BMC Gastroenterology

RESEARCH

Open Access

Application of finite mixture models to explore subpopulations in Crohn's disease patients

Mehari Gebre Teklezgi[1*], Gebru Gebremeskel Gebrerufael[2] and Hirut Teame Gebru[1]

Abstract

Background Inflammatory bowel disease (IBD) commonly refers to ulcerative colitis (UC) and Crohn disease (CD), which are chronic inflammatory diseases of the gastrointestinal (GI) tract of unknown etiology. This study has been conducted to examine whether there are different components in the data, and if these components related to the treatment and the Inflammatory bowel disease (IBD) score at baseline.

Methodology This is a clinical study which consisted of 291 subjects, who divided over four treatment arms and were measured during a seven-week period. The number of weeks in the period Week 0ne through Week seven was considered as the outcome of interest, as well treatment and IBD score at baseline were considered as predictors. Different statistical methods such as explanatory data analysis and finite mixture model were employed to explore the outcome of interest.

Results From the finite mixture model, two components were obtained. Most of the patients, 196(67.4%), were classified in the first component (P1). The deviance for single component of the mixture model corrected for the covariates was 1049.3 and that of the two components was 948.8. The effect of ibdsc0 was significant in both subpopulations with *p*-value = 0.0001 for subpopulation1, and *p*-value = 0.0422 for subpopulation2, and Exp (0.01) = 1.01 and exp(0.087) = 1.09 are the amounts by which the mean count (μ) is multiplied per unit change in the ibdsc0 for subpopulations 1 and 2, respectively.

Conclusions The two components are not related to the treatment, and as a result, the treatment does not completely explain the presence of potential clusters in the outcome. Ibdsc0 partially explains the presence of potential clusters in the outcome.

Keywords Inflammatory bowel disease, Poisson model, Mixture model

*Correspondence:
Mehari Gebre Teklezgi
meharistat@gmail.com
[1]Department of Public Health, College of Medicine and Health Sciences, Adigrat University, P.O.Box. 50, Adigrat, Tigray, Ethiopia
[2]Department of Statistics, Adigrat University, Adigrat, Ethiopia

Introduction

Inflammatory bowel disease (IBD) commonly refers to ulcerative colitis (UC) and Crohn disease (CD), which are chronic inflammatory diseases of the gastrointestinal (GI) tract of unknown etiology. These chronic, life-long situations can be treated but not cured, and can significantly affect a patient's quality of life and may have a high financial burden [1]. Ulcerative colitis, inflammation always starts in the rectum, ranges proximally a certain distance, and then brusquely stops. A clear demarcation exists between involved and uninvolved mucosa. UC mainly encompasses the mucosa and the submucosa, with development of crypt swellings and mucosal ulceration. The mucosa naturally looks granular and friable [2]. Crohn disease, in contrast, involves of segmental connection by a general granulomatous inflammatory progression. The most vital pathologic feature is involvement of all layers of the bowel, not just the mucosa and the submucosa. Unlike UC, CD is discontinuous, with skip areas intermingled between one or more involved areas [3].

Crohn's disease is a long-lasting inflammatory illness of the gastrointestinal tract, with signs like chronic stomach pain, diarrhea, obstruction, and perianal lesions [4]. Globally, the estimated incidence of CD ranges from 0.58 to 20.2 cases per 100,000 person-years, whereas the incidence amount to 50–322 per 100,000 individuals [5]. Medical therapy used to treat CD comprises the categories of 5-aminosalicylates (5-ASA), antibiotics, corticosteroids, immunomodulatory, and biologics. Biologics are by far the strongest treatment for CD and are powerfully suggested for patients with moderate-to-severe CD who failed to respond to conventional therapy [6].

The prominence of finite mixture models in the statistical analysis of data is obvious in the continually growing rate at which researches on theoretical and practical aspects of mixture models look in the statistical and general scientific literature; this is due to the reason that finite mixtures of distributions are extensively used to deliver computationally appropriate representations for modeling complex distributions of data on random phenomena [7]. The primary main analyses regarding the use of mixture models was commenced about 125 years ago, and was fitted a mixture of two normal probability density functions through different means μ1 and μ2 and different variances σ12 and σ22 to some data [7]. Furthermore, the first recorded presence of the Finite mixture models in the recent statistical literature is in 19 century in a paper by [8] that was used it in the framework of modeling outliers. After few years [9], used a mixture of two univariate normal distributions to examine a dataset encompassing ratios of forehead to body lengths for 1,000 crabs [10].

Regarding the modeling of count data, the fitting of a single Poisson distribution regularly forces a lot of structure on the data leading to difficulties such as overdispersion; the usage of mixture model allows a compromise between the homogeneous Poisson model and nonparametric models which, although avoiding strong distributional assumptions, have other drawbacks including high-data dependency of model estimates [11]. Finite mixture models are broadly used to represent data produced by a heterogeneous population, where the observed data can be considered as arising from multiple latent subpopulations or components. Mostly, useful when data is not homogeneous and appears to be generated from multiple underlying groups [12]. They allow for modeling populations with distinct subgroups that have different characteristics, and representing complex distributions as a combination of simpler distributions. Besides, they provide flexibility in approximating arbitrary probability distributions. Instead of relying on a single distribution assumption, a mixture model combines multiple distributions to Model multimodality in data, and Capture skewness, kurtosis, and other deviations from standard distributions [13].

