ORIGINAL ARTICLE

PREDICTION OF INTER EPISODIC TIME FOR RECURRING MENTAL ILLNESS USING ORDER STATISTICS

Alka Sabharwal[1], Sakshi Kaushik[2], and Gurprit Grover[3]

[1]*Department of Statistics, Kirori Mal College, University of Delhi, Delhi, India.*

[2]*Quartesian, Chennai, Tamil Nadu, India.*

[3]*Department of Statistics, Faculty of Mathematical Sciences, University of Delhi, Delhi, India.*

Corresponding author: Sakshi Kaushik
Email: sakshi2007@gmail.com

ABSTRACT

Recurrent episodes are common across various mental disorders. Information on time to next episode, also referred as inter episodic times, provides a valuable tool for planning and evaluating the health outcomes of treatment in patients and developing effective preventive maintenance therapy. The objective is to obtain the prediction interval for the future inter episodic time when the number of previous episodes for a patient is small and inter episodic times are dependent. A data of 28 patients with a history of 3 or more recurring episodes of illness is extracted from a retrospective data of 146 patients diagnosed with mental and behavioral disorders. The prediction interval for time to occurrence of next episode is obtained using order statistics assuming that it will follow the order followed by previous inter episodic times. The validity of the results is verified using simulation studies with data generated using covariance structure of the real dataset. From the simulation studies, we found that more than 80% of the simulated inter episodic times lie in the simulated prediction intervals. This paper is highly beneficial to medical health professionals to predict time to next episode for patients with few previously known episodes of the concerned disease. The study has an implication to rare diseases where generally small database (patients) is available.

KEYWORDS: *Mental disorders, order statistics, recurrent events, time to next episode, waiting time*

INTRODUCTION

Mental disorders are common non-communicable diseases rising with epidemic rates globally. They are associated with a significant burden of morbidity and disability[1]. This burden further increases with recurrence of episodes of illness. Recurrent episodes are common across various mental disorders viz. depression, bipolar affective disorder, and schizophrenia. Depression is highly recurrent and significantly impacts personal and public health[2]. The course of depressive disorder is progressive with an increased risk of recurrence with the number of episodes[3]. Over 90% of patients with bipolar disorder experience recurrences during their lifetimes, often within 2 years of an initial episode and the consequences of recurrent illness for patients are substantial[4]. Thus, knowledge of the risk associated and time to recurrence is of clinical and scientific importance. The time to next episode aids in testing the effectiveness of antipsychotic medications.

Statistical analysis of recurrent event data has received considerable attention recently. A common characteristic among these events is the intrinsic correlation between those occurring in the same patient[5]. The events from the same patient share the information to some extent, resulting in correlated event times within the patient[6].

Recurrent event occurrences are modelled through counts, or numbers of events over specific periods of time, and through waiting times between successive events, also referred as inter episodic times. Counts are often useful with frequently occurring events within individuals while waiting times between events are useful with infrequent events[7]. Numerous survival models have been developed to analyze recurrent event data. However, most of them assume independence between successive event times resulting in biased results.

Prediction problems come up naturally in several real-life situations. With respect to order statistics, Grover, and Sabharwal estimated mean Diabetic Nephropathy onset time for all patients belonging to advanced nephropathy group[8]. They further concluded that application of order statistics for estimation and prediction is particularly beneficial in studies with small sample sizes. The best linear unbiased predictor (BLUP) method is another useful approach for prediction of order statistics and is based on the general linear model for the location and scale family[9].

A lot of work has been done previously to model recurrent event data in terms of waiting times when the number of recurrent events is large[5,7,10].

The objective of this study is to obtain the prediction interval for the future inter episodic time when the number of previous episodes for a patient is small and inter episodic times are dependent. The prediction interval is obtained using order statistics assuming that it will follow the order followed by previous inter episodic times. The retrospective data collected for this study consisted of 146 patients diagnosed with mental and behavioural disorders. Out of this data, patients with a history of 3 or more recurring episodes of illness were extracted resulting in a sample of size 28. The resulting data is further classified in two categories depending on the number of recurrent episodes. First category includes 21 patients with three episodes of illness while second category consists of 7 patients with four episodes of illness. Hence, first and second inter episodic times are available for all 28 patients while third inter episodic time is available for 7 patients only (belonging to second category) out of 28. The aim of this study is to predict the third inter episodic time for 21 patients (belonging to first category) using order statistics and their prediction interval. The validity of the results is verified using simulation studies with samples of sizes 30 and 50. The data generated is based on the covariance structure of the real dataset.

None of the previous studies, to the best of our knowledge, has estimated or predicted the time to next occurrence using order statistics especially in mental and behavioral disorders. They majorly targeted either to estimate the risk associated with recurrent episodes or obtain predictors of recurrence[2,4,11,12].

The outline of the rest of the paper is as follows: In Section 2, material used, and methodology has been discussed. Section 3 presents the results obtained on applying order statistics to the real as well as simulated data. The findings of the study are discussed in Section 4 and the paper is concluded in Section 5.

METHODS AND MATERIAL

Material used

A retrospective data of 146 patients diagnosed with mental and behavioural disorders (in accordance with American Psychiatric Association's Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-V) and the World Health Organization (WHO) International Classification of Diseases (ICD-10)) is collected from Lady Hardinge Medical College & Smt. S.K. Hospital, New Delhi, India. From this data, a sub sample consisting of patients with a history of at least 3 recurring episodes of illness was extracted resulting in a sample of size 28. The resulting sample consisted of patients diagnosed with Bipolar Affective Disorder (BPAD), Schizophrenia, Dementia and Psychotic Disorder: Not otherwise Specified (NOS). Out of these 28 patients only 7 patients had experienced 4 episodes of illness.
In this study information on three inter episodic times, say, t_{i1}, t_{i2} and t_{i3} are available, where t_{i1}, t_{i2} and t_{i3} denote waiting times between 1st and 2nd, 2nd and 3rd and 3rd and 4th episodes for i^{th} patient respectively. t_{i1} and t_{i2} are available for all 28 patients while t_{i3} is available for 7 patients only.

