling Vaccine Adverse Event Count Data	Charles E. Rose, Stacey W. Martin, and Kathleen A. Wannemuehler	2006 Sep 03	USA	To acknowledge the contributions of individuals and organisations involved in a human clinical trial related to vaccine research	Motivating data was from a multicenter, double-blind, randomised, placebo-controlled human anthrax vaccine absorbed (AVA)	Clinical trial study	Compared the fit of Poisson, NB, ZIP, ZINB, Poisson Hurdle (PH), and Negative Binomial Hurdle models.	Logit and Log	Over-dispersion in count data can be caused by unobserved heterogeneity, temporal dependency, and excess zeroes. ZI and hurdle models fit better than standard models (Poisson and NB). However, the ZI models are suitable for count endpoints with both structural and sample zeroes, while hurdle models are preferred when only sample zeroes are present.
11.	Structural zeroes and zero-inflated models	Hua He; Wan Tang; Wenjuan Wang; and Paul Crits-Christoph	2014 Aug	USA	To highlight the importance of addressing the differences between structural and random zeroes in count data analysis, and To introduce the ZIP model as a solution for handling	Data from a study on HIV-risk sexual behaviours among adolescent girls.	Controlled randomised study	The study provides an overview of structural zeroes, some basic concepts of mixture distribution, the limitations of the Poisson model, and the application of the ZIP regression model.	Logit and Log	The ZIP regression model is an effective technique that provides a better fit than the Poisson regression model when working with ZI data that contains both structural and random zeros. Overdispersed count data may not be suitable for the Poisson model. Data with an excessive number of zeros can be interpreted more thoroughly using the ZI

					ZI data in psychosocial and behavioural studies.					models.
12.	Testing the Dual-State-Process assumption in the preventive care services use	Dimitris Zavras	2019 Nov 30	Greece	To investigate the factors influencing preventive care-seeking behaviour, particularly focusing on infrequency and abstention, and To discuss the applicability of the dual-state process assumption in different health care systems.	The Nomenclature of Territorial Units for Statistics II (NUTS II) was used to base the sample selection strata for the 2011 Hellenic Statistical Authority Census. In 2003, people who were at least 18 years old made up the sample. The technique for gathering data was computer-assisted telephone interviewing (CATI).	Panhellenic cross-sectional survey	A ZINB model was compared to a standard NB model using the corrected Vuong test.		The study comes to the conclusion that abstinence and infrequency are important factors in the behaviour of seeking preventive treatment. In many health care systems, the dual-state process assumption might not hold true, and in developed nations, the use of preventive care is still very low.
13.	Untangle the structural and random zeros in statistical modelling.	Wan Tang; Hua He; Wenjuan Wang; and Ding-Geng (Din) Chen	2017 Oct 24	USA	To address the methodological gap in modelling zero-inflated	The National Centre for Health Statistics (NCHS)	Cross-sectional data	The study uses a maximum likelihood approach to develop parametric	Logit and Log	The study found that both models successfully identified significant associations between alcohol use and depression. The findings

					count data as predictors by developing a parametric method to distinguish between structural and random zeros.	performed a survey research programme in 2009–2010 called the National Health and Nutrition Examination Survey (NHANES), which was used to evaluate the health and nutritional status of adults in the United States (He *et al.*, 2014b). The Patient Health Questionnaire (PHQ-9) is used to measure depressive symptoms .		methods for modelling ZI count data as predictors.		highlighted the importance of distinguishing between random and structural zeros in statistical modeling.

14.	Zero-inflated multiscale models for aggregated small area health data.	Mehreteab Aregay;
		Andrew B. Lawson;
		Christel Faes;
		Russell S. Kirby;
		Rachel Carroll;
			2018 Jan 04	Georgia	To develop zero-inflated multiscale models that jointly describe the risk variations at different geographical levels.	Skin cancer data available from the state of Georgia via the Georgia Division	Cross-sectional data	The proposed zero-inflated multiscale models use joint convolution models that describe the Risk variation at multiple scale levels via a	Logit and Log	The proposed zero-inflated multiscale models provide a consistent risk estimate at the fine and coarse levels when high percentages of structural zeros are present in the data.
										The models are flexible