The objective of this study is, to investigate whether there are different components in the data and if these components related to the treatment and to the Inflammatory bowel disease (IBD) score at baseline. This is a longitudinal study which consisted of 291 subjects with 1813 total observations, who divided over four treatment arms (0 = placebo, 1 = 1000 mg, 2 = 2000 mg; 3 = 4000 mg) and were measured during a 7-week period. The outcome of interest is the number of weeks in the period Week 1 through Week 7 in which a value of IBD larger than 100 was observed, as well treatment and IBD score at baseline were considered as predictors. The numbers of subjects who have IBD score larger than 100 are 169, and are presented by the following flow chart.

From Fig. 1 observed that from the 169 number of subjects who have IBDS > 100, there are 37 subjects who took the TA1(0 = P) = Treatment Arm (0 = placebo); 40 subjects who took the TA2(1 = 1000 mg) = Treatment Arm(1 = 1000 mg); 49 subjects who took TA3(2 = 2000 mg) = Treatment Arm (2 = 2000 mg), and 43 number of subjects who took the treatment arm with 4000 mg(3 = 4000 mg).

Methodology

Exploration data analysis (EDA) was done to gain some insight into the distribution of the outcome of interest. For this reason, summary statistics and graphical description were used.

Rationale of using poisson and mixture models Count Data are typically observed in many applied fields such as in actuarial science when assessing risk and the pricing of insurance contracts [14], like in genetics to model the

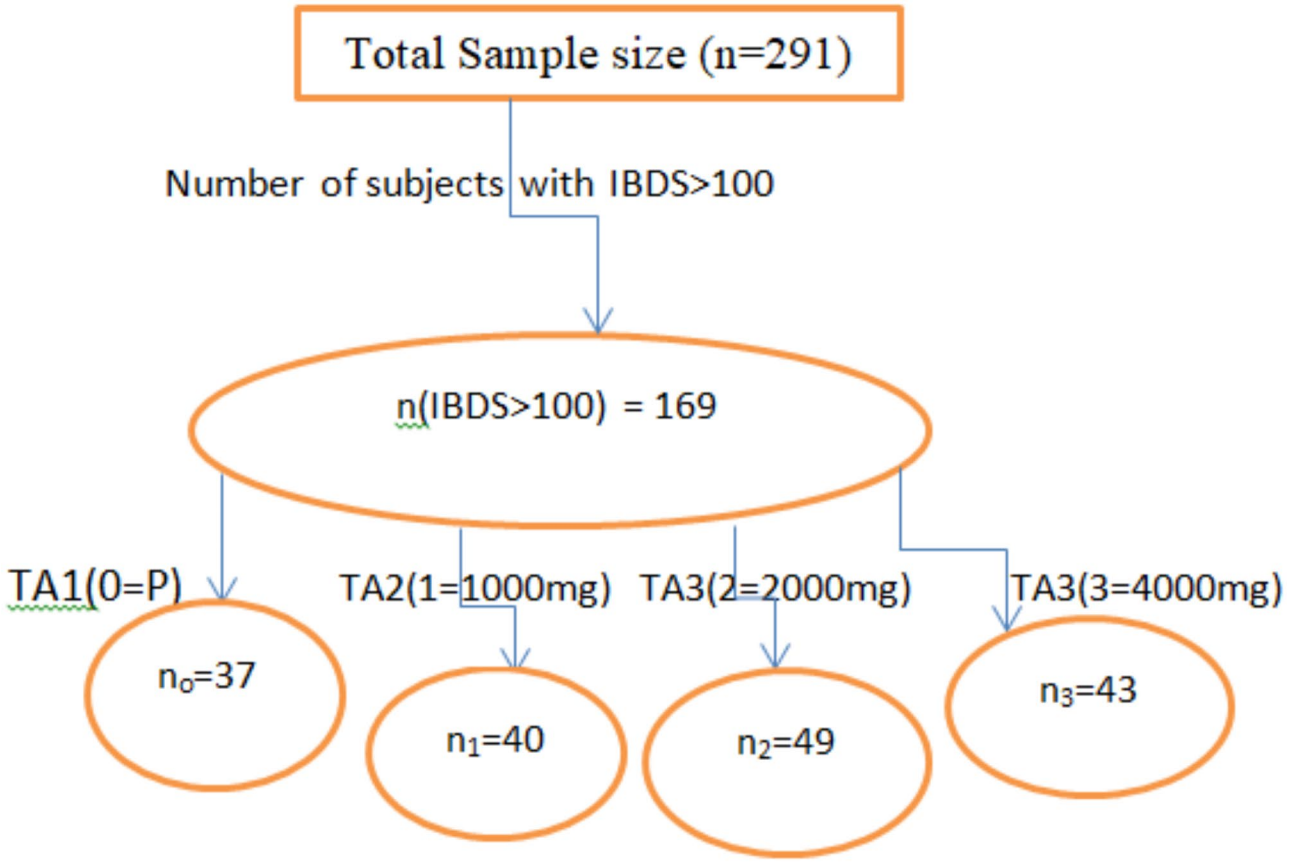

Fig. 1 Flow chart for the total subjects, the subjects with IBD score > 100 with separate subjects who took specific treatment

number of genes involved in phenotype variability [15]. In such case, the Poisson regression model is a widely used statistical tool for analyzing count data, particularly in medical research involving Crohn's disease patients [16, 17]. As a member of the generalized linear model family, Poisson regression provides flexibility in modeling non-normal data distributions, which is advantageous in various research contexts [18]. However, it's important to note that the Poisson model assumes the mean and variance of the count data are equal. In practice, count data often exhibit overdispersion (variance exceeds the mean), which typically occur with an excess of zeroes or extreme large values [19]. As a result, this model is not suitable for overdispersed data which can lead to underestimated standard errors and inflated Type I error rates [20].

