EKOIST Journal of Econometrics and Statistics, 38, 75-86

DOI: 10.26650/ekoist.2023.38.1159543
http://ekoist.istanbul.edu.tr

İSTANBUL ÜNİVERSİTESİ YAYINEVİ

EKOIST Journal of Econometrics and Statistics

Received: 10.08.2022
Revision Requested: 09.12.2022
Last Revision Received: 17.12.2022
Accepted: 21.12.2023
Online yayın: 23.05.2023

RESEARCH ARTICLE / ARAŞTIRMA MAKALESİ

Semi-Parametric Modeling of Churn Confounding Competing Risks Using Time-Dependent Covariates Among Mobile Phone Subscribers in Kenya

Ndilo B. Fwaru*, Leonard K. Alii, Jerita J. Mwambi*****

Abstract
Mobile phone service providers are currently experiencing high churn rates. As a result, service providers are trying to develop ways to predict churn rates and uncover why subscribers' churn occurs. However, the task of predicting churn in the mobile phone industry is complicated due to the large, sparse, and unbalanced nature of the data especially when competing risks are confounded by time-dependent covariates.

This paper aims to develop a semi-parametric model (the adjusted Cox model) by adjusting the extended Cox proportional hazards model to model competing risks confounded by time-dependent covariates and uses data from three mobile phone service providers in Mombasa and Kilifi Counties in Kenya to analyze and evaluate the validity and performance of the model.

The paper establishes that the adjusted Cox model is a better model for predicting subscriber's survival outcomes as well as for detecting the most influential covariates when competing risks are confounded with time-dependent covariates.

Keywords
Semi-parametric, Competing Risks, Time-dependent Covariates, Churn

* Corresponding author: Ndilo B. Fwaru (Mr.), Pwani University, School of Pure and Applied science, Dapartment of Mathematics and Computer Science, Kilifi, Kenya. E-mail: nbfwaru@gmail.com ORCID: 0000-0003-2709-6455

** Leonard K. Alii (Dr.), Pwani University, School of Pure and Applied science, Dapartment of Mathematics and Computer Science, Kilifi, Kenya. E-mail: l.alii@pu.ac.ke ORCID: 0009-0009-5059-1834

*** Jerita J. Mwambi, (Dr.), Pwani University, School of Pure and Applied science, Dapartment of Mathematics and Computer Science, Kilifi, Kenya. E-mail: j.mwambi@pu.ac.ke ORCID: 0000-0001-6506-0635

To cite this article: Fwaru, N.B., Alii, L.K., & Mwambi, J.J. (2023). Semi-Parametric modeling of churn confounding competing risks using time-dependent covariates among mobile phone subscribers in Kenya. *EKOIST Journal of Econometrics and Statistics*, 38, 75-86. https://doi.org/10.26650/ekoist.2023.38.1159543

1. Introduction

For many mobile phone service providers, finding reasons for losing subscribers, measuring subscriber loyalty, and regaining subscribers are issues of concern. The service providers organize a variety of research and campaigns to avoid losing subscribers.

The ability to predict whether a particular subscriber is at high risk of churn when there is still time to act represents a significant additional potential source of revenue for the service providers. (*Customer Churn Prediction & Prevention Model*, 2021, December 22). Predictive churn modeling techniques seek to understand subscriber behavior and attributes that indicate the risk and timing of a customer's churn.

If there are no competing risks and only time-invariant covariates are used, the Cox proportional hazards model (Cox, 1972) is expressed as follows:

$$\lambda(\mathrm{t}) = \lambda_0(\mathrm{t})\exp\{\beta^{\mathrm{T}}\mathbf{X}\} \tag{1}$$

and can be used to estimate the probability of an event occurring at any time. However, the assumption that the hazard is constant, prevents us from integrating time-dependent variables into the model (Austin et al., 2019).