		Kevin Watjou			To study the spatial distribution of disease incidence at different geographical levels, addressing the problems of excessive zeros and scaling effects in aggregated data.	of Public Health OASIS system (http://oasis.state.ga.us) at both the county and public health (PH) district levels in 2008.		shared random effect component.		and can be extended beyond two scale levels, making them a useful tool for studying the spatial distribution of disease incidence at different geographical levels. The study highlights the importance of accounting for excessive zeros and overdispersion when modelling the relative risks of skin cancer at multiple scales. This study concludes that the models that account for scaling effects and excessive zeros are better suited for analyzing small area health data with heavily skewed zeros.
15.	Zero-inflated Poisson model-based likelihood ratio test for drug safety signal detection	Lan Huang; Dan Zheng; Jyoti Zalkikar; and Ram Tiwari	2017 Feb	USA	To develop a method that takes into account the high percentages of observed zero cells in developing statistical signal detection methods.	2006–2011 Adverse Event Reporting System (AERS) data with all drugs and AEs	Clinical trials	The proposed method uses a ZIP model to handle the true zeros and develops the likelihood ratio test (LRT) for signal detection.		The ZIP LRT method is a useful approach for signal detection in adverse event reports that incorporates the effect of binary or categorical factors. The ZIP LRT method is recommended for signal detection in drugs from the AERS database with a large number of zero-count cells, especially when the estimated percentage of true zeros is high.

The presence of zeros may necessitate the use of specialised models, such as ZI models or hurdle models, to accurately capture the data distribution. However, only a few studies have previously been carried out on how to classify the intertwined relationship between structural (or genuine) and sampling (imputed) zeros in the data. Unfortunately, conventional regression techniques frequently produce biased and incorrect results since they are unable to distinguish between various kinds of zeros. More so, the nature of these zeros is not uniform; they can arise from different sources. It is important to recognise that imputed zeros might have different characteristics compared to structural zero counts. For instance, they may depend on the limitations in data collection, missing data mechanisms, reporting, or recording processes, the imputation method used, or potential biases in the imputed values.

The approaches currently employed for handling zeros in individual records are: Zero-inflated models, such as ZIP and ZINB models, account for the presence of structural zero counts and imputed zeros in count data; multiple imputation techniques are used to handle missing data, including imputed zeros, by generating multiple plausible imputations to incorporate uncertainty into the analysis. Sensitivity analysis helps assess the robustness of results based on the classification of zeros and missing data, exploring how different assumptions impact the conclusions.

According to Rose *et al.* (2006), temporal dependency, extra zeroes, and unobserved heterogeneity can all lead to over-dispersion in count data. Over-dispersion brought on by temporal dependency and unobserved heterogeneity can be addressed using the NB model. Compared to standard models (Poisson and NB), ZI and hurdle models fit better. The modelling framework should be taken into consideration if the study's goal is to make inferences; ZI and hurdle models should be used according to the goals and design of the study. For instance, the ZI modelling framework is generally more appropriate if the study design results in the counting of endpoints with both structural and sample zeros. On the other hand, hurdle models are generally favoured when the endpoint of interest exhibits just sample zeros by design. Conversely, the ZI and hurdle modelling frameworks should also be sufficient if the study's main goal is to create a prediction model. The ZI and hurdle models are favoured over standard models by likelihood ratio tests (LRTs), AIC, and BIC; as a result, ZINB and NBH models are suggested for modelling vaccine adverse event count data. There is limited generalizability because the models mentioned were only used with a particular clinical trial dataset. The study emphasises the importance of collaboration between professionals and funding agencies in carrying out effective clinical trials for vaccine development.

Generalised dynamic Poisson models were proposed by Schmidt and Pereira (2011), taking into account ZI versions of the Poisson-gamma and Poisson-log-normal (Poi-LN) mixing models. In terms of model fitting, the models produced findings that were similar. Various models demonstrate good fit to the observed values, with mixture models providing marginally broader credible intervals. According to the study's findings, count data with excess zeros and overdispersion, like dengue fever cases, can be effectively analysed using the proposed models, including zero-inflated versions, to reveal the presence of the illness even in the absence of cases. The ZIP model yields useful estimates for model comparison and forecasting, and it reflects the presence or absence of the observable processes in an effective manner. Models with inflated zeros showed a high likelihood that the observed zeros originated from the Poisson part. The study emphasised how crucial it is to take ZI models for count data into account.