Addressing overdispersion In real-world data often display overdispersion, where the variance exceeds the mean. In such case, Finite Poisson mixture models accommodate this by modeling the data as a combination of multiple Poisson distributions, each with its own mean, effectively capturing the extra variability [7]. Besides, the population consists of unobserved subgroups with distinct characteristics, and the Finite mixture models identify and model these latent subpopulations, providing a more accurate representation of the underlying data structure [3].

A finite mixture model is the probabilistic model that represents the data as a weighted combination of two or more probability density functions (PDFs), and the model assumes the data is generated from a combination of g component distributions [21]. It was fitted since it can be applied to data where observations originated from heterogeneous groups, and is a very popular statistical modeling technique constitutes a flexible and easily extendable model class for counting unobserved heterogeneity and approximating general distribution functions [17]. The general form of the finite mixture model is given as follows.

$$\frac{Y_i}{P} \sim \text{Poisson}(\lambda i), \text{ where } P \sim \begin{vmatrix} \lambda 1 & \lambda 2 \ldots \lambda g \\ p1 & p2 \ldots pg \end{vmatrix}, 1, 2, 3, \ldots, g,$$

Y_i= the outcome of interest. i.e. the number of weeks for the i^{th} subject in the j^{th} $(1 \le j \le g)$ subpopulation. The population from which the response measurement comes consisted of g number of subpopulations with proportion of P_1, P_2, P_3...P_g, and mean values of λ_1, λ_2, λ_3... λ_g

respectively. After we have recognized the sensible number of components (subpopulations), and fitted a plausible mixture model, deciding of interest. Therefore, this is often done based on posterior probabilities density function.

The mixing proportions π_j are nonnegative and sum to one and the $\mathrm{fij}\,(\mathrm{yi})$ are the component densities, and we can refer to the $\mathrm{fi}\,(\mathrm{yi},\ \pi\,\mathrm{i})$ as densities since even if the vector is discrete, we can still view the $\mathrm{fij}\,(\mathrm{yi})$ as densities by the adoption of counting measure [3]. The posterior probability for observation i to belong to the j^{th} component is then given by the formula:

$$\pi\ \mathrm{ij}\ = (\ \pi\ \mathrm{ifij}\,(\mathrm{yi}))\,/\sum_{\mathrm{i}=1}^{\mathrm{g}}\pi\ (\mathrm{fij}(\ \mathrm{yi}\,))$$

Where, π_j and π_{ij} express how likely the i^{th} subject is to belong to component j without taking into account and with taking into account the observed response value y_i for that observation respectively. Classify observation i into component j if and only if.

$$\pi_{\,ij}\ = \mathrm{Maxk}\,(\pi\,\mathrm{ik})\,.$$

i.e., classify into the component to which observation i is most likely to belong [11].

The empirical proportions (P_1, P_2, P_3.P_g) are computed directly from the data, reflecting the observed counts within each group. However, the estimated mixture proportions (π_j) are derived from the model and represent the probabilities of an observation belonging to a particular latent component [22]. If these values are close, it indicates that the model's assumptions about the mixture structure and latent subpopulations are consistent with the actual data distribution. That is the observed data aligns well with the model's estimated parameters. I.e., the mixture model has accurately identified the relative contributions of each subpopulation (or component) to the overall data distribution [23].

Model extension

Separate Poisson Models, adjusting for the given predictors (treatment and ibd) were fitted. For two subpopulations, the following Poisson models were proposed.

$$\log(\mu\,1) = \beta\,10 + \beta\,11\mathrm{dose1i} + \beta\,12\mathrm{dose2i} + \beta\,13\mathrm{dose3i} + \beta\,14\mathrm{ibdsc0i}$$

$$\log(\mu\,2) = \beta\,20 + \beta\,21\mathrm{dose1i} + \beta\,22\mathrm{dose2i} + \beta\,23\mathrm{dose3i} + \beta\,24\mathrm{ibdsc0i}$$

Where Yi ~ Poisson (μk), μk is the expected count of y_i for k = 1, 2. Placebo = 0 is the reference, dose1 = 1, dose2 = 2, dose3 = 3. This model assumes that at each covariate, the population consists of two sub-populations. The proportion of each sub-population depends on the given predictors, and the relation between predictors and the response is different for both sub- populations as well.

The modeling procedure can be summarized by the Fig. 2 flow diagram.

Nonparametric Maximum Likelihood Estimation (NPMLE) is an extension of Maximum Likelihood Estimation (MLE) that estimates distributions or parameters without assuming a strict parametric form [24]. In the context of finite mixture models, NPMLE can estimate the mixing proportions (πg) and component distributions fg(x|θg)) without a parametric assumption, and it provides flexible estimation and works well for complex data structures [24].

Software All analyses were performed using R version 3.2 and SAS version 9.4. In addition, all tests were performed at a 5% level of significance.

Results

Frequencies of the number of weeks in which the ibdsc value greater than 100 were summarized.

As it can be seen from Table 1, the frequency decreases across the weeks, but increased in the 7th week. Moreover, we can observe that the mean is 2.09 and the variance is 4.973, from this we can observe that there may be over-dispersion.

