ISSN 1927-7032 (Print)
ISSN 1927-7040 (Online)

International Journal of Statistics and Probability

Vol. 11, No. 3 May 2022

CANADIAN CENTER OF SCIENCE AND EDUCATION

INTERNATIONAL JOURNAL OF STATISTICS AND PROBABILITY

An International Peer-reviewed and Open Access Journal for Statistics and Probability

International Journal of Statistics and Probability (ISSN: 1927-7032; E-ISSN: 1927-7040) is an open-access, international, double-blind peer-reviewed journal published by the Canadian Center of Science and Education. This journal, published **bimonthly** (January, March, May, July, September and November) in both **print and online versions**, keeps readers up-to-date with the latest developments in all areas of statistics and probability.

The scopes of the journal:

- Computational statistics
- Design of experiments
- Sample survey
- Statistical modelling
- Statistical theory
- Probability theory

The journal is included in:

- BASE
- Google Scholar
- JournalTOCs
- LOCKSS
- SHERPA/RoMEO
- Ulrich's

Submission Policy

Submission of an article implies that the work described has not been published previously (except in the form of an abstract or as part of a published lecture or academic thesis), that it is not under consideration for publication elsewhere, that its publication is approved by all authors and tacitly or explicitly by the authorities responsible where the work was carried out. However, we accept submissions that have previously appeared on preprint servers (for example: arXiv, bioRxiv, Nature Precedings, Philica, Social Science Research Network, and Vixra); have previously been presented at conferences; or have previously appeared in other "non-journal" venues (for example: blogs or posters). Authors are responsible for updating the archived preprint with the journal reference (including DOI) and a link to the published articles on the appropriate journal website upon publication.

iThenticate Professional Plagiarism Prevention

The publisher and journals have a zero-tolerance plagiarism policy. We check the issue using two methods: a plagiarism prevention tool (iThenticate) and a reviewer check. All submissions will be checked by iThenticate before being sent to reviewers.

creative commons

We insist a rigorous viewpoint on the self-plagiarism. The self-plagiarism is plagiarism, as it fails to contribute to the research and science.

IJSP accepts both Online and Email submission. The online system makes readers to submit and track the status of their manuscripts conveniently. For any questions, please contact ijsp@ccsenet.org.

International Journal of Statistics and Probability; Vol. 11, No. 3; May 2022
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Editorial Team

International Journal of Statistics and Probability; Vol. 11, No. 3; May 2022
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Contents

International Journal of Statistics and Probability; Vol. 11, No. 3; May 2022
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Cutoff Value for Wilcoxon-Mann-Whitney Test by Minimum *P*-value: Application to COVID-19 Data

Toru Ogura[1] & Chihiro Shiraishi[2]

[1] Clinical Research Support Center, Mie University Hospital, Mie, Japan

[2] Department of Pharmacy, Mie University Hospital, Mie, Japan

Correspondence: Toru Ogura, Clinical Research Support Center, Mie University Hospital, Mie, Japan

Received: February 4, 2022 Accepted: March 4, 2022 Online Published: March 9, 2022

doi:10.5539/ijsp.v11n3p1 URL: https://doi.org/10.5539/ijsp.v11n3p1

Abstract

Dependent and independent variables may appear uncorrelated when analyzed in full range in medical data. However, when an independent variable is divided by the cutoff value, the dependent and independent variables may become correlated in each group. Furthermore, researchers often convert independent variables of quantitative data into binary data by cutoff value and perform statistical analysis with the data. Therefore, it is important to select the optimum cutoff value since performing statistical analysis depends on the cutoff value. Our study determines the optimal cutoff value when the data of dependent and independent variables are quantitative. The piecewise linear regression analysis divides an independent variable into two by the cutoff value, and linear regression analysis is performed in each group. However, the piecewise linear regression analysis may not obtain the optimal cutoff value when data follow a non-normal distribution. Unfortunately, medical data often follows a non-normal distribution. We, therefore, performed the Wilcoxon-Mann-Whitney (WMW) test with two-sided for all potential cutoff values and adopted the cutoff value that minimizes the *P*-value (called minimum *P*-value approach). Calculating the cutoff value using the minimum *P*-value approach is often used in the log-rank and chi-squared test but not the WMW test. First, using Monte Carlo simulations at various settings, we verified the performance of the cutoff value for the WMW test by the minimum *P*-value approach. Then, COVID-19 data were analyzed to demonstrate the practical applicability of the cutoff value.

Keywords: COVID-19 data, cutoff value, minimum *P*-value approach, non-normal distribution, quantitative data, Wilcoxon-Mann-Whitney test

1. Introduction

The clarified relationship between dependent and independent variables in the medical field can result in optimal patient treatment. These variables may appear uncorrelated when analyzed in the full range. However, when an independent variable is divided by the cutoff value, dependent and independent variables may become correlated in each group. Since the performance of the statistical analysis depends on the cutoff value, selecting the optimum cutoff value is important. The receiver operating characteristic curve is a recognized method for predicting the dependent variable of binary data from the independent variable of quantitative data (Greiner, Pfeiffer, & Smith, 2000; Zou, O'Malley, & Mauri, 2007). The linear regression analysis often predicts the outcome when both dependent and independent variables are quantitative data and show a linear relationship (Shiraishi, Matsuda, Ogura, & Iwamoto, 2021). Although medical data may not be a linear relationship when analyzed in a full range of independent variables, it may possess a linear relationship in each group when an independent variable is divided and grouped into two. The piecewise linear regression analysis is recognized method for predicting the outcome from such data (Nakamura, 1986; Vieth, 1989). However, when the data follow a non-normal distribution, the piecewise linear regression analysis may not obtain the optimal cutoff value.

First reported in Wuhan, China, COVID-19 patients have spread worldwide (World Health Organization, 2020, 2021). Many clinical trials are conducted to discover an effective treatment for COVID-19 patients (Capra et al, 2020; Hogan II et al., 2020; Aiswarya et al., 2021). Using Supplementary data of Hogan II et al. (2020), we investigated the relationship between the age and days to discharge in COVID-19 data. The data of the days to discharge were considered to follow a non-normal distribution. We then searched for the optimum cutoff value in this situation. We perform the Wilcoxon-Mann-Whitney (WMW) test with two-sided for all potential cutoff values and adopted the cutoff value that minimizes the *P*-value (called minimum *P*-value approach). Calculating the cutoff value using the minimum *P*-value approach showed excellent results in the log-rank and chi-squared tests (Altman, Lausen, Sauerbrei, & Schumacher, 1994; Mazumdar & Glassman, 2000; Liu et al., 2020) but not the WMW test. First, using Monte Carlo simulations (MCSs) at various settings, we verified the performance of the cutoff value for the WMW test by the minimum *P*-value approach. Then, COVID-19 data were analyzed to demonstrate the practical applicability of the cutoff value.

In Section 2, we described the cutoff value for the WMW test by the minimum P-value approach, while Section 3 verified the performance of the cutoff value using MCSs. Additionally, in Section 4, we presented an attempt to calculate the cutoff value using COVID-19 data and finally concluded conclude the research in Section 5.

2. Cutoff Value by Minimum P-Value Approach

Let $(\boldsymbol{x}, \boldsymbol{y}) = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ be two-dimensional random vectors of sample size $n \geq 2$, where $\boldsymbol{x}$ and $\boldsymbol{y}$ are independent and dependent variables, respectively. Let $x_{(i)}$ denote the i-th order statistics, $x_{(1)} \leq \cdots \leq x_{(n)}$. The potential cutoff value is written as $c_{(j)} = (x_{(j)} + x_{(j+1)})/2$, $j = 1, \ldots, n-1$. The data are divided into two groups: $\{(x_{(1)}, y_{(1)}), \ldots, (x_{(j)}, y_{(j)})\}$ and $\{(x_{(j+1)}, y_{(j+1)}), \ldots, (x_{(n)}, y_{(n)})\}$, depending on whether $x_{(i)} < c_{(j)}$ or $x_{(i)} \geq c_{(j)}$, where $y_{(i)}$ is the data paired with $x_{(i)}$ ($y_{(i)}$ is not the order statistic of y_i). We performed the WMW test between $\{y_{(1)}, \ldots, y_{(j)}\}$ and $\{y_{(j+1)}, \ldots, y_{(n)}\}$ in sequence from $j = 1$ to $n-1$, and the P-value was written as $\{P_{(1)}, \ldots, P_{(n-1)}\}$. The optimal cutoff value was $c = c_{(j)}^{\min}$ corresponding to $P_{(j)}^{\min} = \min(P_{(1)}, \ldots, P_{(n-1)})$. Since there is almost no advantage of dividing by the cutoff value when the sample size of one group is small, we used the cutoff value where each group has five or more patients in this manuscript.

3. MCSs

We verified the effectiveness of the cutoff value using MCSs. The population cutoff value was set to 50. In Patterns 1–9, $\{x_1, \ldots, x_n\}$ were generated from a normal distribution $\mathrm{N}(\mu, \sigma^2)$ and $\{y_1, \ldots, y_n\}$ were generated from a three-parameter gamma distribution $\mathrm{Ga}(\alpha, \beta, \gamma)$, where μ, σ^2, α, β, and γ are the mean, variance, shape, scale, and location parameters, respectively. In Patterns 10–18, both $\{x_1, \ldots, x_n\}$ and $\{y_1, \ldots, y_n\}$ were generated from $\mathrm{Ga}(\alpha, \beta, \gamma)$. Also, the parameters of $\mathrm{Ga}(\alpha, \beta, \gamma)$ where y_i were generated and differed depending on whether $x_i < 50$ or $x_i \geq 50$. Our simulation settings are summarized in Table 1. Although data are expected to be heavily biased in the cases of x_i generated from $\mathrm{N}(40, 10^2)$ and $\mathrm{N}(60, 10^2)$, it is necessary to have high estimation accuracy of the cutoff value even in such settings. The sample size is set to $n = 50, 100, 150$. We used the cutoff value where the sample size of one group is at least 5. The replication size used in this study is 1 000 000. We used the software R version 4.1.1 (R core team, 2021) for the MCSs. The MCS was conducted using the following procedure:

1. Generate random samples $\{x_1, \ldots, x_n\}$ from distribution in Table 1.
2. Generate random samples $\{y_1, \ldots, y_n\}$ from distribution in Table 1 (The distribution used depends on whether $x_i < 50$ or $x_i \geq 50$).
3. Combine $\{x_1, \ldots, x_n\}$ and $\{y_1, \ldots, y_n\}$ into two-dimensional random vectors $(\boldsymbol{x}, \boldsymbol{y}) = \{(x_1, y_1), \ldots, (x_n, y_n)\}$.
4. Sort $\{x_1, \ldots, x_n\}$ in ascending order, $x_{(1)} \leq \cdots \leq x_{(n)}$.
5. Set potential cutoff value $c_{(j)} = (x_{(j)} + x_{(j+1)})/2$, $j = 5, \ldots, (n-5)$.
6. Divide into two groups, $\{(x_{(1)}, y_{(1)}), \ldots, (x_{(j)}, y_{(j)})\}$ and $\{(x_{(j+1)}, y_{(j+1)}), \ldots, (x_{(n)}, y_{(n)})\}$, depending on whether $x_{(i)} < c_{(j)}$ or $x_{(i)} \geq c_{(j)}$.
7. Perform the WMW test between two groups for each $c_{(j)}$ and express the P-value as $P_{(j)}$.
8. Repeat steps 5–7 from $j = 5$ to $n-5$.
9. Decide optimal cutoff value $c = c_{(j)}^{\min}$ that satisfies $P_{(j)}^{\min} = \min(P_{(5)}, \ldots, P_{(n-5)})$.
10. Independently, repeat steps 1–9 1 000 000 times.
11. Calculate summary statistics and proportion of cutoff value in range.

Table 1. Distributions of generating random samples of x and y in MCSs

Pattern	x	y ($x_i < 50$)	y ($x_i \geq 50$)	Pattern	x	y ($x_i < 50$)	y ($x_i \geq 50$)
1	$\mathrm{N}(40, 10^2)$	Ga(1.5,10,10)	Ga(2.5,10,10)	10	Ga(1.5,10,30)	Ga(1.5,10,10)	Ga(2.5,10,10)
2	$\mathrm{N}(50, 10^2)$	Ga(1.5,10,10)	Ga(2.5,10,10)	11	Ga(1.5,10,35)	Ga(1.5,10,10)	Ga(2.5,10,10)
3	$\mathrm{N}(60, 10^2)$	Ga(1.5,10,10)	Ga(2.5,10,10)	12	Ga(1.5,10,40)	Ga(1.5,10,10)	Ga(2.5,10,10)
4	$\mathrm{N}(40, 10^2)$	Ga(1.5,10,10)	Ga(1.5,15,15)	13	Ga(1.5,10,30)	Ga(1.5,10,10)	Ga(1.5,15,15)
5	$\mathrm{N}(50, 10^2)$	Ga(1.5,10,10)	Ga(1.5,15,15)	14	Ga(1.5,10,35)	Ga(1.5,10,10)	Ga(1.5,15,15)
6	$\mathrm{N}(60, 10^2)$	Ga(1.5,10,10)	Ga(1.5,15,15)	15	Ga(1.5,10,40)	Ga(1.5,10,10)	Ga(1.5,15,15)
7	$\mathrm{N}(40, 10^2)$	Ga(1.5,10,10)	Ga(1.5,10,20)	16	Ga(1.5,10,30)	Ga(1.5,10,10)	Ga(1.5,10,20)
8	$\mathrm{N}(50, 10^2)$	Ga(1.5,10,10)	Ga(1.5,10,20)	17	Ga(1.5,10,35)	Ga(1.5,10,10)	Ga(1.5,10,20)
9	$\mathrm{N}(60, 10^2)$	Ga(1.5,10,10)	Ga(1.5,10,20)	18	Ga(1.5,10,40)	Ga(1.5,10,10)	Ga(1.5,10,20)

We use the cutoff values calculated by the Student's t-test and Welch's t-test for comparison in this manuscript. They were obtained by changing the WMW test in step 7 of the MCS procedure to the Student's t-test and Welch's t-test. Tables 2–4 show the summary statistics (mean, standard deviation (SD), first quartile (Q1), median, and third quartile (Q3)) for the

cutoff value and the proportion of the cutoff value that fall into five ranges ($49 \leq c \leq 51$, $48 \leq c \leq 52$, $47 \leq c \leq 53$, $46 \leq c \leq 54$, and $45 \leq c \leq 55$). Within the five ranges set, the proportion of cutoff value calculated by the WMW test was the highest of the three tests, except for Patterns 1, 3, 4, and 6 at $n = 50$.

Table 2. Summary of cutoff values in MCSs ($n = 50$)

		Summary statistics					Proportion of cutoff value in range				
Pattern		Mean	SD	Q1	Median	Q3	49-51	48-52	47-53	46-54	45-55
1	WMW	44.845	7.432	40.675	47.912	50.083	27.2%	41.6%	51.2%	57.9%	62.8%
	Student's	45.442	7.581	41.392	48.658	50.567	27.1%	42.3%	52.6%	59.7%	64.8%
	Welch's	41.310	8.553	33.518	43.026	49.221	17.7%	27.4%	34.3%	39.5%	43.6%
2	WMW	49.794	4.620	48.121	49.941	51.489	37.4%	54.5%	65.2%	72.6%	78.2%
	Student's	50.899	4.999	48.920	50.368	52.953	33.7%	50.0%	60.7%	68.4%	74.2%
	Welch's	47.817	5.682	44.235	49.085	50.658	29.2%	43.2%	52.5%	59.3%	64.9%
3	WMW	54.925	7.451	49.792	51.817	58.974	27.5%	42.3%	52.2%	59.0%	63.9%
	Student's	56.305	7.714	50.246	53.455	61.661	22.8%	35.5%	44.4%	50.9%	55.9%
	Welch's	54.461	7.446	49.684	51.540	57.377	27.0%	42.7%	53.8%	61.6%	67.1%
4	WMW	45.660	6.938	42.961	48.545	50.150	31.6%	47.2%	57.1%	63.8%	68.6%
	Student's	46.635	6.967	44.581	49.441	50.829	32.1%	48.9%	59.7%	67.0%	72.0%
	Welch's	41.484	8.477	33.773	43.533	49.255	19.1%	29.1%	35.9%	41.1%	45.1%
5	WMW	49.715	3.998	48.443	49.925	51.067	43.3%	61.2%	71.7%	78.5%	83.4%
	Student's	51.253	4.443	49.450	50.489	52.851	38.2%	55.4%	66.0%	73.3%	78.7%
	Welch's	47.502	5.224	44.367	48.983	50.347	32.9%	47.4%	56.5%	63.1%	68.3%
6	WMW	54.033	6.959	49.664	51.148	56.551	31.9%	48.0%	58.2%	65.0%	69.7%
	Student's	55.983	7.475	50.262	53.073	60.693	24.9%	38.1%	47.0%	53.5%	58.4%
	Welch's	53.586	6.926	49.587	51.057	54.989	30.9%	48.1%	59.7%	67.7%	73.2%
7	WMW	46.042	6.394	44.301	48.674	50.032	35.9%	51.9%	61.5%	67.9%	72.4%
	Student's	45.274	7.360	41.553	48.355	50.256	28.5%	43.4%	53.3%	60.1%	65.0%
	Welch's	42.919	8.309	35.783	46.550	49.758	26.3%	38.4%	45.8%	51.0%	54.7%
8	WMW	49.226	3.515	48.270	49.778	50.519	48.5%	66.4%	76.3%	82.5%	86.7%
	Student's	49.855	4.436	48.325	49.965	51.352	40.0%	57.3%	67.9%	75.0%	80.2%
	Welch's	48.061	5.026	45.638	49.431	50.513	36.3%	51.6%	60.8%	67.3%	72.3%
9	WMW	53.151	6.558	49.409	50.582	54.515	35.9%	53.1%	63.6%	70.4%	74.9%
	Student's	54.292	7.107	49.730	51.311	57.145	31.2%	46.9%	56.8%	63.6%	68.3%
	Welch's	54.455	7.371	49.718	51.621	57.176	26.9%	42.5%	53.4%	61.3%	67.0%
10	WMW	47.853	6.651	44.559	49.241	51.266	25.3%	39.6%	49.7%	57.4%	63.5%
	Student's	48.995	7.177	45.593	49.940	52.801	23.0%	36.4%	46.3%	54.1%	60.4%
	Welch's	44.571	7.750	37.131	45.950	50.137	17.8%	28.0%	35.4%	41.3%	46.2%
11	WMW	49.692	5.446	47.249	49.797	51.572	32.0%	48.1%	58.7%	66.5%	72.4%
	Student's	50.980	6.173	48.154	50.278	53.303	28.7%	43.7%	54.1%	61.9%	68.0%
	Welch's	47.355	6.241	42.257	48.248	50.495	24.0%	36.3%	44.9%	51.5%	57.0%
12	WMW	50.916	5.173	48.490	49.999	51.740	38.3%	55.7%	66.7%	74.4%	80.3%
	Student's	52.269	6.132	49.179	50.464	53.563	34.3%	50.7%	61.3%	68.9%	74.7%
	Welch's	49.663	5.502	46.228	49.423	50.894	30.8%	46.0%	56.5%	65.1%	73.0%
13	WMW	48.274	5.947	45.960	49.461	51.076	30.0%	45.7%	56.3%	64.0%	70.0%
	Student's	50.011	6.487	47.691	50.271	53.153	26.8%	41.6%	51.9%	60.0%	66.3%
	Welch's	44.450	7.393	37.374	46.098	49.995	19.9%	30.6%	38.1%	44.0%	48.8%
14	WMW	49.670	4.717	47.835	49.827	51.178	37.6%	54.9%	65.7%	73.2%	78.7%
	Student's	51.510	5.650	49.084	50.492	53.369	32.9%	49.0%	59.6%	67.2%	72.9%
	Welch's	47.018	5.666	42.426	48.171	50.235	26.9%	39.9%	48.6%	55.1%	60.4%
15	WMW	50.537	4.367	48.704	49.969	51.204	44.3%	62.4%	73.1%	80.2%	85.3%
	Student's	52.384	5.733	49.584	50.572	53.357	38.7%	55.7%	66.0%	73.0%	78.0%
	Welch's	49.154	4.682	46.290	49.345	50.519	34.9%	50.7%	61.3%	69.6%	77.1%
16	WMW	48.006	5.122	46.289	49.296	50.489	35.3%	52.1%	62.8%	70.3%	75.7%
	Student's	48.197	6.582	45.136	49.474	51.464	26.2%	40.7%	50.8%	58.5%	64.6%
	Welch's	45.615	7.008	39.802	48.082	50.201	26.5%	39.4%	47.8%	53.9%	58.5%
17	WMW	49.113	3.889	47.769	49.658	50.544	43.3%	61.2%	71.8%	78.7%	83.6%
	Student's	49.826	5.331	47.546	49.875	51.545	33.9%	50.2%	60.9%	68.5%	74.3%
	Welch's	47.683	5.365	44.148	49.042	50.381	32.6%	46.9%	55.9%	62.2%	67.1%
18	WMW	49.875	3.541	48.490	49.823	50.598	48.8%	67.2%	77.6%	84.3%	89.1%
	Student's	50.833	4.963	48.647	50.007	51.497	41.3%	59.1%	69.9%	77.3%	82.7%
	Welch's	49.556	4.752	46.901	49.617	50.753	36.4%	52.6%	63.1%	71.3%	78.3%

In Patterns 1, 3, 4, and 6 at $n = 50$, the data were biased due to the generation of $\boldsymbol{x}$ from $N(40, 10^2)$ or $N(60, 10^2)$. When the sample size increased to $n = 100$ and 150, the cutoff value calculated by the WMW test was the best even when the data were biased.

Table 3. Summary of cutoff values in MCSs ($n = 100$)

		Summary statistics					Proportion of cutoff value in range				
Pattern		Mean	SD	Q1	Median	Q3	49-51	48-52	47-53	46-54	45-55
1	WMW	47.507	6.239	47.012	49.582	50.486	41.3%	58.4%	68.5%	75.0%	79.5%
	Student's	48.237	6.634	47.747	49.996	51.415	37.0%	53.8%	64.5%	71.9%	77.5%
	Welch's	42.970	9.499	35.445	47.563	50.019	28.0%	40.1%	47.5%	52.8%	56.7%
2	WMW	49.865	3.067	49.112	49.972	50.705	56.0%	74.0%	82.9%	88.1%	91.4%
	Student's	50.781	3.912	49.493	50.203	51.566	49.8%	67.6%	77.0%	82.8%	86.7%
	Welch's	47.633	5.383	45.968	49.518	50.318	42.9%	57.8%	66.0%	71.3%	75.2%
3	WMW	52.248	6.276	49.293	50.284	52.705	40.5%	57.8%	68.1%	74.9%	79.5%
	Student's	53.763	7.331	49.724	50.884	55.153	35.4%	51.4%	61.1%	67.7%	72.3%
	Welch's	51.711	6.917	48.356	50.153	52.283	33.7%	50.7%	62.0%	70.4%	76.7%
4	WMW	48.194	5.121	47.946	49.695	50.381	48.5%	66.3%	75.9%	81.6%	85.4%
	Student's	49.365	5.327	49.012	50.165	51.577	43.0%	60.9%	71.6%	78.8%	83.9%
	Welch's	43.138	9.232	36.428	47.710	49.961	30.7%	42.9%	50.1%	55.0%	58.8%
5	WMW	49.839	2.327	49.287	49.965	50.501	63.6%	80.9%	88.7%	92.7%	95.1%
	Student's	50.969	3.289	49.755	50.258	51.451	56.0%	73.6%	82.3%	87.2%	90.3%
	Welch's	47.544	4.959	46.347	49.499	50.180	47.6%	62.3%	69.8%	74.6%	78.0%
6	WMW	51.528	5.259	49.339	50.163	51.784	46.8%	64.8%	74.8%	81.0%	85.1%
	Student's	53.455	6.787	49.858	50.830	54.197	39.9%	56.4%	65.8%	72.0%	76.2%
	Welch's	51.010	6.025	48.448	50.084	51.567	38.8%	56.8%	68.3%	76.4%	82.2%
7	WMW	48.398	4.071	48.209	49.627	50.137	56.4%	73.5%	81.8%	86.6%	89.5%
	Student's	47.957	6.112	47.522	49.783	50.722	41.9%	59.0%	69.1%	75.7%	80.4%
	Welch's	45.160	8.432	43.617	49.253	50.095	43.7%	57.5%	64.5%	68.7%	71.6%
8	WMW	49.592	1.901	49.221	49.901	50.251	69.2%	85.2%	91.7%	94.9%	96.7%
	Student's	49.932	3.067	49.202	49.990	50.658	58.5%	76.1%	84.4%	89.1%	92.0%
	Welch's	48.277	4.545	47.869	49.784	50.320	52.6%	68.0%	75.4%	79.8%	82.9%
9	WMW	50.832	4.590	49.109	50.016	51.100	50.3%	68.6%	78.5%	84.5%	88.4%
	Student's	51.711	5.879	49.219	50.134	51.831	45.1%	62.7%	72.6%	78.9%	83.1%
	Welch's	51.904	6.892	48.585	50.232	52.640	33.3%	50.2%	61.7%	70.2%	76.7%
10	WMW	49.205	5.021	47.863	49.794	50.932	40.0%	57.5%	68.1%	75.1%	80.1%
	Student's	50.613	6.198	48.633	50.215	52.472	34.7%	51.0%	61.3%	68.5%	73.9%
	Welch's	45.789	7.541	40.340	48.487	50.248	29.2%	42.3%	50.4%	56.1%	60.4%
11	WMW	49.892	3.755	48.764	49.928	50.815	49.4%	67.7%	77.6%	83.7%	87.8%
	Student's	51.181	5.255	49.276	50.227	51.982	43.3%	60.7%	70.9%	77.3%	81.9%
	Welch's	47.599	5.562	44.662	49.232	50.295	37.4%	51.9%	60.3%	66.1%	70.4%
12	WMW	50.318	3.394	49.215	49.990	50.757	56.8%	74.9%	83.7%	88.9%	92.1%
	Student's	51.494	5.155	49.551	50.213	51.648	50.6%	68.2%	77.4%	83.0%	86.6%
	Welch's	49.030	4.523	46.938	49.618	50.389	43.6%	59.4%	68.4%	74.8%	79.9%
13	WMW	49.326	3.941	48.362	49.820	50.650	47.1%	65.6%	75.9%	82.3%	86.6%
	Student's	51.235	5.303	49.379	50.376	52.479	40.1%	57.4%	67.8%	74.7%	79.5%
	Welch's	45.614	7.013	40.942	48.448	50.093	32.4%	45.8%	53.9%	59.3%	63.4%
14	WMW	49.807	2.830	49.011	49.926	50.564	56.9%	75.4%	84.4%	89.5%	92.7%
	Student's	51.413	4.646	49.651	50.311	51.889	49.1%	66.9%	76.5%	82.3%	86.1%
	Welch's	47.386	4.970	45.020	49.187	50.143	41.4%	56.0%	64.1%	69.4%	73.3%
15	WMW	50.112	2.458	49.364	49.983	50.537	64.4%	81.7%	89.3%	93.3%	95.6%
	Student's	51.490	4.632	49.781	50.263	51.506	56.7%	73.9%	82.1%	86.7%	89.7%
	Welch's	48.776	3.763	47.213	49.611	50.235	48.9%	64.4%	72.9%	78.7%	83.3%
16	WMW	49.035	2.990	48.405	49.708	50.260	54.8%	73.1%	82.4%	87.8%	91.2%
	Student's	49.589	5.104	48.205	49.919	51.114	41.2%	58.7%	69.1%	75.9%	80.7%
	Welch's	47.101	6.163	46.135	49.441	50.227	43.9%	59.2%	67.3%	72.2%	75.5%
17	WMW	49.491	2.113	48.975	49.844	50.261	63.7%	81.1%	89.0%	93.1%	95.5%
	Student's	50.076	3.929	48.927	49.978	50.844	51.2%	69.4%	79.0%	84.7%	88.5%
	Welch's	48.276	4.432	47.515	49.681	50.275	50.1%	65.6%	73.4%	78.0%	81.1%
18	WMW	49.790	1.821	49.291	49.920	50.286	69.5%	85.7%	92.3%	95.6%	97.4%
	Student's	50.332	3.541	49.284	49.997	50.665	59.7%	77.3%	85.6%	90.2%	93.0%
	Welch's	49.242	3.743	48.063	49.825	50.418	51.1%	67.2%	75.6%	81.2%	85.4%

As n increases, the proportion of the cutoff value calculated by the WMW test in each range increases. In the range of $45 \leq c \leq 55$, the proportion of cutoff value calculated by the WMW test was greater than 90% in many patterns at $n = 150$.