He *et al.* (2014b) highlight the prevalence of structural zeros in count data and show how the ZIP model can handle zero-inflated data with structural zeros in an efficient manner. It also describes the components of the ZIP model for modelling Poisson means and structural zeros, as well as the methods underlying it. The study made clear how crucial it is to distinguish between structural and random zeros when analysing count data, since, if improperly handled, structural zeros in count data might result in false interpretations. When dealing with ZI data that contains both structural and random zeros, the ZIP regression model is a useful tool that offers a better fit than the Poisson regression model. The Poisson model might not be appropriate for count data that has overdispersion. The ZI models offer a more comprehensive interpretation of data that contain an excessive number of zeros and can be used to assess how interventions affect outcomes such as adolescent girls' sexual behaviour. Depression was not significantly associated with the outcome at three months, and higher baseline sexual behaviour was associated with being a structural zero in the outcome at three months. HIV knowledge and baseline sexual behaviour were highly associated with the number of vaginal sex encounters using condoms. A prominent constraint is that the Poisson and NB models are unable to distinguish between random and structural zeroes in count data. Furthermore, there can be restrictions on how the ZIP model can address particular kinds of overdispersion.

Structural zeros in response variables have been the subject of much research, however, Gupta *et al.* (2015) used a ZIP-mixed regression model to examine the variables influencing female condom failure and demonstrate the ability of the model to distinguish between those who are genuinely problem-free (structural zeros) and those who may be unable or reluctant to acknowledge their issues (reporting zeros). When used in epidemiologic research, ZI models offer better data fit and a more comprehensive understanding of the factors that contribute to condom use problems. In comparison to Poisson and NB models, the model fits the data better; nevertheless, if excess zero counts in skewed data are ignored, the variability may be underestimated, and wrong conclusions may be drawn. When evaluating female condom failure, the ZIP-mixed regression model provided more information than traditional models, emphasising the importance of age and beliefs as factors. The structural zero group was considerably more likely to include older women. However, the assumptions made about missing data due to attrition after the third follow-up visit and self-reported data on condom use issues that could be biased were noticeable limitations in their study.

To determine the true prevalence of contamination and the within-lot distribution of concentrations, Duarte *et al.* (2015) developed a method that fits a discrete zero-inflated Poisson-lognormal (PLN) distribution to raw plate count data without assuming a limit of quantification (LOQ). The technique is useful for evaluating quantitative microbiological risk assessment enumeration data since it can accurately estimate genuine prevalence and concentration characteristics. The study advises against using LOQ in the analysis of microbial data, citing the possibility of incorrect interpretations of zero counts due to the concept's theoretical establishment. Instead, a more accurate estimate can be obtained by using a method that can distinguish between genuine and artificial zeroes in microbial counts; While underestimating σ in the majority of scenarios, the model performs better when there are lower standard deviations or higher means for μ. It also produced accurate estimates of mean, standard deviation, and prevalence, particularly when there are low true prevalence levels and low expected standard deviation.

Greene *et al.* (2017) address misreporting in survey data on social bad (focused on cannabis usage), comprising sensitive responses like illegal drug intake by proposing a Double Zero-Inflated Ordered Probit (DZIOP) model applied to illicit drug use recorded on an ordinal scale. The main focus of their work is on the idea of "misreporting," which occurs when people give false or misleading information in surveys. Traditional models frequently fail to account for this phenomenon, which results in conclusions regarding drug usage that are not true. Three tiers make up the model: participation determination, misreporting identification, and an ordered probit model to ascertain consumption levels for accurately reporting participants. This method can be used with different health models that have sensitive responses, though. The data set consists of a sample of 50,153 Australian citizens, including information on cannabis consumption over the previous 12 months gathered using specific questions. According to the model, misreporting has a significant impact on the incidence of cannabis use. Participant misreporters, nonparticipants, and infrequent users can all contribute to zero observations in the data; misreporters account for 17% of reported zeros, infrequent users for 1%, and true nonparticipants for 72% of the observations. Because accurate data on drug use is essential for policy analysis, particularly when taking into account the potential misreporting due to legal risks and social stigma associated with drug consumption, the findings highlight the significance of taking misreporting into account when modelling drug consumption and its implications for policy-making.