To model competing risks, Prentice et al. (1978) adapted Equation 1 to

$$\lambda_{\mathrm{j}}(\mathrm{t}|\mathbf{X}) = \lambda_{j0}(\mathrm{t})\exp\{\beta_{\mathrm{j}}{}^{\mathrm{T}}\mathbf{X}\} \tag{2}$$

where *j* represents a competing risk, to analyze the cause-specific hazard functions based on different event types. The cause-specific approach, however, suffers from the assumption of independent censoring for subjects that are not censored but failed due to competing events.

To overcome this, Fine and Gray (1999) developed the following equation:

$$\tilde{\lambda}_{\mathrm{j}}(\mathrm{t}|\mathbf{X}) = \tilde{\lambda}_{j0}(\mathrm{t})\exp\{\beta_{\mathrm{j}}{}^{\mathrm{T}}\mathbf{X}\} \tag{3}$$

to model the cumulative incidence function by imposing the proportional hazards assumption onto the sub-distribution hazards. The Fine and Gray model, however, can only model competing risks using time-invariant covariates.

For time-dependent covariates, Therneau and Grambsch (2000) developed the extended Cox model as follows:

$$\lambda_j(t|\boldsymbol{X}(\boldsymbol{t})) = \lambda_0(t)exp\{\beta_j{}^T\boldsymbol{X}(\boldsymbol{t}) + \gamma_j{}^T\boldsymbol{X}\} \tag{4}$$

The model, however, produces inaccurate estimates in the presence of competing risks and therefore cannot be used to model competing risks.

Beyersmann and Schumacher (2008) proposed ad hoc approaches within the sub-distribution in a time-dependent framework to extrapolate the internal time-dependent covariates. However, the use of such techniques leads to an implicit definition of the sub-distribution hazard that may be difficult to interpret.

This paper develops a semi-parametric model of churn for when competing risks are confounded with time-dependent covariates.

The paper also contributes to the existing literature on modeling churn when competing risks are confounded with time-dependent covariates and also forms the basis for further study in other areas such as employee turnover within a business, components, and equipment longevity, duration of unemployment, and cause of death among patients when competing risks are confounded with time-dependent covariates.

The paper is organized into five sections as follows: Section 1 provides the introduction, Section 2 presents the methodology, Section 3 explains the data and empirical results, and Section 4 shows the findings. Section 5 discusses the conclusions.

2. Methodology

To develop a semi-parametric model (-adjusted Cox model-) for modeling, detecting the most influential time-dependent covariates and predicting subscriber's survival outcomes when competing risks are confounded with time-dependent covariates, Equation (1) is adjusted as:

$$\lambda_j(t|\mathbf{X}(t)) = \lambda_j(t)\frac{\exp\left\{(\beta_j+\gamma_j)^T\mathbf{X}(t)\right\}}{\exp\left(\beta_j{}^T\mathbf{X}(t)\right)F(t)+\exp\left(\gamma_j{}^T\boldsymbol{X}(t)\right)S(t)} \tag{5}$$

where

X(t) represents the time-dependent covariates of age, marital status, occupation, education level, and residence; *j* represents competing risks such as network quality churn, service quality churn, price sensitivity churn, carrier responsiveness churn, and fraud churn; β_j and γ_j are regression coefficients; $S(t)$ is the baseline survival function calculated as:

$$S(t) = \prod_{j=1}^{5} S_j(t) \tag{6}$$

$F(t)$is the baseline cumulative distribution function calculated as:

$$F(t) = \sum_{j=1}^{5}\left(\int_0^t \Lambda_j(u)S(u)du\right) \tag{7}$$

where $\Lambda_j(t)$ is the baseline cumulative hazard for the j^{th} cause, and Equation (5) is the adjusted Cox model.