Methodology

Suppose that n subjects may experience an episode of illness and let $N_i(t)$ denote the number of events over the time period $(0, t]$ for the i^{th} patient. The times of occurrence of episodes for i^{th} patient are denoted by $x_{i,1} \leq x_{i,2} \leq \cdots$ and let $t_{ij} = x_{i,j+1} - x_{i,j}$ (with $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots$) denote waiting or gap time between successive episodes referred as inter episodic times. It is observed that t_{ij} follows lognormal distribution (as per BIC) with parameters μ_j and σ_j with the probability density function (pdf)

$$f(t_{ij}) = \frac{1}{t_{ij}\,\sigma_j\sqrt{2\pi}} exp\left(-\frac{(\ln t_{ij} - \mu_j)^2}{2\,\sigma_j{}^2}\right) \quad (1)$$

and cumulative distribution function (cdf)

$$F(t_{ij}) = \frac{1}{2}\left[1 + erf\left(\frac{\ln t_{ij} - \mu_j}{\sigma_j\sqrt{2}}\right)\right] \quad (2)$$

where erf is an error function defined as[13]:

$$erf(z) = \frac{2}{\sqrt{\pi}}\int_0^z e^{-t^2} dt$$

Estimation of ordered inter episodic times
Let $\tau_{1j} \leq \tau_{2j} \leq \cdots$ be the ordered inter episodic times of patients. The pdf and expected value of r^{th} $(r = 1, 2, \ldots, n)$ order statistic is given as:

$$f_{\tau_{rj}}(x) = \frac{n!}{(r-1)!(n-r)!}[F(x)]^{(r-1)}[1 - F(x)]^{(n-r)} f(x) \quad (3)$$

$$E_{\tau_{rj}}(x) = \int x f_{\tau_{rj}}(x) dx \quad (4)$$

on substituting (1) and (2) in (3),

$$f_{\tau_{rj}}(x = \frac{n!}{(r-1)!\,(n-r)!}\left[\frac{1}{2}\left\{+erf\left(\frac{\ln t_{ij} - \mu_j}{\sigma_j\sqrt{2}}\right)\right\}\right]^{(r-1)}\left[-\frac{1}{2}\left\{1 + erf\left(\frac{\ln t_{ij} - \mu_j}{\sigma_j\sqrt{2}}\right)\right\}\right]^{(n-r)}\left[\frac{1}{t_{ij}\,\sigma_j\sqrt{2\pi}} exp\left(-\frac{(\ln t_{ij} - \mu_j)^2}{2\,\sigma_j{}^2}\right)\right]$$

$$\Longrightarrow f_{\tau_{rj}}(x) = \frac{(-1)^{n-r}}{\beta(r, n-r+1)\cdot 2^{n-1}}\left[1erf\left(\frac{\ln t_{ij} - \mu_j}{\sigma_j\sqrt{2}}\right)\right]^{(n-1)}\left[\frac{1}{t_{ij}\,\sigma_j\sqrt{2\pi}} exp\left(-\frac{(\ln t_{ij} - \mu_j)^2}{2\,\sigma_j{}^2}\right)\right] \quad (5)$$

Simulation of multivariate distribution
For simulating multivariate lognormal inter episodic times t_{i1}, t_{i2}, …, the exponential

relationship between lognormal and normal distributions is considered, i.e., if $U \sim N(\mu, \Sigma)$ follows multivariate normal distribution (here trivariate), then $Y = \exp(U)$ follows multivariate lognormal distribution[14,15]. The density of U is given by

$$f_U(u) = |2\pi det\Sigma|^{-1/2} exp\left\{-\frac{1}{2}(u-\mu)'\Sigma^{-1}(u-\mu)\right\}$$

where μ is the mean and Σ is the $m \times m$ positive-definite variance-covariance matrix of U. The simulation technique is based on factorization of the variance-covariance matrix as $\Sigma = PP'$ for some $m \times m$ matrix P. The Cholesky decomposition can be used for this factorisation[16]. Let $Z_1, Z_2, \ldots, Z_m$ be independent univariate $N(0,1)$, and $Z' = (Z_1, \ldots, Z_m)$. Then U can be simulated as[17] $U = \mu + PZ$. Then $U \sim N(\mu, \Sigma)$.

Steps for estimation of inter episodic times

The following steps are used to estimate inter episodic times:

1. Appropriate distribution is fitted to inter episodic times t_{i1}, t_{i2} and t_{i3} and parameters are estimated using method of maximum likelihood estimation (mle)
2. Estimation of ordered inter episodic times τ_{i1}, τ_{i2} and τ_{i3} using order statistics
3. Obtaining prediction interval for τ_{i3}
4. Comparison of results from real data with simulation study

All calculations were performed in R version 3.4.0.

RESULTS

The application and results of the methodology followed to estimate and predict inter episodic times (described in section 2.2.3) is discussed stepwise as follows.

Step 1: Distribution Selection and Parameter Estimation

Various distributions are fitted to inter episodic times t_{i1} and t_{i2} resulting in log normal to be the most appropriate distribution based on minimum BIC[18] values of 136.2362 and 113.9542 for t_{i1} and t_{i2}, respectively. Table 1 presents the BIC values calculated for various distributions fitted to t_{i1} and t_{i2} $(i = 1, 2, \ldots, 28)$. Figure 1 presents the lognormal pdf plots of t_{i1} and t_{i2}. Further, the validity of fitted distribution is verified using Anderson–Darling (AD) goodness of fit test[19]. The Anderson-Darling statistic and highly significant p-values (t_{i1}: AD=0.46152, p-value = 0.7849; t_{i2}: AD=0.15414, p-value = 0.9984) confirms that log normal distribution appropriately fits t_{i1} and t_{i2}. Since the number of patients with 4 episodes is 7 i.e. t_{i3} is available for 7 patients only, thus no distribution based on BIC values was performed for t_{i3}. The distribution of t_{i3} is assumed to be log normal assuming that it will follow the order of either t_{i1} or t_{i2}. The scale and shape parameters of t_{i1} and t_{i2} are estimated using the method of maximum likelihood estimation (mle) are also presented in Table 1. For t_{i3}, bootstrapping procedure was adopted to generate a sample and then the parameters of t_{i3} are estimated using mle[20,21].