The distinct effects of structural zeros and random zeros in variables related to alcohol use and depression were addressed by Tang *et al.* (2017) through the development of statistical models, predicting the difference between structural and random zeros is essential to prevent biased estimations, and the novel method enables modelling of ZI count data used as predictors. This methodological difficulty is addressed by the developed parametric technique. The developed method can be used in a variety of fields, including social science, behavioural studies, and public health, to analyse the relationship between traits and outcomes, particularly when differentiating between different types of zeros (structural and random), and to accurately model ZI count data as predictors, improving the validity of statistical inferences. The model can be used with a variety of response variables, including count, binary, continuous, and survival data. When the main model is properly stated, estimates are relatively close to true values. The estimations of some parameters are impacted by misspecification of the main model, especially when structural zeros are neglected.

Based on assumptions about conditional independence and the comprehensiveness of the main model, as well as an auxiliary zero-inflated model to handle zero-inflated count data, the study found that both models successfully detected substantial relationships between alcohol use and depression. Moderate alcohol use or abstinence from alcohol was protective against depression. It was proven that the approach was not resistant against departures from assumed parametric models and that it was only applicable to cross-sectional investigations. The results made clear how crucial it is for statistical modelling to distinguish between random and structural zeros. The study found that the differential impact of structural zeros and random zeros in variables linked to alcohol use and depression was successfully addressed by the proposed approach. However, structural zeros in count data are often latent and not directly observable, posing a challenge in modelling, and the method's performance is evaluated through simulations, but may not fully capture real-world complexities. It's only applicable to cross-sectional studies and lacks robustness against parametric models, requiring further research.

The ZIP likelihood ratio test (ZIP LRT) was proposed by Huang *et al.* (2017) as a statistical approach for signal detection in adverse event reports. Based on the ZIP model, which assumes that the observed cell counts are either true zeros with probability ω_j or Poisson modelled counts with probability (1-ω_j), the ZIP LRT method uses a Monte Carlo simulation to find the distribution of the test statistic under the null hypothesis (i.e., the LRT for signal detection). The expectation and maximisation (EM) procedure yields the maximum likelihood estimates of the model parameters p and q under both the null and alternative hypotheses, as well as the estimation of the true zero probability ω_j. A number of parameters are used to assess the method's performance, including power, sensitivity, false discovery rate (FDR), and type I error rate.

The approach is used with suspect and concurrent drug data from the Adverse Event Reporting System (AERS) for the years 2006–2011. The findings demonstrate that by adjusting for potential genuine zeros, the ZIP LRT approach controls the type I error as well as the FDR. On the other hand, the ZIP LRT method, which takes into account the impact of binary or categorical factors, is a helpful strategy for signal discovery in adverse event reports. When detecting signals in pharmaceuticals from the AERS database that have a high proportion of zero-count cells—that is, when the expected percentage of genuine zeros is high—the ZIP LRT approach is suggested. Large clinical trial safety datasets and longitudinal postmarket safety data collected over time can both be analysed using the ZIP LRT approach. A ZINB model can also be used to modify the procedure to account for the dependence and heterogeneity among the n_{ij}'s.

Ye *et al.* (2018) explore the structural zero issue associated with count predictor variables, clarifying their potentially misleading impact when they are found within the predictors themselves, which is a crucial component of regression analysis in medical and public health studies. To simultaneously model the ZI count predictor and the response variable of interest, a Generalised Estimating Equation (GEE)-type mixture model was utilised. The study's limitations were that it only considered cross-sectional data and that the method relied on linear functions linked to the means rather than a full specification of the ZI model used (ZIP) for the auxiliary model (i.e. they only use the information about the mean of the distribution). To account for the extra zeros in the longitudinal data on issues with female condom use reported by women at high risk of sexually transmitted diseases (STDs). To facilitate the efficient computation of permutation tests and the estimation of optimal weights through the use of a least squares objective function, Shestopaloff *et al.* (2018) suggested a mixture model that specifies the components generated from the posterior distribution of the Poisson rate conditional on the count. By emphasising low counts and zero inflation in particular, the strategy tackles some of the drawbacks of the current methods for evaluating microbiome data. Overall, this approach has benefits for microbiome data analysis: it can distinguish between true absences (or structural zeros) and non-structural zeros (or samples that are missed). This helps to provide more accurate estimates of true zeros, especially in cases where the proportion of zeros is high. It also incorporates flexibility by using a mixture distribution and effectively models low counts.