The unknown parameters $\theta=\{(\beta_j,\gamma_j,\Lambda_j),j=1,\ldots,5\}$ in Equation (5) are estimated as follows: For n = 6000 observations, let t_i=830 weeks be the observation time, d_i the churn indicator (1 if a subscriber churned, 0 if censored), j_i j_i the cause of churn index (takes a value of between 1 and m = 5 for subscriber churn and is undefined for censored cases), and $x_i(t)$ the vector of time-dependent covariates, then the likelihood function for the unknown parameters of

$\theta=\{(\beta_j,\gamma_j,\Lambda_j),j=1,\ldots,5\}$ would be:

$$L_n(\theta)=\prod_{i=1}^{6000}\prod_{j=1}^{5}\lambda_j(t_i,x_i(t))^{d_{ij}}exp\{-\Lambda_j(t_i,x_i(t))\} \tag{8}$$

where

$$\Lambda_j(t_i,x_i(t))=\int_0^t\frac{exp\{(\beta_j+\gamma_j)^T\boldsymbol{X}(s)\}}{exp(\beta_j{}^T\boldsymbol{X}(s))F(s)+exp(\gamma_j{}^T\boldsymbol{X}(s))S(s)}d(s) \tag{9}$$

The logarithm of the likelihood function then becomes:

$$\ell_n(\theta)=\sum_{i=1}^{6000}\sum_{j=1}^{5}\lambda_j(t)^{d_{ij}}+d_{ij}(\beta_j+\gamma_j)^T\boldsymbol{X}(\boldsymbol{t})-\ell_n\left(exp(\beta_j{}^T\boldsymbol{X}(\boldsymbol{t}))F(t)+exp(\gamma_j{}^T\boldsymbol{X}(\boldsymbol{t}))S(t)\right)-60000t \tag{10}$$

with the vector θ being obtained by maximizing the likelihood function as

$$\frac{\partial}{\partial\lambda_j}\ell_n(\theta)=\sum_{i=1}^{6000}\sum_{j=1}^{5}d_{ij}\lambda_j(t)^{d_{ij}-1}=0, \tag{11}$$

$$\frac{\partial}{\partial\beta_j}\ell_n(\theta)=\sum_{i=1}^{6000}\sum_{j=1}^{5}d_{ij}X(\boldsymbol{t})-\left(\frac{exp(\beta_j{}^T\boldsymbol{X}(\boldsymbol{t}))F(t)}{exp(\beta_j{}^T\boldsymbol{X}(\boldsymbol{t}))F(t)+exp(\gamma_j{}^T\boldsymbol{X}(\boldsymbol{t}))S(t)}\right)=0 \tag{12}$$

and

$$\frac{\partial}{\partial\gamma_j}\ell_n(\theta)=\sum_{i=1}^{6000}\sum_{j=1}^{5}d_{ij}\boldsymbol{X}(\boldsymbol{t})-\left(\frac{exp(\gamma_j{}^T\boldsymbol{X}(\boldsymbol{t}))S(t)}{exp(\beta_j{}^T\boldsymbol{X}(\boldsymbol{t}))F(t)+exp(\gamma_j{}^T\boldsymbol{X}(\boldsymbol{t}))S(t)}\right)=0 \tag{13}$$

The estimates of the unknown parameters, $\hat{\theta}=(\hat{\lambda_j},\hat{\beta_j},\hat{\gamma_j})$ are obtained using the Newton-Raphson method. The estimators $\hat{\lambda_j}$, $\hat{\beta_j}$ and $\hat{\gamma_j}$ are asymptotically normal with the asymptotic mean respectively equal to λ_{j0}, β_{j0} and γ_{j0} with the asymptotic variance shown respectively as follows:

$$-\sum_{i=1}^{6000}\sum_{j=1}^{5}\frac{\lambda_0(t)}{d_{ij}\lambda_{j0}(t)^{d_{ij}}}, \tag{14}$$

$$\frac{1}{6000}\sum_{j=1}^{5}\left(\frac{exp\left(\beta_{j0}{}^{T}\boldsymbol{X(t)}\right)F(t)+exp\left(\gamma_{j}{}^{T}\boldsymbol{X(t)}\right)S(t)}{exp\left(\beta_{j0}{}^{T}\boldsymbol{X(t)}\right)F(t)}\right) \tag{15}$$

and

$$\frac{1}{6000}\sum_{j=1}^{5}\left(\frac{exp\left(\beta_{j}{}^{T}\boldsymbol{X(t)}\right)F(t)+exp\left(\gamma_{j0}{}^{T}\boldsymbol{X(t)}\right)S(t)}{exp\left(\gamma_{j0}{}^{T}\boldsymbol{X(t)}\right)S(t)}\right) \tag{16}$$

Thus, the distribution of the maximum likelihood estimators $\widehat{\lambda_j}, \widehat{\beta_j}$ and $\widehat{\gamma_j}$ can be estimated using normal distributions with the respective means λ_{j0}, β_{j0} and γ_{j0} with the variance being defined by Equations 14, 15, and 16.