Below in Table 2 is given the number of observations with respect to the distribution of weeks across treatment groups. It was revealed that each treatment has no the same number of observations through all the seven weeks. In 0 doses, there are 37 observations in the first week, but there are decreasing numbers of observations throughout the seven weeks; in dose 1, starting 40 numbers of observations in the first week, and then different number of observations in each of the weeks were observed, and the same is true in all the treatment groups. It can be also revealed that there are highest (170) total number of observations in all treatments in the first week as compared to the other weeks, whereas there are smallest (94) number of observations in the last week. This indicated that the total number of observations was increasing missing their follow-ups as weeks increase. Moreover, the total number of observations in treatment group1 is the highest (203), whereas in the treatment group3 are the smallest (176). This difference can happen due to the reason that the number of subjects having IBD score > 100 are higher in treatment group1 and smallest in treatment group3.

Histogram of the number of weeks for which the ibdsc is greater than 100 was given (Fig. 3). And it was observed that the distribution of the number of weeks seems to have multiple modal values which might not be easily described by standard distributions. In addition, the

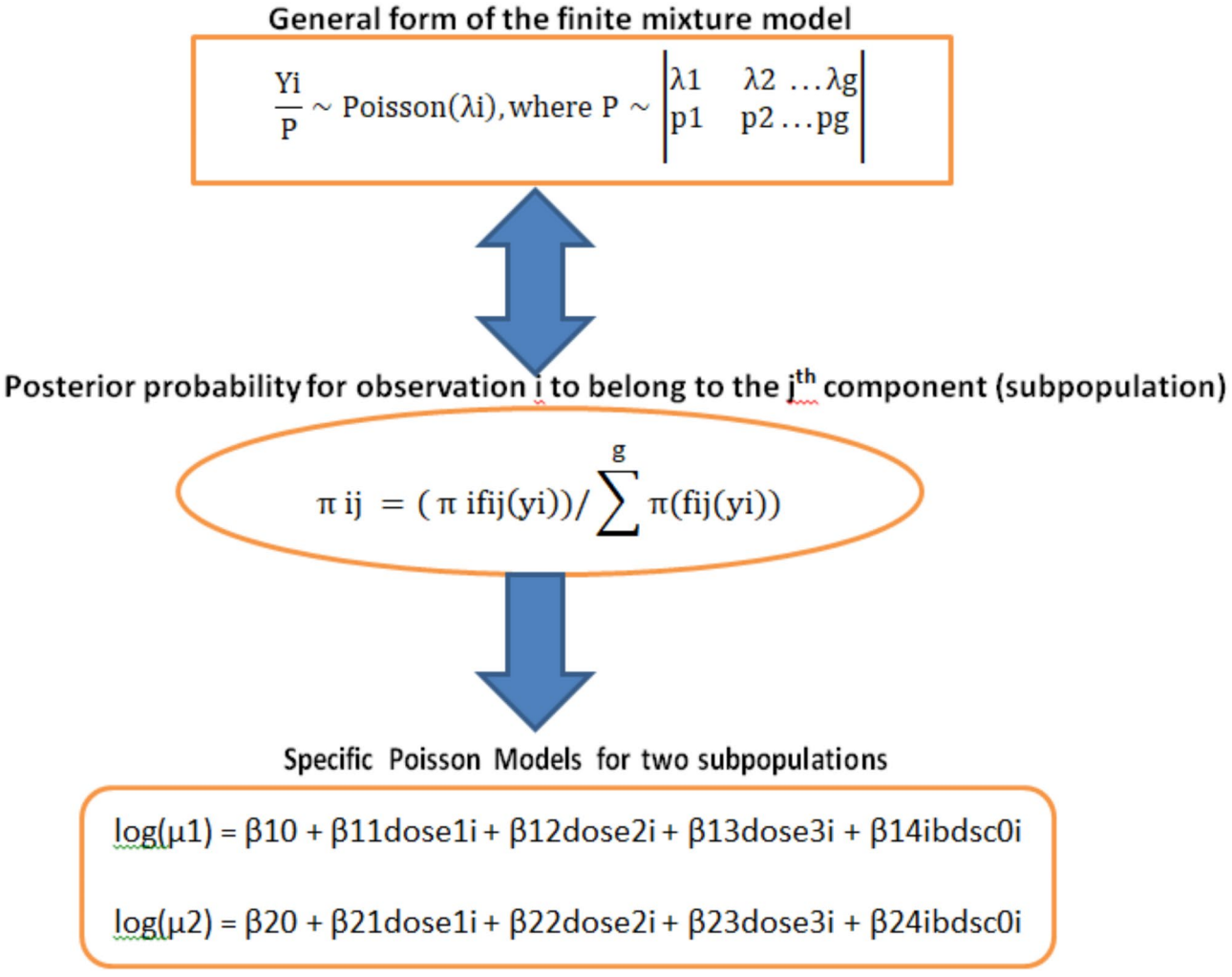

Fig. 2 Flow diagram of the modeling steps

Table 1 Frequency and summary measures of weeks

Weeks(Count)	0	1	2	3	4	5	6	7
Frequency	85	75	36	28	19	13	8	27
N	Mean	Std Dev	Min	Max				
291	2.09	2.23	0	7				

Table 2 Number of observations in each week across treatment groups for IBD score > 100

Dose	Week							Total (%)
	1	2	3	4	5	6	7	
0	37	28	26	21	25	28	18	183 (23.88)
1	40	25	24	27	28	28	31	203 (28.17)
2	50	24	22	25	28	26	25	200 (26.05)
3	43	28	23	22	18	22	20	176 (21.9)
Total (%)	170(23.3)	105(13.8)	95(12.2)	95(12.2)	99(13.0)	104(13.7)	94(12.3)	762 (100)

variance is higher than the mean (Table 1) which might indicate the presence of over-dispersion. Therefore, one way to take into account these problems is modeling underling heterogeneity using a finite mixture model.