Table 4. Summary of cutoff values in MCSs ($n = 150$)

		Summary statistics					Proportion of cutoff value in range				
Pattern		Mean	SD	Q1	Median	Q3	49-51	48-52	47-53	46-54	45-55
1	WMW	48.692	4.732	48.514	49.828	50.458	52.6%	70.2%	79.2%	84.5%	88.0%
	Student's	49.394	5.291	49.032	50.093	51.310	46.4%	63.9%	73.7%	79.9%	84.4%
	Welch's	44.404	9.233	41.138	48.949	50.101	38.1%	51.4%	58.7%	63.4%	66.8%
2	WMW	49.909	2.056	49.432	49.983	50.455	68.0%	84.3%	91.1%	94.5%	96.4%
	Student's	50.557	2.899	49.671	50.128	50.975	61.8%	78.7%	86.5%	90.7%	93.2%
	Welch's	48.095	4.710	47.819	49.762	50.243	54.6%	69.3%	76.4%	80.6%	83.5%
3	WMW	51.106	4.829	49.352	50.095	51.321	51.3%	69.1%	78.3%	83.9%	87.6%
	Student's	52.407	6.238	49.681	50.402	52.517	45.9%	62.9%	72.2%	78.0%	81.8%
	Welch's	50.530	5.972	48.164	50.009	51.238	40.3%	57.5%	67.9%	75.2%	80.6%
4	WMW	49.116	3.501	48.897	49.854	50.318	60.8%	78.1%	86.0%	90.2%	92.7%
	Student's	50.170	3.892	49.573	50.195	51.338	53.1%	70.9%	80.1%	85.7%	89.4%
	Welch's	44.596	8.835	42.322	48.959	50.031	41.4%	54.6%	61.4%	65.8%	69.0%
5	WMW	49.893	1.477	49.540	49.978	50.323	75.4%	89.8%	95.0%	97.2%	98.4%
	Student's	50.631	2.361	49.833	50.158	50.891	68.3%	84.2%	90.6%	93.8%	95.6%
	Welch's	48.133	4.310	48.085	49.749	50.143	59.7%	73.5%	79.7%	83.3%	85.8%
6	WMW	50.621	3.667	49.441	50.053	50.915	58.4%	76.1%	84.6%	89.3%	92.1%
	Student's	52.147	5.542	49.829	50.402	52.089	51.7%	68.8%	77.5%	82.5%	85.8%
	Welch's	50.064	4.991	48.434	50.002	50.906	46.3%	64.2%	74.3%	80.9%	85.6%
7	WMW	49.167	2.486	48.984	49.795	50.112	69.2%	84.5%	90.7%	93.8%	95.6%
	Student's	49.039	4.587	48.792	49.934	50.614	53.6%	71.0%	79.8%	85.0%	88.4%
	Welch's	46.692	7.506	47.913	49.683	50.114	57.7%	71.2%	76.8%	80.0%	81.9%
8	WMW	49.739	1.203	49.502	49.937	50.168	80.3%	92.6%	96.6%	98.2%	99.0%
	Student's	49.933	2.040	49.489	49.993	50.417	70.7%	86.2%	92.3%	95.2%	96.8%
	Welch's	48.822	3.743	48.974	49.909	50.260	65.3%	79.3%	85.1%	88.2%	90.1%
9	WMW	50.176	3.117	49.279	49.977	50.580	61.4%	78.8%	86.9%	91.2%	93.9%
	Student's	50.777	4.533	49.318	50.035	50.919	55.9%	73.5%	82.0%	86.9%	90.0%
	Welch's	50.827	5.999	48.507	50.073	51.547	39.7%	57.1%	67.9%	75.4%	81.0%
10	WMW	49.608	3.629	48.743	49.896	50.670	51.2%	69.7%	79.5%	85.3%	89.0%
	Student's	50.846	5.064	49.260	50.194	51.765	44.7%	62.5%	72.5%	78.8%	83.2%
	Welch's	46.765	6.628	44.903	49.281	50.235	39.8%	54.7%	63.0%	68.3%	72.0%
11	WMW	49.918	2.565	49.223	49.960	50.541	61.2%	78.9%	87.2%	91.7%	94.3%
	Student's	50.901	4.138	49.556	50.159	51.283	54.6%	72.3%	81.3%	86.4%	89.7%
	Welch's	48.103	4.620	47.138	49.624	50.241	49.0%	64.3%	72.1%	76.9%	80.3%
12	WMW	50.113	2.184	49.488	49.992	50.476	68.9%	85.0%	91.6%	94.9%	96.7%
	Student's	50.925	3.921	49.700	50.129	50.991	62.6%	79.2%	86.7%	90.6%	93.0%
	Welch's	49.050	3.543	48.107	49.805	50.284	54.9%	70.2%	78.0%	82.9%	86.6%
13	WMW	49.626	2.654	49.004	49.899	50.448	59.2%	77.4%	86.1%	90.9%	93.7%
	Student's	51.139	4.265	49.650	50.274	51.670	51.0%	69.0%	78.5%	84.1%	87.8%
	Welch's	46.667	6.101	45.258	49.239	50.105	43.8%	58.4%	66.1%	71.0%	74.5%
14	WMW	49.865	1.830	49.374	49.955	50.373	69.1%	85.6%	92.3%	95.5%	97.2%
	Student's	50.961	3.541	49.776	50.202	51.174	61.0%	78.3%	86.2%	90.4%	92.9%
	Welch's	48.038	4.135	47.407	49.599	50.127	53.7%	68.4%	75.5%	79.9%	82.8%
15	WMW	50.013	1.501	49.585	49.988	50.343	76.1%	90.3%	95.3%	97.5%	98.6%
	Student's	50.878	3.370	49.847	50.159	50.898	68.9%	84.3%	90.4%	93.4%	95.1%
	Welch's	48.974	3.005	48.368	49.805	50.183	60.7%	75.1%	81.9%	86.1%	89.2%
16	WMW	49.401	1.932	49.003	49.821	50.180	66.9%	83.7%	90.8%	94.4%	96.4%
	Student's	49.857	3.760	48.947	49.970	50.763	52.8%	71.0%	80.5%	86.0%	89.6%
	Welch's	48.158	4.906	48.360	49.748	50.200	57.8%	73.2%	80.1%	83.8%	86.1%
17	WMW	49.660	1.368	49.341	49.901	50.175	75.3%	89.8%	95.0%	97.4%	98.5%
	Student's	50.010	2.716	49.323	49.988	50.541	63.5%	80.8%	88.5%	92.5%	94.9%
	Welch's	48.882	3.425	48.852	49.859	50.227	63.7%	78.4%	84.5%	87.7%	89.7%
18	WMW	49.827	1.131	49.536	49.948	50.186	80.4%	93.0%	97.0%	98.6%	99.3%
	Student's	50.104	2.357	49.527	49.996	50.413	71.8%	87.0%	93.0%	95.8%	97.3%
	Welch's	49.380	2.890	48.952	49.925	50.324	62.8%	77.9%	84.6%	88.6%	91.3%

4. COVID-19 Data

We demonstrated the cutoff value calculated by the WMW test using COVID-19 data. We utilized the clinical outcomes data in 110 hospitalized COVID-19 patients treated with famotidine and cetirizine for a minimum of 48 h (Hogan II et al., 2020), as shown in Table 5. The data are presented by Supplementary data of their paper. The dosage and administration route were famotidine 20 mg intravenously (IV) and cetirizine 10 mg IV (or oral) at 12 h intervals. The duration of the clinical trials was from April 3, 2020, to June 13, 2020. Recently, it was revealed that cetirizine (Histamine-1 blocker) (Freedberg, et al., 2020; Janowitz et al., 2020) and famotidine (Histamine-2 blockers) (Blanco et al., 2021) showed a significant effect as an anti-SARS-CoV-2 which is the name of the pathogen that causes COVID-19.

Table 5. Clinical outcomes in 110 hospitalized COVID-19 patients (x: age (years old), y: days to discharge (day))

x	79	53	34	64	78	50	83	71	85	91	73	65	81	57	93	79	71	59	50	43
y	5	6	2	32	18	5	11	4	5	33	35	14	18	8	12	8	9	4	5	7
x	80	58	39	46	41	60	68	89	83	39	72	45	63	87	43	92	22	92	64	72
y	20	29	7	8	6	7	11	-	-	18	16	15	11	-	7	12	10	11	10	21
x	92	72	51	81	56	74	64	58	57	70	17	38	81	69	51	51	80	61	80	25
y	6	5	11	20	5	6	8	6	13	7	7	-	6	42	9	11	4	25	11	10
x	63	89	76	24	71	69	97	27	71	76	66	60	79	84	63	49	94	79	68	63
y	-	-	-	7	10	19	-	6	9	5	9	-	7	7	-	6	17	5	30	13
x	69	91	79	61	48	33	76	50	37	21	53	73	56	67	45	73	75	73	43	55
y	-	-	-	-	7	15	19	4	3	4	-	13	8	5	11	8	8	5	12	-
x	68	63	48	38	70	60	73	57	75	72										
y	16	8	-	6	5	13	14	7	4	8										

-: Died without recovery.

The independent variable x is the age (years old), and the dependent variable y is the days to discharge (day). Patients whose dependent variable was listed as hyphens in Table 5 died without recovery. In this manuscript, we used 93 patients that have recovered and were discharged. We also used the software R to calculate the cutoff value by the WMW test, and the sample code was presented in Appendix. Figure 1 is a scatter plot of the age and days to discharge, and the dashed line shows the cutoff value calculated by the WMW test. The days to discharge of all young patients were short. On the other hand, the days to discharge of many elderly patients were short, but the days to discharge of some elderly patients were long. Therefore, the scatter plot looked like a lower right triangle. There was no linear relationship between x and y, and y that followed a non-normal distribution. The cutoff value calculated by the WMW test was 59.5 years old, and the P-value using that cutoff value was 0.011. Since we set the potential cutoff value as $c_{(j)} = (x_{(j)} + x_{(j+1)})/2$ to accommodate a variety of quantitative data, the cutoff value was output as 59.5 years old. Because the age data were in 1-year increments, the two groups of less than 59.5 years old and greater than or equal to 59.5 years old were the same as the two groups of less than 60 years old and greater than or equal to 60 years old. Considering the scatter plot, the patient of $(x, y) = (58, 29)$ may seem better in the right-hand group. However, if the cutoff value was 57.5 years old, the patients of $(x, y) = (58, 6)$ and (59,4) would also move to the right-hand group. Additionally, since even a large value for only one patient has a little effect on the WMW test, it is believed that 59.5 years old was selected as the cutoff value.

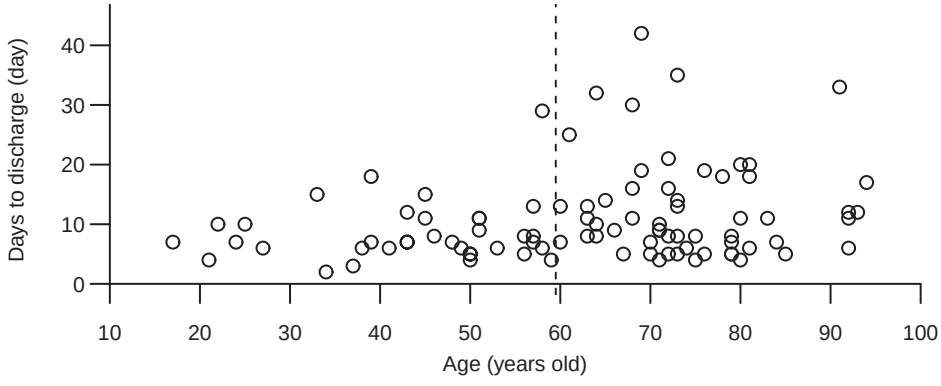

Figure 1. Scatter plot of the age and days to discharge in COVID-19 data of 93 recovered patients. The dashed line shows the cutoff value

Reznikov et al. (2021) identified antihistamine candidates for repurposing by mining electronic health records of usage in a population of more than 219 000 subjects tested for SARS-CoV-2. They concluded that prior usage of loratadine, diphenhydramine, cetirizine, hydroxyzine, and azelastine was associated with a reduced incidence of positive SARS-CoV-2 test results in the group of greater than or equal to 61 years old. It is believed that the cutoff value calculated by the WMW test obtained a good result, because there was a report that the cutoff value set at age 61 years old provided beneficial effects.

5. Conclusions

This study divides the COVID-19 data, which followed a non-normal distribution, into two cutoff values. In the log-rank and chi-squared tests, the method of calculating the cutoff value by the minimum *P*-value approach was well established. However, because there was no application of cutoff value for the WMW test by the minimum *P*-value approach, we verified the performance when the method was applied to the WMW test using MCSs at various settings. The MCS results at various settings showed high performance of the cutoff value calculated by the WMW test. Furthermore, in COVID-19 data, when the data were divided into two groups with the cutoff value calculated by the WMW test, it was confirmed that they were split into two groups with different characteristics. Therefore, we concluded that the cutoff value for the WMW test by the minimum *P*-value approach is valid.

References

Aiswarya, D., Arumugam, V., Dineshkumar, T., Gopalakrishnan, N., Lamech, T. M., Nithya, G., ... Sakthirajan, R. (2021). Use of remdesivir in patients with COVID-19 on hemodialysis: a study of safety and tolerance. *Kidney international reports, 6*(3), 586-593. https://doi.org/10.1016/j.ekir.2020.12.003

Altman, D. G., Lausen, B., Sauerbrei, W., & Schumacher, M. (1994). Dangers of using "optimal" cutpoints in the evaluation of prognostic factors. *Journal of the National Cancer Institute, 86*(11), 829-835. https://doi.org/10.1093/jnci/86.11.829

Blanco, J. I. M., Bonilla, J. A. A., Homma, S., Suzuki, K., Fremont-Smith, P., & de Las Heras, K. V. G. (2021). Antihistamines and azithromycin as a treatment for COVID-19 on primary health care – A retrospective observational study in elderly patients. *Pulmonary pharmacology & therapeutics, 67*, 101989. https://doi.org/10.1016/j.pupt.2021.101989

Capra, R., De Rossi, N., Mattioli, F., Romanelli, G., Scarpazza, C., Sormani, M. P., & Cossi, S. (2020). Impact of low dose tocilizumab on mortality rate in patients with COVID-19 related pneumonia. *European journal of internal medicine, 76*, 31-35. https://doi.org/10.1016/j.ejim.2020.05.009

Freedberg, D. E., Conigliaro, J., Wang, T. C., Tracey, K. J., Callahan, M. V., Abrams, J. A., ... Landry, D. W. (2020). Famotidine use is associated with improved clinical outcomes in hospitalized COVID-19 patients: a propensity score matched retrospective cohort study. *Gastroenterology, 159*(3), 1129-1131. https://doi.org/10.1053/j.gastro.2020.05.053

Greiner, M., Pfeiffer, D., & Smith, R. D. (2000). Principles and practical application of the receiver-operating characteristic analysis for diagnostic tests. *Preventive veterinary medicine, 45*(1-2), 23-41. https://doi.org/10.1016/S0167-5877(00)00115-X

Hogan II, R. B., Hogan III, R. B., Cannon, T., Rappai, M., Studdard, J., Paul, D., & Dooley, T. P. (2020). Dual-histamine receptor blockade with cetirizine-famotidine reduces pulmonary symptoms in COVID-19 patients. *Pulmonary pharmacology & therapeutics, 63*, 101942. https://doi.org/10.1016/j.pupt.2020.101942

Janowitz, T., Gablenz, E., Pattinson, D., Wang, T. C., Conigliaro, J., Tracey, K., & Tuveson, D. (2020). Famotidine use and quantitative symptom tracking for COVID-19 in non-hospitalised patients: a case series. *Gut, 69*(9), 1592-1597. http://dx.doi.org/10.1136/gutjnl-2020-321852

Liu, H., Yang, Y., Chen, C., Wang, L., Huang, Q., Zeng, J., ... Liu, J. (2020). Reclassification of tumor size for solitary HBV-related hepatocellular carcinoma by minimum *p* value method: a large retrospective study. *World journal of surgical oncology, 18*(1), 1-10. https://doi.org/10.1186/s12957-020-01963-z

Mazumdar, M., & Glassman, J. R. (2000). Categorizing a prognostic variable: review of methods, code for easy implementation and applications to decision]making about cancer treatments. *Statistics in medicine, 19*(1), 113-132. https://doi.org/10.1002/(SICI)1097-0258(20000115)19:1<113::AID-SIM245>3.0.CO;2-O

Nakamura, T. (1986). BMDP program for piecewise linear regression. *Computer Methods and Programs in Biomedicine, 23*(1), 53-55. https://doi.org/10.1016/0169-2607(86)90080-5

R Core Team. (2021). *R: A language and environment for statistical computing* (Version 4.1.1). Vienna, Austria: R

Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Reznikov, L. R., Norris, M. H., Vashisht, R., Bluhm, A. P., Li, D., Liao, Y. S. J., ... Ostrov, D. A. (2021). Identification of antiviral antihistamines for COVID-19 repurposing. *Biochemical and biophysical research communications, 538*, 173-179. https://doi.org/10.1016/j.bbrc.2020.11.095

Shiraishi, C., Matsuda, H., Ogura, T., & Iwamoto, T. (2021). Factors affecting serum phenobarbital concentration changes in pediatric patients receiving elixir and powder formulations. *Journal of Pharmaceutical Health Care and Sciences, 7*(1), Article number: 7. https://doi.org/10.1186/s40780-021-00190-2

Vieth, E. (1989). Fitting piecewise linear regression functions to biological responses. *Journal of applied physiology, 67*(1), 390-396. https://doi.org/10.1152/jappl.1989.67.1.390

World Health Organization. (2020). Coronavirus disease 2019 (COVID-19): situation report, 94.

World Health Organization. (2021). *Coronavirus disease (COVID-19)*. Retrieved from https://www.who.int/

Zou, K. H., O'Malley, A. J., & Mauri, L. (2007). Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. *Circulation, 115*(5), 654-657. https://doi.org/10.1161/CIRCULATIONAHA.105.594929

Appendix

Sample code of the software R

We presented a sample code of the software R for calculating the cutoff value using COVID-19 data. Another practical example can be calculated by replacing two vectors of x and y with suitable ones.

```
library(exactRankTests)
x<-c(79,53,34,64,78,50,83,71,85,91,73,65,81,57,93,79,71,59,50,43,80,58,39,46,41,60,68,
     39,72,45,63,43,92,22,92,64,72,92,72,51,81,56,74,64,58,57,70,17,81,69,51,51,80,61,
     80,25,24,71,69,27,71,76,66,79,84,49,94,79,68,63,48,33,76,50,37,21,73,56,67,45,73,
     75,73,43,68,63,38,70,60,73,57,75,72)
y<-c(5,6,2,32,18,5,11,4,5,33,35,14,18,8,12,8,9,4,5,7,20,29,7,8,6,7,11,18,16,15,11,7,
     12,10,11,10,21,6,5,11,20,5,6,8,6,13,7,7,6,42,9,11,4,25,11,10,7,10,19,6,9,5,9,7,
     7,6,17,5,30,13,7,15,19,4,3,4,13,8,5,11,8,8,5,12,16,8,6,5,13,14,7,4,8)
n<-length(x); dat0<-data.frame(x,y); dat1<-dat0[order(dat0[,1]),]; res<-NULL
for(j in 5:(n-5)){cj<-(dat1$x[j]+dat1$x[j+1])/2; y1<-y[x<cj]; y2<-y[x>=cj]
res<-rbind(res,c(Cutoff=cj,Pvalue=wilcox.exact(y1,y2)$p))}; res[order(res[,2]),][1,]
```

International Journal of Statistics and Probability; Vol. 11, No. 3; May 2022
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Bayesian Bivariate Cure Rate Models Using Copula Functions

Jie Huang[1] & Haiming Zhou[1] Nader Ebrahimi[1]

[1] Division of Statistics, Northern Illinois University, DeKalb, IL, USA

Correspondence: Jie Huang, Division of Statistics, Northern Illinois University, DeKalb, IL, USA

Received: January 10, 2022 Accepted: March 2, 2022 Online Published: March 9, 2022

doi:10.5539/ijsp.v11n3p9 URL: https://doi.org/10.5539/ijsp.v11n3p9

Abstract

Bivariate survival cure rate models extend the understanding of time-to-event data by allowing for a cured fraction of the population and dependence between paired units and make more accurate and informative conclusions. In this paper, we propose a Bayesian bivariate cure rate mode where a correlation coefficient is used for the association between bivariate cure rate fractions and a new generalized Farlie Gumbel Morgenstern (FGM) copula function is applied to model the dependence structure of bivariate survival times. For each marginal survival time, we apply a Weibull distribution, a log normal distribution, and a flexible three-parameter generalized extreme value (GEV) distribution to compare their performance. For the survival model fitting, DIC and LPML are used for model comparison. We perform a goodness-of-fit test for the new copula. Finally, we illustrate the performance of the proposed methods in simulated data and real data via Bayesian paradigm.

Keywords: bivariate cure rate, Copulas, Goodness-of-fit, Bayesian approach, survival analysis

1. Introduction

In survival analysis, it is of primary interest to measure the association between two time-to-event random variables associated with one individual. In many cases, the lifetime of paired unites from same individual would affect each other. For instance, visual loss for one eye could affect another eye for the same patient. Another case in cancer study, for example, a fraction of patents may response positively to the treatment. On the other word, this fraction of patients will not experience death within the follow-up period, and they have long term survival times. In the literature, frailty models with a single shared frailty were popular. They account for unobserved heterogeneity by including random effect. The main feature of the shared frailty models is all units share the same frailty. Because of this limitation, they have been extended to model data with more complex dependence relations. Yashin and Iachine (1999) involved correlated stochastic hazard in a given frailty of survival distribution. Peng and Taylor (2011) applied different random effects to model the correlations for cure patients and uncured patients, respectively. Gallardo, Gómez, and de Castro (2018) proposed a cure rate model and applied the competing risks approach to the latent causes of the event of interest.

An alternative is the use of copula functions. Unlike the frailty approaches, the copula approach models the joint distribution by connecting the two marginal distributions through a copula function. Modelling dependence is one of the primary interests in multivariate analysis. The advantages of the copula are as follows. First, the copula models the marginal distributions and the dependence parameters separately which allows flexibility in marginal models and straightforward construction of covariate effects. Secondly, the copula can handle the censoring through the marginal distributions. Thirdly, the conditional distributions can be obtained through the copula model. Romeo, Tanaka, and Pedroso-de Lima (2006) introduced the Archimedian copula family for modeling the dependence of bivariate lifetime components where a Weibull distribution is considered as the marginal distribution. Louzada, Suzuki, and Cancho (2013) proposed an FGM long-term bivariate survival copula model. They assumed a mixture cure rate model for the marginal distribution of each lifetime and assumed fixed cure fraction for the entire population. C.-M. Chen, Lu, and Hsu (2013) applied the pairwise odds ratio to the association of the insusceptibility of the individuals and adopted Clayton copula to measure the association of susceptible individuals with a semiparametric distribution as marginal regression model. Lakhal-Chaieb and Duchesne (2017) introduced a link function to relax subject-specific-effect assumption which improves the range of potential association and add flexibility to dependence structure.

The literature has introduced many other modelling approaches for bivariate long term data using copula functions, as for example the paper introduced by Louzada et al. (2013). But in that paper, the authors only present more simple cure fraction survival model situation assuming dependence with a FGM copula function structure. In this paper, a mixture model is applied to analyze a bivariate censored data with different susceptibilities. A correlation coefficient is applied for cure rates and a generalized bivariate Farlie Gumbel Morgenstern (FGM) copula, proposed by Bekrizadeh and Jamshidi (2017), is applied for the association of bivariate failure times. As studied in Louzada et al. (2013), it showed a weak

dependence between tow lifetimes in the diabetic retinopathy study data. The FGM commonly used to model very weak linear dependences and give more reliable estimates. One of the advantages of the generalized FGM copula has wider range of correlation compared to regular FGM copula. According to Bekrizadeh and Jamshidi (2017), the estimated correlations based on the generalized FGM copula is closer to the actual correlation of the observed data compared to FGM copula. This conclusion is also shown in diabetic retinopathy study data. Also, the Spearmans ρ of two lifetimes in the diabetic retinopathy study data is 0.376 which is greater than the upper bound of ρ for the FGM copula. Whereas the generalized FGM copula has one more parameter with upper bound of ρ up to 0.385. In Bayesian analysis of proposed model, we perform standard MCMC method in consideration of cure fractions and censoring for both lifetimes. We employ the flexible three-parameter generalized extreme value (GEV) distribution on the marginal survival time. We also apply Weibull distribution and log normal distribution for model comparison.

Misspecifying the copula model may have impact on the inference procedure. Therefore, it may be necessary for our proposed methodology to use a goodness-of-fit test for adequacy check. In the literature, specification tests have been extensively investigated such as rank based test as in Wang and Wells (2000), kernels as in Fermanian (2005), and blanket tests. Genest, Rémillard, and Beaudoin (2009) show that all of blank tests have no power in differentiating some copulas such as Gaussian copula and Student's t copula. Also, it's difficult for deriving analytically in the test statistics of Student's t copula and vine copula since blank tests require certain probability integral transformation. S. Zhang, Okhrin, Zhou, and Song (2016) propose an alternative specification test for semi-parametric copulas which does not require any probability integral transformation. The proposed test is referred to as pseudo-in-and-out-sample test (PIOS) which takes a form of ratio constructed via in sample pseudo-likelihood and out of sample pseudo-likelihood.

Compared to the POIS test proposed by S. Zhang et al. (2016), our work makes the following new contributions. First of all, the test is extended to be applicable to a parametric copula model of right-censored survival times. Secondly, the test is extended to the case of mixture cure rate model for individual survival function. Finally, we discuss how to identify the susceptible subjects in the mixture cure rate model in order to produce a PIOS test statistic.