Equation (5) has now been fully specified and is usable for modeling and predicting subscriber's survival outcomes when competing risks are confounded by time-dependent covariates. The most influential time-dependent covariates will have a $p<0.05$.

3. Data and Empirical Results

To assess the model's adequacy and performance, a churn dataset is used with five time-dependent variables (age, marital status, occupation, education level, and residence) and 5 competing risks (network quality, service quality, price sensitivity, carrier responsiveness, and fraud) from November 2003 - July 2019. The weekly churn rate for this period is used to determine the extent of customer churn mong Safaricom PLC, Airtel Networks Limited, and Telkom Kenya Limited.

The population size for the study includes all present and past active mobile subscriptions in Kenya. The sample size (n=6000) was calculated using the stratified sampling technique as well as the random sampling method of Yamane (1967). The latest national census data from 2019 was used as a sampling framework for identifying the subscribers.

Primary data were gathered through close-ended questionnaires to find out if subscriber churn had occurred based on: subscriber, residence, age group, marital status, occupation, and education.

The customers were monitored for 830 weeks. The minimum follow-up time was 0 weeks and the maximum was 830 weeks. Of the subscribers, 1128 churn events (18.80%) occurred during the monitoring, with 678 (60.11%) happening within 207 weeks, 956 (84.75%) within 265 weeks, and 1093 (96.90%) within 623 weeks of line activation.

4. Findings

Table 1

Demographic factors of categorical covariates.

Demographic factors		Status of censoring or event				
		Total	Churn	Censored	Event/Churn Percentage	Weeks to churn
Marital status	Married	3,371	633	2,738	19%	202
	Single	2,629	495	2,134	19%	207
Level of Education	Tertiary	1,417	285	1,132	20%	199
	Secondary	746	111	635	15%	219
	Primary	3,620	684	2,936	19%	204
	None	217	48	169	22%	204
Age	18 – 25 years	778	146	632	19%	165
	26 – 35 years	2,416	438	1,978	18%	205
	36 – 45 years	2,241	441	1,800	20%	208
	46+ years	565	103	462	18%	240
Occupation	Employed	2,841	617	2,224	22%	206
	Unemployed	771	144	627	19%	201
	Self-employed	2,388	367	2,021	15%	203
Place of residence	Urban	2,169	416	1,753	19%	199
	Rural	3,831	712	3,119	19%	208

Table 1 shows, the average time until churn to have been 202 weeks for married subscribers and 207 weeks for single subscribers. Both subscribers had equal odds of churning.

Of the total sample population 20% with tertiary education, 15% with secondary education, 19% with primary education, and 22% with no formal education, churned. The average time until churning was 199 weeks for subscribers with tertiary education, 219 for those with secondary education, 204 for those with primary education, and 204 for those with no formal education. Subscribers with no formal education had a higher propensity to churn followed by subscribers with tertiary education. Subscribers with secondary education had the least probability of churning.

Of the total sample population, 19% of those between 18-25 years, 18% of those between 26-35 years, 20% of those between 36-45 years, and 18% of those 46 years or older had churned. The average number of weeks to churn was 165 for subscribers between 18-25 years, 205 for those between 26-35 years, 208 for those between 36-45 years, and 240 for those 46 years or older. Subscribers between 36-45 years old,

had a higher propensity to churn, while subscribers between 26-35 years, and those 46 years or older had the least probability of churning.

Of the total sample population, 22% of employed, 19% of unemployed, and 15% of self-employed subscribers churned. The average number of weeks until churning was 206 for employed, 201 for the unemployed, and 203 for the self-employed subscribers. Employed subscribers had a higher propensity to churn, while self-employed subscribers had the lowest.