Since the number of weeks is count data, it was reasonable to assume Poisson distribution. Nevertheless, as the assumption is not met (variance is higher than the mean, Table 1), there might be over dispersion, and then this model could not be appropriate. As it was depicted

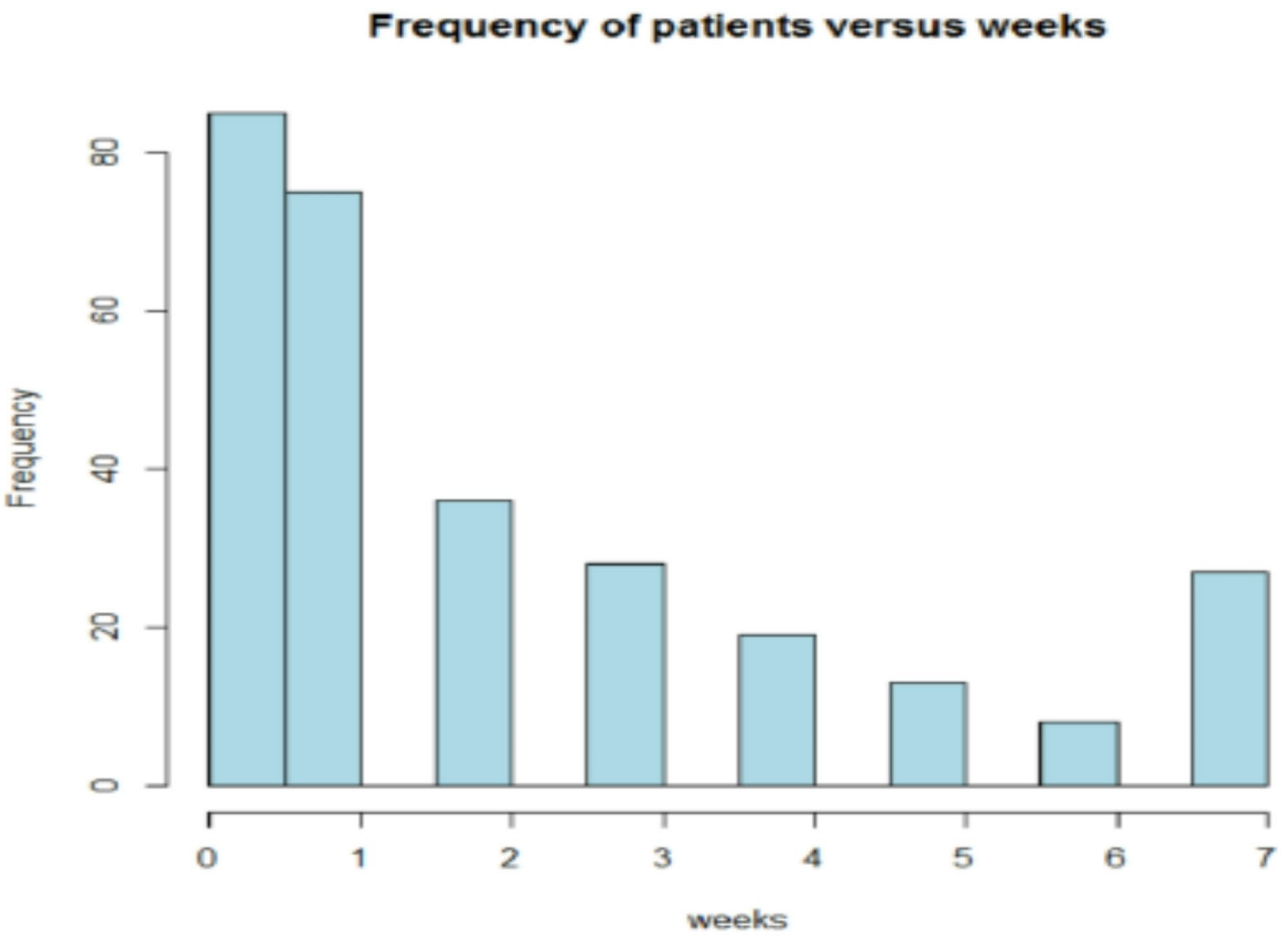

Fig. 3 Histogram of number of weeks in which ibdsc is greater than 100

Table 3 AIC and BIC values of three fitted models

Model	AIC	BIC
Model With 1 component	506.857	559.945
Model With 2 components	468.338	503.776
Model With 3 component3	524.642	606.114

from the histogram as well as the summary statistics, it was observed heterogeneity as well as seems there is over dispersion among the weeks. For this reason, it is sensible to consider a model which accounts the multi-modality and over-dispersion problems, and then finite Poisson mixture model was fitted.

Finite mixture model fitting

After observing the variance becomes larger than the mean, and that of Fig. 3 which looks having multimodality, we went to the model checking. We fitted three models, model one with a single component (assuming unimodal), model two is a mixture model with two components, and the third model is a mixture model with three components. AIC and BIC of all the three models were determined as given Table 3.

As it was revealed from Table 3, the AIC as well as the BIC values of the model with two components are the smallest, and which is the indication that the model with two components has the best fit. Therefore, all analysis was done using this model.

