Exploring survival in Early Childhood Educators of British Columbia(ECEBC):Comparison between members and across British Columbia

Tanvira Chaudhury

Abstract

This paper examines the survival of 3957 Early Childhood Educators of British Columbia(ECEBC) members who held organizational membership between 1st January, 2007 and 31st May, 2018. I use both non-parametric and semi-parametric methods to analyze membership survival in ECEBC and regional variation in their survival. Using ECEBC's databases along with census subdivision-level data from government databases, I find that being part of a local ECEBC branch and being a full-time member relative to being an associate member are significantly associated with lower risk of membership termination. I find that unobserved factors pose substantial excess risk of membership termination and it is not analogous to having higher terminations in a census subdivision(CSD).CSDs such as Prince Rupert, Chetwynd and Vancouver show lowest excess risk of membership termination relative to average while North Vancouver city and Maple Ridge show highest excess risk relative to average.

Introduction

Early Childhood Educators of British Columbia(ECEBC) is a non-profit organization that has created a sense of community among British Columbia's early childhood educators (ECE's) through local ECEBC branches in various census subdivisions(CSDs), through organizational membership and through organizing various professional training workshops. In addition, it also campaigns for the wages of ECE's to be raised to the living wage level. However, ever since 2008, its membership has hovered around 800 members or so. ECEBC does not have the tools to identify the causes of low membership or which of its members could be most at risk of membership termination.In addition, ECEBC also wishes to know the effect its local branches might have had on risk of membership termination. Thus the main objective of this paper is to analyze membership survival in ECEBC. More specifically, the objectives are:

- to pinpoint which ECEBC members are most at risk of membership termination,
- to identify factors which might be significantly associated with membership termination,
- to analyze regional variation in membership termination to identify regions with higher and lower excess risk of membership termination than average

I explore the above mentioned issues through survival analysis- I consider an extended Cox model with a) clustering and b) cluster-specific random effects.
In economics, survival analysis has been used extensively to analyze various aspects of patents and innovation. For example, Helmers and Rogers(2008) analyzed the relationship between survival time of more than 162,000 British companies and their IP activity as well as explored regional variation in survival using a proportional hazards model as well as Kaplan-Meier estimation. Nakata and Zhang(2011) used Weibull regression and a mixture regression technique to analyze the relationship between patent filing activity and patent value as well as patent strategy for 214 Japanese electrical and electronic firms.Survival analysis has also been used broadly in demographic and biostatistics literature to analyze time-to-event data. For example, Guo and Rodriguez (1992) analyzed the impact of a mother's birth history on child mortality in Guatemala using a Cox proportional hazards model with gamma frailty at the family level. Sastry(1997) included both family and village effects to analyze risk of child mortality in Northeast Brazil using a piecewise exponential model with gamma frailty. Both papers found substantial random effects that were significantly different from zero. I apply the shared frailty model in a new context-to analyze ECEBC membership survival in different BC census subdivisions with the following findings- a significant positive(negative) association with risk in membership termination for student(full-time) ECEBC members relative to associate members and also a lower risk of membership termination for ECEBC members with local branch membership. I also found certain census subdivisions to have much higher and lower excess risk of membership termination than average.

Unlike papers on churn analysis which have very detailed consumer and product characteristics, I do not have access to detailed ECEBC member characteristics(detailed explanation in the data section-section II)nor information of all-inclusive membership benefits that could potentially promote retention or help predict churn likelihood.However, a model accounting for heterogeneity in survival outcomes allows me to analyze both member-specific and census-division specific variation in survival outcomes.

The paper is organised as follows.In Section I, the data is introduced.In Section II, I discuss the model and estimation strategy. In Section III some descriptive statistics are provided. Sections IV and V present results of non-parametric and semi-parametric estimation respectively while section VI concludes.

Section I-Data

Member-specific data

Member-specific data comes from ECEBC's personal database- the membership database includes members' names,member identification number, join date and last renewal date(by day, month and year),membership status at last renewal,annual amount paid for ECEBC subscription,[1] and location(city of residence, address and zip code).

ECEBC membership is of three types:student members(those pursuing a post-secondary degree in early childhood education or have 500 hours of relevant work experience), full-time members(those with an ECE certificate from the BC government) and associate members(those who are assistant ECE's or have an operating license or have switched from full-time membership to associate membership or have retired from the field).Join dates range from 1969 to 2018 while renewal dates are from 2008 onwards.Due to very few observations of members joining prior to 2007 and absence of termination data before 2008, I focus my attention only on the sample of members between 2007 and 2018.