Aregay *et al.* (2018) addressed the issues of excessive zeros and scaling effects in aggregated data by proposing zero-inflated multiscale models to examine the spatial distribution of disease incidence at various geographical levels. The proposed models make use of joint convolution models with a shared random effect component to explain the risk variations at various scale levels. The models, which include correlated and uncorrelated heterogeneity, are an extension of the convolution model put forth by Besag *et al.* (1991). When the data has a significant percentage of structural zeros, the proposed zero-inflated multiscale models offer a consistent risk estimate at the fine and coarse levels. Because of their adaptability and ability to be extended beyond two scale levels, the models are a valuable resource for exploring the regional distribution of disease incidence at various levels. The models are applied to Georgian skin cancer data, and the findings demonstrate a favourable correlation between daily sunshine exposure and the incidence of skin cancer at the county and public health district levels. The proposed models are useful for studying the regional distribution of disease incidence at various levels, especially when it comes to rare or low-incidence cancers. Decisions about public health policy can be informed by the models, which can be used to identify regions with a high risk of illness occurrence.

The study emphasises how crucial it is to take overdispersion and excessive zeros into consideration when simulating relative risks of skin cancer at different scales. The study's results suggest that models with scaling effects and excessive zeros included are more appropriate for analysing health data from small areas with heavily skewed zeros. The study might not apply to data from the real world because it is restricted to simulated data. The intricacies of the data, such as non-linear correlations and interactions, might not be fully captured by the models. To get around the problem of utilising the data twice, the study also recommends expanding the models to incorporate numerous covariate effects and employing multilevel models with spatial interaction effects.

The dual-state-process assumption in the usage of preventive care services is examined by Zavras (2019), who concludes that the assumption is invalid because preventive services are not used frequently. The study examines data from different health care systems on preventative care-seeking behaviour, abstinence factors, and infrequency. It also looks at how often various preventive treatments should be retested and how often preventive care is used in developed countries. Using the corrected Vuong test (Shankar *et al.*, 1995; Solinas *et al.*, 2009), a ZINB model was compared to a normal NB model. It's possible that healthcare systems other than the Greek system do not adhere to the dual-state process premise. Gender and the percentage of monthly household income allocated to debt repayment and bills have an impact on the usage of preventive health care in Greece. Two important factors influencing the behaviour of people seeking preventive care are abstinence and infrequency. Even in free or heavily discounted services, few wealthy nations adopt preventive care. Because preventive services are rarely used, most zeros are sample zeros. Just 1.91% of those surveyed said they had never had a routine check-up. An inferential technique known as LUMINATE (longitudinal microbiome inference and zero detection) was developed by Joseph *et al.* (2020) and used on the C. diff dataset, which comprises both sequencing counts using 16S sequencing. They discovered that LUMINATE performs better in terms of efficiency and accuracy than the other models: Temporal Gaussian Process Model for Compositional Data Analysis (TGP-CODA; Äijö *et al.,* 2018); and multinomial logistic-normal dynamic linear models (MALLARDs; Silverman, 2019). It can also reliably differentiate between technical zeros, which occur when a taxon is below the detection threshold, and biological zeros, which occur when a taxon is absent from the community.

Using datasets from the American Gut Project (AGP), Deek and Li (2021) develop a zero-inflated Latent Dirichlet Allocation model (zinLDA) within a Bayesian framework that can accurately identify representative taxa and subcommunity structures. They claim that a zinLDA for sparse count data observed in microbiome studies provides better fits to the data and can separate structural zeros from sampling zeros with reasonable sensitivity and specificity. Determining the number of clusters or subcommunities is still one of the study's difficult tasks, though.

Model fit criteria alone shouldn't be the sole reason to choose ZI mixture models over the conventional Poisson model. A ZI mixed spatially varying coefficient (SVC) modelling framework was applied to the precipitation and cholera datasets in Ghana by Osei *et al.* (2022). A noteworthy finding of the study was that while the Poisson SVC model performed pretty well in terms of fit and could capture over-dispersion, it was unable to distinguish between zeros in susceptible and non-susceptible populations. They argue that there is no need to compare the model fit of a ZI model to the conventional Poisson model since the goal of the ZI model is to differentiate between the various sources of zero incidences.