2. The Model

Suppose two lifetimes T_1 and T_2 associated to the same subject. Let d_k be an indicator variable showing a subject is susceptible to the kth event and a corresponding cure probability is $P_k = Pr(d_k = 0)$, $k = 1, 2$. We assume mixture models for T_1 and T_2, given, respectively, by

$$S_1(t_1) = Pr(T_1 > t_1) = P_1 + (1 - P_1)S_{10}(t_1)$$
$$S_2(t_2) = Pr(T_2 > t_2) = P_2 + (1 - P_2)S_{01}(t_2),$$

where $S_{10}(t_1) = Pr(T_1 > t_1|d_1 = 1)$ and $S_{01}(t_2) = Pr(T_2 > t_2|d_2 = 1)$ are survival functions associated with T_1 and T_2 when the subject is susceptible for the underlying event.

The joint survival function for T_1 and T_2 is given by

$$S(t_1, t_2) = \sum_{d_1,d_2} S(t_1, t_2|d_1, d_2)Pr(d_1, d_2). \tag{1}$$

where $S(t_1, t_2|d_1 = 1, d_2 = 1)$, for example, is the joint survival function of T_1 and T_2 for the susceptible individuals.

Assuming that covariance between d_1 and d_2 is ρ, we have

$$Pr(d_1 = 1, d_2 = 1) = (1 - P_1)(1 - P_2) + \rho \triangleq \varphi_{11}, \tag{2}$$
$$Pr(d_1 = 1, d_2 = 0) = (1 - P_1)P_2 - \rho \triangleq \varphi_{10}, \tag{3}$$
$$Pr(d_1 = 0, d_2 = 1) = (1 - P_2)P_1 - \rho \triangleq \varphi_{01}, \tag{4}$$
$$Pr(d_1 = 0, d_2 = 0) = P_1P_2 + \rho \triangleq \varphi_{00}. \tag{5}$$

Now, using (1) - (5) we get

$$S(t_1, t_2) = \varphi_{11}S(t_1, t_2|d_1 = 1, d_2 = 1) + \varphi_{10}S_{10}(t_1) + \varphi_{01}S_{01}(t_2) + \varphi_{00},$$

where $S(t_1, t_2|d_1 = 1, d_2 = 1)$, S_{10} and S_{01} are defined as above. Now one possibility is to use different parametric distributions for $S(t_1, t_2|d_1 = 1, d_2 = 1)$. Another possibility is to use copula functions which link marginal distributions with a joint distribution. Throughout this paper, we use copula functions for joint survival of susceptible individuals. Thus, the joint survival function for T_1 and T_2 can be written as,

$$S(t_1, t_2) = \varphi_{00} + \varphi_{10}S_{10}(t_1) + \varphi_{01}S_{01}(t_2) + \varphi_{11}C(S_{10}(t_1), S_{01}(t_2)), \tag{6}$$

where $C(\cdot, \cdot)$ is a bivariate copula function.

2.1 Distributional Assumptions on $S_{01}(\cdot)$ and $S_{10}(\cdot)$

Let $\{S_{\boldsymbol{\gamma}}(\cdot)\}$ denote a parametric family of survival functions with support on $\mathbb{R}^+$, where $\boldsymbol{\gamma}$ is a vector of parameters. In this paper, we consider three families: Weibull, log normal and log generalized extreme value (GEV). The Weibull distribution has survival function

$$S_{\boldsymbol{\gamma}}(t) = \exp\Big[-\Big(\frac{t}{\mu}\Big)^{\lambda}\Big], \text{ where } \boldsymbol{\gamma} = (\mu, \lambda) \in \mathbb{R}^+ \times \mathbb{R}^+. \tag{7}$$

The log normal has survival function

$$S_{\boldsymbol{\gamma}}(t) = 1 - \Phi\left(\frac{\log t - \mu}{\sigma}\right), \text{ where } \boldsymbol{\gamma} = (\mu, \sigma) \in \mathbb{R} \times \mathbb{R}^+. \tag{8}$$

The log GEV has survival function

$$S_{\boldsymbol{\gamma}}(t) = \begin{cases} 1 - \exp\Big\{-\Big(1 + \xi\frac{\log t - \mu}{\sigma}\Big)_+^{-\frac{1}{\xi}}\Big\} & \text{if } \xi \neq 0 \\ 1 - \exp\Big\{-\exp\Big(-\frac{\log t - \mu}{\sigma}\Big)\Big\} & \text{if } \xi = 0, \end{cases} \tag{9}$$

where $\boldsymbol{\gamma} = (\mu, \sigma, \xi) \in \mathbb{R} \times \mathbb{R}^+ \times \mathbb{R}$ and $x_+ = \max(0, x)$. Now, given each of the above distribution families, we assume $S_{01}(t_1) = S_{\boldsymbol{\gamma}_1}(t_1)$ and $S_{10}(t_2) = S_{\boldsymbol{\gamma}_2}(t_2)$.

The Weibull distribution can produce only monotonic hazard rates. In contrast, the shape of the hazard function for a logGEV is various such as U-shaped, or bell shaped, or a combination of both.

2.2 A Generalized FGM Bivariate Copula

In this paper, we use the generalized class of Farlie-Gumbel-Morgenstern (FGM) copula proposed by Bekrizadeh and Jamshidi (2017) which is given by

$$C_{\theta}^{p}(s,t) = st[1 + \theta(1-s)(1-t)]^p, p \in [1, \infty], \ \theta \in [-p^{-1}, p^{-1}], \ \forall (s,t) \in [0,1]^2. \tag{10}$$

When $p = 1$, it reduces to the symmetric FGM copula. The Spearman's ρ can be written as $\rho = 12\sum_{k=1}^{p}\binom{p}{k}\theta^k\Big[\frac{1}{(k+1)(k+2)}\Big]^2$, where the upper bound of ρ is up to 0.3805 approximately, and the lower bound is equal to -0.3333 which is same as that of symmetric FGM. Thus, the range of ρ in this generalized FGM is $[-0.3333, 0.3805]$. A good example is when $p = 3$, $\theta = 0.33 < \frac{1}{3}$, the estimated Spearmans ρ, $\rho \approx 0.3583$, which is out of the range of Spearmans ρ for FGM copula, that is $[-1/3, 1/3]$. The Kendall's τ can be written as $\tau = 4\int_0^1\int_0^1 c(s,t)C(s,t)dsdt - 1$. The estimated Kendals $\tau \approx 0.2397$ which is out of the range of τ for FGM copula, that is $[-2/9, 2/9]$. As we can see that the generalized FGM improve the correlation range which can be applied for more data. Also, the Spearmans ρ increases as p increases and θ is fixed. And Spearmans ρ increases as θ increases and p is fixed. As discussed in Bekrizadeh and Jamshidi (2017), if true correlation is within the range of FGM, the estimated correlations based on the generalized FGM copula showed strong consistency and was closer to the correlations which come from the observed data compared to the regular FGM copula.

2.3 Likelihood and MCMC

Denote (T_{i1}, T_{i2}) and (C_{i1}, C_{i2}) as bivariate lifetimes and corresponding censored bivariate times, respectively, for $i = 1, \cdots, n$. For each individual i, observed time can be denoted as $t_{ij} = \min(T_{ij}, C_{ij})$ by assuming (T_{i1}, T_{i2}) and (C_{i1}, C_{i2}) are independent. Denote $\delta_{ij} = I(t_{ij} = T_{ij})$ as a censoring indicator, $j = 1, 2$.

Let $\boldsymbol{\theta} = (\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \boldsymbol{\varphi}, \theta, p)$ denote the set of model parameters, where $\boldsymbol{\varphi} = (\varphi_{00}, \varphi_{10}, \varphi_{01}, \varphi_{11})$, and $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$ are parameters for $S_{01}(\cdot)$ and $S_{10}(\cdot)$ respectively. Considering the joint survival function, $S(t_{i1}, t_{i2})$, given by the equation (6) with the copula function given by the equation (10), the log-likelihood of i-th individual is given by

$$\begin{aligned} l_i(\boldsymbol{\theta}) = {} & \delta_{i1}\delta_{i2}\log f(t_{i1}, t_{i2}) + \delta_{i1}(1-\delta_{i2})\log\left(-\frac{\partial S(t_{i1}, t_{i2})}{\partial t_{i1}}\right) \\ & + \delta_{i2}(1-\delta_{i1})\log\left(-\frac{\partial S(t_{i1}, t_{i2})}{\partial t_{i2}}\right) + (1-\delta_{i1})(1-\delta_{i2})\log S(t_{i1}, t_{i2}), \end{aligned} \tag{11}$$

where

$$\begin{aligned} f(t_{i1}, t_{i2}) =& \frac{\partial^2 S(t_{i1}, t_{i2})}{\partial t_{i1} \partial t_{i2}} \\ =& \varphi_{11} f_1 f_2 [1 + \theta F_1 F_2]^{p-2} \Big\{ (1 + \theta F_1 F_2)[1 + \theta F_1 (F_2 - pS_2)] \\ &+ \theta p S_1 [(1 - 2F_2)(1 + \theta F_1 F_2) + \theta(p-1) F_1 F_2 (1 - F_2)] \Big\}, \end{aligned}$$

$$-\frac{\partial S(t_{i1}, t_{i2})}{\partial t_{i1}} = f_1 \Big\{ \varphi_{10} + \varphi_{11} S_2 (1 + \theta F_1 F_2)^{p-1} [1 + \theta F_2 (F_1 - pS_1)] \Big\},$$

$$-\frac{\partial S(t_{i1}, t_{i2})}{\partial t_{i2}} = f_2 \Big\{ \varphi_{01} + \varphi_{11} S_1 (1 + \theta F_1 F_2)^{p-1} [1 + \theta F_1 (F_2 - pS_2)] \Big\}, \text{ and}$$

$$S(t_{i1}, t_{i2}) = \varphi_{00} + \varphi_{10} S_1 + \varphi_{01} S_2 + \varphi_{11} S_1 S_2 (1 + \theta F_1 F_2)^p.$$

Here $S_1 = S_{10}(t_{i1})$, $S_2 = S_{01}(t_{i2})$, $F_1 = 1 - S_{10}(t_{i1})$, $F_2 = 1 - S_{01}(t_{i2})$, $f_1 = -\partial S_{10}(t_{i1})/\partial t_{i1}$ and $f_2 = -\partial S_{10}(t_{i2})/\partial t_{i2}$. Then the likelihood function of $\boldsymbol{\theta}$ for entire population is given by

$$L(\boldsymbol{\theta}) = \exp\Big(\sum_{i=1}^{n} l_i(\boldsymbol{\theta}) \Big). \tag{12}$$

We assume independent priors on the model parameters as

$$\pi(\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \boldsymbol{\varphi}, \theta, p) = \pi(\boldsymbol{\gamma}_1)\pi(\boldsymbol{\gamma}_2)\pi(\boldsymbol{\varphi})\pi(\theta)\pi(p), \tag{13}$$

where a Dirichlet prior for $\boldsymbol{\varphi} \overset{set}{=} (\varphi_1, \varphi_2, \varphi_3, \varphi_4)$ with hyperparameter value $w_1 = w_2 = w_3 = w_4$ is

$$\pi(\boldsymbol{\varphi}) = \frac{\Gamma(\sum_{i=1}^{4} w_i)}{\prod_{i=1}^{4} \Gamma(w_i)} \prod_{i=1}^{4} \varphi_i^{w_i - 1}. \tag{14}$$

Also a $Beta(a_\theta, b_\theta)$ distribution is assigned to $\frac{1}{2}(1 - \theta)$ and an inverse gamma distribution $IG(a_p, b_p)$ is assigned to $p - 1$. Therefore, the joint posterior distribution can be written as

$$\pi(\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \boldsymbol{\varphi}, \theta, p | \{t_{ij}\}) \propto L(\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \boldsymbol{\varphi}, \theta, p) \times \pi(\boldsymbol{\gamma}_1)\pi(\boldsymbol{\gamma}_2)\pi(\boldsymbol{\varphi})\pi(\theta)\pi(p).$$

In order to guarantee proper posteriors, we adopt proper priors with known hyper-parameters. Thus, the following prior distributions are assigned to parameters of marginal distributions (1) for the log GEV distribution, we assume $\pi(\boldsymbol{\gamma}_1)\pi(\boldsymbol{\gamma}_2) = \pi(\mu_1)\pi(\sigma_1)\pi(\xi_1)\pi(\mu_2)\pi(\sigma_2)\pi(\xi_2)$, where $\mu_1, \mu_2 \sim N(0, \sigma^2_{\mu_j})$, $\sigma_1, \sigma_2 \sim IG(a_{\sigma_j}, b_{\sigma_j})$ and $\xi_1, \xi_2 \sim N(0, \sigma^2_{\xi_j})$; (2) for the Weibull distribution, we assume $\pi(\boldsymbol{\gamma}_1)\pi(\boldsymbol{\gamma}_2) = \pi(\mu_1)\pi(\mu_2)\pi(\lambda_1)pi(\lambda_2)$, where $\mu_j \sim Gamma(a_{\mu_j}, b_{\mu_j})$ and $\lambda_j \sim Gamma(a_{\lambda_j}, b_{\lambda_j})$, $j = 1, 2.$; (3) for the log normal distribution, we assume $\pi(\boldsymbol{\gamma}_1)\pi(\boldsymbol{\gamma}_2) = \pi(\mu_1)\pi(\sigma_1)\pi(\mu_2)\pi(\sigma_2)$, where $\mu_1, \mu_2 \sim N(0, \sigma^2_{\mu_j})$ and $\sigma_1, \sigma_2 \sim IG(a_{\sigma_j}, b_{\sigma_j})$.

Since its integration is not easy to perform, we use MCMC techniques to construct sample chains which are progressively more likely realizations of the distribution of the target distribution. Specifically, we simulate samples of parameters via Metropolis-Hastings (HM) steps within the Gibbs sampler. More details on the algorithm can be found in Web Appendix A

2.4 Model Comparison Criteria

To set notation, let $\mathcal{D}$, $\mathcal{D}_i$ and $\mathcal{D}_{-i}$ be the observed dataset, the ith data point, and the dataset with $\mathcal{D}_i$ removed, respectively, $i = 1, \ldots, n$. Let $L(\mathcal{D}|\boldsymbol{\theta})$ be the likelihood function based on observed data $\mathcal{D}$, and $L_i(\cdot|\boldsymbol{\theta})$ be the likelihood contribution based on $\mathcal{D}_i$ where $\boldsymbol{\theta}$ is the entire collection of model parameters under a particular model. Suppose $\{\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(\mathcal{L})}\}$ are random samples drawn from the full posterior $p_{post}(\boldsymbol{\theta}|\mathcal{D})$ and $\hat{\boldsymbol{\theta}} = \sum_{l=1}^{\mathcal{L}} \boldsymbol{\theta}^{(l)}/\mathcal{L}$ is the posterior mean estimate for $\boldsymbol{\theta}$.

Several model comparison methods are proposed in the literature. In the paper we will consider the following criteria.

(1) The deviance information criterion (DIC), a generalization of the Akaike information criterion (AIC), is commonly used for comparing complex hierarchical models for which the asymptotic justification of AIC is not appropriate

(Burnham & Anderson, 2004; Vaida & Blanchard, 2005). This criterion can be incorporated during the Monte Carlo simulation. Lower values of DIC indicate better adjustment. The expression of DIC can be written as

$$\text{DIC} = -2\log L(\mathcal{D}|\hat{\boldsymbol{\theta}}) + 2p_D, \tag{15}$$

where

$$p_D = 2\left(\log L(\mathcal{D}|\hat{\boldsymbol{\theta}}) - \frac{1}{\mathcal{L}}\sum_{l=1}^{\mathcal{L}}\log L(\mathcal{D}|\boldsymbol{\theta}^{(l)})\right).$$

(2) The conditional predictive ordinate (CPO) method represents a posterior predictive approach that has proven useful in Bayesian model selection Box (1980); M.-H. Chen, Ibrahim, and Sinha (2002); Gelfand and Dey (1994). CPO provides a useful cross-validation approach that is computationally efficient, requiring only a sample from the posterior distribution. Larger values for the CPO_i imply better models and lower values indicate influential observations. The conditional predictive ordinate (CPO) statistic for data point $\mathcal{D}_i$ is given by

$$\text{CPO}_i = f(\mathcal{D}_i|\mathcal{D}_{-i}) = \int L_i(\mathcal{D}_i|\boldsymbol{\theta})p_{post}(\boldsymbol{\theta}|\mathcal{D}_{-i})d\boldsymbol{\theta},$$

where $p_{post}(\cdot|\mathcal{D}_{-i})$ is the posterior density of $\boldsymbol{\theta}$ give $\mathcal{D}_{-i}$. As noted by (Gelfand & Dey, 1994), one can use importance sampling to estimate CPO_i by

$$\widehat{\text{CPO}}_i = \left\{\frac{1}{\mathcal{L}}\sum_{l=1}^{\mathcal{L}}\frac{1}{L_i(\mathcal{D}_i|\boldsymbol{\theta}^{(l)})}\right\}^{-1}. \tag{16}$$

The LPML can be written in terms of CPO as

$$\text{LPML} = \sum_{i=1}^{n}\log\widehat{\text{CPO}}_i. \tag{17}$$

3. Goodness of Fit Test

In this section, we are going to extend the pseudo in-and-out-of-sample test so called PIOS test proposed by J. Zhang and Peng (2007), which perform a goodness-of-fit on the hypotheses given that the marginal distribution for susceptible subjects are fully specified. The hypotheses are stated as below.

$$\begin{aligned} H_0 &: C_0 \in C = \{C(\cdot;\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta | F_{01}(t_1); F_{10}(t_2)\} \\ &\quad vs \\ H_1 &: C_0 \notin C = \{C(\cdot;\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta | F_{01}(t_1); F_{10}(t_2)\}, \end{aligned}$$

where $C_0(\cdot)$ is the true copula function, $\Theta \subset \Re^2$ is a 2-dimensional parameter space, and $F_{01}(t_1), F_{10}(t_2)$ are the CDF for the susceptible individuals in the lifetimes T_1 and T_2, respectively.

To derive the goodness-of-fit test statistic, two-step estimation technique was applied. According to Shih and Louis (1995), the first step is to estimate parameters $\mathbf{P}_j$, $\boldsymbol{\gamma}_j$ and $\boldsymbol{\varphi}$ by maximizing the marginal likelihood function $\sum_{i=1}^{n}\Big[\delta_{ij}\log f_j(t_{ij}) + (1-\delta_{ij})\log S_j(t_{ij})\Big]$, where $f_j(t) = -\partial S_j(t)/\partial t$, and denote the estimates as $\hat{\mathbf{P}}_j$, $\hat{\boldsymbol{\gamma}}_j$ and $\hat{\boldsymbol{\varphi}}$. The second step is to obtain a pseudo maximum likelihood estimates (PMLE) of (θ, p) by maximizing $\sum_{i=1}^{n} l_i(\hat{\boldsymbol{\gamma}}_1, \hat{\boldsymbol{\gamma}}_2, \hat{\boldsymbol{\varphi}}, \theta, p)$, and denote the estimates by $\hat{\theta}$ and $\hat{p}$. The PMLE $(\hat{\theta}, \hat{p})$ in "in-sample" pseudo log likelihood is obtained using the full data. And the PMLE $(\hat{\theta}_{(-i)}, \hat{p}_{(-i)})$ in "out-sample" pseudo log likelihood is obtained using the subset of data with ith observation removed. Then test statistic can be written as

$$T_n = \hat{l}_{in} - \hat{l}_{out} = \sum_{i=1}^{n} l_i(\hat{\boldsymbol{\gamma}}_1, \hat{\boldsymbol{\gamma}}_2, \hat{\boldsymbol{\varphi}}, \hat{\theta}, \hat{p}) - \sum_{i=1}^{n} l_i(\hat{\boldsymbol{\gamma}}_1, \hat{\boldsymbol{\gamma}}_2, \hat{\boldsymbol{\varphi}}, \hat{\theta}_{(-i)}, \hat{p}_{(-i)}), \tag{18}$$

where the log-likelihood function of *ith* data can be written as

$$\begin{aligned} l(\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \boldsymbol{\varphi}, \theta, p; D_i) &= \delta_{i1}\delta_{i2}\log f(t_{i1}, t_{i2}) + (1-\delta_{i1})\delta_{i2}\log\left\{-\frac{\partial S(t_{i1}, t_{i2})}{\partial t_{i2}}\right\} \\ &\quad + (1-\delta_{i2})\delta_{i1}\log\left\{-\frac{\partial S(t_{i1}, t_{i2})}{\partial t_{i1}}\right\}. \end{aligned} \tag{19}$$

There are several good properties regarding to this test statistic. First, T_n in Equation (18) converges in probability to the dimension of the parameter vector of the copula under the null hypothesis. Secondly, T_n is asymptotically subject to normal distribution under null hypothesis. The proofs can be found in J. Zhang and Peng (2007).

Instead of estimating the asymptotic variance analytically, we use the following bootstrap technique to approximates the asymptotic variance of the test statistic in finite samples.

1: Calculate test statistic T_n for the bivariate survival times using the original n pairs of observations.

2: Sample with replacement with size n.

3: Calculate the test statistic, denoted as $T_n^{(b)}$, using the sampled data from step 2.

4: Repeat 2 and 3 B times so have B test statistics, denoted as T_n, $T_n^B = \{T_n^{(b)}, b = 1, \cdots, B\}$. Calculate the standard deviation, $sd\{T_n^B\}$. Finally, calculate the p-value which is $2\left[1 - \Phi\left(\left|\frac{T_n - m}{sd\left(T_n^B\right)}\right|\right)\right]$, where Φ is the cdf of a standard normal distribution and m is the dimension of the parameter vector of copula function ($m = 2$ in our case).

There is a practical issue of choosing B, the number of bootstrap samples. Scholars have recommended more bootstrap samples as available computing power has increased. However, increasing the number of samples cannot increase the amount of information in the original data. It can only reduce the effects of random sampling errors which can arise from a bootstrap procedure itself (Kloke, McKean, & McKean, 2015). Racine and MacKinnon (2007) discuss this issue at length and proposed a method for choosing the number of bootstrap samples. Theoretical results derived by Olive (2017) suggest using $B \geq [nlogn]$. We choose the number of bootstrap samples same as sample size because of computing power.

One thing I need to mention here is before we obtain a PMLE of copula parameters (θ, p) in step 2, we need to estimate $\boldsymbol{\varphi}$ using equations in (1) which is same to estimate $\rho = cov(d_{i1}, d_{i2})$, where d_{ij} is an indicator variable showing the ith subject is susceptible for the jth event and $P(d_{ij} = 1) = 1 - P_j$. For that we have to identify the subjects that are susceptible for the events and we should know how many subjects are actually susceptible for the jth event according to our model. Since we have cure rate estimates $\hat{P}_j$, it indicates $n(1 - \hat{P}_j)$ subjects are susceptible for the jth event. Note that the subject with an uncensored observation is susceptible. Let D_j be the number of uncensored observations for the jth event. Then we need to choose $n(1 - \hat{P}_j) - D_j \triangleq n_j$ observations from censored observations. Actually, any subject having censored lifetime observation could be susceptible for the event. However, the smaller the value of observation, the smaller the value of the survival function, the more likely the subject having this survival time is susceptible. Thus, we can make $d_{ij} = 1$ for the subjects having the first n_j smallest survival times among those censoring times. Once we have d_{ij}, we use the sample covariance, $\hat{\rho}$, between d_{i1} and d_{i2} to estimate ρ. Plugging $\hat{\rho}$ and $\hat{P}_j$ into equations in (1), we can get an estimate $\hat{\boldsymbol{\varphi}}$ for $\boldsymbol{\varphi}$.

4. Simulation Studies

4.1 Estimation

In this section, we are going to use the results from simulation studies to illustrate the performance of the proposed methodology. The following four models are selected for model comparison by choosing different survival functions $S_{01}(\cdot) = S_{\gamma_1}(\cdot)$ and $S_{10}(\cdot) = S_{\gamma_2}(\cdot)$ (Weibull, log normal or log GEV) and the copula functions $C(\cdot, \cdot)$ (FGM or generalized FGM) in the model specification of (6):

Model 1: $S_{\gamma_1}(\cdot)$ and $S_{\gamma_2}(\cdot)$ are from Weibull; $C(s,t) = st[1 + \theta(1-s)(1-t)]$;

Model 2: $S_{\gamma_1}(\cdot)$ and $S_{\gamma_2}(\cdot)$ are from Weibull; $C(s,t) = st[1 + \theta(1-s)(1-t)]^p$;

Model 3: $S_{\gamma_1}(\cdot)$ and $S_{\gamma_2}(\cdot)$ are from log GEV; $C(s,t) = st[1 + \theta(1-s)(1-t)]^p$;

Model 4: $S_{\gamma_1}(\cdot)$ and $S_{\gamma_2}(\cdot)$ are from log normal; $C(s,t) = st[1 + \theta(1-s)(1-t)]^p$.

The simulation study includes a total of 48 simulated data sets based upon the four models, four sample sizes ($n = 50, 100, 200, 100$) and three censoring rates (L: 15% to 20%, M: 30% to 40%, and H: 45% to 60%). Once the setting is fixed, follow the steps below to get one simulated data set.

Step 1: Draw two independent uniform random variables (u_{i1}, v_{i2}).

Step 2: Set $u_{i2} = C_{2|1}^{-1}(u_{i1}, v_{i2})$, where $C_{2|1}^{-1}$ denotes the pseudo-inverse of $C_{2|1}$. More specifically, solve the following equation for u_{i2} when $C(s,t) = st[1+\theta(1-s)(1-t)]^p$: $u_{i2}[1+\theta(1-u_{i1})(1-u_{i2})]^{p-1}[1+\theta(1-u_{i1})(1-u_{i2}) - \theta p u_{i1}(1-u_{i2})] - v_{i2} = 0$.

Step 3: Generate a bivariate survival times (T_{i1}, T_{i2}) from (u_{i1}, u_{i2}) via $T_{i1} = F_{\gamma_1}^{-1}(u_{i1})$ and $T_{i2} = F_{\gamma_2}^{-1}(u_{i2})$, where $F_{\gamma_j}^{-1}(\cdot)$ is the quantile function of the distribution corresponding to $S_{\gamma_j}(\cdot)$.

Step 4: Generate latent indicator values (d_{i1}, d_{i2}) according to the distribution of $P(d_1 = i, d_2 = j) = \varphi_{ij}$, where i, j is 0 or 1 and d_{ij} is an indicator variable with 1 indicating that the ith subject is susceptible for the jth event. If $d_{ij} = 0$, we change T_{ij} to be a big number, say $10,000$, since the ith subject is cured for the jth event.

Step 5: Simulate the censoring time C_{ij} from Weibull distributions, which results in censoring rate for different levels (L, M, H) where $i = 1, \cdots, n$ and $j = 1, 2$.

Step 6: Obtain the observed data $D = \{(t_{i1}, t_{i2}, \delta_{i1}, \delta_{i2}), i = 1, \cdots, n\}$, where $t_{i1} = \min(T_{i1}, C_{i1})$ and $t_{i2} = \min(T_{i2}, C_{i2})$, $\delta_{i1} = I(T_{i1} \le C_{i1})$ and $\delta_{i2} = I(T_{i2} \le C_{i2})$.