For residence, 416 subscribers living in urban areas and 712 subscribers living in rural areas had churned. The average time to churn was 199 weeks for urban subscribers, and 208 weeks for rural subscribers, showing both to have equal odds of churning.

Table 2 shows that, of the total sample population, 17% of Safaricom, 22% of Airtel, and 24% of Telkom subscribers had churned, with their respective average times until churning being 203, 210, and 193 weeks. Telkom subscribers had a higher propensity to churn while Safaricom subscribers had the least.

Table 2

Status of the censoring or event for the mobile phone providers.

Subscriber	**Status of censoring or event**				
	Total	**Churn**	**Censored**	**Event/Churn Percentage**	**Time to churn**
Safaricom	3,871	657	3,214	17%	203
Airtel	1,728	374	1,354	22%	210
Telkom	401	97	304	24%	193

Table 3 shows, 502 subscribers churned due to network quality, 350 due to service quality, 113 due to price sensitivity and carrier responsiveness, and 50 due to fraud.

Table 3

Number of Subscribers who churned based on competition

Event	**Censored**	**Network quality**	**Service quality**	**Price Sensitivity**	**Carrier responsiveness**	**Fraud**
Number of Subscribers	4872	502	350	113	113	50

Figure 1 reveals most churn events to have occurred between 150-250 weeks after line activation. Safaricom subscribers had the lowest chance of churn, while Telkom subscribers had relatively a higher chance of churn.

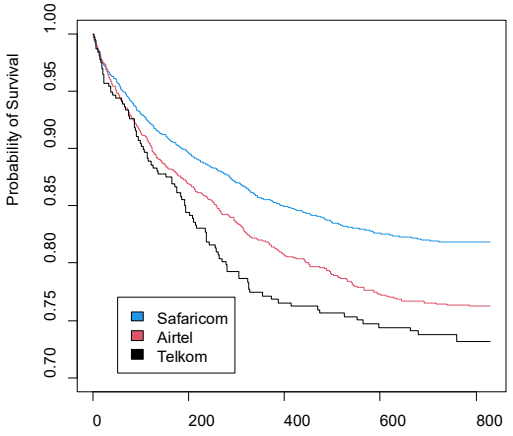

Figure 1 Kaplan-Meier estimates for s

Figure 2(a) shows those living in urban areas to have stayed loyal longer than those living in rural areas.

Figure 2(b) shows subscribers between 18-25 years of age to have been less loyal compared to those 46 years or older.

Figure 2(c) shows no statistically significant difference to have occurred regarding churn rates between married and single subscribers.

Figure 2(d) shows self-employed subscribers to have stayed loyal longer compared to employed subscribers.

Figure 2(e) shows subscribers with secondary education to have remained loyal subscribers longer compared to subscribers with no formal education.