Here to estimate the number of components, nonparametric maximum likelihood estimation (NPMLE) was performed. The final fitted mixture model was with two components, and log likelihood value of -556.2931, and given as follows:

$$\frac{\mathrm{Yi}}{\mathrm{P}} \sim \text{Poisson}\,(\lambda\,\mathrm{i})\,,\ \text{where}\,\mathrm{P} \sim \begin{vmatrix} 0.789 & 4.430 \\ 0.637 & 0.363 \end{vmatrix},\ \text{for}\,\mathrm{i} = 1,\ 2.$$

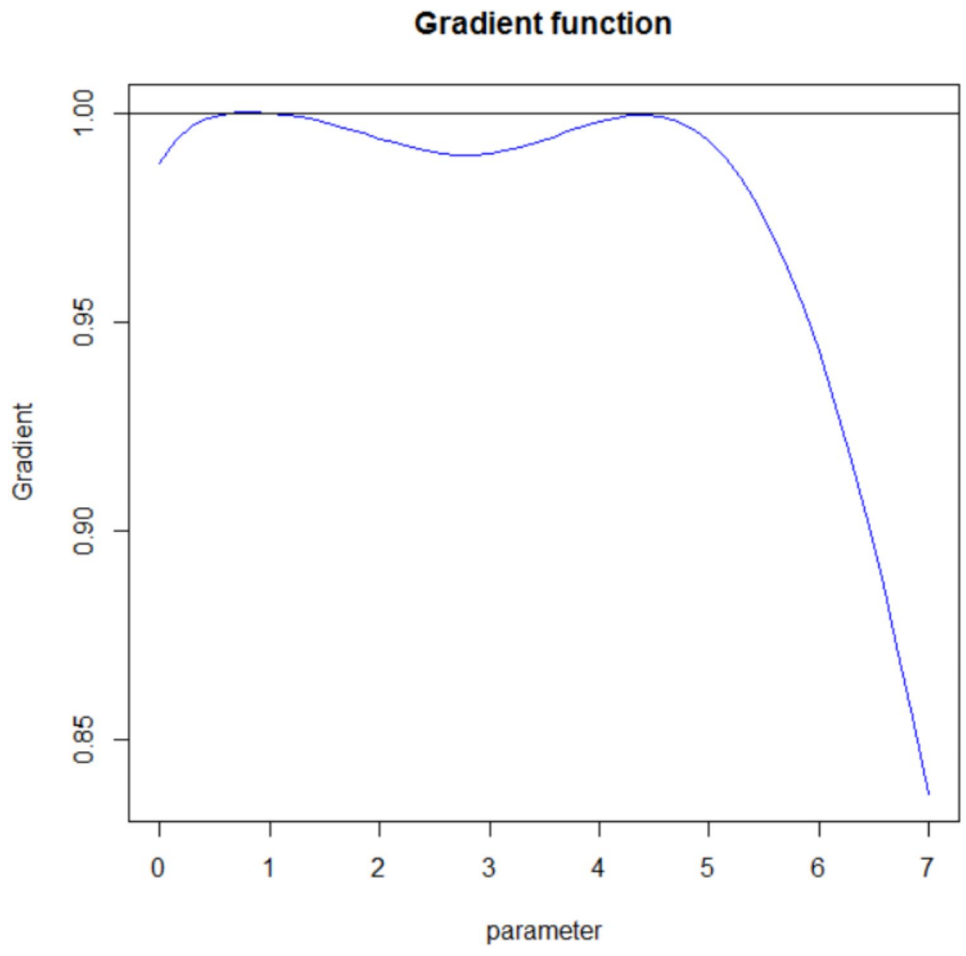

Fig. 4 Plot of Gradient function

Table 4 Final number of components and classification of observations; estimates of covariates

Component	Count	Freq	$\widehat{\lambda}$	π^	Pro	Effect	Estimate	S.Error	*P*-value
						Intercept	-4.335	0.492	< 0.0001
1	0 1 2	196	0.789	0.637	0.674	1	-0.157	0.237	0.507
							0.008	0.204	0.970
						3	-0.250	0.203	0.218
						ibdsc	00.042	0.004	0.0001
						Intercept	0.537	0.464	0.248
2	3 4 5 6 7	95	4.430	0.363	0.326	1	0.138	0.203	0.496
							0.008	0.008	0.973
						3	0.160	0.227	0.481
						ibdsc0	0.008	0.004	0.0422
Total		291	1 1						

Placebo = 0 is reference. 1 = 1000 mg, 2 = 2000 mg, count = weeks, freq = frequency, pro = proportion

The plot of gradient function versus the parameter value (lambda) is presented (Fig. 4). From this figure, it can be clearly observed that the gradient function was less than or equal to one, which pointed that the estimated value of the parameters of the distribution function were the non-parametric maximum likelihood estimates(NPMLE), and also these estimates are unique as the gradient function is identically one. Classification was done based on the fitted mixture model, and as we can see from Table 4, the proportions are very close to the estimated components (π^). Most of the patients, 196(67.4%), were classified in the first component.

Model extension

The deviance for single component of the mixture model corrected for the covariates was 1049.3 and that of the two components was 948.8. The change in the deviance of these mixture models is large, and which revealed that there might be evidence for the presence of mixture after correcting for the patient characteristics (covariates) in the model.