For true values of parameters, we have $(\mu_1, \lambda_1) = (30, 5)$, $(\mu_2, \lambda_2) = (20, 4)$,
$\theta = 0.6$ for model 1 and 2, $p = 1.5$ for model 2, 3 and 4 and $(\mu_1, \sigma_1, \xi_1) = (3, 0.2, 0.1)$, $(\mu_2, \sigma_2, \xi_2) = (2, 0.3, 0.2)$ for model 3. We have $(\mu_1, \sigma_1) = (2.5, 1)$,
$(\mu_2, \sigma_2) = (2, 1.5)$ for model 4. We have $(\varphi_{11}, \varphi_{10}, \varphi_{01}, \varphi_{00}) = (0.70, 0.15, 0.10, 0.05)$, $(\varphi_{11}, \varphi_{10}, \varphi_{01}, \varphi_{00}) = (0.40, 0.30, 0.20, 0.10)$ and $(\varphi_{11}, \varphi_{10}, \varphi_{01}, \varphi_{00}) = (0.15, 0.40,$
$0.25, 0.20)$ for low censoring, medium censoring and high censoring, respectively.

MATLAB 2017b is our first choice for all the computation and R is used to generate the tables and graphs of the results. We plot running means of variables of interest vs iteration number such as sample trace plots. By visual inspection, the chains are began from over-dispersed starting points. Therefore, we decided to discard the first 5,000 iterations which contributes the burn-in phase and run another 50,000 iterations. To reduce the correlations of successive samples, we store every 10th values of the chain after burn-in phase which result in 5000 samples for the posterior analysis.

Table 1 provides summary results of model comparison for the cases with $n = 200$ and $n = 500$. For each sample size, a total 12 data sets are generated from each one of the four models and three censoring rates. The bold entries indicate the best fitted model according to DIC and LPML. The effective number of parameter p_D indicates the model complexity. According to DIC and LPML in Table 1, the best model and true model are consistent except for one case with $n = 200$ and medium censoring rate. However, Model 3 performs better than any other models even the true model is not Model 3 in small sample sizes such as $n = 50$ and $n = 100$. This information is available in Table 1 in Web Appendix B.

The posterior mean, standard deviation (SD) and %95 HPD interval under different scenarios are available in Tables 2 - 5 in Web Appendix B. As we can see that the true values of parameters are inside the 95% highest posterior density (HPD) interval. The posterior mean of association parameters are essentially unbiased under the true model. The estimated standard deviation is fairly small. Figure 1 shows the estimated survival function based upon Kaplan-Meier estimator and four different models when the Model 3 is true model. It shows that the estimated survival function based on Model 3 is closer to Kaplan-Meier estimates compared to other three models. However, the estimated survival function of Model 3 does not show big difference from Kaplan-Meier estimate, even Model 3 is not true model. This information can be seen in Figures 1 - 3 in Web Appendix B. The examination of these tables and figures shows misspecification of model can lead to significantly biased estimates which result in inaccurate inference and incorrect conclusions.

Table 1. Model comparison results for n=200 and 500

True Model	Censoring rate	Fitted Model	n = 200			n = 500		
			DIC	LPML	p_D	DIC	LPML	p_D
1	L	1	**2433**	**-1216**	7.78	**6040**	**-3020**	7.87
		2	2434	-1217	7.82	6041	-30201	7.95
		3	2453	-1226	8.70	6050	-3025	9.32
		4	2478	-1239	7.91	6119	-3060	7.94
	M	1	**2104**	**-1052**	7.77	**5343**	**-2672**	7.72
		2	2105	-1053	7.85	5344	-2673	7.92
		3	2108	-1055	8.39	5394	-2697	9.25
		4	2132	-1066	7.72	5415	-2708	7.67
	H	1	**1749**	**-875**	7.45	**4243**	**-2121**	7.27
		2	1750	-876	7.62	4244	-2122	7.58
		3	1754	-877	8.48	4258	-2129	10.98
		4	1790	-895	7.43	4332	-2166	7.44
2	L	1	2411	-1206	7.51	6017	-3009	7.57
		2	**2410**	**-1205**	7.86	**6016**	**-3008**	7.70
		3	2414	-1207	8.31	6030	-3018	9.15
		4	2450	-1225	7.67	6118	-3059	7.83
	M	1	2114	-1057	7.13	5231	-2615	7.44
		2	2113	-1056	7.45	**5230**	**-2614**	7.69
		3	**2109**	**-1054**	8.69	5262	-2630	9.36
		4	2150	-1075	7.87	5323	-2661	7.65
	H	1	1702	-852	6.96	4287	-2144	7.56
		2	**1700**	**-850**	8.54	**4286**	**-2143**	7.76
		3	1709	-855	8.57	4298	-2149	9.79
		4	1731	-866	7.66	4365	-2182	7.76
3	L	1	2465	-1235	6.97	6038	-3022	7.47
		2	2463	-1234	8.37	6037	-3021	7.81
		3	**2187**	**-1093**	9.91	**5473**	**-2736**	9.82
		4	2258	-1129	7.78	5651	-2826	7.90
	M	1	2169	-1089	7.12	5314	-2662	7.06
		2	2168	-1088	7.48	5313	-21661	7.76
		3	**1974**	**-987**	9.42	**4861**	**-2430**	9.71
		4	2019	-1010	7.65	4973	-2487	7.67
	H	1	1752	-883	7.00	4215	-2110	7.29
		2	1751	-882	7.22	4214	-2009	7.47
		3	**1596**	**-798**	8.87	**3944**	**-1972**	9.41
		4	1069	-535	7.68	3978	-1990	7.87
4	L	1	2594	-1297	6.80	6486	-3243	6.97
		2	2595	-1298	7.30	6487	-3244	7.48
		3	2581	-1290	8.89	6452	-3226	9.11
		4	**2575**	**-1288**	7.25	**6437**	**-3219**	7.43
	M	1	2283	-1142	7.50	5708	-2854	7.68
		2	2283	-1142	8.05	5709	-2855	8.24
		3	2271	-1136	9.80	5678	-2839	9.76
		4	**2266**	**-1133**	7.99	?**5665**	?**-2833**	8.19
	H	1	1826	-925	7.29	4566	-2312	7.47
		2	1827	-925	7.82	4567	-2312	8.01
		3	1817	-920	9.52	4542	-2299	9.49
		4	?**1813**	?**-918**	7.77	?**4532**	?**-2295**	7.96

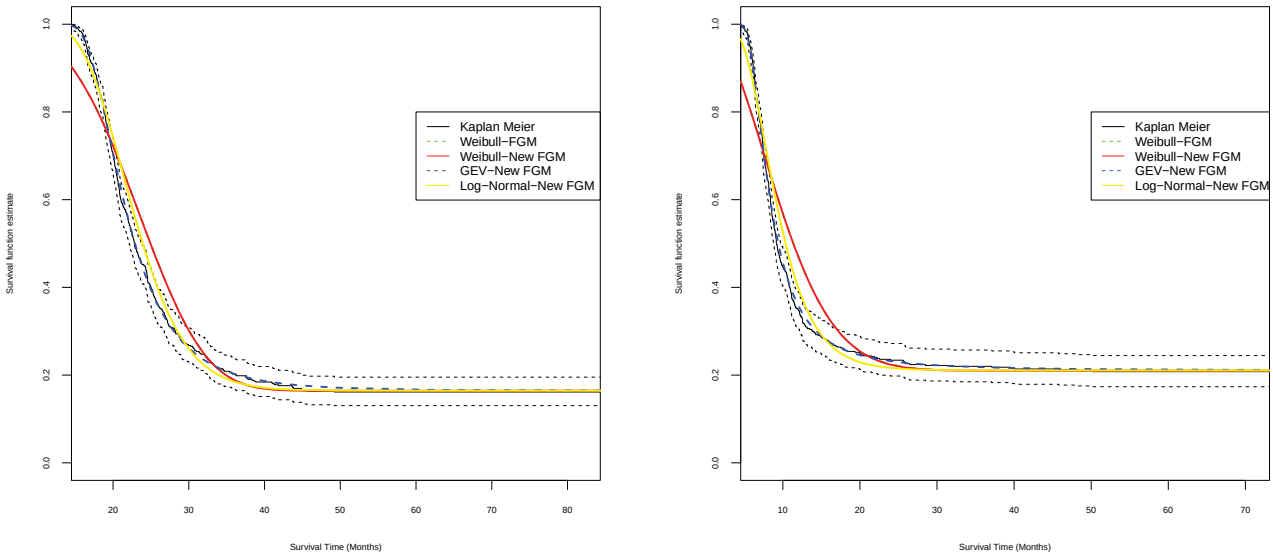

Figure 1. The plots of estimated survival functions of different models based upon simulated data sets from model 3 with n = 500 and low censoring rate. Left: Survival function estimate for treated eye; Right: Survival function estimate for untreated eye

4.2 Goodness of Fit Performance

In this section, we perform goodness of fit test for our proposed model rely on empirical type I error and test power. The following copulas are considered: (1) the generalized FGM, (2) Clayton, and (3) Gaussian. We generate a total of 36 simulated datasets for each different marginal distribution (Weibull, log GEV and log normal) based on three copulas, with four different sample sizes (100, 200, 500, 1000), under three levels of censoring rate (20%, 40%, 60%), and three levels of Kendalls τ, ($\tau = 0.16, 0.20, 0.24$). For each dataset, the empirical p-value is calculated using bootstrap samples.

First, we need to generate the data based on three copulas, four different sample sizes, three levels of censoring rate and three levels of Kendall's τ.

Secondly, we can follow the steps in Section 3 to calculate the test statistics and produce the p-value through bootstrap samples. Recall in our bootstrap steps, we state that the subject associated with smaller censored lifetime is more likely to be the susceptible subject. On the other hand, the subject associated with bigger censored lifetime is more likely to be the cured subject. Note that the likelihood function in Equation (18) is based upon the entire population which implies this test is to perform the goodness-of-fit for the model, not only for the copula. We can limit the population to the susceptible subjects which is the subset of the entire population. In this way, we can test the goodness-of-fit test only for copula. We can still follow the step in Section 3 to identify the susceptible subjects that have censored lifetime, and change the joint survival function and joint density function involved in Equation (18) to the ones of susceptible subjects. The type I errors and test powers are available in Tables 9 - 11 and Tables 15 - 17, respectively, in Web Appendix C.

Table 6 - 8 in Web Appendix C report the empirical type I error under different sample sizes and with three different marginal distributions for susceptible individuals. The type I error is empirical proportions to reject the null hypothesis when the null hypothesis is assumed to be true at significance level equal to 5%. As we can see from these tables, the overall performance of our test is good, especially for sample size equal to 1000. Regardless of marginal distributions and censoring rates, the type I error decreases as sample size increases. Type I error does not show certain trend when Kendall's τ increases.

Table 12 - 14 in Web Appendix C report the empirical power under different sample sizes and with three different marginal distributions for susceptible individuals. As we can see from these tables, the overall performance of our test in differentiating among new FGM, Clayton and Gaussian copula is good. However, n its hard to differentiate the Gaussian copula and Clayton copula as the power of rejecting the Clayton copula is low, even with big sample data simulated from Gaussian. This might be due to the right censoring of simulated data, which leads to insufficient information of the upper tail dependence from the data. It is noted that the test power increases as the sample sizes increases, or the censoring rate

decreases, but does show certain trend when Kendall's τ increases.

It is also noted that type I error and test power do not change significantly when marginal distributions are changed. It implies that our test works regardless of marginal distribution.

5. Real Data Analysis

In this section, we demonstrate our proposed model by using the Diabetic Retinopathy Study (DRS) data which was first considered by Huster, Brookmeyer, and Self (1989). There are 162 patients and each patient received laser treatment for one eye and no treatment for the other eye. In the analysis considered here, the time to blindness for the eye randomized to laser treatment and not received the treatment are denoted as T_1 and T_2, respectively. The blindness is defined as the time from initiation of treatment to the time when visual acuity dropped below 5/200 two visits in row. To check the data, there are 75 censored observations in T_1 and 84 censored observations in T_2. For those censored observations, some patients will not have the occurrence of blindness in the period of study because of drop-out and end of study.

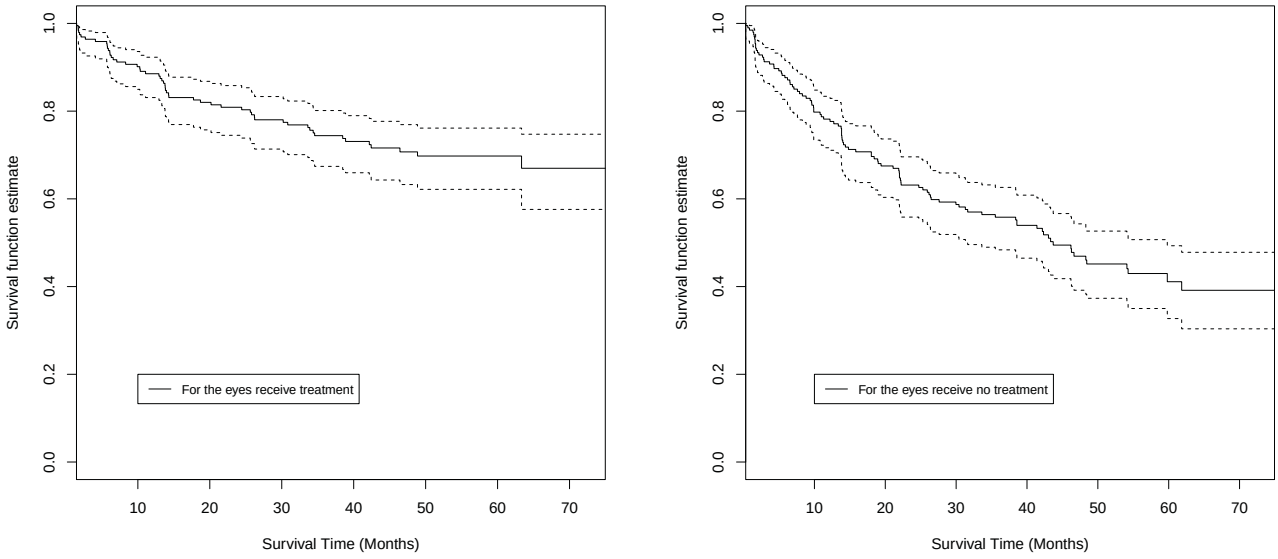

Figure 2. Kaplan-Meier nonparametric estimates for the survival function

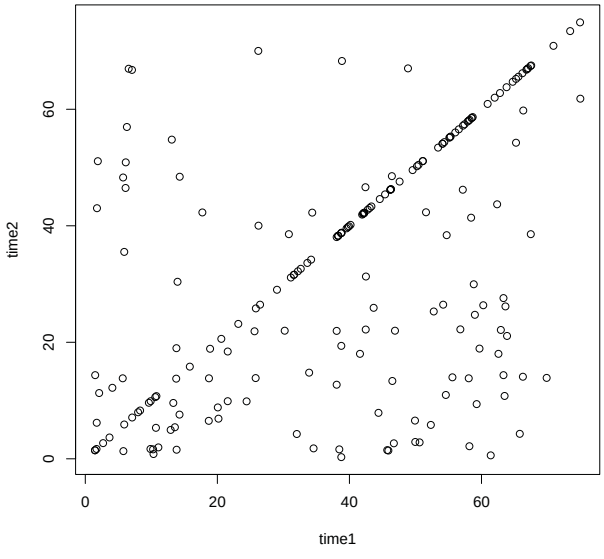

Figure 3. Scatter Plot for DRS data

As a preliminary analysis of the data, Figure 2 gives estimated survival function of T_1 and T_2 based upon the Kaplan Meier estimates. From this plot, we have some indication of cure fraction. Also, it seems to have better results for the eye that received treatment which has larger time to blindness in comparison with the eye that received no treatment. Figure 3 gives the scatter plot for T_1 and T_2. From this plot, we observe the two lifetimes spread out all over the place which means there is no certain trend for the relationship between T_1 and T_2. In other word, the correlation between T_1 and T_2 is weak. From Figure 2 and Figure 3, cure rate model should be considered when a significant proportion of patients are "cured" and copulas with week dependence should be considered.

We fit the following six models such as log GEV, log normal and Weibull with symmetric FGM and generalized FGM copulas. The model of log GEV with generalized FGM copula gives lowest DIC (1367.31) and highest LPML (-683.61) so that our proposed model with log GEV with generalized FGM copula is the best for fitting the data.

As we discussed in the simulation study, the first 5,000 iterations will be ignored, and another 50,000 iterations will be used to consider the simulation of each parameters. To get approximated uncorrelated values, we store every 10th values of the chain after burn-in phase which gives a final chain of size 5,000 for the posterior analysis. We also apply our test procedure to this data. The empirical estimate of Kendalls rank correlation is 0.376. The corresponding p-value of our test for the generalized FGM is 0.296. At the significant level 0.05, we failed to reject these this copula.

Table 2. Posterior summaries: Bivariate log GEV distribution based upon a generalized FGM copula

Parameter	Mean	SD	95% Credible Interval	DIC	LPML
μ_1	3.15	0.40	(2.37 , 3.91)		
μ_2	2.55	0.21	(2.15 , 2.96)		
σ_1	1.72	0.42	(1.01 , 2.59)		
σ_2	1.41	0.15	(1.15 , 1.71)		
ξ_1	-0.21	0.30	(-0.82 , 0.35)		
ξ_2	-0.60	0.17	(-0.92 , -0.26)	1367.31	-683.61
φ_{00}	0.23	0.08	(0.06 , 0.36)		
φ_{01}	0.15	0.10	(0.00 , 0.33)		
φ_{10}	0.06	0.05	(0.00 , 0.15)		
φ_{11}	0.57	0.12	(0.32 , 0.78)		
p	1.68	0.76	(1.09 , 2.79)		
θ	0.35	0.23	(-0.08 , 0.78)		
θ	0.42	0.20	(0.05 , 0.8)		

Table 3. The DRS data: Bayesian criteria for models proposed in Louzada et al. (2013)

	DIC	LPML
FGM Weibull	1522	-761.81
FGM Exponential	1525	-763.46
PSF Weibull	1527	-764.89
PSF Exponential	1524	-762.87
Frank Weibull	1522	-762.04
Frank Exponential	1525	-763.69
Clayton Weibull	1523	-762.47
Clayton Exponential	1525	-763.93
Independence Weibull	1528	-764.55
Independence Exponential	1530	-765.93

Table 2 shows the posterior summaries of interest assuming log GEV as marginal distribution for the lifetime T_1 and T_2 for susceptible subjects using a generalized FGM. The posterior mean of all the parameters are in the %95 HPD interval. The standard deviation is relatively small. The time to blindness of untreated eye has a lower cure rate compared to that of treated eye. The posterior estimates for other models are shown in tables 18 - 22 in Web Appendix D. Table 3 presents the model comparison criteria for the same DRS data which is discussed in Louzada et al. (2013). The results show that our method performs better than all the models proposed in Louzada et al. (2013).

The 3D plot of the joint survival function and corresponding contour plot for model of log GEV with generalized FGM copula presented in Figure 4. It shows that the joint survival functions are decreasing as time t_1 goes up or time t_2 goes

up. It also shows that the joint survival function decreases slowly when t_1 goes up compared to when t_2 goes up. That implies that the treatment actually has positive effect on the eye.

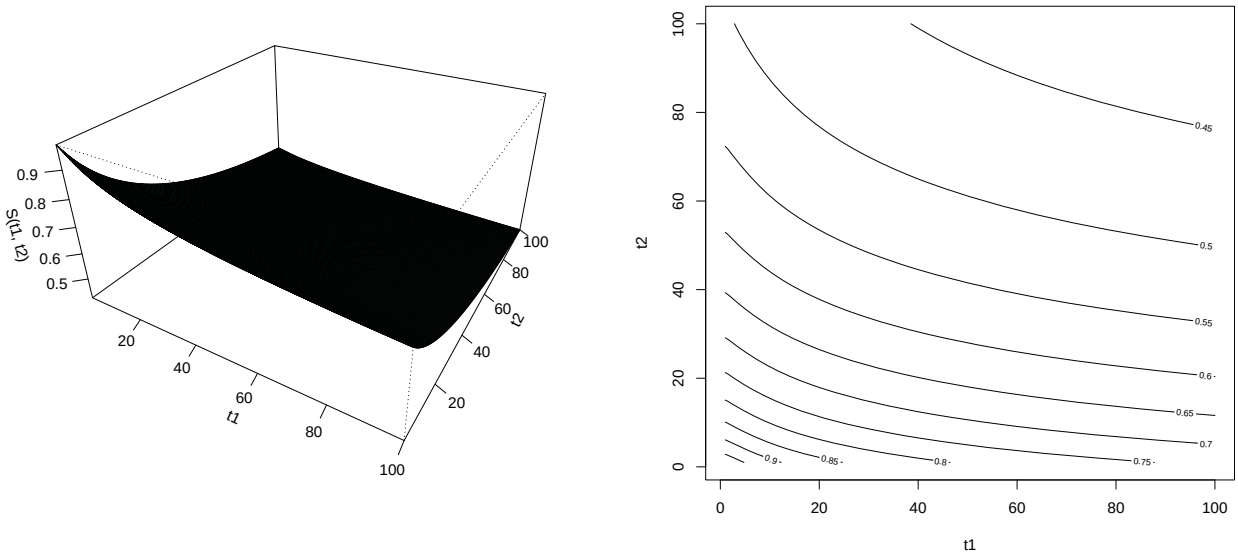

Figure 4. The 3D plot of joint survival function and corresponding contour plot under Model 3. Left: Joint survival function; Right: Contour plot

6. Discussion

The use of copula functions could be a good alternative way to analyze bivariate lifetime data. Observe that in many applications of lifetime modeling we could have individuals that are "long term survivors" or "cure individuals". An analytical structure of the statistical methodology was developed to model the dependence between cure rate fractions, and an extremely flexible generalized extreme value distribution was employed to model the logarithm of the survival time. It is very useful to use copulas to avoid the problem of the marginal distributions depending on the dependence structure, especially the use of the generalized FGM copula. This allows a broader range of correlations than the typical FGM copula indicating that the methods can be applied to more data contexts.

To check for adequacy of the generalized copula for our situation, we have extended the PIOS test to the new proposed test for right-censored bivariate survival times where the possibility of cure must be incorporated. The fact that the performance of our test does not depend on the choice of marginal distributions provides a lot of flexibility and avoids model misspecification issues. Also, the test is computationally straightforward and easily constructed.

References

Bekrizadeh, H., & Jamshidi, B. (2017). A new class of bivariate copulas: dependence measures and properties. *METRON*, *75*(1), 31–50.

Box, G. E. (1980). Sampling and bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society: Series A (General)*, *143*(4), 383–404.

Burnham, K., & Anderson, D. (2004). Model selection and multi-model inference. *Second. NY: Springer-Verlag*, *63*(2020), 10.

Chen, C.-M., Lu, T.-F. C., & Hsu, C.-M. (2013). Association estimation for clustered failure time data with a cure fraction. *Computational Statistics & Data Analysis*, *57*(1), 210–222.

Chen, M.-H., Ibrahim, J. G., & Sinha, D. (2002). Bayesian inference for multivariate survival data with a cure fraction. *Journal of Multivariate Analysis*, *80*(1), 101–126.

Fermanian, J.-D. (2005). Goodness-of-fit tests for copulas. *Journal of multivariate analysis*, *95*(1), 119–152.

Gallardo, D. I., Gómez, Y. M., & de Castro, M. (2018). A flexible cure rate model based on the polylogarithm distribution. *Journal of Statistical Computation and Simulation*, *88*(11), 2137–2149.

Gelfand, A. E., & Dey, D. K. (1994). Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society: Series B*, *56*(3), 501–514.

Genest, C., Rémillard, B., & Beaudoin, D. (2009). Goodness-of-fit tests for copulas: A review and a power study. *Insurance: Mathematics and economics*, *44*(2), 199–213.

Huster, W. J., Brookmeyer, R., & Self, S. G. (1989). Modelling paired survival data with covariates. *Biometrics*, 145–156.

Kloke, J., McKean, J. W., & McKean, J. W. (2015). *Nonparametric statistical methods using r*. CRC Press Boca Raton.

Lakhal-Chaieb, L., & Duchesne, T. (2017). Association measures for bivariate failure times in the presence of a cure fraction. *Lifetime Data Analysis*, *23*(4), 517–532.

Louzada, F., Suzuki, A. K., & Cancho, V. G. (2013). The fgm long-term bivariate survival copula model: modeling, bayesian estimation, and case influence diagnostics. *Communications in Statistics-Theory and Methods*, *42*(4), 673–691.

Olive, D. J. (2017). Theory for linear models. In *Linear regression* (pp. 313–342). Springer.

Peng, Y., & Taylor, J. M. (2011). Mixture cure model with random effects for the analysis of a multi-center tonsil cancer study. *Statistics in medicine*, *30*(3), 211–223.

Racine, J. S., & MacKinnon, J. G. (2007). Simulation-based tests that can use any number of simulations. *Communications in StatisticsSimulation and Computation®*, *36*(2), 357–365.

Romeo, J. S., Tanaka, N. I., & Pedroso-de Lima, A. C. (2006). Bivariate survival modeling: a bayesian approach based on copulas. *Lifetime Data Analysis*, *12*(2), 205–222.

Shih, J. H., & Louis, T. A. (1995). Inferences on the association parameter in copula models for bivariate survival data. *Biometrics*, 1384–1399.

Vaida, F., & Blanchard, S. (2005). Conditional akaike information for mixed-effects models. *Biometrika*, *92*(2), 351–370.

Wang, W., & Wells, M. T. (2000). Model selection and semiparametric inference for bivariate failure-time data. *Journal of the American Statistical Association*, *95*(449), 62–72.

Yashin, A. I., & Iachine, I. A. (1999). Dependent hazards in multivariate survival problems. *Journal of Multivariate Analysis*, *71*(2), 241–261.

Zhang, J., & Peng, Y. (2007). A new estimation method for the semiparametric accelerated failure time mixture cure model. *Statistics in medicine*, *26*(16), 3157–3171.

Zhang, S., Okhrin, O., Zhou, Q. M., & Song, P. X.-K. (2016). Goodness-of-fit test for specification of semiparametric copula dependence models. *Journal of Econometrics*, *193*(1), 215–233.

International Journal of Statistics and Probability; Vol. 11, No. 3; May 2022
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Interactions Between Water Level, Crude Oil, and Hydroelectric Power Generation in Ghana; A Structured Vector Auto Regression Approach

Smart Asomaning Sarpong[1], Akwasi Agyei[2]

[1'2] Institute of Research, Innovation and Development – IRID, Kumasi Technical University, Ghana

Correspondence: Smart Asomaning Sarpong, Institute of Research, Innovation and Development – IRID, Kumasi Technical University, Ghana. E-mail: smart.asarpong@kstu.edu.gh / smartsarpong2015@gmail.com

Received: February 13, 2022 Accepted: April 12, 2022 Online Published: April 20, 2022

doi:10.5539/ijsp.v11n3p22 URL: https://doi.org/10.5539/ijsp.v11n3p22

Abstract

Countries that suffer disturbances in their power generation are less likely to meet many of the sustainable development goals and general economic growth. This study used a three-variable SVAR model to examine the interactions of water level, crude oil and power generated from the Akosombo hydroelectric generation Dam in Ghana. Data used for this span from January 2010 to December 2019. From the results, none of the three important variables studied was found to be completely independent; dam level and crude oil are adjusted to absorb power generation shocks. All three variables drift away from their normal levels to contain shock before returning to their desired levels at varying time points. It has also been established that Dam water level shocks leads to a negative response in both power generation and crude oil in the short run. Overall, shocks to crude oil explains much of the variability in power generation than shocks to dam water level. These findings convey that there is exist very useful interactions among the three-variables studied in this paper. Policy makers should institute effective measures for early detection and intervention of the short-term power disturbance that characterizes the hydroelectric power generation to ensure a sustainable power and growth of the Ghanaian economy.