4. CONCLUSION

Conclusively, for accurate statistical analyses and decision-making methods in a variety of fields, including healthcare, environmental sciences, economics, and social sciences, actual (structural) zero counts and imputed zeros must be accurately represented in individual records. The interconnectedness of these numbers makes it difficult to discern between real zero counts and imputed zeros, though. In general, ZI models for public health research studies can be thought of as permitting zeroes to have originated from both at-risk (vulnerable to the effects) and non-at-risk groups. Hurdle models, on the other hand, could be conceived as consisting of zeros from only the population that is at vulnerability (at-risk).

REFERENCES

Äijö, T., Müller, C.L., & Bonneau, R. (2018). Temporal Probabilistic Modelling of Bacterial Compositions Derived From 16s Rrna Sequencing. *Bioinformatics*, *34*(3), 372-380. https://doi.org/10.1093/bioinformatics/btx549

Aregay, M., Lawson, A.B., Faes, C., Kirby, R.S., Carroll, R., & Watjou, K. (2018). Zero-Inflated Multiscale Models for Aggregated Small Area Health Data. *Environmetrics*, *29*(1), e2477. https://doi.org/10.1002/env.2477

Besag, J., York, J., & Mollié, A. (1991). Bayesian Image Restoration, with Two Applications in Spatial Statistics. *Annals of the Institute of Statistical Mathematics*, *43(1)*, 1-20. https://doi.org/10.1007/bf00116466

Chase, J., Fouquier, J., Zare, M., Sonderegger, D.L., Knight, R., Kelley, S.T., & Caporaso, J.G. (2016). Geography and Location are the Primary Drivers of Office Microbiome Composition. *MSystems*, *1*(2), e00022-16. https://doi.org/10.1128/mSystems.00022-16

Deek, R.A., & Li, H. (2021). A Zero-Inflated Latent Dirichlet Allocation Model for Microbiome Studies. *Frontiers in Genetics*, *11*, 602594, ISSN 1664-8021, https://doi.org/10.3389/fgene.2020.602594

Duarte, A.S.R., Stockmarr, A., & Nauta, M.J. (2015). Fitting a Distribution to Microbial Counts: Making Sense of Zeroes. *International Journal of Food Microbiology*, *196*, 40-50, ISSN 0168-1605, https://doi.org/10.1016/j.ijfoodmicro.2014.11.023

Greene, W., Harris, M.N., Srivastava, P., & Zhao, X. (2017). Misreporting and Econometric Modelling of Zeros in Survey Data on Social Bads: An Application to Cannabis Consumption. *Health Economics*, *27*(2), 372-389. https://doi.org/10.1002/hec.3553

Gupta, R., Szczesniak, R.D., & Macaluso, M. (2015). Modeling Repeated Count Measures with Excess Zeros in an Epidemiological Study. *Annals of Epidemiology*, *25*(8),583-589. https://doi.org/10.1016/j.annepidem.2015.03.011

He, H., Tang, W., Wang, W., & Crits-Christoph, P. (2014a). Structural Zeroes and Zero- Inflated Models. *Shanghai Archives of Psychiatry*, *26*(4), 236-242. https://doi.org/10.3969/j.issn.1002-0829.2014.04.008

He, H., Wang, W., Crits-Christoph, P., Gallop, R., Tang, W., Chen, D.G.D., & Tu, X.M. (2014b). On the Implication of Structural Zeros as Independent Variables in Regression Analysis: Applications to Alcohol Research. *Journal of Data Science: JDS*, *12*(3), 439-460. https://doi.org/10.6339/JDS.201407_12(3).0004

Huang, L., Zheng, D., Zalkikar, J., & Tiwari, R. (2017). Zero-Inflated Poisson Model-Based Likelihood Ratio Test for Drug Safety Signal Detection. *Statistical Methods in Medical Research*, *26*(1), 471-488. https://doi.org/10.1177/0962280214549590

Joseph, T.A., Pasarkar, A.P., & Pe'er, I. (2020). Efficient and Accurate Inference of Mixed Microbial Population Trajectories from Longitudinal Count Data. *Cell Systems*, *10*(6), 463-469.e6, ISSN 2405-4712. https://doi.org/10.1016/j.cels.2020.05.006