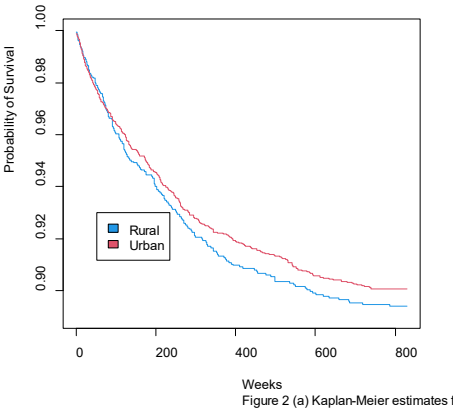

Figure 2 (a) Kaplan-Meier estimates f

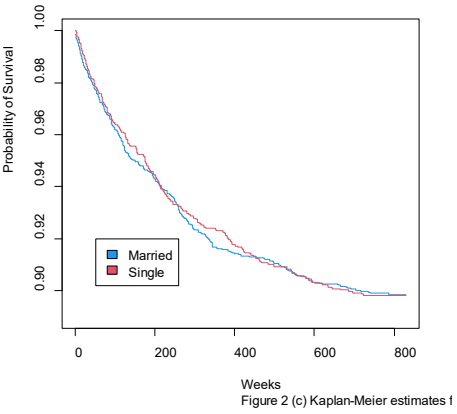

Figure 2 (c) Kaplan-Meier estimates f

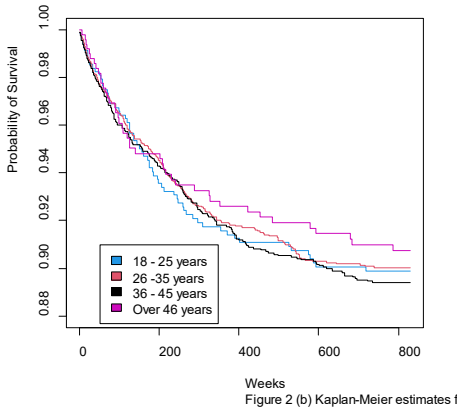

Figure 2 (b) Kaplan-Meier estimates f

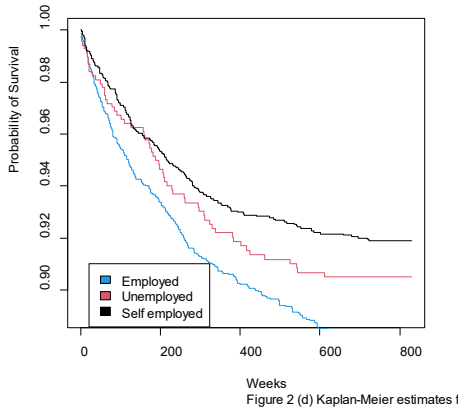

Figure 2 (d) Kaplan-Meier estimates f

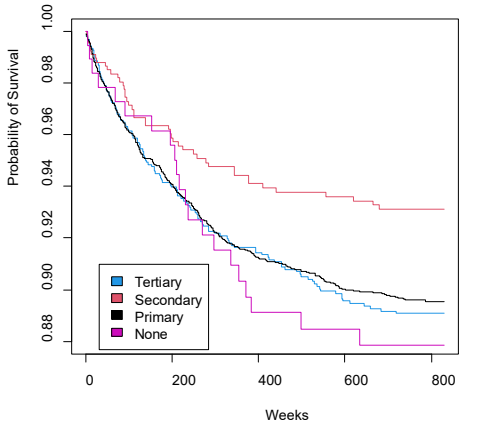

Figure 2 (e) Kaplan-Meier estimates f

Table 4 and Figure 3 show the cumulative incidence function for network quality churn (the risk of network quality churn in the presence of service quality churn, price sensitivity churn, carrier responsiveness churn, and fraud churn) after 600 weeks to be 0.08.

Table 4

Summary of the results of the CIF for the adjusted Cox model

Cumulative incidence function - Estimates				
Cause	**200**	**400**	**600**	**800**
Network quality churn	0.0483	0.0705	0.0800	0.0837
Service quality churn	0.0350	0.0495	0.0562	0.0583
Price sensitivity churn	0.0110	0.0152	0.0185	0.0188
Carrier responsiveness churn	0.0112	0.0152	0.0180	0.0188
Fraud churn	0.0047	0.0073	0.0078	0.0083

Similarly, the cumulative incidence function for service quality churn (the risk of service quality churn in the presence of network quality churn, price sensitivity churn, carrier responsiveness churn, and fraud churn) after 800 weeks was shown to be 0.0583.

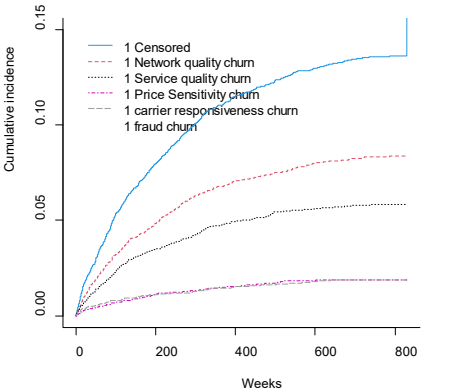

Figure 3 Overall cumulative incidence

Table 5 gives a comparison of the cause-specific hazards model and the adjusted Cox model.