Keywords: SVAR, Hydroelectric, Ghana, long run shocks, impulse response

1. Introduction

If the world can achieve a high percentage of the targets set under the sustainable development goals, stable electricity supply will play a crucial role (Owusu & Asumadu-Sakodie, 2016; Owusu et al., 2016). About 10% of the population of the world still do are yet to be hooked onto electricity making it a matter of global concern (IEA, 2015). Among all the energy production mix, hydropower remains the largest renewable energy resource due to its cost-effectiveness and reliability (IEA, 2015). According to Benefoh and Ackom (2016), electricity supply that is both reliable and inexpensive is crucial to any country's development.

The Akosombo dam is a hydroelectric power generating station on the Volta River in the south-eastern of Ghana and it is managed by the Volta River Authority. It serves as the major source of electricity in Ghana. It has a powerplant which contains six turbine generator units, and it operates between 276ft maximum and 248ft minimum headwater level. Currently, the Akosombo dam produces 1000MW electricity at its maximum operating capacity (Miescher, 2021). According to Smokorowski (2021), the peak of the hydro is the only reliable flexible method of producing electricity besides fossil fuels. The Akosombo hydroelectric project was meant among others to open Ghana to industrialization and hence modern development. Fishing, farming, transportation, and tourism are some of the other good effects (Gyau-Boakye, 2001). The availability of water resources usually determines when and how much energy the hydroelectric plant will generate on a seasonal and annual time frame (Carpentier et al., 2017).

Long-run shocks in power generation are the unanticipated changes in power generations over long time. The shocks trigger the operation of the powerplants in production of electricity. Because the dam's primary source of water is rain, which is unpredictable and dependent on weather conditions (Mensah, 2013), there are a lot of factors that causes a disturbance either to increase or decrease the water level. During the dry season, the level of water in the reservoir and the surrounding area reduces, while during the rainy season, it swells. As a result, power generation becomes unstable which affect the growth and sustainability of a country. Ghana's industrial and economic growth has resulted in a rising demand for power that exceeds the capacity of the Akosombo dam power plant. Part of the reason for the limited producing capacity is a lack of fuel supply to existing thermal power plants (Kemausuor & Ackom, 2017).

Russ (2020) studied the effects of runoff shocks on general growth of the economy. His suggested that rainfall should

not be considered a good determinant of water availability for power generation. According to Taghizadeh-hesary and Yoshino (2013), Oil-producing countries gain from shocks in oil price.

According to United State Agency for International development (2017), drought and reduced rainfall threaten access to reliable sources of power. Silver et al. (2012) showed that, apart from USA, an increasing Renewable Energy Source on Electricity share has economic cost on GDP per capita. Kumi (2017) indicated that despite the increased in electricity in Ghana and the doubling of installed generation capacity from 2006-2016, the country still suffers from inadequate electricity supply.

Ashong (2016) pointed out that poor rainfall and its resulting impact on hydropower generation are to blame for Ghana's lack of renewable energy. Boadi and Owusu (2019) found out that, changes in rainfall patterns accounts for 21% of year-on-year fluctuations in hydroelectric power generated from Akosombo. Kabo-Bah et al. (2016) indicated that temperature negatively correlate with hydropower generation while humidity and rainfall positively affect power generation. Eshun and Amoako-Tuffour (2016) pointed out that prolong drought usually is the root cause of unstable power supply.

Many previous studies, most of them focused on factors affecting renewable energy sources (Ashong, 2016; Kabo-Bah et al., 2016; Salub et al., 2020). This may be due to the environmental and climate factors in which power production from both renewable and non-renewable sources depend on. As such, any change in those factors also affect the production process. Others looked at shocks from either water level or oil price in relation to the growth of an economy (Russ, 2020; Taghizadeh-hesary & Yoshino, 2013; Lorde et al., 2009).

There exist a growing body of literature on the two key components of hydroelectric power generation; water level, and Crude oil used to power the turbines (Miescher, 2021; Dehghani et al., 2019; Mekonnen & Hoekstra, 2012; Harrison & Whittington, 2002). That notwithstanding, the question that needs to be addressed is, if we hold environmental, geographical and climate conditions constant, how does power generation respond to shocks in dam level and crude oil? This study therefore attempts to address some three critical issues. First, it seeks to empirically examine the joint dependence of water level of dam, crude oil used, and amount of power generated and to establish whether these variables help explain one another. Second, the study examines whether any of the variables has a higher (or lower) influence on other variables. Third, how is a shock in one of the variables absorbed by the other variables.

To achieve the study objectives, we model water level, crude oil use and power generated in a three-variable structured vector autoregression (SVAR) framework. SVAR is a very useful method developed by Sims (1989) and remains the preferred method of many researchers investigating interactions between structured variable (Mertens & Olea, 2018; Mumtaz & Theodoridis, 2020). Flexibility in allowing variables to be determined endogenously, and ability to reveal a theoretical model closely related to empirical reality are some advantages SVAR has over other methods. To the best of our knowledge, no previous study exists in Ghana that examines joint dependence of dam level shocks, crude oil shocks and power generation. This study may therefore contribute to knowledge in this regard and form a good basis for policy formulation and decision making.

2. Materials and Methods

This study used a monthly secondary data on dam from January 2010 to December 2019. The data consist of three variables namely, power generation, dam level and crude oil for the sample period. In this study, we analyze the relationship between power generation, dam level and crude oil in the context of Ghana in Structural Vector Autoregression (SVAR) framework. The model building involves four steps to obtain an appropriate model that will help develop the relationship among the variables. The Eviews version 11 (Eviews11) statistical software would be used to analyze the data to achieve the aim of the study.

2.1 Series Transformation

The SVAR model provides an avenue to transform all the series into their natural logarithm form. This will minimize fluctuations in the data set (Tiwari, 2011). Detailed overview of the SVAR model is available in (Sims, 1989) and (Christiano, 2012).

2.1.1 Test for Stationarity

To identify the order of integration, the Augmented Dickey-Fuller (ADF) test of unit root will be used to access the stationarity of the series. The ADF test is a regression test that analyze a series stationarity under the null hypothesis; there is a unit root in the series. The regression equation of the model is given by:

$$\Delta y_t = \alpha + \beta t + \gamma y_{(t-1)} + \delta_1 \Delta y_{t-1} + \cdots + \delta_p \Delta y_{t-p} + \varepsilon_t$$

Where,

y_t is the observed time series

α is constant

β is the coefficient of the time trend

p is the order of AR process.

If $\gamma = 0$, the series is random walk and if -1 < 1+ γ < 1, the series is stationary.

2.2 Model Estimation Using SVAR

Before estimating the model, model order p to be used in the study must be determined using the Akaike Information Criteria (AIC). The AIC has been proved to perform better especially when the sample size is small (Liew, 2004).

2.3 Structural Vector Autoregressive (SVAR) Model

The structural VAR model helps to impose long-run restrictions on the variables. This study make use of three variables namely, power generation, dam level and crude oil. By following the Kandil and Trabelsi (2012) estimation procedure, the representation of the variables in SVAR framework are as follows: Using the log transform of the variables, let $\Delta x_t = [\Delta P_t,\ \Delta D_t,\ \Delta O_t]$' and $\varepsilon_t = [\varepsilon_{pt}, \varepsilon_{dt}, \varepsilon_{ot}]$ where Δ represent the first order differenced operator $\varepsilon_{pt}, \varepsilon_{dt}, \varepsilon_{ot}$ represent power, dam level and crude oil shocks. The structural VAR model can be written as:

$$\Delta x_t = B_0\varepsilon_t + B_1\varepsilon_{t-1} + B_2\varepsilon_{t-2} + \cdots = B(L)\varepsilon_t$$

$$\text{where } B(L) = \begin{pmatrix} B_{11}(L) & B_{12}(L) & B_{13}(L) \\ B_{21}(L) & B_{22}(L) & B_{23}(L) \\ B_{31}(L) & B_{32}(L) & B_{33}(L) \end{pmatrix}$$

The matrix B is a 3×3 matrix that provides the impulse responses of endogenous variables to structural shocks and L is the lag operator. It is assumed that the shocks $\varepsilon_t = [\varepsilon_{pt}, \varepsilon_{dt}, \varepsilon_{ot}]$ are serially uncorrelated and have a covariance matrix normalized to the identity matrix. This implies that power generation is subjects to three structural shocks.

To compute the above model, restrictions must be imposed on the parameter matrices. The restrictions can either be of contemporaneous type or long-run type. This study applies the long-run restrictions method proposed in (Herwartz, 2019). Similar to Kim and Chow (2003), the following restrictions are imposed:

- Dam level shocks and crude oil shocks will have no effect on power generation in the long-run. This is equivalent to $B_{12}(L)$= $B_{13}(L)$=0. Thus, the cumulative effects of dam level shock and crude oil shock on power generation will be zero (0).
- Crude oil shocks have no long-run effects on dam level. This amount to $B_{23}(L) = 0$.

The long-run restrictions amount to $B_{12}(L) = B_{13}(L) = B_{23}(L) = 0$ which are enough to identify matrix B_i.

2.4 Impulse Response Function

After estimation of the model, we then obtain the impulse response functions of the variables. The impulse response functions assess the dynamic effect of a structural shock of one standard deviation on the variable over a given period [35]. Using Kandil and Trabelsi (2012) SVAR methodology, let us consider a reduced VAR model

$$w_t = \mu + \prod_1 w_{t-1} + \prod_2 w_{t-2} + \cdots + \prod_j w_{t-j} + \varepsilon_t$$

Where w_t represent the endogenous variables (power-generation, damlevel and crude oil) $\prod_j$ for j = 1… represent coefficient matric and μ represent the error term

To interpret the coefficients (μ, $\prod_1, \ldots, \prod_j$), we then employ the impulse response analysis. We make a shock to ε_t and look at the dynamic propagation based on the MA representation:

$$w_t = \varepsilon_t + C_1\varepsilon_{t-1} + C_2\varepsilon_{t-2} + \cdots + C_j\varepsilon_{t-j} + C_0$$

where, $\frac{\partial(w_t)}{\partial\varepsilon_t'} = I_p, \frac{\partial(w_{t+1})}{\partial\varepsilon_t'} = C_1, \frac{\partial(w_{t+2})}{\partial\varepsilon_t'} = C_2 \ldots$

constants $C_1 \ldots C_{t-j}$ are matrices. These derivatives are represented in a graph called impulse response function.

2.5 Variance Decomposition

It is very useful that we determine variations among the variables in terms of percentages. Variance decomposition will help determine the amount of variation in the dependent variable that is explained by the independent variables. This

study placed more emphasis on power generation in response to one standard deviation shocks in dam level and crude oil.

3. Results

3.1 Graphs of the Series

To better understand the series, we obtained the graph of the series and found out that, series exhibit changes in mean over time which suggest a nonstationary nature of the series. This movements in the series indicates a presence of shocks over a given time. It can be observed that, all the time series show possible change in mean which suggest that the series is possibly non-stationary. This can further be approved using ADF statistical test.

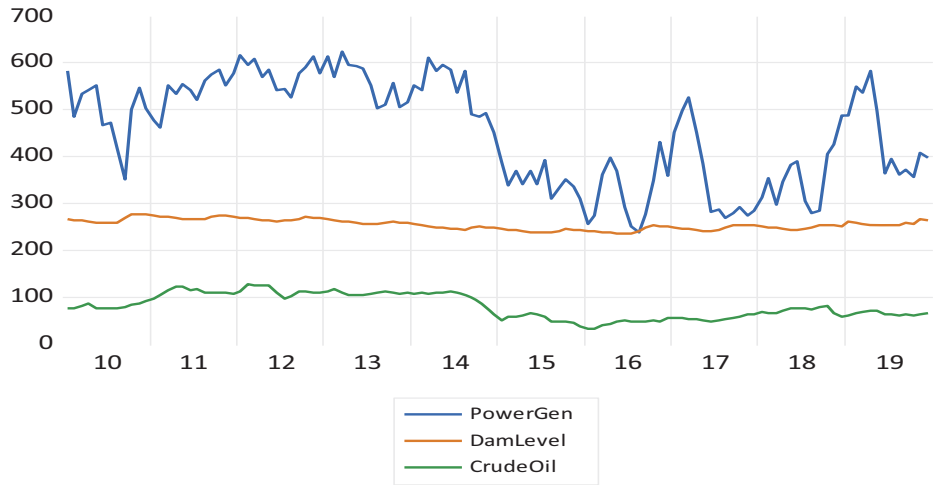

Figure 1. Graph of the series

The ADF statistical test was used to test stationarity of the series. All series were nonstationary at their levels since the p-values for power generation dam level and crude oil were greater than 5% level of significance. The series then became stationary after first differencing which satisfy the stationarity requirement of the underlying model. Table 1 shows the results of the ADF test at 5% level of significance.

Table 1. ADF test results

At their levels	**ADF-Stat**	**P-value**
Powergen	-2.077335	0.2542
Damlevel	-1.013997	0.7565
Crudeoil	-1.561538	0.4990
At first difference		
Powergen	-8.010330	0.0000
Damlevel	-6.524328	0.0000
Crudeoil	-8.111935	0.0000

It can be observed that the log transform of all the variables were non-stationary at levels and became stationary after first difference. This is because all the p-values are less than 5% level of significance after first difference.

3.3 Model Estimation

Now that all the variables followed a unit root but are stationary at first difference, we proceed to estimate the model. Before that, we need to determine the optimum lag order for the model. Table 2 below displays the results of the optimum lag selection.

Table 2. Optimum lag selection

Lag	LogL	AIC	SC	HQ
0	561.0366	-10.05471	-9.981483	-10.02501
1	589.4967	-10.40535*	-10.11242*	-10.28652*
2	595.4129	-10.34978	-9.837168	-10.14183
3	601.1303	-10.29064	-9.558331	-9.993562
4	607.3263	-10.24011	-9.288117	-9.853917
5	615.5950	-10.22694	-9.055249	-9.751618
6	625.8826	-10.25014	-8.858757	-9.685695
7	637.4988	-10.29727	-8.686203	-9.643711
8	646.4955	-10.29722	-8.466452	-9.554529

SVAR model demands that we obtain an appropriate lag for the model. This prevents the model from giving a spurious result. We obtained the lag using the AIC and the results suggested lag 1 for the model. This was the lag order used throughout the model estimation.

3.4 Variance Decomposition

From the above table, all the information criteria returned lag 1 as the optimum lag for the model to be estimated. To estimate the structural vector autoregressive model, we impose long-run restrictions. This will help know the effect of shocks on the variables in the long-run. The table below summarizes the parameter estimates of the SVAR model. The results of fitted model tell us that, a cumulative shocks of dam level and crude oil are zero. Thus, they have no long-run effects on power generation.

Table 3. SVAR model estimates

Structural VAR is just-identified

Model: e = Phi*Fu where E[uu]=1				
F =				
C(1)	0	0		
C(2)	C(4)	0		
C(3)	C(5)	C(6)		
	Coefficient	Std. Error	z-Statistics	Prob.
C(1)	0.112565	0.007327	15.36229	0.0000
C(2)	0.002955	0.001444	2.046518	0.0407
C(3)	0.039659	0.008791	4.511396	0.0000
C(4)	0.015548	0.001012	15.36229	0.0000
C(5)	-0.018361	0.008318	-2.207454	0.0273
C(6)	0.089417	0.005821	15.36229	0.0000
Log likelihood	617.1500			
Estimated S matrix:				
0.094532	-0.042454	-0.038000		
0.004263	0.008280	0.001760		
0.030850	-0.020085	0.064167		
Estimated F matrix:				
0.112565	0.000000	0.000000		
0.002955	0.015546	0.000000		
0.039659	-0.018361	0.089417		

3.5 Impulse-Response Function

In the SVAR model, the impulse response functions trace the effects of a shock on endogenous variables. Figure 2 displays the impulse-response function of the variables. Shock 1 represent the shocks in power generation, shock 2 represent shocks in dam level and shock 3 represent the shocks in crude oil. It can be observed that the effect of a shock in dam level and crude oil vanishes over a short period. Also, the effect of a power generation shock on both dam level and crude oil is immediate. Thus, sudden and permanent increase can be seen.

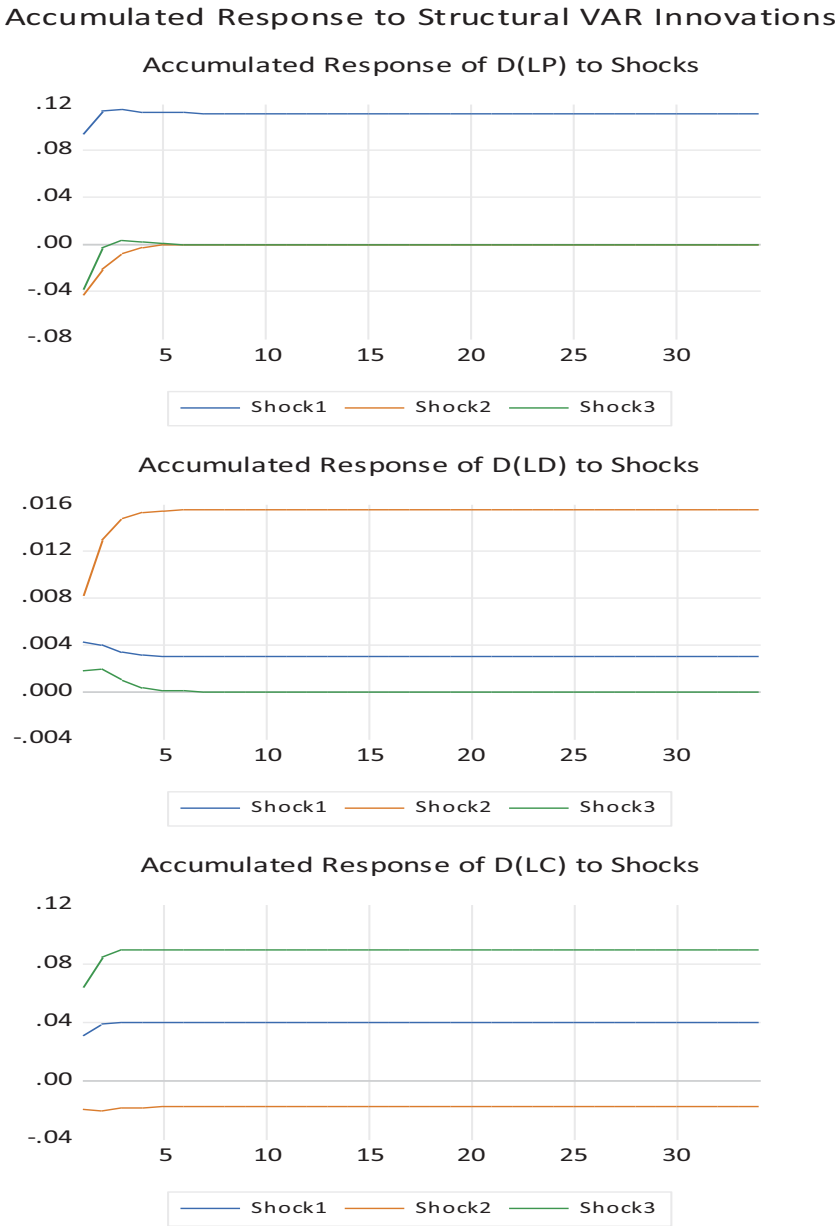

Figure 1. Accumulated impulse response functions (D(LP) = Power Generated, D(LD) = Dam level, D(LC) = Crude oil)

To better understand the response behavior to shock, Figure 3 (the orthogonal inpulse response function (OIRF)) is constructed. OIRF assumes that, the hydroelectric production system is in a steady state prior to any shock and that shocks apply to one variable only at any given time. Each row of the OIRF plot explains how the hydroelectric production system absorbes a one-standard-deviation of orthogonal shock and the length of time it takes for these variable to return to a steady state.

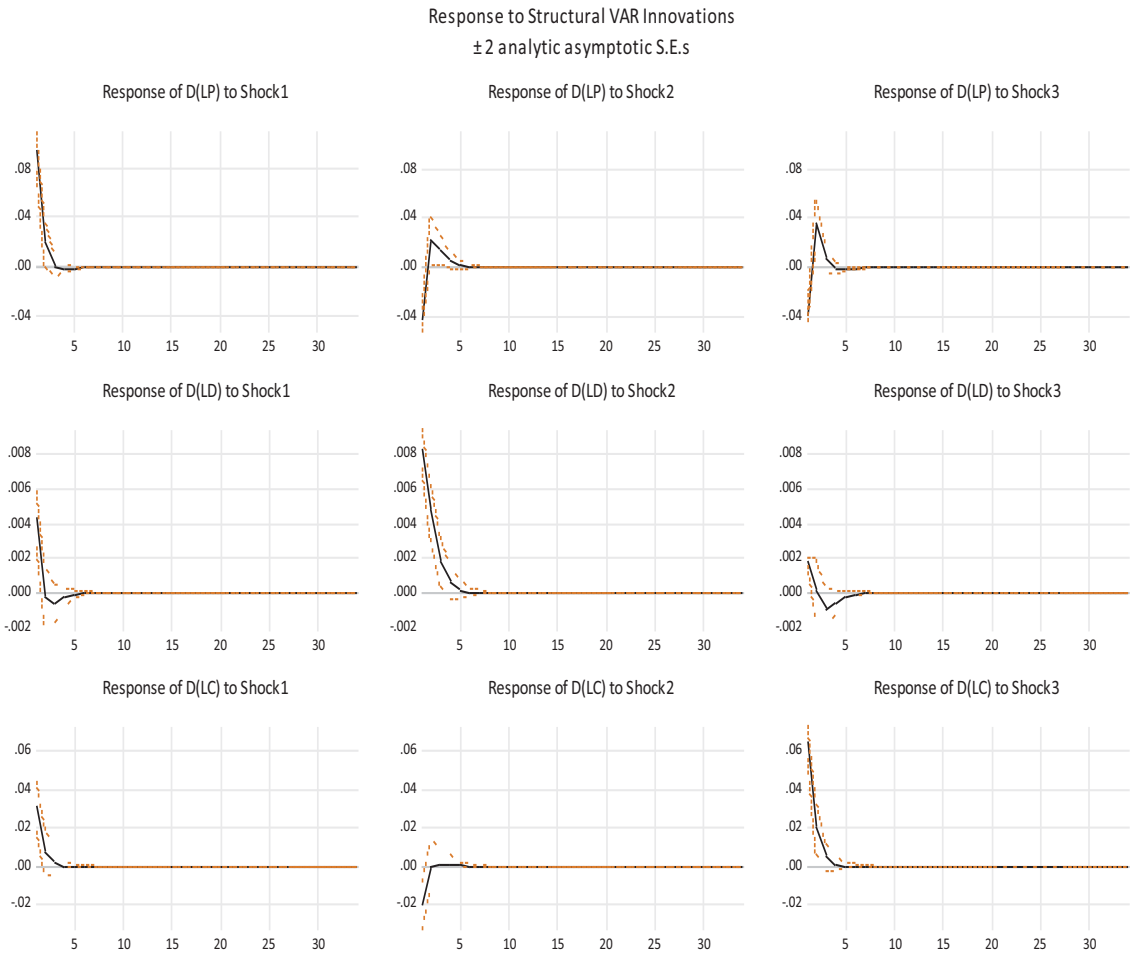

Figure 2. orthogonalized impulse response function. * broken lines denotes a 95% confidence interval; *x-axis = forecast in months; y-axis = magnitude of response; * D(LP) = Power Generated = shock 1, D(LD) = Dam level = shock 2, D(LC) = Crude oil = shock 3; * Graph D(LP) to Shock A = response of variable Power to a shock in variable A; * Graph A:A shows how the shock is absorbed)

The impulse response function as shown in Figure 3 dipicts a one-standard deviation of shock to the dam level, crude oil and power generated. First row of shows that, dam level and crude oil are adjusted to absorbe power generation shocks. As revealed by Graph (D(LP) to shock 2), a positive shock in power generation is followed by a positive response in dam water level which remains statistically significant for close to 8 months. Also, crude oil response positively to power generation siginificantly for almost 8 months. It is clear from the second row that Dam water level shocks leads to a negative response in both power generation and crude oil while crude oil shocks leads to responses in dam level and power generation. It can be observed from figure 3 also that, as dam level and crude oil increases, power generation increases in the short run but returns to a steady state in the long-run. An increase in dam level will decrease crude oil in the short run. Also, an increase in power generation will increase crude oil in the short run, but it will remain steady in the long-run.

3.6 Variance Decomposition

The variance decomposition identifies which shock is more important in accounting for variability in power generation after presenting the contribution of each shock to the movement in power generation. From table 4 below, it can be observed that power generation explains much of the variations by itself (about 64%) while dam level explains about 17% and crude oil explains about 19% of the variations in power generation in the long-run. This shows that crude oil contributes more in explaining the variability in power generation.

Table 4. Variance decomposition results

Variance Decomposition of D(LP)				
Period	S.E.	Shock1	Shock2	Shock3
1	0.110374	73.35282	14.79435	11.85282
2	0.119646	65.17989	15.78681	19.03330
3	0.120583	64.17588	16.80794	19.01619
4	0.120716	64.04831	16.96176	18.98992
5	0.120740	64.02735	16.97390	18.99875
6	0.120744	64.02443	16.97429	19.00127
7	0.120744	64.02412	16.97426	19.00162
8	0.120744	64.02409	16.97425	19.00165
9	0.120744	64.02409	16.97425	19.00166
10	0.120744	64.02409	16.97425	19.00166
11	0.120744	64.02409	16.97425	19.00166
12	0.120744	64.02409	16.97425	19.00166

4. Discussion

This study aimed at analyzing the response of power generation to shocks in dam level and crude oil. The variables included in the study were monthly data for power generation, dam level and crude oil. The structural vector autoregressive methodology was used in the study to obtain an appropriate model that analyzed the relationship among the variables in terms of shocks. The variance decomposition analysis reveals shocks to dam level and crude oil only account for about 36% of total variability in power generation from the Akosombo dam in the short run but in the long run, these variations will become steady. This finding is supported by (Miescher, 2021; Eshun & Amoako-Tuffour, 2016; Dehghani et al., 2019). One interesting finding of this study is the revelation that shocks to crude oil explains much of the variability in power generation than the water level of the Akosombo dam. This finding is supported by (Russ, 2020; Taghizadeh-hesary & Yoshino, 2013; Lorde et al., 2009). Again, both dam level and crude oil cumulative shocks to power generation vanishes in a relatively short period and returns to desired levels in the long run. The short-run decrease in power generation in Ghana usually results in power rationing which affect the economy (Ashong, 2016; Owusu & Owusu, 2019; Kabo-Bah et al., 2016; Eshun & Amoako-Tuffour, 2016; Sulub et al., 2020).