Step 2 and 3: Estimation of ordered inter episodic times τ_{i1}, τ_{i2} and τ_{i3} using order statistics and Prediction interval for τ_{i3}

Thirdly, ordered inter episodic times for t_{i1} and t_{i2}, denoted as τ_{i1} and τ_{i2}, are estimated for all 28 patients using order statistic from equations (3), (4) and (5) with the parameters estimated above.

TABLE 1: Bayesian information criteria (BIC) values and maximum likelihood estimates (mle) of scale and shape parameters of different distributions for inter episodic times

Variables	Distribution	BIC Values	Selected Distribution (mle parameters)
t_{i1}*a	Birnbaum-Saunders	139.8074	Lognormal
	Gamma	142.4370	$(\widehat{\mu_{i1}} = 1.3299, \widehat{\sigma_{i1}} = 0.6706)$
	Lognormal	136.2362	
	Weibull	144.9766	
	Normal	160.1432	
t_{i2}*a	Birnbaum-Saunders	117.6582	Lognormal
	Gamma	118.7630	$(\widehat{\mu_{i1}} = 0.9254, \widehat{\sigma_{i1}} = 0.6750)$
	Lognormal	113.9542	
	Weibull	121.1714	
	Normal	135.5846	
t_{i3}*bc	NA	NA	Lognormal
			$(\widehat{\mu_{i1}} = 0.6283, \widehat{\sigma_{i1}} = 0.5626)$

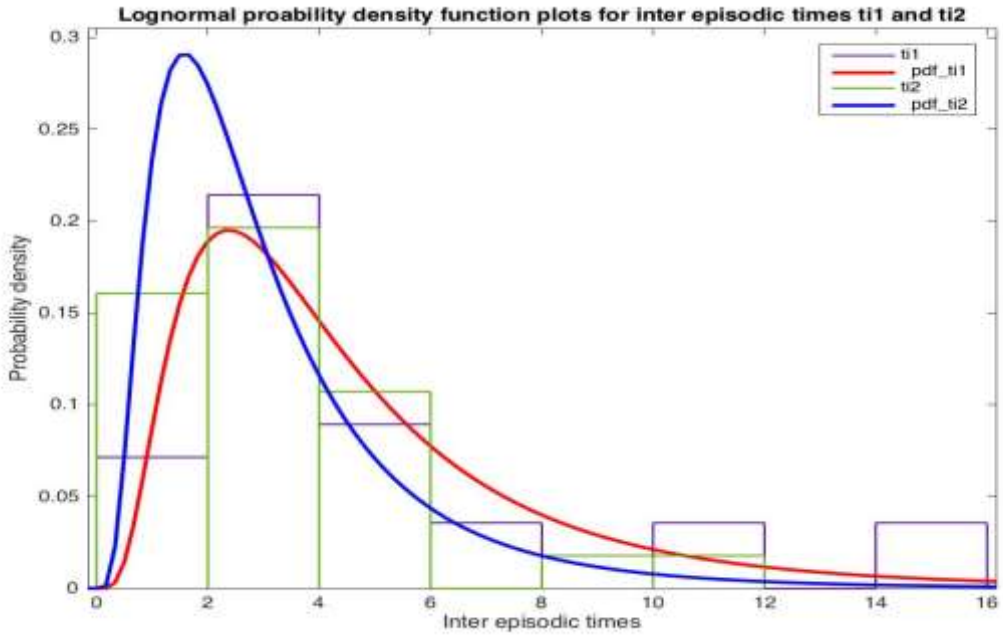
BIC: Bayesian information criteria, mle: maximum likelihood estimates

* $t_{i1} = x_{i,2} - x_{i,1}$, $t_{i2} = x_{i,3} - x_{i,2}$ *and* $t_{i3} = x_{i,4} - x_{i,3}$ *denote waiting or gap time between successive episodes referred as inter episodic times for* i^{th} *patient.*

a *Distributions for* t_{i1}*and* t_{i2} *are obtained using Bayesian information Criterion (BIC).*

b *Distribution for* t_{i3} *is assumed to be lognormal based on of* t_{i1}*and* t_{i2}.

c *Parameters for* t_{i3} *are estimated from the sample generated using bootstrapping technique.*



$t_{i1} = x_{i,2} - x_{i,1}$ *and* $t_{i2} = x_{i,3} - x_{i,2}$ *denote waiting or gap time between successive episodes referred as inter episodic times for* i^{th} *patient.*

FIGURE 1 Lognormal probability density function plots for inter episodic times t_{i1} and t_{i2}.