Liberati, A., Altman, D.G., Tetzlaff, J., Mulrow, C., Gøtzsche, P.C., Loannidis, J.P., Clarke, M., Devereaux, P.J., Kleijnen, J., & Moher, D. (2009). The PRISMA Statement for Reporting Systematic Reviews and Meta-Analyses of Studies that Evaluate Health Care Interventions: Explanation and Elaboration. *Annals of Internal Medicine*, 151(4), W-65- W-94. https://doi.org/10.7326/0003-4819-151-4-200908180-00136

Mullahy, J. (1986). Specification and Testing of Some Modified Count Data Models. *Journal of Econometrics*, *33*(3), 341-365.

Osei, F.B., Stein, A., & Andreo, V. (2022). A Zero-Inflated Mixture Spatially Varying Coefficient Modeling of Cholera Incidences. *Spatial Statistics*, 48, 100635, ISSN 2211-6753. https://doi.org/10.1016/j.spasta.2022.100635

Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016). Rayyan—a web and Mobile App for Systematic Reviews. *Systematic Reviews*, 5, 1-10. https://doi.org/10.1186/s13643-016-0384-4

Rose, C.E., Martin, S.W., Wannemuehler, K.A., & Plikaytis, B.D. (2006). On the Use of Zero-Inflated and Hurdle Models for Modeling Vaccine Adverse Event Count Data. *Journal of Biopharmaceutical Statistics*, *16*(4), 463-481. https://doi.org/10.1080/10543400600719384

Schmidt, A.M., & Pereira, J.B.M. (2011). Modelling Time Series of Counts in Epidemiology. *International Statistical Review*, *79*(1), 48-69. https://doi.org/10.1111/j.1751-5823.2010.00123.x

Shankar, V., Mannering, F., & Barfield, W. (1995). Effect of Roadway Geometrics and Environmental Factors on Rural Freeway Accident Frequencies. *Accident Analysis & Prevention*, *27*(3), 371-389. https://doi.org/10.1016/0001-4575(94)00078-Z

Shestopaloff, K., Escobar, M.D., & Xu, W. (2018). Analyzing Differences Between Microbiome Communities Using Mixture Distributions. *Statistics in Medicine*, *37*(27), 4036-4053. https://doi.org/10.1002/sim.7896

Silverman, J.D., Roche, K., Holmes, Z.C., David, L.A., & Mukherjee, S. (2019). Bayesian Multinomial Logistic Normal Models Through Marginally Latent Matrix-T Processes. *Journal of Machine Learning Research*, *23*(7), 1-42. https://doi.org/doi:10.48550/arXiv.1903.11695

Solinas, G., Campus, G., Maida, C., Sotgiu, G., Cagetti, M.G., Lesaffre, E., & Castiglia, P. (2009). What Statistical Method Should be Used to Evaluate Risk Factors Associated with DMFs Index? Evidence from the National Pathfinder Survey of 4- year- old Italian Children. *Community Dentistry and Oral Epidemiology*, *37*(6), 539-546. https://doi.org/10.1111/j.1600-0528.2009.00500.x

Tang, W., He, H., Wang, W.J., & Chen, D.G. (2017). Untangle the Structural and Random Zeros in Statistical Modelings. *Journal of Applied Statistics*, *45*(9), 1714-1733. https://doi.org/10.1080/02664763.2017.1391180

Turpin, W., Espin-Garcia, O., Xu, W., Silverberg, M.S., Kevans, D., Smith, M.I., ... & Croitoru, K. (2016). Association of Host Genome with Intestinal Microbial Composition in a Large Healthy Cohort. *Nature Genetics*, *48*(11), 1413-1417. https://doi.org/10.1038/ng.3693

Ye, P., Tang, W., He, J., & He, H. (2018). A GEE-Type Approach to Untangle Structural and Random Zeros in Predictors. *Statistical Methods in Medical Research*, *28*(12), 3683-3696. https://doi.org/doi:10.1177/0962280218812228

Zavras, D. (2019). Testing the Dual-State-Process Assumption in the Preventive Care Services Use. *Journal of Health Social Science*, *5*(1), 127-140. https://doi.org/10.19204/2019/tstn4