Ghana experience longer periods of rainy season than dry season (USAID, 2017). During rainy seasons, dam level increases which provide much water to regulate the operation of the hydropower plants, thereby increasing power generation. Reduction in water levels negatively affect hydropower generation as supported by (USAID, 2017). Availability of crude oil also improve the operations of the powerplants thereby increasing power generation (Russ, 2020; Taghizadeh-hesary & Yoshino, 2013; Lorde et al., 2009).

The variance decomposition result has assured that, dam level and crude oil donate approximately 17% and 19% in total variation in power generation respectively. This may be because of climate variability and fuel supply challenges as indicated by (Kumi, 2017; Boadi & Owusu, 2019).

5. Conclusions

This study used a three-variable SVAR model to examine the interactions of water level, crude oil and power generated. From the results, none of these three important variables are completely independent; dam level and crude oil are adjusted to absorb power generation shocks. It has also been established that Dam water level shocks leads to a negative response in both power generation and crude oil while crude oil shocks lead to responses in dam level and power generation. With the aid of the orthogonal impulse response function this study can confirm that all three variables deviate from their desired levels to absorb shock before returning to their desired levels at varying time points.

The impulse response identifies a decrease in power generation in the short run. This means that increase in dam level and crude oil negatively affects power generation in the short run. In the long-run, shocks to dam level and crude oil will have no effect on power generation. Therefore, policy makers should institute effective measures that will detect and avert the short-term power disturbance to ensure power sustainability and growth of the economy. The energy sector should also explore in alternative ways of obtaining fuel, such as, regasification, to reduce fuel supply challenges. This paper draws conclusions based on a single model without controlling for other possible factors that influence hydroelectric power generation and hence recommend that further research considers that.

Author Contributions: Conceptualization, S.A.S. and A.A.; methodology, S.A.S., and A.A; software, S.A.S.; validation, S.A.S.; formal analysis, S.A.S. and A.A; investigation, S.A.S; resources, S.A.S.; data curation, S.A.S.; writing—original draft preparation, S.A.S., and A.A, writing—review and editing, S.A.S., and A.A; supervision, S.A.S.; project administration, S.A.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to privacy reasons.

Acknowledgments: Staff of the Institute of Research, Innovation and Development - IRID, Kumasi Technical University—KsTU, Ghana.

Conflicts of Interest: The authors declare no conflict of interest.

References

Ashong, J. N. T. (2016). How Effectively Has Ghana Implemented Its Policy for Large-Scale Renewable Electricity Deployment: A Qualitative Assessment. *Renewable Energy Law and Policy Review*, *7*(2), 133–144. https://www.jstor.org/stable/26256493

Benefoh, D. T., & Ackom, E. K. (2016). Energy and low carbon development efforts in Ghana: institutional arrangements, initiatives, challenges and the way forward. *AIMS Energ*, *4*, 481-503. https://doi.org/10.3934/energy.2016.3.481

Boadi, S. A., & Owusu, K. (2019). Impact of climate change and variability on hydropower in Ghana. *African Geographical Review*, *38*(1), 19–31. https://doi.org/10.1080/19376812.2017.1284598

Carpentier, D., Haas, J., Olivares, M., & de la Fuente, A. (2017). Modeling the multi-seasonal link between the hydrodynamics of a reservoir and its hydropower plant operation. *Water (Switzerland)*, *9*(6). https://doi.org/10.3390/w9060367

Christiano, L. J. (2012). Christopher A. Sims and vector autoregressions. *Scand. J. Econ., 114*(4), 1082–1104. https://doi.org/10.1111/j.1467-9442.2012.01737.x

Christopher, N., & Kenneth, U. (2018). *Munich Personal RePEc Archive On The Feasibility of a Common Currency in West Africa : Evidence from a Multivariate Structural VAR*. *88357*.

Dehghani, M., Riahi-Madvar, H., Hooshyaripor, F., Mosavi, A., Shamshirband, S., Zavadskas, E. K., & Chau, K. Wing. (2019). Prediction of hydropower generation using Grey wolf optimization adaptive neuro-fuzzy inference system. *Energies*, *12*(2), 1–20. https://doi.org/10.3390/en12020289

Eshun, M. E., & Amoako-Tuffour, J. (2016). A review of the trends in Ghana's power sector. *Energy, Sustainability and Society*, *6*(1), 1-9. https://doi.org/10.1186/s13705-016-0075-y

Gyau-Boakye, P. (2001). Environmental impacts of the Akosombo dam and effects of climate change on the lake levels. *Environment, Development and Sustainability*, *3*(1), 17-29. https://doi.org/10.1023/A:1011402116047

Harrison, G. P., & Whittington, H. B. W. (2002). Susceptibility of the Batoka Gorge hydroelectric scheme to climate change. *Journal of hydrology*, *264*(1-4), 230-241. https://doi.org/10.1016/S0022-1694(02)00096-3

Herwartz, H. (2019). Long-run neutrality of demand shocks: Revisiting Blanchard and Quah (1989) with independent structural shocks. *Journal of Applied Econometrics*, *34*(5), 811-819. https://doi.org/10.1002/jae.2675

IEA. (2015). Situational Analysis of Energy Industry, Policy and Strategy for Kenya. Available at: https://www.africaportal.org/dspace/articles/situational-analysis-energy-industry-policy-and-strategy-kenya. Accessed November 12th, 2021

Kabo-Bah, A. T., Diji, C., Nokoe, K., Mulugetta, Y., Obeng-Ofori, D., & Akpoti, K. (2016). Multiyear rainfall and temperature trends in the Volta River basin and their potential impact on hydropower generation in Ghana. *Climate*, *4*(4), 49. https://doi.org/10.3390/cli4040049

Kandil, M., & Trabelsi, M. (2012). Is the Announced Monetary Union in GCC Countries Feasible? A Multivariate Structural Var Approach. *Middle East Development Journal*, *4*(1). https://doi.org/10.1142/S1793812012500010

Kemausuor, F., & Ackom, E. (2017). Toward universal electrification in Ghana. *Wiley Interdisciplinary Reviews: Energy and Environment*, *6*(1), 1–14. https://doi.org/10.1002/wene.225

Kim, Y., & Chow, H. K. (2003). Optimum currency area in Europe: An alternative assessment. *Economics Letters*, *81*(3), 297–304. https://doi.org/10.1016/S0165-1765(03)00196-4

Kumi, E. N. (2017). The Electricity Situation in Ghana: Challenges and Opportunities. *CGD Policy Paper*, *September*, 30. https://www.cgdev.org/publication/electricity-situation-ghana-challenges-and-opportunities

Liew, V. K.-S. (2004). Which Lag Length Criteria Should We Employ? *Economic Bulletin*, *3*(33), 1–9.

Lorde, T., Jackman, M., & Thomas, C. (2009). The macroeconomic effects of oil price fluctuations on a small open oil-producing country: The case of Trinidad and Tobago. *Energy Policy*, *37*(7), 2708–2716. https://doi.org/10.1016/j.enpol.2009.03.004

Mekonnen, M. M., & Hoekstra, A. Y. (2012). The blue water footprint of electricity from hydropower. *Hydrology and Earth System Sciences*, *16*(1), 179–187. https://doi.org/10.5194/hess-16-179-2012

Mensah, D. (2013). *Application of Time Series in Predicting the Water Levels of the Akosombo Dam*.

Mertens, K., & Olea, J. L. M. (2018). Marginal tax rates and income: new time series evidence. *Q. J. Econ., 133*(4), 1803–1884. https://doi.org/10.1093/qje/qjy008

Miescher, S. F. (2021). Ghana's Akosombo Dam, Volta Lake Fisheries & Climate Change. *Daedalus*, *150*(4), 124–142. https://doi.org/10.1162/daed_a_01876

Mumtaz, H., & Theodoridis, K. (2020). Dynamic effects of monetary policy shocks onmacroeconomic volatility. J. Monet. *Econ., 114*, 262–282. https://doi.org/10.1016/j.jmoneco.2019.03.011

Owusu, P. A., Asumadu-Sarkodie, S., & Ameyo, P. (2016). A review of Ghana's water resource management and the future prospect. Cogent. Eng. https://doi.org/10.1080/23311916.2016.1164275

Owusu, P., & Asumadu-Sarkodie, S. (2016). A review of renewable energy sources, sustainability issues and climate change mitigation. *Cogent. Eng., 3,* 1167990. https://doi.org/10.1080/23311916.2016.1167990

Russ, J. (2020). Water runoff and economic activity : The impact of water supply shocks on growth. *Journal of Environmental Economics and Management*, *101*, 102322. https://doi.org/10.1016/j.jeem.2020.102322

Silva, S., Soares, I., & Pinho, C. (2012). The impact of renewable energy sources on economic growth and co 2 emissions-A svar approach. *European Research Studies Journal*, *15*(SPECIAL ISSUE), 133–144. https://doi.org/10.35808/ersj/374

Sims, C. A. (1989). Models and their uses. *Am. J. Agric. Econ., 71*(2), 489–494. https://doi.org/10.2307/1241619

Smokorowski, K. E. (2021). The ups and downs of hydropeaking: a Canadian perspective on the need for, and ecological costs of, peaking hydropower production. *Hydrobiologia*, *0123456789*. https://doi.org/10.1007/s10750-020-04480-y

Sulub, Y. A., Hamid, Z., & Nazri, M. N. M. (2020). Renewable energy supply and economic growth in malaysia: An application of bounds testing and causality analysis. *International Journal of Energy Economics and Policy*, *10*(3), 255–264. https://doi.org/10.32479/ijeep.8980

Taghizadeh-hesary, F., & Yoshino, N. (2013). *An Estimation of the Impact of Oil Shocks on Crude Oil Exporting Economies An Estimation of the Impact of Oil Shocks on Crude Oil Exporting Economies and Their Trade Partners*. *January*. https://doi.org/10.3868/s060-002-013-0029-3

Tiwari, A. K. (2011). A structural VAR analysis of renewable energy consumption, real GDP and CO2 emissions: evidence from India. *Economics Bulletin*, *31*(2), 1793-1806.

USAID. (2017). Ghana: Power Africa Fact Sheet. Retrieved from USAID: https://www.usaid.gov/powerafrica/ghana

Zhou, F., Li, L., Zhang, K., Trajcevski, G., Yao, F., Huang, Y., Zhong, T., Wang, J., & Liu, Q. (2020). Forecasting the Evolution of Hydropower Generation. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2861–2870. https://doi.org/10.1145/3394486.3403337

International Journal of Statistics and Probability; Vol. 11, No. 3; May 2022
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Approximation of the Binomial Probability Function Using the Discrete Normal Distribution

Mohammad Fraiwan Al-Saleh [1] & Doaa Suhail Obeidat [2]

[1] Department of Statistics, Yarmouk University, Jordan. Currently at Department of Mathematics and Statistics, Jordan University of Science and Technology, Jordan

[2] Department of Statistics, Yarmouk University, Jordan

Correspondence: Mohammad Fraiwan Al-Saleh, Department of Statistics, Yarmouk University, Jordan

Received: March 17, 2022 Accepted: April 19, 2022 Online Published: April 21, 2022

doi:10.5539/ijsp.v11n3p32 URL: https://doi.org/10.5539/ijsp.v11n3p32

Abstract

A new method of approximating the Binomial probability function is introduced. The method is based on the discrete normal distribution. In particular, the discrete normal probability function is used to approximate the binomial probability function. The new approximation is compared with the exact values and the approximation based on Central limit theorem. The maximum absolute error of the approximation is used to measure the accuracy of the method. It turned out that this method of approximation is useful and easy to use in practice. Also, the result can be an important theoretical statistical result that can be used in educational statistics.

Keywords: binomial probability function, discrete normal distribution, continuity correction; maximum absolute error, central limit theorem, maximum absolute error

1. Introduction and Some Closely Related Works

The Normal distribution is extremely important in statistics and is often used in practice as an approximate distribution for the distributions of real-valued random variables whose distributions are unknown. X is normally distributed with mean μ and variance σ^2, $X \sim N(\mu,\sigma^2)$, if it has the probability density function (pdf):

$$f(x;\mu,\sigma^2)=\frac{1}{\sigma\sqrt{2\pi}}\,e^{\frac{-(x-\mu)^2}{2\sigma^2}},\ -\infty<x<\infty,\quad -\infty<\mu<\infty,\quad 0<\sigma^2<\infty.$$

The curve is bell-shaped, symmetric; mean=median= mode. If $\mu=0$ & $\sigma=1$, then X is a standard normal random variable; ϕ & Φ denotes pdf and the cumulative distribution function(cdf) of the N(0,1), respectively.

Using the normal distribution to approximate other distribution is a very old topic; it goes back to more than 300 years. In 1730, DeMoivre consider the problem of approximating the binomial probability function(pf) by the normal distribution. For more details, more explanation and references about the work on this topic see Govindarajulu (1965). Normal approximation of the main discrete distributions (binomial, Poisson, negative binomial, hypergeometric, etc.) was considered by Govindarajulu (1965).

The main comments that usually raised by users of statistics about the approximation of the binomial probability function(bbf) using the normal distribution are:

(1) We are approximating a discrete probability function using a continuous one;

(2) The bbf is skewed (except for p=0.5), while the normal pdf is symmetric.

(3) We need a correction factor to use this method of approximation.

Motivated by maximum entropy distribution, the "Discrete Normal Distribution" was first introduced by Kemp (1997). Consider the probability function(pf) of a discrete random variable X given by:

$$f(x;\lambda,q)=P(X=x)=\frac{\lambda^x q^{x(x-1)/2}}{\sum_{x=-\infty}^{\infty}\lambda^x q^{x(x-1)/2}},\qquad x=0,\pm1,\pm2,\ldots$$

where λ and q are parameters, $\lambda > 0$, $q \in (0,1)$. **Szablowski (2001)** took $q = e^{\frac{-1}{\sigma^2}}$ & $\lambda = e^{\left(\frac{-1}{2\sigma^2}+\frac{\alpha}{\sigma^2}\right)}$, $\alpha \in \Re$,

which gives the discrete normal pdf. Thus, X is discrete normal, $X \sim DN(\mu,\sigma^2)$, if its pdf is

$$f(x) = P(X = x) = \frac{\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sum_{i=-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(i-\mu)^2}{2\sigma^2}}}, \quad x = 0,\pm1,\pm2,...$$

For another discrete analogue of the Laplace distribution, see **Seidu et al. (2004).**

A Binomial r.v. X, $b(n,p)$, has the pf:

$$f(x;n,p) = P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, x = 0,1,...,n \quad \& \ 0 \text{ O.w.}$$

$\mu = np$ & $\sigma^2 = np(1-p)$. Clearly, the variance is smaller than the mean with maximum value of $n/4$ occurs at $p = 0.5$.

It is easy nowadays to find the value of the pf at any value of X using calculators. However, there is a theoretical interest, when teaching statistical courses, to approximate the binomial using some well-known distribution such as normal, Poisson, etc. These approximations are mentioned in almost any text book in statistics and probability. Several ways have been used to approximate the binomial pf. The most popular way is by using the normal distribution with $\mu = np$ & $\sigma^2 = np(1-p)$. This is justified by the central limit theorem (CLT). The following is the simplest form of CLT:

If $X_1, X_2, ..., X_n$ are iid with $E(X_i) = \mu$ & $Var(X_i) = \sigma^2$, $-\infty < \mu < \infty$, $0 < \sigma^2 < \infty$, then as $n \to \infty$, $\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \xrightarrow{d} N(0,1)$.

Now, if $X \sim b(n,p)$, then $X = \sum_{i=1}^{n} X_i$, where X_i's are iid $b(1,p)$. Thus, for large n we have

$$\frac{X - E(X)}{\sqrt{\text{var}(X)}} = \frac{X - np}{\sqrt{np(1-p)}} \approx N(0,1),$$

i.e. for large n,

$$X \approx N(np, np(1-p)).$$

As mentioned above, one problem of this approximation is that a continuous distribution is being used to approximate a discrete one. To get around this problem, what is called "continuity correction" has been used:

$$\Pr(X = k) = \Pr(k - 0.5 \le X \le k + 0.5), k = 0,1,...,n.$$

There are some restrictions on the use of CLT approximation. The two rules of thumb for the approximation to be accurate were introduced and investigated by **Schader & Schmid (1989),** these two rules are:

- $np(1-p) > 9$.
- $np > 5$ for $0 < p \le 0.5$ and $n(1-p) > 5$ for $0.5 < p < 1$.

The shape of the normal pdf is bell shaped. Thus, normal approximation of the binomial pf may not be suitable to use if the shape of the binomial distribution is very skewed; i.e. p is closed to zero or to one. However, it easier to use the standard normal distribution tables especially by beginner students than to use the exact formula of the binomial pf.

Another way to approximate the Binomial distribution was proposed by **Chang et al. (2008)** based on the skew normal distribution:

$$P(X \le k) = F_{n,p}(k) \approx \Psi_{\lambda}\left(\frac{k + 0.5 - \mu}{\sigma}\right),$$

where, $\Psi_\lambda(c)=\int_{-c}^{c}2\varphi(z)\Phi(\lambda z)dz$. The values of the parameters σ^2, μ and λ are chosen suitably based on n & p.

A discrete random variable X has a Poisson distribution if its pf is:

$$f(x;\lambda)=p(X=x)=\frac{\lambda^x e^{-\lambda}}{x!},\quad x=0,1,2,\ldots\ ,\ \lambda>0.$$

$\mu=\sigma^2=\lambda$. If n is large and p is small, with $np=\lambda$ fixed, the terms of $b(n,p)$ are found to be near the Poisson distribution, $P(\lambda)$;

$$\binom{n}{x}p^x(1-p)^{n-x}\approx\frac{(np)^x e^{-np}}{x!}\quad\textbf{(See Filler (1968)).}$$

(For more details see Hogg and Tanis (1996)).

In this paper, the discrete normal is used instead of the normal. In section 2, the approximation of the binomial probability function, bpf, using discrete normal is investigated. Conclusions and some suggested future works are the content of section 3.

2. Approximation of bpf Using Discrete Normal pdf

In this section, we investigate the appropriateness of using the general discrete normal distribution to approximate the bpf. The approximate value is compared with the exact value and with normal approximation.

Let X be a random variable with discrete normal distribution, $DN(\mu,\sigma^2)$. The pdf of X is:

$$f(x;\mu,\sigma^2)=\frac{\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sum_{i=-\infty}^{\infty}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(i-\mu)^2}{2\sigma^2}}},\quad x=0,\pm1,\pm2,\ldots;\quad \sigma>0,\ -\infty<\mu<\infty.$$

Clearly, the most inconvenient part of this pf is its denominator.

Zsablowski (2001) introduced the following general formula:

$$\text{For } 0<q<1,\ \sum_{k=-\infty}^{\infty}q^{k^2/2}=\sqrt{\frac{2\pi}{\ln(1/q)}}\left(1+2\sum_{k=1}^{\infty}\exp\left(-\frac{2\pi^2k^2}{\ln(1/q)}\right)\right).$$

In particular, for $q\ge\exp(-2\pi^2/(7\ln10-\ln5))=0.25653$, he showed that

$2\exp\left(\frac{-2\pi^2}{\ln(1/q)}\right)\le10^{-6}$. So, with high accuracy level (up to the fifth digit), we have

$$\sum_{k=-\infty}^{\infty}q^{k^2/2}=\sqrt{\frac{2\pi}{\ln(1/q)}}.$$

By taking $q=\exp(-1)\cong0.36788>0.25658$, he got with accuracy up to 10^{-6}:

$$\sum_{k=-\infty}^{\infty}\frac{\exp(-k^2/2)}{\sqrt{2\pi}}\approx1,\ \ \&\ \ \sum_{k=-\infty}^{\infty}k^2\frac{\exp(-k^2/2)}{\sqrt{2\pi}}\approx1.$$

Thus, based on these results, if X is $DN(\mu,\sigma^2)$, then for $\sigma^2>0.73$, we have:

$$f(x;\mu,\sigma^2)=\frac{\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sum_{i=-\infty}^{\infty}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(i-\mu)^2}{2\sigma^2}}},\approx\frac{1}{\sigma\sqrt{2\pi}}e^{\frac{-(x-\mu)^2}{2\sigma^2}},\quad x=0,\pm1,\pm2,\ldots,$$

$$E(X) \approx \mu,\ Var(X) \approx \sigma^2 .$$

Let $Y \sim b(n,p)$, as we mentioned in the previous section, by CLT and under some conditions on n and p, the distribution of Y is approximately normal with mean np and variance $np(1-p)$. If $X \sim DN(\mu,\sigma^2)$, by using the methods of moment similar to the work of **Chang et al. (2008),** we have $E(Y)=np$, $Var(Y)=np(1-p)$ equating them to μ and σ^2, we obtained an approximate value of (μ,σ^2) as $\mu^*=np,\ \sigma^{*2}=np(1-p)$. Thus, we can use the $DN(np,np(1-p))$ to approximate the bpf. Thus, for $k=0,1,2,...,n$, we have

$$P(Y=k)=\binom{n}{k}p^k(1-p)^{n-k} \approx f(k;np,np(1-p))=\frac{1}{\sqrt{2\pi np(1-p)}}e^{\frac{-(k-np)^2}{2np(1-p)}}.$$

For different values for n & p, we computed the approximate value for bpf using discrete normal probability function at $X=k$, and compared it with the exact value. The maximum absolute error, $MABE(n,p)$, is used to measure the accuracy of the approximation:

$$MABE(n,p)=\max_{k\in\{0,1,2,...,n\}}\left|P(Y=k)-f(k;np,np(1-p))\right|.$$

The smaller the $MABE(n,p)$, the better the approximation. Also, this approximation is compared with the CLT approximation.

Tables (1-3) contain the numerical calculations for the exact value and the values of the discrete normal and normal approximations for selected values of (p,n); p =0.1, 0.3, 0.5, $n=10$, 30, 50. Since $MABE(n,p)$ is symmetric around $p=0.5$; i.e. $MABE(n,p)=MABE(n,1-p)$, for $DN(np,np(1-p))$ & $N(np,np(1-p))$, there is no need to consider the values of p that are larger than 0.5.

Table (1a). The exact and the approximate values of $p(X=k)$ when $n=10, p=0.1$

k	$B(n,p)$	$DN(np,np(1-p))$	$N(np,np(1-p))$
0	0.3487	0.2413	0.2422
1	0.3874	0.4205	0.4018
2	0.1937	0.2413	0.2422
3	0.0575	0.0456	0.0527
4	0.0112	0.0029	0.0041
≥ 5	0.0016	0.0001	0.0001
$MABE(n,p)$		**0.1074**	**0.1065**

Table (1b). The exact and the approximate values of $p(X=k)$ when $n=30, p=0.1$

k	$B(n,p)$	$DN(np,np(1-p))$	$N(np,np(1-p))$
0	0.0425	0.0458	0.0475
1	0.1413	0.1158	0.1166
2	0.2277	0.2018	0.1998
3	0.2361	0.2428	0.2391
4	0.1771	0.2018	0.1998
5	0.1023	0.1158	0.1166
6	0.0474	0.0459	0.0475
≥7	0.0258	0.0202	0.0166
$MABE(n,p)$		**2. 5911×10⁻²**	**2. 7858×10⁻²**

Table (1c). The exact and the approximate values of $p(X = k)$ when $n = 50, p = 0.1$

k	$B(n, p)$	$DN(np, np(1-p))$	$N(np, np(1-p))$
≤ 2	0.1117	0.1232	0.1145
3	0.1386	0.1206	0.1204
4	0.1809	0.1683	0.1671
5	0.1849	0.1881	0.1863
6	0.1541	0.1683	0.1671
7	0.1076	0.1206	0.1204
≥8	0.1222	0.1298	0.1193
$MABE(n, p)$		**1. 7983×10⁻²**	**1. 8112×10⁻²**

Table (2a). The exact and the approximate values of $p(X = k)$ when $n = 10, p = 0.3$

k	$B(n, p)$	$DN(np, np(1-p))$	$N(np, np(1-p))$
10	0.1398	0.1411	0.1403
11	0.1271	0.1325	0.1319
12	0.1033	0.1098	0.1096
13	0.0755	0.0804	0.0804
14	0.0499	0.0519	0.05216
≥15	0.0607	0.0549	0.0558
$MABE(n, p)$		**7. 0740×10⁻³**	**7. 3598×10⁻³**

Table (2b). The exact and the approximate values of $p(X = k)$ when $n = 30, p = 0.3$

k	$B(n, p)$	$DN(np, np(1-p))$	$N(np, np(1-p))$
≤ 6	0.1595	0.1579	0.1595
7	0.1218	0.1157	0.1154
8	0.1501	0.1468	0.1460
9	0.1573	0.1589	0.1579
10	0.1416	0.1468	0.1460
11	0.1103	0.1157	0.1154
12	0.0748	0.0778	0.0780
≥13	0.0845	0.0802	0.0802
$MABE(n, p)$		**6. 1444×10⁻³**	**6. 4243×10⁻³**

Table (2c). The exact and the approximate values of $p(X = k)$ when $n = 50, p = 0.3$

k	$B(n, p)$	$DN(np, np(1-p))$	$N(np, np(1-p))$
≤ 12	0.2229	0.2193	0.2202
13	0.1050	0.1018	0.1015
14	0.1190	0.1174	0.1170
15	0.1224	0.1231	0.1226
16	0.1147	0.1174	0.1170
17	0.0983	0.1018	0.1015
18	0.0772	0.0802	0.0801
≥19	0.1406	0.1391	0.1400
$MABE(n, p)$		**3. 6270×10⁻³**	**3. 6730×10⁻³**

Table (3a). The exact and the approximate values of $p(X = k)$ when $n = 10, p = 0.5$

k	$B(n, p)$	$DN(np, np(1-p))$	$N(np, np(1-p))$
≤ 2	0.0547	0.0537	0.0567
3	0.1172	0.1134	0.1145
4	0.2051	0.2066	0.2045
5	0.2461	0.2523	0.2482
6	0.2051	0.2066	0.2045
7	0.1172	0.1134	0.1145
≥ 8	0.0547	0.0537	0.0567
$MABE(n, p)$		**6. 2195×10⁻³**	**2. 7198×10⁻³**

Table (3b). The exact and the approximate values of $p(X = k)$ when $n = 30, p = 0.5$

k	$B(n, p)$	$DN(np, np(1-p))$	$N(np, np(1-p))$
≤ 11	0.1002	0.0994	0.1006
12	0.0805	0.0799	0.0800
13	0.1115	0.1116	0.1113
14	0.1354	0.1363	0.1356
15	0.1445	0.1457	0.1449
16	0.1354	0.1363	0.1356
17	0.1115	0.1116	0.1113
18	0.0806	0.0800	0.0800
≥ 19	0.1002	0.0994	0.1006
$MABE(n, p)$		**1. 2087×10⁻³**	**5. 1919×10⁻⁴**

Table (3c). The exact and the approximate values of $p(X = k)$ when $n = 50, p = 0.5$

k	$B(n, p)$	$DN(np, np(1-p))$	$N(np, np(1-p))$
≤ 22	0.2399	0.2390	0.2398
23	0.0960	0.1084	0.0959
24	0.1080	0.1128	0.1081
25	0.1123	0.1084	0.1125
26	0.1080	0.0962	0.1081
27	0.0960	0.0787	0.0959
≥ 28	0.2399	0.2390	0.3357
$MABE(n, p)$		**5. 6274×10⁻⁴**	**2. 4528×10⁻⁴**

Comparison Exact value and Discrete Normal Approximation:

Based on Table 1(a), ($n = 10, p = 0.1$), that *the exact,* $B(n, p)$, and approximate value, $DN(np, np(1-p))$, are close for almost all values of k. The maximum absolute error is 0.1074. The approximation is not so good for small values of k. Similar comments can be said based on Table 1(b), ($n = 30, p = 0.1$), Table 1(c), ($n = 50, p = 0.1$); $MABE(50, 0.1)$ is 0.02591 and $MABE(30, 0.1)$ is 0.01798. It can be seen that $MABE(n, p)$ is decreasing in n for fixed p.