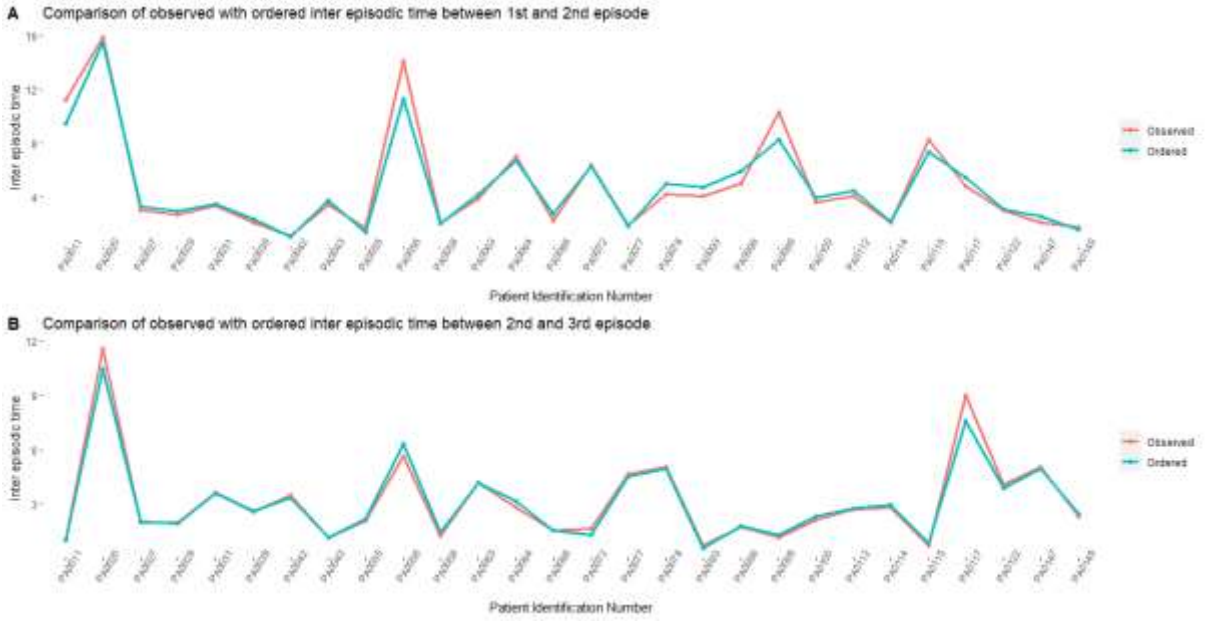


FIGURE 2 A: Comparison of observed inter episodic time t_{i1} with estimated ordered inter episodic time τ_{i1} between 1st and 2nd episodes. B: Comparison of observed inter episodic time t_{i2} with estimated ordered inter episodic time τ_{i2} between 2nd and 3rd episodes.

$t_{i1} = x_{i,2} - x_{i,1}$ *denote waiting or gap time between 1st and 2nd episode referred as first inter episodic time for* i^{th} *patient and* τ_{i1} *is the ordered inter episodic times for* t_{i1}.

$t_{i2} = x_{i,3} - x_{i,2}$ *denote waiting or gap time between 2nd and 3rd episode referred as first inter episodic time for* i^{th} *patient and* τ_{i2} *is the ordered inter episodic times for* t_{i2}.

Figure 2 presents the comparison of t_{i1} with τ_{i1} and t_{i2} with τ_{i2}. From Figures 2A and 2B, it is evident that the estimated ordered inter episodic times almost overlap with the observed inter episodic times for all 28 patients. This further strengthens our estimation approach. In this study, it is assumed that the waiting time to next episode will follow the order followed by the previous waiting times. Hence, the ordered inter episodic times for t_{i3} are estimated twice using order statistics assuming that t_{i3} will follow the order of either t_{i1} or t_{i2}, denoted as $\tau_{i3\_t_{i1}}$ and $\tau_{i3\_t_{i2}}$,. Further, using these estimates, a prediction interval for t_{i3} is obtained by taking the

lower and upper limits as minimum ($\tau_{i3\_}t_{i1}, \tau_{i3\_}t_{i2}$) and maximum ($\tau_{i3\_}t_{i1}, \tau_{i3\_}t_{i2}$), respectively. Thus, the time to next episode of illness is expected to lie in this interval. Table 2 presents the ordered inter episodic times τ_{i1}, τ_{i2}, $\tau_{i3\_}t_{i1}$ and $\tau_{i3\_}t_{i2}$ and prediction interval for t_{i3}. It is evident from the Table that for patient with ptid PA0039, $\tau_{i3\_}t_{i1}$ and $\tau_{i3\_}t_{i2}$ are estimated to be 1.2247 and 1.8614 with respect to the order followed by t_{i1} and t_{i2}. The prediction interval is obtained as (1.2247, 1.8614). The observed inter episodic time t_{i3} is 1.3333 which lies in the prediction interval. Thus, it is predicted that the time to 4th episode for all patients may lie within this predicted interval.

Step 5: Comparison of results with simulation studies

Simulation studies with sample sizes 30 and 50 are performed to compare the results obtained in this study for generalisation. The procedure followed to perform simulation study is explained as:

1. t_{i1}, t_{i2} and t_{i3} are converted to normal variates by applying logarithm transformation[14,15].
2. The mean and covariance structure of the transformed t_{i1}, t_{i2} and t_{i3} are obtained.

Predicated on this mean and covariance structure, multivariate normal distribution (here tri variate) are generated using "rmvnorm" function from "dmutate" package in R software. Simulation studies with sample sizes 30 and 50 are performed to compare the results obtained in this study for generalisation. The procedure followed to perform simulation study is explained as:

3. t_{i1}, t_{i2} and t_{i3} are converted to normal variates by applying logarithm transformation[14,15].
4. The mean and covariance structure of the transformed t_{i1}, t_{i2} and t_{i3} are obtained.
5. Predicated on this mean and covariance structure, multivariate normal distribution (here tri variate) are generated using "rmvnorm" function from "dmutate" package in R software.

TABLE 2: Ordered Inter episodic times and prediction interval estimated using order statistics