The result was given in Table 4, and adjusting for the treatment and ibd score at baseline, the fitted models for population 1(component 1) and population 2 (component 2), respectively were given bellow:

$$\begin{aligned}&\log(\mu\,1)\\&=-4.335\;+\;0.10\text{dose1i}\\&+0.0086\text{dose2}\;-0.250\text{dose3}\\&+0.042\text{ibdsc0}\end{aligned}$$

$$\begin{aligned}&\log(\mu\,2)\\&=\;0.537\;+\;0.138\text{dose1i}\\&+0.008\text{dose2}\;+\;0.160\text{dose3}\\&+0.0087\text{ibdsc0}\end{aligned}$$

In two of the components, the effect of treatment was insignificant (*p*-values for all the treatment groups are larger than 0.05 significance level, Table 4). Therefore, the treatment does not completely explain the presence of potential clusters in the outcome. In the other hand, the effect of ibdsc0 was significant in both subpopulations, *P*-values < 0.05 significance level, Table 4). Exp(0.042) = 1.043 and exp(0.008) = 1.008 are the amounts by which the mean count (μ) is multiplied per unit change in the ibdsc0 for subpopulations 1 and 2, respectively. This showed that the patient characteristics (ibdsc0) completely explain the presence of potential clusters in the outcome.

Components, their relationship, and covariates

As the variance is higher than the mean look at Table 1(4.973 > 2.09), and this indicates the presence of over-dispersion. Therefore, using the mixture model to account a data with such potential heterogeneity problem is very important. As a result, this is the justification that using the poison mixture model is needed.

The Poisson fitted models for the two components showed that though the effect of ibdsc0 is significant on both components, its effect is higher in component 1 as its *p*-value = 0.0001 which is much more higher than the *p*-value = 0.0422 for the fitted model of component 2. Besides, the multiplying factor of the mean count (μ) for component 1 is higher(1.043) as compared with the multiplying factor of the mean count (μ) for component 2(1.008) per unit change for Inflammatory bowel disease score at baseline (ibdsc0).

We can also observe that the total subjects in subpopulation 1 are higher ($n_1 = 196$) as compared to that of the subpopulation 2 which are only ($n_2 = 96$). However, there are only the weeks 1 and 2 included in this component, but the rest (weeks 3, 4, 5, 6 and 7) were included in component 2. As of the number of subjects, the proportion is also higher for the component 1 which is 0.674 as compared to that of the component 2 which is 1-0.674 = 0.326.

It should also be mentioned that the empirical proportions (Pro) for both components are close to the estimated mixture proportions (π^). This indicated that the model's assumptions about the mixture structure and latent subpopulations are consistent with the actual data distribution. That is the observed data aligns well with the model's estimated parameters.

Clinical implication of the predictors

Significance of the inflammatory bowel disease at baseline (ibdsc0) suggested that the initial severity of the ibd symptoms at baseline intensely influences disease evolution or outcomes over the 7-week period. Higher baseline IBD scores may predict more weeks with severe disease (IBD > 100), highlighting the importance of early and accurate assessment of baseline disease severity in clinical settings. Moreover, patients with higher baseline IBD scores may require closer monitoring and possibly more aggressive or tailored therapeutic interventions.

The finding that the different treatment doses did not have a statistically significant effect on the number of weeks with IBD > 100 raises concerns about the efficacy of these treatments in the context of this study. This can be due to the reason that the treatment may not be effective in altering the course of the disease within the 7-week period. The study design, duration, or sample size may not have been sufficient to detect true effects.

Discussion

Using the Crohn Inflammatory bowel disease patients' dataset, this study empirically explored, and determined whether there were different components in the data and if these components were related to the treatment (dose) and to the Inflammatory bowel disease (IBD) scores at baseline (ibdsc0) in Adigrat University, Adigrat, Tigray, Ethiopia. The Crohn Inflammatory bowel disease was examined over seven weeks along with the mentioned related risk factors, using Poisson mixture model analysis. The empirical proportions (Pro) for both components are close to the estimated mixture proportions (π^). This is the indication that the observed data aligns well with the model's estimated parameters, and this result is coincided with the research studied by McLachlan, G., and Karlis, D., & Xekalaki, E. [11 & 22].

The predictor variable dose has insignificant effect on the number of weeks, and this is due to the reason that

the effect of each treatment group (0, 1, 2, and 3) on the number of weeks in which the patients stay can't that much variability. Moreover, the length of time (seven weeks) may be short period of time to show the effect of the treatment groups, and there may be unmeasured factors. Of course, this result opposed for the result found from [4]. However, the Inflammatory bowel disease (IBD) scores at baseline (ibdsc0) has significant effect on the number of weeks, and this result was coincided with [6]. The significant effect of ibdsc0 on the number of weeks suggested that the severity of the disease at the start of treatment (follow-up) plays an important role in determining how long patients remain on therapy.

Conclusion

This study was consisted of 291 subjects and was divided over four treatment arms (doses). The measurement was taken during a seven week periods. The number of weeks in this period was considered as response variable; and treatment groups (doses) and inflammatory bowel disease score at baseline (ibdsc0) were taken as predictor variables. Finite Poisson Mixture Model, which takes into account heterogeneity of the outcome were fitted. And two different subpopulations (components) were identified. Patients were classified into the subpopulations for which they belong to base on the posterior probability.

Moreover, the relationship between the two components and covariates were investigated by fitting Poisson mixture model adjusting to the covariates. Treatment effect was insignificant in both components, and this might be due to the short period of follow-up times. However, inflammatory bowel disease score at baseline (ibdsc0) was significant in both components, and this showed that the severity of the disease at treatment starting time plays a vital role in exploring for how long patients can remain on treatments. Therefore, the treatment does not completely explain the presence of potential clusters in the outcome of interest. Whereas, ibdsc0 completely explains the presence of potential clusters in the outcome.

Since baseline IBD severity predicts outcomes, Baseline IBD scores could be used to stratify patients into different risk categories, allowing clinicians to prioritize resources and interventions for patients at higher risk of prolonged disease activity.