Based on Table 2 (a,b,c), it can be seen that the $MABE(n, 0.2)$ is decreasing in n : $MABE(10, 0.3) = 0.01685$, $MABE(30, 0.3) = 0.00614$, $MABE(50, 0.3) = 0.00363$. Clearly the approximation is very accurate. Based on Table 3(a,b,c); *MABE*(10,0.5)=0.00622, *MAB(*30,0.5)=0.00121, *MAB(*50,0.5)=0.000562. Clearly the approximation is extremely accurate.

Comparison of Discrete Normal and Central Limit Theorem Approximations:

Several ways have been used to approximate the Binomial distribution. The most popular way is the one using the Normal distribution with $\mu = np$ and $\sigma^2 = np(1-p)$. This approximation is justified by the central limit theorem, one of the most important theorems in probability and statistics:

$$\frac{X - E(X)}{\sqrt{\text{var}(X)}} = \frac{X - np}{\sqrt{np(1-p)}} \approx N(0,1),$$

The approximation of the binomial by normal probability function is suitable if $np > 5$, **(See Schader and Schmid (1989)).** Since the normal distribution is continuous and the binomial is discrete and takes only nonnegative integers, the correction factor for continuity is being used:

$$P(X = k) = P(k - 0.5 \le X \le k + 0.5).$$

For example, if X is $b(10,0.1)$, then the exact value of $P(X = 1)$ is

$$P(X = 1) = \binom{10}{1}(0.1)^1(0.9)^9 = 0.3874;$$

while the approximate value using CLT is

$$P(X = 1) = P(0.5 \le X \le 1.5) \cong \int_{0.5}^{1.5} \frac{1}{\sqrt{1.8\pi}} e^{\frac{-1}{1.8}(x-1)^2} = 0.40184.$$

Based on the above tables, we can say that except for very few cases of fluctuations, the approximation using the discrete normal distribution is satisfactory and recommended for use. Furthermore, using the discrete normal probability function is more reasonable and easier.

3. Conclusions and Some Suggested Future Works

In this paper, we considered the discrete normal pdf to approximate the binomial probability function. The choice of this pdf is motivated by CLT and the fact that the binomial random variable is a discrete one. Taking into account the accuracy and easiness as well, the discrete normal strongly recommended for use in practice.

It might of interest to go for some theoretical results. The approximate discrete probability function can be also used to make inference about *p*, when it is unknown. For example, when the distribution of X is approximated by $N(np, np(1-p))$. Based on this approximated distribution an approximate a $(1-\alpha)$% C.I. can be obtained for *p*. Similar things can be done using the discrete pdf.

Other methods of discretizing the normal distribution can be used. For example:

$$p(Y = k) = p(X \le k) - p(X \le k-1) = \Phi\left(\frac{k-\mu}{\sigma}\right) - \Phi\left(\frac{k-1-\mu}{\sigma}\right)$$

is an approach **(See Gomes-Deniz, Vazquez-Polo, Garcia-Garcia (2012)).** Another approach is by using the greatest integer of the continuous random variable; $Y = [X]$. These approaches can be used to obtain discrete normal distributions, which can be investigated for the approximation of the binomial function. Truncated discrete normal is another choice motivated by the binomial r.v. being nonnegative. Since the Poisson pf is the limit of the binomial pf as $n \to \infty, p \to \infty$ and $np = \lambda$, it can be used when p is very small.

References

Chang, C. H., Lin, J. J, Pal, N., & Chiang, M. C. (2008). A Note on Improved Approximation of the Binomial Distribution by the Skew-Normal Distribution. *The American Statistician, 62*(2), 167-170. https://doi.org/10.1198/000313008X305359

Filler, W. (1968). *An Introduction to Probability Theory and its Applications*. (3rd ed.) Wiley Eastern Limited.

Gomiz-Deniz, E., Vazquez-Polo, F., & Garcia-Garcia, V. (2012). A Discrete Version of the Half-Normal Distribution and its Generalization with applications. *Statistical Papers, 55,* 497-511. https://doi.org/10.1007/s00362-012-0494-6

Govindarajulu, Z. (1965). Normal Approximations to the Classical Discrete Distributions. *Sankhyā: The Indian Journal of Statistics, Series A, 27,* 143–72. http://www.jstor.org/stable/25049363

Hogg, R., & Tanis, E. (1996). Probability and Statistical Inference. (5th ed.). Prentice Hall.

Kemp, A. W. (1997). Characterizations of a Discrete Normal Distribution. *Journal of Statistical Planning and Inference, 63*, 223 -229. https://doi.org/10.1016/S0378-3758(97)00020-7

Schader, M., & Schmid, F. (1989). Tow Rules of Thumb for the Approximation of the Binomial Distribution by the Normal Distribution. *The American Statistician, 43*(1), 23-24. https://doi.org/10.1080/00031305.1989.10475601

Seidu, I., & Kozubowski, T. J. (2004). A Discrete analogue of the Laplace distribution. *Journal of Statistical Planning and Inference, 136,* 1090 -1102. https://doi.org/10.1016/j.jspi.2004.08.014

Szablowski, P. J. (2001). Discrete Normal Distribution and its Relationship with Jacobi Theta Function. *Statistics & Probability Letters, 52*, 289-299. https://doi.org/10.1016/S0167-7152(00)00223-6

International Journal of Statistics and Probability; Vol. 11, No. 3; May 2022
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Statistical Reliability of a Diet-Disease Association Meta-analysis

S. Stanley Young[1] and Warren B. Kindzierski[2]

[1] CGStat, Raleigh, NC, USA

[2] Independent consultant, St Albert, Alberta, Canada

Correspondence: Warren B. Kindzierski, 12 Hart Place, St Albert, Alberta, T8N 5R1, Canada.

Received: April 4, 2022 Accepted: April 26, 2022 Online Published: April 27, 2022

doi:10.5539/ijsp.v11n3p40 URL: https://doi.org/10.5539/ijsp.v11n3p40

Abstract

Risk ratios or p-values from multiple, independent studies – observational or randomized – can be pooled to address a common research question in meta-analysis. However, reliability of independent studies should not be assumed as claimed risk factor−disease relationships may fail to reproduce. An independent evaluation was undertaken of a published meta-analysis of cohort studies examining diet−disease associations; specifically between red and processed meat and six disease outcomes (all-cause mortality, cardiovascular mortality, all cancer mortality, breast cancer incidence, colorectal cancer incidence, type 2 diabetes incidence). The number of hypotheses examined were counted in 15 random base papers (14%) of 105 used in the meta-analysis. Test statistics (relative risk values with 95% confidence limits) for 125 results used in the meta-analysis were converted to p-values; p-value plots were used to examine the effect heterogeneity of the p-values. The possible number of hypotheses examined in the 15 base papers was large, median = 20,736 (interquartile range = 1,728–331,776). Each p-value plot for selected health effects showed either a random pattern (p-values > 0.05), or a two-component mixture (small p-values < 0.001 while other p-values appeared random). Given potentially large numbers of hypotheses examined in the base studies, questionable research practices cannot be ruled out as explanations for some test statistics with small p-values. Like the original findings of the published meta-analysis, our independent evaluation concludes that base papers used in the meta-analysis do not support evidence for an association between red and processed meat and the six health effects investigated.

Keywords: cohort studies, red meat, health effects, meta-analysis, multiple testing bias

1. Introduction

Food and dietary intake habits represent a complex system of interacting components that may affect health status and disease over an individual's lifetime. Nutritional epidemiology uses methods to study how diet might affect health status and disease. These methodologies require a strong statistical component to develop useful and interpretable diet−disease associations (Prentice & Huang, 2018). The semi-quantitative food frequency questionnaire (FFQ) – a self-administered dietary assessment instrument – is commonly used to assess dietary intake (Boeing, 2013). A FFQ distributes a structured food list and a frequency response section to study participants, who indicate their usual frequency of intake of each food over a set period of time (Satija et al., 2015).

Causal criteria in nutritional epidemiology include (Potischman & Weed, 1999): consistency, strength of association, dose response, plausibility, and temporality. A longstanding critique of nutritional epidemiology in establishing causality is that it relies predominantly on observational study data, which researchers generally judge to be less reliable than experimental data (Satija et al., 2015). Bias – systematic alteration of research findings due to factors related to study design, data acquisition, statistical analysis, or reporting of results (Boffetta et al., 2008; NASEM, 2016, 2019; Randall & Welser, 2018) – can undermine a study's reliability to apply these causal criteria. Further, selective reporting occurs in published observational studies with researchers routinely testing many hypotheses during a study and then only reporting results that are interesting (i.e., statistically significant) (Gotzsche, 2006; Frieden, 2017).

One aspect of reproducibility – the performance of another study statistically confirming the same hypothesis or claim – is a cornerstone of science and reproducibility of research findings is needed before causal inference can be made (Moonesinghe et al., 2007). However, irreproducible published studies reportedly occur in a wide range of scientific disciplines – including general medicine, clinical sciences, oncology, nutrition, biology, psychological sciences (Young et al., 2022). Incomplete reporting occurs in biomedical research (Dickersin & Chalmers, 2011; Frieden, 2017). These types of situations can lead to an inability to reproduce research claims (Sarewitz, 2012). Part of the problem may arise from researchers examining large numbers of hypotheses and using multiple statistical models without statistical

correction – referred to as multiple testing and multiple modelling or multiple testing bias (Westfall & Young, 1993; Young & Kindzierski, 2019; Young et al., 2022).

Meta-analysis is a systematic procedure for statistically combining data (test statistics) from multiple studies that address a common research question (Egger et al., 2001), such as whether a particular food has an association with a disease. Meta-analysis has been placed at the top of the medical evidence-based pyramid – above case–control and cohort studies, and randomized trials (Murad et al., 2016). However, questions remain about whether the test statistics themselves being combined in meta-analysis may be derived using imperfect or limited statistical methodologies.

As a case in point, Peace et al. (2018) recently examined aspects of multiple testing associated with test statistics combined from ten base papers in a Malik et al. (2010) meta-analysis of sugar-sweetened beverage intake and risk of metabolic syndrome and type 2 diabetes. Peace et al. (2018) observed that none of the base papers in the Malik et al. meta-analysis corrected for multiple testing bias. Given the importance of statistics in developing useful and interpretable risk factor−disease associations, we were interested in understanding whether multiple testing bias might be occurring elsewhere in diet−disease association meta-analysis studies. Specifically, we randomly selected and independently evaluated base studies in a meta-analysis of the association between red and processed meat and selected human chronic effects.

2. Method

2.1 Data Sets

Vernooij et al. (2019) – herein referred to as Vernooij – published a meta-analysis of cohort studies relating to health claims from red and processed meat in the journal *Annals of Internal Medicine*. We selected six of 30 health effects that they examined for further independent evaluation – those that combined the largest number of base papers. These health effects included: all-cause mortality, cardiovascular mortality, all cancer mortality, breast cancer incidence, colorectal cancer incidence, type 2 diabetes incidence. Upon request, one of the Vernooij researchers provided data we used for our evaluation. We then used search space analysis (counting of the numbers of hypotheses examined in base studies) (Peace et al., 2018) and p-value plots (Schweder & Spjøtvoll, 1982) to evaluate the six diet−disease association claims.

Vernooij systematically reviewed 1,501 papers and selected 105 primary papers for further analysis. Their data set included 70 different population cohorts. They used GRADE (*Grading of Recommendations Assessment, Development and Evaluation*) criteria (Guyatt et al., 2008) – which do not assess multiple testing bias – to select base papers for their meta-analysis. Their study complied with recommendations of PRISMA (*Preferred Reporting Items for Systematic Reviews and Meta-Analyses*) (Moher et al., 2009).

Vernooij stated that the base papers used for meta-analysis, which were observational studies, provided low- or very-low-certainty evidence according to GRADE criteria. They concluded "…*dietary patterns with less red and processed meat intake may result in very small reductions in adverse cardiometabolic and cancer outcomes*." Numerous nutritional epidemiologists reacted to their research with some asking the editor of *Annals of Internal Medicine* to withdraw the paper before publication (Monaco, 2019; Arends, 2020).

2.2 Numbers of Hypotheses Tested in Single Studies (Counting)

One needs to estimate the number of hypotheses examined in a single study to assess the potential for multiple testing bias. We selected a subset of studies from Vernooij and counted the possible hypotheses examined in these studies. A 5 to 20% sample from a population whose characteristics are known is considered acceptable for most research purposes as it provides an ability to generalize for the population (Creswell, 2003). We believed the Vernooij judgment that their systematic review (screening) process selected 105 base papers with sufficiently consistent (known) characteristics for meta-analysis. We then randomly selected 15 of the 105 base papers (14%) for counting purposes.

The number of hypotheses considered in an individual base paper used by Vernooij was estimated as follows. Cohort studies generally use a direct statistical analysis strategy on data collected – e.g., what causes or risk factors are related to what outcomes (health effects). If a data set contains "C" causes and "O" outcomes, C × O possible hypotheses can be investigated. An adjustment factor "A" (also called a covariate) can be included as a yes/no adjustment – such as income or education – to see how it can modify each of the C × O hypotheses. Here an adjustment factor is included or excluded; and a multiplier of 2 is assumed for each adjustment factor considered. We counted causes (C), outcomes (O), and yes/no adjustment factors (A); where the number of hypotheses can be approximated as $= C \times O \times 2^{A}$.

We then specifically examined the 15 random base papers for evidence of whether a paper: i) mentioned multiple testing bias in different forms (i.e., multiple hypotheses or hypothesis, multiple testing, multiple comparisons, multiplicity) and/or, ii) made any mention of correcting for this bias.

2.3 P-value Plots

Epidemiologists traditionally use risk statistics (e.g., risk ratios or odds ratios) and confidence intervals instead of p-values from a hypothesis test to establish statistical significance. Given that researchers can estimate risk statistics, confidence intervals and p-values from the same data (Altman & Bland, 2011a,b), one can be estimated from the other. We estimated p-values from risk statistics and confidence intervals for all data used by Vernooij using JMP statistical software (SAS Institute, Cary, NC). We then developed p-value plots (Schweder & Spjøtvoll, 1982) to inspect the distribution of the set of p-values – i.e., the test statistics used by Vernooij.

The p-value is a random variable derived from a distribution of the test statistic used to analyze data and to test a null hypothesis. In a well-designed and conducted study, the p-value is distributed uniformly over the interval 0 to 1 regardless of sample size under the null hypothesis and a distribution of true null hypothesis points plotted against their ranks in a p-value plot should form a 45-degree line when there are no effects (Schweder & Spjøtvoll, 1982; Hung et al., 1997; Bordewijk et al., 2020). Researchers can use the plot to assess the heterogeneity of the test statistics combined in meta-analyses.

The p-value plots were constructed and interpreted as follows:

- Computed p-values were ordered from smallest to largest and plotted against the integers, 1, 2, 3,…
- If p-value points on the plot followed an approximate 45-degree line, we concluded that test statistics resulted from a random (chance) process and the data supported the null hypothesis of no significant association.
- If p-value points on the plot followed approximately a line with a flat/shallow slope, where most (the majority) of p-values were small (< 0.05), then test statistic data set provided evidence for a real, statistically significant, association.
- If numbers of possible hypotheses tested were high in the base studies and p-value points on the plot exhibited a bilinear shape (divided into two lines), the data set of test statistics used for meta-analysis is consistent with a two-component mixture and a general (over-all) claim is not supported. In addition, a small p-value reported for the overall claim in the meta-analysis may not be valid (Schweder & Spjøtvoll, 1982).

Questionable research practices (QRP) involve approaches used by researchers during data collection, analysis, and reporting that may increase false-positive findings in published literature (Ware & Munafò, 2015; Kunert, 2016). P-value plotting is a useful tool to detect the possibility that QRP may have affected test statistics drawn into meta-analysis and rendered the meta-analysis unreliable.

2.4 Numbers of Hypotheses Tested on Cohort Population Data Sets

An interesting problem of multiple testing bias may exist with cohort population data sets. While it is time-consuming and expensive to set up and follow a new cohort, it can be relatively inexpensive to add new measurements and research questions (hypotheses) to an existing cohort. For these reasons, it is possible to have many hypotheses examined on a given cohort as data for the cohort can be used repeatedly. A single published study of a particular cohort data set may only address the tip of the iceberg in terms of numbers of hypotheses examined and multiple testing bias. Collectively there may be numerous other hypotheses at issue when one considers that the same cohort data set can be used many times over for research. Many published papers in literature based on a single cohort data set imply large number of hypotheses examined overall with the possibility of large numbers of false positive (chance) results reported in literature.

First, we wanted to show how common FFQ data is used by researchers investigating health effects. A potential problem is that researchers using FFQs – which are typically utilized in cohort studies – can examine many hypothesis and produce large numbers of false positive (chance) results. We did a Google Scholar (GS) database search to record the approximate number of articles in Web literature with the exact phrase “food frequency questionnaire” and a “[health effect]” mentioned anywhere in an article. We looked at 18 health effects: obesity, inflammation, depression, mental health, all-cause mortality, high blood pressure, lung and other cancers, metabolic disorders, low birth weight, pneumonia, autism, suicide, COPD (chronic obstructive pulmonary disease), ADHD (attention-deficit/hyperactivity disorder), miscarriage, atopic dermatitis, reproductive outcomes, erectile dysfunction.

Second, we did another GS database search to record the approximate number of articles in Web literature using food frequency questionnaire (FFQ) data for each cohort indicated in the 15 selected base papers from Section 2.2. We used the exact phrase “[cohort name]” and the term “FFQ” mentioned anywhere in an article for the search.

3. Results

3.1 Research Questions Asked in Single Studies (Counting)

Table 1 shows the count characteristics of 15 random papers selected from Vernooij. While early food frequency

questionnaire (FFQ) studies used only 61 foods (Willett et al., 1985), these 15 base papers include FFQ−cohort populations examining as many as 280 foods and 32 different health outcomes (Table 1). Summary statistics of the 15 base papers are presented in Table 2. The median number of causes (predictors) was 15 and the median number of adjustment factors (covariates) was 9 in Table 2. These numbers suggest a great scope of the numbers of hypotheses examined (search space).

Table 1. Characteristics of 15 randomly selected papers from Vernooij

Citation#	Base Paper 1st Author	Year	Foods	Outcomes	Causes (Predictors)	Adjustment Factors (Covariates)	Tests	Models	Search Space
8	Dixon	2004	51	3	51	17	153	131,072	20,054,016
31	McNaughton	2009	127	1	22	3	22	8	176
34	Panagiotakos	2009	156	3	15	11	45	2,048	92,160
38	Héroux	2010	18	32	18	9	576	512	294,912
47	Akbaraly	2013	127	5	4	5	20	32	640
48	Chan	2013	280	1	34	10	34	1,024	34,816
49	Chen	2013	39	4	12	5	48	32	1,536
53	Maruyama	2013	40	6	30	11	180	2,048	368,640
56	George	2014	122	3	20	13	60	8,192	491,520
57	Kumagai	2014	40	3	12	8	36	256	9,216
59	Pastorino	2016	45	1	10	6	10	64	640
65	Lacoppidan	2015	192	1	6	16	6	65,536	393,216
80	Lv	2017	12	3	27	8	81	256	20,736
92	Chang-Claude	2005	14	5	3	7	15	128	1,920
99	Tonstad	2013	130	1	4	10	4	1,024	4,096

Note: Citation# is Vernooij reference number; Author name is first author listed for reference; Year = publication year; Foods = # of foods used in Food Frequency Questionnaire; Tests = Outcomes × Causes; Models = 2^A where A = number of Adjustment Factors; Search Space = Tests × Models = approximation of number of hypotheses examined.

Researchers may believe they gain advantage by studying large numbers of outcomes, causes, and adjustment factors (i.e., testing many hypotheses), on the presumption that this maximizes their chances of discovering risk factor−health outcome associations (Willett et al., 1985). However, what they may have maximized is their likelihood of registering a false positive. Given that the conventional threshold for statistical significance in most disciplines is a p-value of less than 0.05, a false positive result should occur 5% of the time by chance alone in a multiple testing setting (Young et al., 2021). The median count of the 15 base papers was 20,736 (refer to Table 2). Five percent of 20,736 possible hypotheses examined in a single FFQ−cohort data set equals 1,037 chance findings that may be mistaken for real results.

Table 2. Characteristics of 15 randomly selected papers from Vernooij

Statistic	Foods	Outcomes	Causes (Predictors)	Adjustment Factors (Covariates)	Tests	Models	Search Space
minimum	12	1	3	3	4	8	176
lower quartile	40	1	8	7	18	96	1,728
median	51	3	15	9	36	512	20,736
upper quartile	129	5	25	11	71	2,048	331,776
maximum	280	32	51	17	576	131,072	20,054,016
mean	93	5	18	9	86	14,149	1,451,216

Note: Foods = # of foods used in Food Frequency Questionnaire; Tests = Outcomes × Causes; Models = 2^A where A = number of Adjustment Factors; Search Space = Tests × Models = approximation of number of hypotheses examined.

In our review of the 15 base papers for evidence of correction for multiple testing bias, thirteen of the papers made no mention of this bias. One paper (Panagiotakos et al., 2009) stated… '*multiple comparisons are made and consequently the probability of false positives findings (i.e., p-value) increases*'. Another paper (George et al., 2014) stated… '*All statistical tests were based on a priori hypotheses; therefore, no adjustment was performed for multiple testing*'.

However, the estimated search space (number of hypotheses examined) is > 490,000 for this paper (refer to Table 1). The only apparent a priori hypotheses stated in their paper were '*how scores on 4 commonly used diet quality indices – the Healthy Eating Index 2010, the Alternative Healthy Eating Index 2010, the Alternate Mediterranean Diet, and the Dietary Approaches to Stop Hypertension – are related to the risks of death from all causes, cardiovascular disease (CVD), and cancer among postmenopausal women*'.

3.2 P-value Plots

The p-value plots for six health outcomes are presented in Figure 1. Each of the six images in Figure 1 indexes rank order (the x axis) and p-value (the y axis). The p values – symbols (circles or triangles) in the body of the six images – are ordered from smallest to largest. The number of p-values in each plot corresponds to the number of studies (base papers) for each of the six outcomes. As noted in the Methods, if there is no effect the p-values will form roughly a 45º line. If the line is flat/shallow with most of the p-values small, then it supports a real effect. Finally, if the shape of the points is bilinear and the counts are high, then the result, i.e., claim, is ambiguous (uncertain) at best.

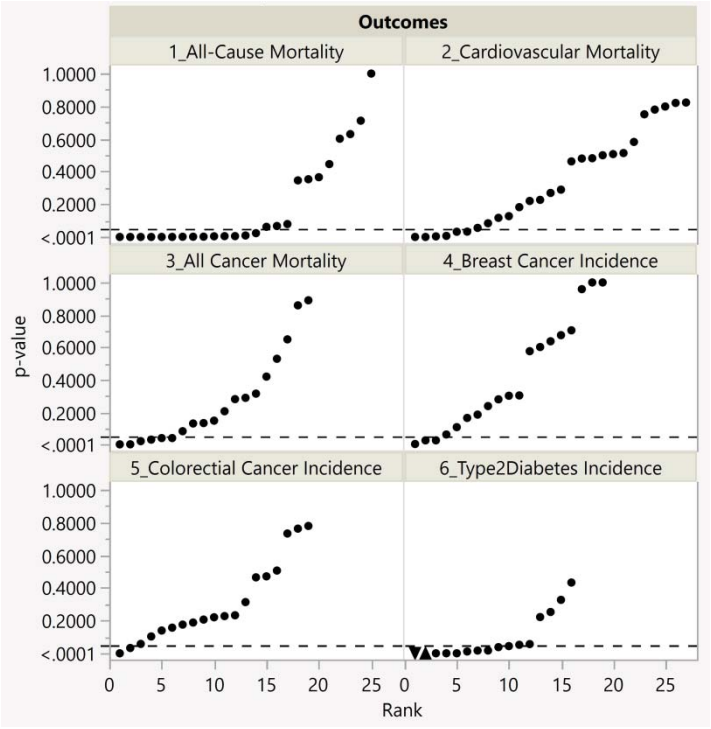

Figure 1. P-value plots (p-value versus rank) for meta-analysis of six health outcomes from Vernooij. Symbols (circles or triangles) are p-values ordered from smallest to largest; triangle pointing downwards (upwards) represents decreasing (increasing) effect

The p-value plots for all-cause mortality, cardiovascular mortality, and all cancer mortality appear bilinear, hence ambiguous. The p-value plots for breast cancer incidence and colorectal cancer incidence appear as 45º lines, suggesting a likelihood of no effect.

The p-value plot for colorectal cancer incidence (bottom left-hand side) is unusual, with the seven largest p-values on a roughly 45º line, two below the 0.05 threshold, and one extremely small p-value (6.2×10^{-5}). Researchers usually take a p-value less than 0.001 as very strong evidence of a real effect (Boos & Stefanski, 2011). Others suggest that small p-values indicate failures of research integrity (Al-Marzouki et al., 2005; Roberts et al., 2007). If the small p-values indicates a real effect, then p-values larger than 0.05 should be rare.

The p-value plot for Type 2 diabetes incidence (bottom right-hand side) has an appearance of a real effect – most of the p-values are small. However, the two smallest p-values – 4.1×10^{-9} and 1.7×10^{-7}, shown as triangles – have conflicting results. The first is for a decrease of effect and the second is for an increase of effect. Our plot might suggest some support for a real association between red or processed meat and Type 2 diabetes—but with a sensible warning of conflicting results of the two smallest p-values. Here we would note a caution about possible failures of research integrity related to the base papers with small p-values as suggested by others (Roberts et al., 2007; Redman, 2013).

Each health outcome presented in Figure 1 displays a wide range of p-value results – refer to Table 3. In the

meta-analysis of breast cancer incidence (middle right-hand side), for example, p-values ranged from < 0.005 to 1 across 19 base papers (> 2 orders of magnitude). In the meta-analysis of Type 2 diabetes incidence (bottom right-hand side), the p-values ranged from < 5 x 10^{-9} to 0.43 (> 7 orders of magnitude) – which suggests possible research integrity issues associated with small p-value results.

Table 3. Minimum and maximum p-values for six health outcomes shown in Figure 1 from Vernooij.