S No	Pt Id	t_{i1}*	τ_{i1}**	t_{i2}*	τ_{i2}**	t_{i3}*ɑ	$\tau_{i3_}t_{i1}$ᵇ	$\tau_{i3_}t_{i2}$ᶜ	PIᵭ
1	PA0011	11.1667	9.4296	1.0833	1.0544		3.7800	0.8856	(0.8856, 3.7800)
2	PA0020	15.8333	15.4901	11.5833	10.4383		5.6310	5.6310	(5.6310, 5.6310)
3	PA0027	3.0000	3.2950	4.0833	3.8750		1.6079	1.5292	(1.5292, 1.6079)
4	PA0029	2.6667	2.9090	2.0000	1.9373		1.4522	1.4522	(1.4522, 1.4522)
5	PA0031	3.3333	3.4926	3.6667	3.6050		1.6890	2.4024	(1.6890, 2.4024)
6	PA0039	2.0833	2.3599	2.8333	3.1654	1.3333	1.2247	1.8614	(1.2247, 1.8614)
7	PA0042	1.0833	1.0186	3.5000	3.3731		0.5240	2.2749	(0.5240, 2.2749)
8	PA0043	3.4167	3.7070	1.1667	1.1947		1.7732	0.9810	(0.9810, 1.7732)
9	PA0055	1.6667	1.3499	2.0833	2.1969		0.7622	1.6079	(0.7622, 1.6079)
10	PA0056	14.0833	11.2789	5.6667	6.3317		4.3680	3.7800	(3.7800, 4.3680)
11	PA0058	2.0833	1.9927	1.2500	1.4484	0.9500	1.0666	1.1471	(1.0666, 1.1471)
12	PA0063	3.8333	4.2130	4.1667	4.1854	2.4167	1.9546	2.7044	(1.9546, 2.7044)
13	PA0064	7.0000	6.6670	2.8333	2.9731	2.1000	2.8904	2.1598	(2.1598, 2.8904)
14	PA0066	2.2500	2.7223	1.5833	1.5700		1.3763	1.2247	(1.2247, 1.3763)
15	PA0072	6.3333	6.2934	1.6667	1.6911		2.7044	1.0666	(1.0666, 2.7044)
16	PA0077	1.9167	1.7990	4.6667	4.5376		0.9810	2.8904	(0.9810, 2.8904)
17	PA0078	4.1667	4.9900	5.0833	5.5546		2.2749	3.3993	(2.2749, 3.3993)
18	PA0093	4.0833	4.6980	2.7500	2.7957		2.0536	0.5240	(0.5240, 2.0536)
19	PA0096	5.0000	5.8848	1.7500	1.8131		2.5447	1.3763	(1.3763, 2.5447)
20	PA0099	10.2500	8.2695	1.1667	1.3243		3.3993	1.3007	(1.3007, 3.3993)
21	PA0100	3.5833	3.9510	2.1667	2.3347		1.8614	1.6890	(1.6890, 1.8614)
22	PA0112	4.0833	4.4616	0.7500	0.8851		2.0536	1.9546	(1.9546, 2.0536)
23	PA0114	2.0833	2.1783	2.5833	2.6321	2.0033	1.1471	2.0536	(1.1471, 2.0536)
24	PA0115	8.2500	7.3290	0.7500	0.6005	2.0000	3.1173	0.7622	(0.7622, 3.1173)
25	PA0117	4.7500	5.4037	9.0000	7.5830		2.4024	4.3680	(2.4024, 4.3680)
26	PA0122	3.0000	3.1009	2.0000	2.0648		1.5292	2.5447	(1.5292, 2.5447)
27	PA0147	2.0833	2.5403	5.0833	4.9747		1.3007	3.1173	(1.3007, 3.1173)
28	PA0149	1.7500	1.5900	2.3333	2.4793		0.8856	1.7732	(0.8856, 1.7732)

S No: serial number, Pt Id: patient identification number, PI: Prediction interval.

* $t_{i1} = x_{i,2} - x_{i,1}$, $t_{i2} = x_{i,3} - x_{i,2}$ *and* $t_{i3} = x_{i,4} - x_{i,3}$ *denote waiting or gap time between successive episodes referred as inter episodic times for* i^{th} *patient.*

** τ_{i1}, τ_{i2} *and* τ_{i3} *denote ordered inter episodic times for* i^{th} *patient estimated for* t_{i1}, t_{i2} *and* t_{i3}.

ɑ *inter episodic times* t_{i3} *are available for only 7 patients. For other 21 patients,* t_{i3} *is unknown indicated by blanks.*

$\tau_{i3_}t_{i1}$ᵇ τ_{i3} *is estimated with respect to the pattern followed by* t_{i1}.

$\tau_{i3_}t_{i2}$ᶜ τ_{i3} *is estimated with respect to the pattern followed by* t_{i2}.

ᵭ *The waiting time for next episode is predicted to lie in this interval, derived on the basis of* $\tau_{i3\_}t_{i1}$ *and* $\tau_{i3\_}t_{i2}$.

TABLE 3: Inter episodic times estimated using order statistics based on simulated data of size 30

S. No.	Pt. Id_s[a]	$t_{i1\_s}^{*}$	$\tau_{i1\_s}^{**}$	$t_{i2\_s}^{*}$	$\tau_{i2\_s}^{**}$	$t_{i3\_s}^{*}$	PI[b]	$\tau_{i3\_s}^{**}$
1	PA0004	0.9870	0.9871	1.7802	1.6197	0.7788	(0.9871, 1.6197)	1.1314
2	PA0010	1.1949	1.3001	6.9840	7.2282	2.8131	(1.3001, 7.2282)	2.5273
3	PA0001	1.5936	1.5232	2.3747	2.5587	1.8380	(1.5232, 2.5587)	1.8922
4	PA0022	1.7379	1.7159	2.7029	2.9296	1.7954	(1.7159, 2.9296)	1.7489
5	PA0079	1.8234	1.8931	2.9574	3.0816	2.1621	(1.8931, 3.0816)	2.2763
6	PA0091	1.8762	2.0617	2.4141	2.6803	1.7545	(2.0617, 2.6803)	1.5896
7	PA0048	2.2982	2.2257	6.3955	6.2419	4.4114	(2.2257, 6.2419)	4.1837
8	PA0074	2.3652	2.3878	2.0245	1.9755	2.9175	(1.9755, 2.3878)	3.0068
9	PA0109	2.4671	2.5492	2.1411	2.2027	2.0528	(2.2027, 2.5492)	2.3258
10	PA0042	2.5783	2.7118	3.3552	3.4157	3.1962	(2.7118, 3.4157)	3.3015
11	PA0047	2.5865	2.8794	5.7411	5.6230	2.8149	(2.8794, 5.6230)	2.6519
12	PA0119	2.7374	3.0541	2.3664	2.4364	2.3496	(2.4364, 3.0541)	2.5003
13	PA0055	2.7886	3.2321	1.9301	1.8602	4.6699	(1.8602, 3.2321)	4.7262
14	PA0096	3.2307	3.4077	2.0530	2.0893	2.1332	(2.0893, 3.4077)	2.1530
15	PA0143	3.2577	3.5869	2.4661	2.8005	3.1858	(2.8005, 3.5869)	3.3574
16	PA0016	3.9670	3.7913	4.6176	4.7441	2.9081	(3.7913, 4.7441)	2.8764
17	PA0082	4.0774	4.0341	11.2184	9.3642	4.9511	(4.0341, 9.3642)	4.7709
18	PA0037	4.1295	4.2912	4.4452	4.5234	2.8711	(4.2912, 4.5234)	2.7663
19	PA0088	4.1980	4.5184	3.7785	3.7041	4.5701	(3.7041, 4.5184)	4.0327
20	PA0076	4.8335	4.7183	2.1981	2.3176	1.4740	(2.3176, 4.7183)	1.3997
21	PA0073	4.9706	4.9765	1.0633	1.1734	3.4436	(1.1734, 4.9765)	3.6672
22	PA0035	5.1790	5.3810	3.7158	3.5542	5.1478	(3.5542, 5.3810)	4.9576
23	PA0044	5.3113	5.8798	1.9262	1.7423	2.9862	(1.7423, 5.8798)	3.1747
24	PA0003	6.1031	6.2946	4.3551	4.2357	5.5460	(4.2357, 6.2946)	5.5363
25	PA0077	6.6991	6.6050	0.8806	0.9251	3.8791	(0.9251, 6.6050)	3.8556
26	PA0135	8.7212	7.1536	1.5925	1.4890	3.3960	(1.4890, 7.1536)	3.5847
27	PA0011	8.9001	8.1923	2.9790	3.2527	8.1529	(3.2527, 8.1923)	7.6428
28	PA0123	10.9601	9.2023	4.7715	5.0559	6.7841	(5.0559, 9.2023)	6.6668
29	PA0051	12.1950	10.9410	4.1495	3.9306	8.4693	(3.9306, 10.9410)	9.7210
30	PA0111	12.8083	14.8790	1.1959	1.3444	6.3466	(1.3444, 14.8790)	6.2384