The lack of significance of the treatment effects shows the need of for further investigation. This is needed to reevaluate the efficacy of the treatment over a longer period; to explore whether specific subgroups may be benefited from the treatment, and to investigate the potential reasons for the treatment inefficacy. Therefore, Clinicians should consider these insights to optimize care for patients with Crohn's Disease, particularly by focusing on early identification and stratification of high-risk patients. It can also be recommended that the treatment should be researched by including other factors.

Acknowledgements
The authors would like to thanks for all study participants, and our friends for their unreserved efforts and willingness to take part in this study.

Author contributions
Mehari Gebre Teklezgi: Analyzed the data, interpreted, discussed, and write the draft of the Manuscript. Gebru Gebremeskel Gebrerufael: collect and organize the data, and conceptualize the theme of the study. Hirut Teame Gebru: interpreted and drafted the manuscript. All authors have read and approved the final.

Funding
This study has no fund.

Data availability
The dataset will be shared up on request and will be obtained through contacting emailing to the corresponding author (Mehari Gebre Teklezgi) using "meharistat@gmail.com".

Declarations

Ethics approval and consent to participate
Adigrat University, College of Natural and Computational Science ethical review office granted ethical approval for this study. However, since the need for the ethical approval was waived by the ethical review office, we do not have a reference number. Moreover, since the study relies on secondary data, no individual informed consent is needed. Therefore, the need for consent to participate was waived by this ethics committee. The University of Adigrat's research regulations were followed in every way during conducting the study.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 19 April 2024 / Accepted: 11 February 2025
Published online: 25 February 2025

References

1. Kessler D, McDowell A. Introducing the FMM procedure for finite mixture models. InProceedings of the SAS Global Forum. 2012.
2. Villanacci V, Reggiani-Bonetti L, Salviato T, Leoncini G, Cadei M, Albarello L, Caputo A, Aquilano MC, Battista S, Parente P. Histopathology of IBD Colitis. A practical approach from the pathologists of the Italian Group for the study of the gastrointestinal tract (GIPAD). Pathologica. 2021;113(1):39.
3. Zhang H, Huang Y. Finite mixture models and their applications: a review. Austin Biometrics Biostatistics. 2015;2(1):1–6.
4. Hazlewood GS, Rezaie A, Borman M, Panaccione R, Ghosh S, Seow CH, Kuenzig E, Tomlinson G, Siegel CA, Melmed GY, Kaplan GG. Comparative effectiveness of immunosuppressants and biologics for inducing and maintaining remission in Crohn's disease: a network meta-analysis. Gastroenterology. 2015;148(2):344–54.
5. Torres J, Mehandru S, Colombel JF, Peyrin-Biroulet L. Crohn's disease. Lancet. 2017;389(10080):1741–55.
6. Yu B, Zhao L, Jin S, He H, Zhang J, Wang X. Model-based meta-analysis on the efficacy of biologics and small targeted molecules for Crohn's disease. Front Immunol. 2022;13:828219.
7. McLachlan GJ, Lee SX, Rathnayake SI. Finite mixture models. Annual Rev Stat its Application. 2019;6(1):355–78.
8. Newcomb S. A generalized theory of the combination of observations so as to obtain the best result. Am J Math 1886;1:343–66.
9. Pearson P K. Contributions to the mathematical theory of evolution. Philos Trans R Soc Lond A. 1894;185:71–110.
10. Melnykov V, Maitra R. Finite mixture models and model-based clustering.

11. McLachlan G. Finite mixture models. A wiley-interscience publication. 2000.
12. Redner RA, Walker HF. Mixture densities, maximum likelihood and the EM algorithm. SIAM Rev. 1984;26(2):195–239.
13. McLachlan GJ, Basford KE. Mixture models: inference and applications to clustering. New York: M. Dekker; 1988.
14. Ahlmann-Eltze C, Huber W. glmGamPoi: fitting Gamma-Poisson generalized linear models on single cell count data. Bioinformatics. 2020;36(24):5701–2.
15. Vu M, Ghosh S, Umashankar K, Weber L, Landis C, Candela N, Chastek B. Comparison of surgery rates in biologic-naïve patients with Crohn's disease treated with vedolizumab or ustekinumab: findings from SOJOURN. BMC Gastroenterol. 2023;23(1):87.
16. Neuman-Boone EE. An examination of factors that contribute to binge eating among bariatric patients. Duquesne University; 2015.
17. Lord D, Park BJ, Levine N. Poisson regression modeling. CrimeStat IV: A spatial statistics program for the analysis of crime incident locations. 2013.
18. Coxe S, West SG, Aiken LS. The analysis of count data: a gentle introduction to Poisson regression and its alternatives. J Pers Assess. 2009;91(2):121–36.
19. Ryan WH, Evers ER, Moore DA. Poisson regressions: a little fishy. Collabra: Psychol. 2021;7(1):27242.
20. Hougaard P, Lee ML, Whitmore GA. Analysis of overdispersed count data by mixtures of Poisson variables and Poisson processes. Biometrics. 1997;1:1225–38.
21. Burgess-Hull AJ. Finite mixture models with student t distributions: an applied example. Prev Sci. 2020;21(6):872–83.
22. Karlis D, Xekalaki E. Mixed poisson distributions. Int Stat Review/Revue Int De Statistique 2005;35–58.
23. Lindsay BG. Mixture models: theory, geometry, and applications. Ims.
24. Laird N. Nonparametric maximum likelihood estimation of a mixing distribution. J Am Stat Assoc. 1978;73(364):805–11.

Publisher's note