Health outcome	Number of p-values	Minimum p-value	Maximum p-value
All-cause mortality	25	1.6E-08	1
Cardiovascular mortality	27	6.4E-06	0.82
Overall cancer mortality	19	0.00032	0.89
Breast cancer incidence	19	0.0024	1
Colorectal cancer incidence	19	6.2x 10^{-5}	0.78
Type 2 diabetes incidence	16	4.1 x 10^{-9}	0.43

The smallest p-value from Table 6 is 4.1 x 10^{-9} – a value small enough to imply certainty (Boos & Stefanski, 2011). A p value this small may register a true finding – and small p-values are more likely in studies with large sample sizes (Young, 2008). But the wide range of p-values in studies asking the exact same research question – including several studies which register results far weaker than $p < 0.05$ – suggests that alternative explanations cannot be ruled out. These explanations may include some form of QRP – ranging from bias (e.g., alteration of research findings due to factors related to study design, data acquisition, and/or analysis or reporting of results) (Ioannidis, 2008) all the way to data fraud and fabrication (Mojon-Azzi & Mojon, 2004; Eisenach, 2009; George & Buyse, 2015).

3.3 Research Questions Asked of Cohort Populations

Table 4 shows how common FFQ data is used by researchers investigating health outcomes in the Google Scholar literature for 18 health effects we selected (search performed 22 March 2021). Obesity associated with foods is a particular topic of interest with researchers. However, outcomes less commonly expected to be related to foods, e.g., reproductive outcomes and erectile dysfunction, have been investigated.

Table 5 presents the 15 cohorts and an estimate of the number of articles in Google Scholar literature for each cohort using FFQs (search performed 27 May 2021). From Table 5 we suggest that researchers overall may examine many hypotheses on a single cohort−FFQ data set and possibly without proper attention to multiple testing bias. We use the example of the Adventist Health Study-2 cohort data set from Table 5 to demonstrate the potential problem. If 653 studies were published on this cohort population data set using FFQs and each study examined approximately 20,000 hypotheses (i.e., similar to the median number of hypotheses in Table 2), 5% of 653 × 20,000 hypotheses equals 653,000 chance findings that may be mistaken for real results across these studies.

4. Discussion

Regarding red meat−disease association studies, others report that red and processed meat consumption is associated with adverse health effects (e.g., Battaglia et al., 2015; Ekmekcioglu et al., 2015). The International Agency for Research on Cancer (IARC), the cancer agency of the World Health Organization, has classified red meat as probably carcinogenic to humans and processed meat as certainly carcinogenic to humans (WHO, 2015). We have stated previously that performance of another study statistically confirming the same hypothesis or claim is a cornerstone of science. The Vernooij meta-analysis offered scientific explanations against red and processed meat–health effect claims. Our independent findings suggest that the base papers used in Vernooij, properly examined statistically for false positives and possible evidence of QRP (i.e., counting of hypotheses and p-value plots), do not support the reliability of red and processed meat–health effect claims.

Examining large numbers of hypotheses without offering all findings (now possible with supplemental material and web posting) makes it challenging to discover how many true or false-positive versus null findings might exist in a single study (or indeed multiple studies using the same cohort data set). A proposal for reporting meta-analysis of observational studies in epidemiology was provided for researchers in the *Journal of the American Medical Association* (Stroup et al., 2000). This proposal is frequently acknowledged in published literature (15,612 Google Scholar citations as of 1 May 2021). However, this proposal makes no mention of multiple testing bias in observational studies, and it offers no recommendations to control for this bias. Procedures to control multiple testing bias are well-established in literature (some examples include Westfall & Young, 1993; Benjamini & Hochberg, 1995; Schaffer, 1995).

Table 4. Google Scholar search of health effects associated with foods in Web literature

RowID	Outcome (effect) of interest	# of citations
1	obesity	42,600
2	inflammation	23,100
3	depression	18,000
4	mental health	10,900
5	all-cause mortality	10,700
6	high blood pressure	9,470
7	lung and other cancers	7,180
8	metabolic disorders	5,480
9	low birth weight	4,630
10	pneumonia	2,140
11	autism	2,080
12	suicide	1,840
13	COPD	1,800
14	ADHD	1,370
15	miscarriage	1,240
16	atopic dermatitis	938
17	reproductive outcomes	537
18	erectile dysfunction	359

Note: Performed on 22 March 2021; Google Scholar search is only an approximation as Web literature changes rapidly, small changes in search specifications can change results.

Meta-analyses may provide greater evidentiary value if they combine test statistics from base papers that use reliable data and analysis procedures and, crucially, all studies are responding to the same process (Fisher, 1950; DerSimonian & Laird, 1986). Base papers that examine many hypotheses and do not correct for multiple testing bias cannot be considered reliable data for meta-analyses. Furthermore, meta-analyses that combine test statistics from base papers that do and do not correct for this bias are not combining comparable statistics.

Bilinear p-value plots in Figure 1 suggest evidence that nutritional epidemiological meta-analyses have combined test statistics from base studies that do not use comparable methods. Alternately, the bilinear plots may register the existence of one or more powerful covariates correlated with a cause (predictor variable) in some of the studies – that, for example, cardiorespiratory fitness is confounded with dietary risk of mortality (Héroux et al., 2010). However, the existence of an unrecognized covariate would also render meta-analysis' results unreliable.

Large numbers of hypotheses examined in the 15 random base papers of Vernooij – refer to Tables 1 and 2 – make it plausible to infer that some test statistics with small p-values among the base papers may be derived from some form of QRP. The large number of articles resulting from these cohort data sets (Tables 5) supports this.

Epidemiology studies that examine many hypotheses tend to provide results of limited quality for each association due to limited exposure assessment and inadequate information on potential confounders (Savitz & Olshan, 1995). These studies are prone to seek out small but (nominally) significant risk factor–health outcome associations (i.e., those that are less than 0.05) in multiple testing environments. These practices may render research susceptible to reporting false-positives as real results, and to risk mistaking an improperly controlled covariate for a positive association. A set of base studies in a meta-analysis where possible numbers of hypotheses examined are large and whose p-values demonstrate bilinearity in a p-value plot should be regarded as questionable.

We note the following limitations of our methods: counting of the possible number of hypotheses examined is not easy as the statistical details of a base study may be presented anywhere in the article or not at all; the counting formula is only an approximation; we did not include possible interactions among the variables; the use of a p-value plot for evaluation of a meta-analytic result is relatively new; and the Google Scholar searches are only approximations of numbers of articles in Web literature as the Web literature changes rapidly and small changes in search specifications can change results.

Table 5. Cohort study names and estimate of papers in Web literature using FFQs for the 15 randomly sampled base papers of Vernooij

Citation#	Author	Year	Cohort Study Name	Papers, Cohort+FFQ
48	Chan	2013	Mr. Os and Ms Os (Hong Kong)	8
56	George	2014	WHI Women's Health Initiative Observational Study	1,520
49	Chen	2013	HEALS and 'Bangladesh'	1,080
53	Maruyama	2013	JACC Japan Collaborative Cohort	758
57	Kumagai	2014	NHI Ohsaki National Health Insurance Cohort	122
47	Akbaraly	2013	Whitehall II study	1,800
99	Tonstad	2013	Adventist Health Study-2	653
80	Lv	2017	China Kadoorie Biobank	143
59	Pastorino	2016	MRC National Survey of Health and Development	148
31	McNaughton	2009	Whitehall II study	1,800
34	Panagiotakos	2009	ATTICA Study	1,650
8	Dixon	2004	DIETSCAN (Dietary Patterns and Cancer Project)	1,080
38	Héroux	2010	ACLS (Aerobics Center Longitudinal Study)	167
65	Lacoppidan	2015	Diet, Cancer, and Health (DCH) cohort	116
92	Chang-Claude	2005	German vegetarian study	13

Note: Google Scholar search performed 17 May 2021; Citation# = Vernooij reference number; Author name = first author listed for reference; Year = publication year; Cohort Name = name of study cohort; Papers, Cohort + FFQ = # of papers in literature mentioning study cohort using a Food Frequency Questionnaire (FFQ); Google Scholar search is only an approximation as Web literature changes rapidly, small changes in search specifications can change results.

5. Findings

We independently evaluated the Vernooij meta-analysis. Specifically, we examined properties of the test statistics that were combined to derive meta-analytic statistical associations between red and processed meat and all-cause mortality, cardiovascular mortality, all cancer mortality, breast cancer incidence, colorectal cancer incidence, type 2 diabetes incidence. The possible number of hypotheses examined in 15 random base papers we evaluated was large, median = 20,736 (interquartile range = 1,728–331,776). Each p-value plot of the test statistics for selected health effects we evaluated showed either a random pattern (p-values > 0.05), or a two-component mixture with small p-values < 0.001 while other p-values appeared random. Given potentially large numbers of hypotheses examined in the base papers, questionable research practices cannot be ruled out as explanations for test statistics with small p-values. Given this evidence, we conclude that: i) our statistical examination does not support the reliability of red meat−negative health claims, and ii) the Vernooij finding – *…dietary patterns with less red and processed meat intake may result in very small reductions in adverse cardiometabolic and cancer outcomes* – is reliable.

Acknowledgments

We thank Bradley Johnston (Department of Nutrition, College of Agriculture and Life Sciences, Texas A&M University, College Station, TX) for providing us with the datasets used in our research. No external funding was provided for this study. The study was conceived based on previous work undertaken by CG Stat for the National Association of Scholars (nas.org), New York, NY.

References

Al-Marzouki, S., Evans, S., & Marshall, T., et al. (2005). Are these data real? Statistical methods for the detection of data fabrication in clinical trials. *British Medical Journal*, *331*, 267. https://doi.org/10.1136/bmj.331.7511.267

Altman, D. G., & Bland, J. M. (2011a). How to obtain a confidence interval from a P value. *British Medical Journal, 343*, d2090. https://doi.org/10.1136/bmj.d2090

Altman, D. G., & Bland, J. M. (2011b). How to obtain the P value from a confidence interval. *British Medical Journal, 343*, d2304. https://doi.org/10.1136/bmj.d2304

Arends, B. (2020). *'Totally bizarre!'—nutritionists see red over study downplaying the serious health risks of red meat*. MarketWatch. https://www.marketwatch.com/story/nutritionists-see-red-over-study-downplaying-the-health-risks-of-red-meat-20

19-10-02

Battaglia Richi, E., Baumer, B., & Conrad, B., et al. (2015). Health risks associated with meat consumption: A review of epidemiological studies. *International Journal for Vitamin and Nutrition Research*, *85*(1−2), 70−78. http://dx.doi.org/10.1024/0300-9831/a000224

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, *57*(1), 125−133. http://dx.doi.org/10.1111/j.2517-6161.1995.tb02031.x

Boeing H. (2013). Nutritional epidemiology: New perspectives for understanding the diet-disease relationship? *European Journal of Clinical Nutrition*, *67*(5), 424−429. http://dx.doi.org/0.1038/ejcn.2013.47

Boffetta, P., McLaughlin, J. K., & Vecchia, C. L., et al. (2008). False-positive results in cancer epidemiology: A plea for epistemological modesty. *Journal of the National Cancer Institute*, *100*, 988−995. http://dx.doi.org/10.1093/jnci/djn191

Boos, D. D., & Stefanski, L. A. (2011). P-value precision and reproducibility. *The American Statistician*, *65*(4), 213−221. https://doi.org/10.1198/tas.2011.10129

Bordewijk, E. M., Wang, R., & Aski, L. M., et al. (2020). Data integrity of 35 randomised controlled trials in women' health. *European Journal of Obstetrics & Gynecology and Reproductive Biology*, *249*, 72−83. http://dx.doi.org/10.1016/j.ejogrb.2020.04.016

Creswell, J. (2003). *Research Design-Qualitative, Quantitative and Mixed Methods Approaches, 2nd ed*. Thousand Oaks, CA: Sage Publications.

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, *7*, 177–188. https://doi.org/10.1016/0197-2456(86)90046-2

Dickersin, K., & Chalmers, I. (2011). Recognizing, investigating and dealing with incomplete and biased reporting of clinical research: From Francis Bacon to the WHO. *Journal of the Royal Society of Medicine*, *104*, 532−538. http://dx.doi.org/10.1258/jrsm.2011.11k042

Egger, M., Davey Smith, G., & Altman, D. G. (2001). Problems and limitations in conducting systematic reviews. In: Egger, M., Davey Smith, G., & Altman, D. G. (eds.) *Systematic reviews in health care: Meta−analysis in context, 2nd ed*. London: BMJ Books. https://doi.org/10.1002/9780470693926

Eisenach, J. C. (2009). Data fabrication and article retraction: how not to get lost in the woods. *Anesthesiology*, *110*(5), 955−956. http://dx.doi.org/10.1097/ALN.0b013e3181a06bf9

Ekmekcioglu, C., Wallner, P., & Kundi, M., et al. (2018). Red meat, diseases, and healthy alternatives: A critical review. *Critical Reviews in Food Science and Nutrition*, *8*(2), 247−261. http://dx.doi.org/10.1080/10408398.2016.1158148

Fisher, R. A. (1950). *Statistical Methods for Research Workers, 11th ed*. pp 99−101. Edinburgh: Oliver and Boyd.

Frieden, T. R. (2017). Evidence for health decision making – beyond randomized, controlled trials. *New England Journal of Medicine*, *377*(5), 465−475. http://dx.doi.org/10.1056/NEJMra1614394

George, S. L., & Buyse, M. (2015), Data fraud in clinical trials. *Clinical Investigation (London)*, *5*(2), 161−173. http://dx.doi.org/10.4155/cli.14.116

George, S. M., Ballard-Barbash, R., & Manson, J. E., et al. (2014). Comparing indices of diet quality with chronic disease mortality risk in postmenopausal women in the Women's Health Initiative Observational Study: Evidence to inform national dietary guidance. *American Journal of Epidemiology*, *180*(6), 616−625. http://dx.doi.org/10.1093/aje/kwu173

Gotzsche, P. C. (2006). Believability of relative risks and odds ratios in abstracts: Cross sectional study. *British Medical Journal, 333*, 231−234. http://dx.doi.org/10.1136/bmj.38895.410451.79

Guyatt, G. H., Oxman, A. D., & Vist, G. E., et al; GRADE Working Group. (2008). GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *British Medical Journal, 336*, 924. http://dx.doi.org/10.1136/bmj.39489.470347.AD

Héroux, M., Janssen, I., & Lam, M., et al. (2010). Dietary patterns and the risk of mortality: Impact of cardiorespiratory fitness. *International Journal of Epidemiology*, *39*(1), 197−209. http://dx.doi.org/10.1093/ije/dyp191

Hung, H. M. J., O'Neill, R. T., & Bauer, P., et al. (1997). The behavior of the p-value when the alternative hypothesis is true. *Biometrics*, *53*, 11–22. https://doi.org/10.2307/2533093

Ioannidis, J. P. (2008). Why most discovered true associations are inflated. *Epidemiology*, *19*(5), 640−648. http://dx.doi.org/10.1097/EDE.0b013e31818131e7

Kunert, R. (2016). Internal conceptual replications do not increase independent replication success. *Psychonomic Bulletin & Review*, *23*(5), 1631−1638. http://dx.doi.org/10.3758/s13423-016-1030-9

Malik, V. S., Popkin, B.,& Bray, G., et al. (2010). Sugar-sweetened beverages and risk of metabolic syndrome and type 2 diabetes. *Diabetes Care*, *33*, 2477−2483. http://dx.doi.org/10.2337/dc10-1079

Moher, D., Liberati, A., & Tetzlaff, J., et al. (2009). The PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine*, *6*, e1000097. http://dx.doi.org/10.1371/journal.pmed.1000097

Mojon-Azzi, S. M., & Mojon, D. S. (2004). Scientific misconduct: from salami slicing to data fabrication. *Ophthalmic Research*, *36*(1), 1−3. http://dx.doi.org/10.1159/000076104

Monaco, K. (2019). *Is everything we know about meat consumption wrong?* Medpage Today. https://www.medpagetoday.com/primarycare/dietnutrition/82492

Moonesinghe, R., Khoury, M. J., & Janssens, A. C. J. W. (2007). Most published research findings are false−But a little replication goes a long way. *PLoS Medicine*, *4*(2), e28. http://dx.doi.org/10.1371/journal.pmed.0040028

Murad, M. H., Asi, N., & Alsawas, M., et al. (2016). New evidence pyramid. *BMJ Evidence-Based Medicine*, *21*(4), 125−127. http://dx.doi.org/10.1136/ebmed-2016-110401

National Academies of Sciences, Engineering, and Medicine (NASEM). (2016). *Statistical challenges in assessing and fostering the reproducibility of scientific results: Summary of a workshop*. Washington, DC: The National Academies Press. https://doi.org/10.17226/21915

National Academies of Sciences, Engineering, and Medicine (NASEM). (2019). *Reproducibility and replicability in science*. Washington, DC: The National Academies Press. https://doi.org/10.17226/25303

Panagiotakos, D., Pitsavos, C., & Chrysohoou, C., et al. (2009). Dietary patterns and 5-year incidence of cardiovascular disease: a multivariate analysis of the ATTICA study. *Nutrition, Metabolism & Cardiovascular Diseases*, *19*(4), 253−263. http://dx.doi.org/10.1016/j.numecd.2008.06.005

Peace, K. E., Yin, J. J., & Rochani, H., et al. (2018). A serious flaw in nutrition epidemiology: A meta-analysis study. *International Journal of Biostatistics*, *14*(2), pp. 20180079. http://dx.doi.org/10.1515/ijb-2018-0079

Prentice, R. L., & Huang, Y. (2018). Nutritional epidemiology methods and related statistical challenges and opportunities. Statistical *Theory and Related Fields*, *2*(1), 2−10. http://dx.doi.org/10.1080/24754269.2018.1466098

Potischman, N., & Weed, D. L. (1999). Causal criteria in nutritional epidemiology. *American Journal of Clinical Nutrition*, *69*(6), 1309S−1314S. http://dx.doi.org/10.1093/ajcn/69.6.1309S

Randall, D., & Welser, C. (2018). *The irreproducibility crisis of modern science: Causes, consequences, and the road to reform*. New York, NY: National Association of Scholars. https://www.nas.org/reports/the-irreproducibility-crisis-of-modern-science/full-report

Redman, B. K. (2013). *Research Misconduct Policy in Biomedicine; Beyond the Bad Apple Approach*. Cambridge, MA: The MIT Press.

Roberts, I., Smith, R., & Evans, S. (2007). Doubts over head injury studies. *British Medical Journal*, *334*, 392. http://dx.doi.org/10.1136/bmj.39118.480023.BE

Sarewitz, D. (2012). Beware the creeping cracks of bias. *Nature*, *485*, 149. https://doi.org/10.1038/485149a

Satija, A., Yu, E., Willett, W. C., et al. (2015). Understanding nutritional epidemiology and its role in policy. *Advances in Nutrition*, *6*(1), 5−18. http://dx.doi.org/10.3945/an.114.007492

Savitz, D. A., & Olshan, A. F. (1995). Multiple comparisons and related issues in the interpretation of epidemiologic data. *American Journal of Epidemiology*, *142*, 904−908. http://dx.doi.org/10.1093/oxfordjournals.aje.a117737

Schaffer, J. P. (1995). Multiple hypothesis testing. *Annual Review of Psychology*, *46*, 561−584. http://dx.doi.org/10.1146/annurev.ps.46.020195.003021

Schweder, T., & Spjøtvoll, E. (1982). Plots of p-values to evaluate many tests simultaneously. *Biometrika*, *69*, 493−502. https://doi.org/10.1093/biomet/69.3.493

Stroup, D. F., Berlin, J. A., & Morton, S. C., et al. (2000). Meta-analysis of observational studies in epidemiology: A proposal for reporting. *Journal of the American Medical Association*, *283*(15), 2008−2012.

http://dx.doi.org/10.1001/jama.283.15.2008

Vernooij, R. W. M., Zeraatkar, D., & Han, M. A., et al. (2019). Patterns of red and processed meat consumption and risk for cardiometabolic and cancer outcomes A systematic review and meta-analysis of cohort studies. *Annals of Internal Medicine*, *171*, 732−741. http://dx.doi.org/10.7326/M19-1583

Ware, J. J., & Munafò, M. R. (2015). Significance chasing in research practice: causes, consequences and possible solutions. *Addiction*, *110*(1), 4−8. http://dx.doi.org/10.1111/add.12673

Westfall, P. H., & Young, S. S. (1993). *Resampling-based multiple testing*. New York, NY: John Wiley & Sons.

Willett, W. C., Sampson, L., & Stampfer, M. J., et al. (1985). Reproducibility and validity of a semiquantitative food frequency questionnaire. *American Journal of Epidemiology*, *122*, 51−65. http://dx.doi.org/10.1093/oxfordjournals.aje.a114086

World Health Organization (WHO). (2015). *IARC Monographs evaluate consumption of red meat and processed meat. Press release No. 240*. Lyon, France: International Agency for Research on Cancer (IARC), World Health Organization. https://www.iarc.who.int/wp-content/uploads/2018/07/pr240_E.pdf

Young, S. S. (2008). *Statistical analyses and interpretation of complex studies*. Medscape. https://www.medscape.org/viewarticle/571523

Young, S.S., Cheng, K.-C., & Chen, J. H., et al. (2022). Reliability of a meta-analysis of air quality−asthma cohort studies. *International Journal of Statistics and Probability*, *11*(2), 61−76. https://doi.org/10.5539/ijspv11n2p61

Young, S. S., & Kindzierski, W. B. (2019). Evaluation of a meta-analysis of air quality and heart attacks, a case study. *Critical Reviews in Toxicology*, *49(1)*, 85−94. https://doi.org/10.1080/10408444.2019.1576587

Young, S. S., Kindzierski, W. B., & Randall, D. (2021). *Shifting Sands, Unsound Science and Unsafe Regulation Report 1. Keeping Count of Government Science: P-Value Plotting, P-Hacking, and PM2.5 Regulation*. New York, NY: National Association of Scholars. https://www.nas.org/reports/shifting-sands-report-i

International Journal of Statistics and Probability; Vol. 11, No. 3; May 2022
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Reviewer Acknowledgements

International Journal of Statistics and Probability wishes to acknowledge the following individuals for their assistance with peer review of manuscripts for this issue. Their help and contributions in maintaining the quality of the journal is greatly appreciated.

Many authors, regardless of whether *International Journal of Statistics and Probability* publishes their work, appreciate the helpful feedback provided by the reviewers.

Reviewers for Volume 11, Number 3

Abayneh Fentie, Hawassa University, Ethiopia

Abdullah A. Smadi, Yarmouk University, Jordan

Emmanuel Akpan, Federal School of Medical Laboratory Technology, Nigeria

Frederic Ouimet, McGill University, USA

Gabriel A. Okyere, Kwame Nkrumah University of Science and Technology, Ghana

Nahid Sanjari Farsipour, Alzahra University, Iran

Wendy Smith

On behalf of,

The Editorial Board of *International Journal of Statistics and Probability*

Canadian Center of Science and Education

ACADEMIC AND OPEN ACCESS

Canadian Center of Science and Education

CCSENET.ORG

➢ CALL FOR MANUSCRIPTS

International Journal of Statistics and Probability is a peer-reviewed journal, published by Canadian Center of Science and Education. The journal publishes research papers in all aspects of statistics and probability. The journal is available in electronic form in conjunction with its print edition. All articles and issues are available for free download online.

We are seeking submissions for forthcoming issues. All manuscripts should be written in English. Manuscripts from 3000–8000 words in length are preferred. All manuscripts should be prepared in LaTeX or MS-Word format, and submitted online, or sent to: ijsp@ccsenet.org

Paper Selection and Publishing Process

a) Submission acknowledgement. If you submit manuscript online, you will receive a submission acknowledgement letter sent by the online system automatically. For email submission, the editor or editorial assistant sends an e-mail of confirmation to the submission's author within one to three working days. If you fail to receive this confirmation, please check your bulk email box or contact the editorial assistant.

b) Basic review. The editor or editorial assistant determines whether the manuscript fits the journal's focus and scope. And then check the similarity rate (CrossCheck, powered by iThenticate). Any manuscripts out of the journal's scope or containing plagiarism, including self-plagiarism are rejected.

c) Peer Review. We use a double-blind system for peer review; both reviewers' and authors' identities remain anonymous. The submitted manuscript will be reviewed by at least two experts: one editorial staff member as well as one to three external reviewers. The review process may take four to ten weeks.

d) Make the decision. The decision to accept or reject an article is based on the suggestions of reviewers. If differences of opinion occur between reviewers, the editor-in-chief will weigh all comments and arrive at a balanced decision based on all comments, or a second round of peer review may be initiated.

e) Notification of the result of review. The result of review will be sent to the corresponding author and forwarded to other authors and reviewers.

f) Pay the article processing charge. If the submission is accepted, the authors revise paper and pay the article processing charge (formatting and hosting).

g) E-journal is available. E-journal in PDF is available on the journal's webpage, free of charge for download. If you need the printed journals by post, please order at http://www.ccsenet.org/journal/index.php/ijsp/store/hardCopies.

h) Publication notice. The authors and readers will be notified and invited to visit our website for the newly published articles.

More Information

E-mail: ijsp@ccsenet.org

Website: http://ijsp.ccsenet.org

Paper Submission Guide: http://ijsp-author.ccsenet.org

Recruitment for Reviewers: http://www.ccsenet.org/journal/index.php/ijsp/editor/recruitment

➢ JOURNAL STORE

To order back issues, please contact the journal editor and ask about the availability of journals. You may pay by credit card, PayPal, and bank transfer. If you have any questions regarding payment, please do not hesitate to contact the journal editor or editorial assistant.

Price: $40.00 USD/copy Shipping fee: $20.00 USD/copy

ABOUT CCSE

The Canadian Center of Science and Education (CCSE) is a private for-profit organization delivering support and services to educators and researchers in Canada and around the world.

The Canadian Center of Science and Education was established in 2006. In partnership with research institutions, community organizations, enterprises, and foundations, CCSE provides a variety of programs to support and promote education and research development, including educational programs for students, financial support for researchers, international education projects, and scientific publications.

CCSE publishes scholarly journals in a wide range of academic fields, including the social sciences, the humanities, the natural sciences, the biological and medical sciences, education, economics, and management. These journals deliver original, peer-reviewed research from international scholars to a worldwide audience. All our journals are available in electronic form in conjunction with their print editions. All journals are available for free download online.

Mission

To work for future generations

Values

Scientific integrity and excellence

Respect and equity in the workplace

CONTACT US

1595 Sixteenth Ave, Suite 301,
Richmond Hill, Ontario, L4B 3N9,
Canada
Tel: 1-416-642-2606
E-mail: info@ccsenet.org
Website: www.ccsenet.org

The journal is peer-reviewed
The journal is open-access to the full text
The journal is included in:

Aerospace Database
BASE (Bielefeld Academic Search Engine)
EZB (Elektronische Zeitschriftenbibliothek)
Google Scholar
JournalTOCs
Library and Archives Canada
LOCKSS
MIAR
PKP Open Archives Harvester
SHERPA/RoMEO
Standard Periodical Directory
Ulrich's

International Journal of Statistics and Probability
Bimonthly

Publisher Canadian Center of Science and Education
Address 1595 Sixteenth Ave, Suite 301, Richmond Hill, Ontario, L4B 3N9, Canada
Telephone 1-416-642-2606
E-mail ijsp@ccsenet.org
Website http://ijsp.ccsenet.org

ISSN 1927-7032