S No: serial number, Pt Id: patient identification number, PI: Prediction Interval.

[a] *Pt Id_s denotes randomly generated patient identification number.*

* *t_{i1_s}, t_{i2_s} and t_{i3_s} denote successive inter episodic times simulated with regard to same covariance and parameter structure as that of actual inter episodic times t_{i1}, t_{i2} and t_{i3}.*

** *τ_{i1_s}, τ_{i2_s} and τ_{i3_s} denote ordered inter episodic times derived from t_{i1_s}, t_{i2_s} and t_{i3_s}.*

[b] *The waiting time for next episode is predicted to lie in this interval, derived on the basis of τ_{i1_s} and τ_{i2_s} .*

6. This multivariate normal distribution is then transformed to multivariate log normal distribution using exponential transformation. The resulting tri variate log normal distribution represent the simulated inter episodic times, denoted as t_{i1_s}, t_{i2_s} and t_{i3_s}.

7. Ordered inter episodic times denoted as τ_{i1_s}, τ_{i2_s} and τ_{i3_s} are estimated for simulated inter episodic times t_{i1_s}, t_{i2_s} and t_{i3_s}. The simulated inter episodic times t_{i1_s}, t_{i2_s} and t_{i3_s} along with estimated ordered times τ_{i1_s}, τ_{i2_s} and τ_{i3_s} for sample sizes 30 and 50 are presented in Tables 3 and 4.
8. The simulated results are compared with results from real data. For example, the patient with ptid PA0149 from Table 2 is comparable to patient with ptid PA0022 in Table 3 since t_{i1} of former is 1.7500 while simulated t_{i1} of latter is $t_{i1_s} = 1.7379$. τ_{i3_s} for PA0022 is estimated to be 1.7489 which lies in the prediction interval (0.8856, 1.7732) estimated for PA0149. Similar comparison can be performed for other patients as well from Tables 2, 3 and 4. Thus, simulation studies generalise our results suggesting that order statistics can be used to predict the time to next episode based on the information available for the recent episode and order of previous episodes.

TABLE 4: Inter episodic times estimated using order statistics based on simulated data of size 50