From the table, the cause-specific hazards model for network quality churn inflates the model coefficients for all the covariates except for marital status compared to the adjusted Cox model.

In both models, only subscriber occupation is statistically significant (p<0.05) for network quality churn.

Table 5.

CSH model for Network quality churn and the adjusted Cox model

Covariates	**Cause-specific hazards model for Network quality churn**			**Adjusted Cox model for Network quality churn**		
	Coef	***p*-value**	***SE***	**Coeff.**	***p*-value**	***SE***
Residence	-0.0953	0.3040	0.0928	-0.0650	0.4800	0.0930
Age group	-0.0149	0.7830	0.0540	-0.0098	0.8500	0.0532
Marital status	-0.0125	0.8900	0.0902	-0.0156	0.8600	0.0904
Occupation	-0.2078	0.0000238	0.0492	-0.1795	0.00029	0.0495
Education	-0.0314	0.535	0.0507	0.0292	0.5700	0.0515

The cause-specific hazards model for network quality churn has a likelihood ratio test of 19.02, while the adjusted Cox model for network quality churn has a value of 14.

5. Discussion and Conclusion

The model identified the most influential (statistically significant) of the covariates of residence, age group, marital status, occupation, and education, to be subscriber occupation. One's occupation highly influences their general social status such as place of residence, education, marital status, and age group, and may be significant in determining behaviors such as churn.

Mobile phone service providers should support the government's policy of free day secondary education, as well as encourage and support subscribers to engage in blue-collar jobs due to the subscribers with secondary education levels and those who are self-employed having lower propensities to churn.

In addition, service providers will need to intensify promotional activities geared towards holding onto subscribers during years 3-5 after line activation.

Furthermore, mobile service providers should make deliberate efforts to provide the best network quality and improve service quality.

When testing the overall significance of the model, the null hypothesis that the two models (i.e., adjusted Cox model and cause-specific hazards model) are equal was tested against the alternative hypothesis the adjusted Cox model is better at predicting subscriber's survival outcomes when competing risks are confounded with time-dependent covariates and the most influential covariates are detected. This study asserts the adjusted Cox model to be better at predicting subscribers' survival outcomes compared to the cause-specific hazards model when competing risks are confounded with time-dependent covariates and the most influential covariates have been detected as the adjusted Cox model's computed test statistic was calculated as 10.04 at a 95% confidence level, which is less than the critical value of 11.07.

Peer-review: Externally peer-reviewed.
Conflict of Interest: The author has no conflict of interest to declare.
Grant Support: The author declared that this study has received no financial support.
Author Contributions: Conception/Design of study: N.B.F., L.K.A., J.J.M.; Data Acquisition: N.B.F.; Data Analysis/ Interpretation: N.B.F.; Drafting Manuscript: N.B.F., L.K.A., J.J.M.; Critical Revision of Manuscript: N.B.F., L.K.A., J.J.M.; Final Approval and Accountability: N.B.F., L.K.A., J.J.M

References

Austin, P. C., Latouche, A., & Fine, J. P. (2019). A review of the use of time-varying covariates in the Fine-Gray subdistribution hazard competing risk regression model. *Statistics in Medicine*, *39*(2), 103–113. https://doi.org/10.1002/sim.8399

Beyersmann, J., & Schumacher, M. (2008). Time-dependent covariates in the proportional subdistribution hazards model for competing risks. *Biostatistics, 9(4),* 765-776.

Cox, D. R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society B, 34(2),* 187–220.

Customer Churn Prediction & Prevention Model. (2021, December 22). Optimove. Retrieved January 15, 2022, from https://www.optimove.com/resources/learning-center/customer-churn-prediction-and-prevention#:%7E:text=Churn%20prediction%20modeling%20techniques%20attempt%20to%20understand%20the,to%20the%20success%20of%20any%20proactive%20retention%20efforts.

Fine, J. & Gray R. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association, 94 (446),* 496–509.

Therneau, T. M, & Grambsch, P. M. (2000). *Modeling survival data: Extending the Cox model*. New York: Springer-Verlag. doi: 10.1007/978-1-4757-3294-8