S. No.	Pt. Id_s[a]	$t_{i1\_s}^{*}$	$\tau_{i1\_s}^{**}$	$t_{i2\_s}^{*}$	$\tau_{i2\_s}^{**}$	$t_{i3\_s}^{*}$	PI[b]	$\tau_{i3\_s}^{**}$
1	PA0102	0.4361	0.7454	2.7889	2.8966	1.5347	(0.7454, 2.8966)	1.6224
2	PA0086	0.8126	1.0123	3.1356	3.3687	0.6712	(1.0123, 3.3687)	1.9149
3	PA0053	1.2382	1.1785	2.7367	2.8811	2.5436	(1.1785, 2.8811)	2.5639
4	PA0100	1.3271	1.3164	3.9172	3.7598	1.4731	(1.3164, 3.7598)	1.5422
5	PA0058	1.3482	1.4398	11.4756	8.0058	6.5746	(1.4398, 8.0058)	6.7467
6	PA0134	1.3701	1.5541	2.6796	2.6827	1.3883	(1.5541, 2.6827)	1.8610
7	PA0119	1.6809	1.6625	2.7965	2.8684	3.6062	(1.6625, 2.8684)	3.8355
8	PA0142	1.8250	1.7667	5.0619	4.9882	4.0023	(1.7667, 4.9882)	3.8260
9	PA0105	1.8434	1.8681	2.1462	2.1546	1.8496	(1.8681, 2.1546)	1.9089
10	PA0068	1.9678	1.9675	1.7874	1.7606	3.5619	(1.7606, 1.9675)	3.7779
11	PA0123	2.4028	2.0655	2.4125	2.3308	2.5935	(2.0655, 2.3308)	2.2850
12	PA0085	2.5237	2.1625	2.9322	2.9202	2.0090	(2.1625, 2.9202)	2.2410
13	PA0097	2.6955	2.2592	1.8748	1.8169	1.7729	(1.8169, 2.2592)	1.8414
14	PA0098	2.6958	2.3572	3.9884	4.7104	3.5326	(2.3572, 4.7104)	3.5668
15	PA0011	2.6998	2.4573	2.4133	2.3660	3.6736	(2.3660, 2.4573)	3.7254
16	PA0062	2.7031	2.5584	3.9798	4.2286	2.7162	(2.5584, 4.2286)	2.8895
17	PA0064	2.8135	2.6573	3.5149	3.5854	1.6213	(2.6573, 3.5854)	2.6983
18	PA0003	2.9180	2.7516	2.8145	2.8536	2.8847	(2.7516, 2.8536)	2.8334
19	PA0094	3.0219	2.8447	1.8754	1.8726	2.0523	(1.8726, 2.8447)	2.1106
20	PA0060	3.0250	2.9459	1.6331	1.5292	1.3680	(1.5292, 2.9459)	1.7521
21	PA0143	3.1925	3.0648	1.3399	1.3345	1.4071	(1.3345, 3.0648)	1.4560
22	PA0139	3.3006	3.2010	3.1501	3.6219	2.2282	(3.2010, 3.6219)	2.1825
23	PA0124	3.3488	3.3391	7.1688	6.4564	4.1158	(3.3391, 6.4564)	4.3878
24	PA0020	3.3576	3.4560	2.5905	2.5589	4.9321	(2.5589, 3.4560)	4.6601
25	PA0145	3.4502	3.5379	1.6257	1.4674	1.9891	(1.4674, 3.5379)	1.9747
26	PA0005	3.4863	3.5969	3.2160	3.7524	2.3367	(3.5969, 3.7524)	2.3543
27	PA0093	3.6926	3.6719	2.3598	2.2832	2.5194	(2.2832, 3.6719)	2.4009
28	PA0061	3.8276	3.8091	1.6222	1.4028	1.7226	(1.4028, 3.8091)	1.7712
29	PA0073	4.0779	4.0299	2.5198	2.4030	2.3218	(2.4030, 4.0299)	2.5506
30	PA0112	4.1219	4.3065	1.0573	1.1799	3.4265	(1.1799, 4.3065)	3.2845
31	PA0081	4.3112	4.5664	0.8574	0.9717	1.2550	(0.9717, 4.5664)	1.1180
32	PA0148	4.3534	4.7287	2.3038	2.2218	4.4637	(2.2218, 4.7287)	4.1793
33	PA0107	4.6219	4.7588	1.6662	1.5889	2.3228	(1.5889, 4.7588)	2.3080
34	PA0033	4.6277	4.7073	0.9369	1.0868	2.8063	(1.0868, 4.7073)	2.8585
35	PA0137	4.6463	4.7036	4.0293	4.9035	4.8284	(4.7036, 4.9035)	4.8316
36	PA0077	5.2436	4.8930	2.0475	2.0325	2.7149	(2.0325, 4.8930)	2.8756
37	PA0076	5.3481	5.3404	1.8960	1.9792	3.1819	(1.9792, 5.3404)	2.8775
38	PA0063	5.4179	5.9587	3.6384	3.5444	7.8214	(3.5444, 5.9587)	7.6459
39	PA0037	5.4948	6.5202	5.6794	5.7283	5.9652	(5.7283, 6.5202)	4.9925
40	PA0065	6.3607	6.7815	1.7430	1.6470	3.8176	(1.6470, 6.7815)	3.3887
41	PA0135	6.4703	6.6771	0.8548	0.7858	4.2246	(0.7858, 6.6771)	4.9258
42	PA0010	7.0965	6.4559	3.1125	3.1027	3.7801	(3.1027, 6.4559)	3.5265
43	PA0103	7.9997	6.6054	2.5332	2.4629	6.4447	(2.4629, 6.6054)	5.9931
44	PA0140	8.1946	7.4869	2.7066	2.8030	3.9241	(2.8030, 7.4869)	3.4656
45	PA0079	8.7039	8.8645	4.6201	4.8280	7.8494	(4.8280, 8.8645)	9.5145
46	PA0087	8.9224	9.8694	2.1372	2.0903	3.2401	(2.0903, 9.8694)	3.0317
47	PA0127	8.9365	10.0016	1.7614	1.7040	2.6938	(1.7040, 10.0016)	2.8010
48	PA0024	10.1556	11.2289	1.8865	1.9266	2.5196	(1.9266, 11.2289)	2.4667
49	PA0019	12.7360	13.2355	3.4502	3.7147	6.5934	(3.7147, 13.2355)	6.7707
50	PA0067	15.1628	17.8515	1.1354	1.2610	4.9721	(1.2610, 17.8515)	4.5108

S No: serial number, Pt Id: patient identification number, PI: Prediction Interval.

[a] Pt Id_s denote randomly generated patient identification number.

** t_{i1_s}, t_{i2_s} and t_{i3_s} denote successive inter episodic times simulated with regard to same covariance and parameter structure as that of actual inter episodic times t_{i1}, t_{i2} and t_{i3}.*

*** τ_{i1_s}, τ_{i2_s} and τ_{i3_s} denote ordered inter episodic times derived from t_{i1_s}, t_{i2_s} and t_{i3_s}.*

[b] the waiting time for next episode is predicted to lie in this interval, derived based on τ_{i1_s} and τ_{i2_s} .

DISCUSSION

Occurrence of recurrent episodes of illness in mental disorders worsens the cognitive and psychological impaired state of a patient. It reflects the relapsing nature of illness and narrows the chances of remission. A recurrent events model can help to gain insights into the disease process. Information on time to next episode provides a valuable tool for planning and evaluating the health outcome results of treatment in patients. It is also beneficial in developing an effective preventive maintenance therapy. Thus, we attempted to estimate (observed inter episodic times) and hence predict the time to next episode of a recurring mental disorder.

Incompleteness of data is a common issue across all dimensions of research. In medical research, data on a single parameter for a patient holds a significant value. The omission of participants with missing values can have a big impact on the analysis[22]. Thus, instead of neglecting the information on the inter episodic times of 7 patients, we utilized this incomplete data to estimate the parameters and obtain the prediction interval for the inter episodic times of all 28 patients assuming that the waiting time to next episode will follow the order followed by the previous waiting times. We observed that out of 7 patients, the observed inter episodic times of 5 patients lied in the prediction interval. However, the remaining 2 were close to the lower limit of the prediction interval. Further, from simulation studies, we found that more than 80% (80% and 82% for studies with sample sizes 30 and 50, respectively) of the simulated inter episodic times lie in the simulated prediction intervals.

In this study, we have simulated multivariate lognormal data based on the covariance structure of the real dataset. This approach is advantageous as it captures the interdependence of variables and hence is closer to the real data.

This paper is highly beneficial to medical health professionals to predict time to next episode for patients with very few previously known episodes of the concerned disease. The number of dependent previous episodes may be few since either there is limited information regarding previous disease history, or the patient is under preventive maintenance therapy. The study has an implication to rare diseases as well where generally small database (patients) is available. The study adds new evidence to the literature for prediction of time to future episodes and its application can be extended to other disorders as well.

One of the limitations of the study is the availability of a smaller number of episodes per patient. Owing to this limitation, the current method used in this study could not be compared with other statistical methods. Further, as the data was collected from one hospital only, thus, the results could not be verified with other hospital patient data. Though, the model was validated using simulation.

CONCLUSION

It can be concluded from both real as well as the simulation study that the time to next episodic (known) lied in the estimated prediction interval. Thus, the estimated prediction interval is reliable and can be used for obtaining the time to next episode.

ACKNOWLEDGEMENT

We are grateful to Dr. K. E. Sadanand Unni, Professor, Department of Psychiatry & Drug De Addiction Center, Lady Hardinge Medical College & Smt. S.K, Hospital, New Delhi, India, for providing us the data and assisting us in gaining a deeper understanding of data concerning mental disorders.

CONFLICT OF INTEREST

The authors have no conflicts of interest to declare.

REFERENCES

1. WHO International Consortium in Psychiatric Epidemiology. Cross-national comparisons of the prevalences and correlates of mental disorders. *Bulletin of the World Health Organization : the International Journal of Public Health* 2000;78(4):413–426. https://pubmed.ncbi.nlm.nih.gov/10885160/

2. Burcusa SL, Iacono WG. Risk for recurrence in depression. *Clinical psychology review* 2007;27(8):959-85. doi: 10.1016/j.cpr.2007.02.005.

3. Kessing LV. Severity of depressive episodes during the course of depressive disorder. *The British journal of psychiatry* 2008;192:290-293. doi: 10.1192/bjp.bp.107.038935.

4. Perlis RH, Ostacher MJ, Patel JK, et al. Predictors of recurrence in bipolar disorder: primary outcomes from the Systematic Treatment Enhancement Program for Bipolar Disorder (STEP-BD). *The American Journal of Psychiatry* 2006;163(2):217–224. doi: 10.1176/appi.ajp.163.2.217.

5. Amorim LD, Cai J. Modelling recurrent events: a tutorial for analysis in epidemiology. *International Journal of Epidemiology* 2015;44(1):324-333. doi: 10.1093/ije/dyu222.

6. Smedinga H, Steyerberg EW, Beukers W, et al. Prediction of Multiple Recurrent Events: A Comparison of Extended Cox Models in Bladder Cancer. *American Journal of Epidemiology* 2017;186(5):612-623. doi: 10.1093/aje/kwx133.

7. Cook RJ, Lawless JF. The Statistical Analysis of Recurrent Events (1 ed.). New York: Springer-Verlag New York 2007.

8. Grover G, Sabharwal A. A Parametric Approach to Estimate Survival Time of Diabetic Nephropathy with Left Truncated and Right Censored Data. *International Journal of Statistics and Probability* 2012:1(1):128-137. DOI:10.5539/ijsp.v1n1p128.

9. Arnold BC, Balakrishnan N, Nagaraja HN. A First Course in Order Statistics. Philadelphia: Society for Industrial and Applied Mathematics 1993.

10. Yang W, Jepson C, et al. Statistical Methods for Recurrent Event Analysis in Cohort Studies of CKD. *Clinical Journal of the American Society of Nephrology* 2017;12(12):2066-2073. doi: 10.2215/CJN.12841216.

11. Kessing LV, Olsen EW, Andersen PK. Recurrence in Affective Disorder: Analyses with Frailty Models. *American Journal of Epidemiology* 1999;149(5):404-411. doi: 10.1093/oxfordjournals.aje.a009827.

12. Doesschate MC, Bockting CL, Koeter M, et al. Prediction of Recurrence in Recurrent Depression: A 5.5-Year Prospective Study. *Journal of clinical psychiatry* 2010;71:984-991. doi: 10.4088/JCP.08m04858blu

13. Lebedev NN. Special Functions and Their Applications. New Jersey: Prentice-Hall 1965.

14. Ghasem T. Multivariate Log – Normal Distribution. ISI Proceedings. 53rd Session. Seoul: International Statistical Institute 2001:1000-1001. https://2001.isiproceedings.org/pdf/329.PDF.

15. Halliwell L. The Lognormal Random Multivariate. Casualty Actuarial Society E-Forum 2015:1-5.

16. Golub GH, Van Loan CF. Matrix Computations (2nd ed.). Baltimore: John Hopkins University Press 1989.

17. Ince PJ, Buongiorno J. Multivariate stochastic simulation with subjective multivariate normal distributions. Proceedings of the 1991 Symposium on Systems Analysis in Forest Resources : March 3-6, 1991, Charleston, South Carolina. Asheville, NC : Southeastern Forest Experiment Station. General technical report SE 1991;74:143-150. https://www.fs.usda.gov/treesearch/pubs/5784.

18. Schwarz G. Estimating the dimension of a model. *Ann.Statist* 1978;6:461–464. https://projecteuclid.org/download/pdf_1/euclid.aos/1176344136.

19. Stephens MA. EDF Statistics for Goodness of Fit and Some Comparisons. *Journal of the American Statistical Association* 1974;69:730-737.

20. Efron B, Tibshirani RJ. An Introduction to the Bootstrap. New York: Chapman & Hall 1993.

21. Grover G, Sabharwal A, Mittal J. A Bayesian Approach for Estimating Onset Time of Nephropathy for Type 2 Diabetic Patients Under Various Health Conditions. *International Journal of Statistics and Probability* 2013;2(2):89-101. DOI:10.5539/ijsp.v2n2p89.

22. Ibrahim JG, Chu H, Chen MH. Missing data in clinical studies: issues and methods. Journal of clinical oncology : official journal of the American Society of Clinical Oncology 2012;30(26):3297-303. DOI: 10.1200/JCO.2011.38.7589