The bursary database includes names of bursary recipients,city of residence, the university they are enrolled in as well as the bursary amount they have been awarded for the semester. ECEBC's bursary program started in fall 2014 and ECEBC has been awarding bursaries from 2014 onwards.[2]

Branch information also comes from ECEBC and it includes branch names as well as areas covered. [3]

Census subdivision-specific data

CSDs[4]have been chosen because they act as a good proxy for the geographical location of ECEBC members.[5] CSD level covariates are much more suitable than higher region-level covariates such as school district level or regional district level covariates as members are more likely to be affected by the different childcare and school facilities in their CSDs instead of what is happening in their neighboring CSDs. Higher region level definitions such as school districts, regional districts, local health areas all either nest CSDs(at the district level) or have CSDs overlapping boundaries(for example, local health area boundaries or Human Early Learning Partnership(HELP)-defined neighborhood boundaries). Given the hierarchical structure of the data- that is with members nested within CSDs, which in turn can further be grouped into districts or health areas, failing to account for clustering at the CSD level will lead to underestimated standard errors especially for higher level covariates.

The data for schools comes from BC Ministry of Education's K-12 School and District Contact database[6] which includes the names of all currently operational schools(public, independent and alternative schools-such as distributed learning, special needs, cyber schools,etc) in each school district, the city they are in, grade range as well as Strongstart[7] centres. I pair the Strongstart data from the school district database with Strongstart outreach[8] data from Strongstart BC Early Learning Programs database.[9]. To account for large number of elementary school closures in the observed time period(109 closed elementary schools since 2007)and create number of elementary schools as a time varying covariate, I piece together the school database and British Columbia Teachers' Federation (BCTF) school closures database.[10] BCTF database has names of closed schools, the school district they were under and school closure date(by day,month and year). BCTF only covers

[1]ECEBC has monthly subscription option but ECEBC only has the aggregated year-end total for subscription in its database.The payment amount also only includes the last year a member is part of ECEBC and so does not provide helpful information about membership survival.

[2]The bursary award periods include- 2014(fall only), 2015(fall and winter term only), 2016(winter, summer and fall),2017(winter, summer and fall) and 2018(winter,summer and fall). The bursary database included data up to 2018 winter bursary during the time of this project.2018 winter bursaries are not included as they have been awarded in late May 2018.

[3]For the areas covered under each branch and branch names please refer to data appendix,table 1.

[4]A census subdivision can be a city, a town, a village, a regional district electoral area, an Indian reserve or a district municipality(Statistics Canada).BC electoral areas are unorganized areas that are not part of a municipality or a reserve but municipal services for which are provided by the provincial government or by the local government

[5]All members can be correctly matched to and classified under a category of CSD provided by Statistics Canada.

[6]http://www.bced.gov.bc.ca/apps/imcl/imclWeb/Home.do

[7]Strongstart BC is an early learning school-based development program led by early childhood educators for children under the age of 5 to help them prepare for kindergarten

[8]Strongstart outreach programs operate at a reduced schedule and are offered in remote rural areas either in elementary schools, in community centres or in Strongstart ORCA buses. In the paper, I put weights on the Strongstart programs, putting half the weight on an outreach program

[9]https://www2.gov.bc.ca/gov/content/education-training/early-learning/learn/strongstart-bc

[10]https://bctf.ca/data.cfm?page=SchoolClosures

public school closures. Since no data is available for non-public school closures, I assume that number of non-public schools and alternative programs remains constant over this time period.
For licensed childcare facilities, the data comes from the BC Childcare map data that is available through Data BC. It gives a listing of 4603 ministry- licensed family and group-care facilities with current vacancies for children from 0 to 12 years of age along with location(address, city and zip codes). There is currently no available data on unlicensed and/or unregulated childcare facilities in BC.

Section II-Model and estimation

Both non-parametric and semi-parametric estimation methods have been used to analyze membership survival[11] and regional differences in survival. Kaplan-Meier product limit estimation has been used for non-parametric estimation and extended Cox models with time varying covariates with both clustering as well as shared frailty have been used for semi-parametric estimation.Non-parametric estimation has the advantage of being robust and requiring fewer assumptions while Cox model gives estimates without requiring any necessary distributional assumption about the underlying baseline hazard.[12]

The time period chosen for observation is January 1st,2007-May 31st, 2018(beyond 31st May, 2018, every member is right censored).This cutoff date translates to 22% right censoring in the data set which is not high enough to compromise validity of survival analysis. For Kaplan-Meier estimation, I looked at survival of three ECEBC member types as well as regional variation in survival by school district regions[13])(see appendix A.1 for detailed exposition on Kaplan-Meier estimation)..

For the extended Cox model, I analyse the association of risk of membership termination with individual-specific covariates as well as census subdivision level covariates.Despite the fact that proportional hazards assumption is violated in an extended Cox model with time-varying covariates(see appendix A.2 for proportional hazards assumption and partial likelihood of Cox model),estimation can still be done by partial likelihood. To account for correlation in survival outcomes for members nested within the same CSD due to unobserved effects, I choose a shared frailty model with lognormal frailty.A Gaussian distribution has the advantage in that the assumption of log normal frailty allows the random effect to be easily interpreted like any of the other Cox model variables. Under the shared frailty model, the hazard is given by:

$$\lambda(t) = \lambda_0(t)e^{X\beta+Zb} \tag{1}$$

where $\lambda_0(t)$, the baseline hazard has an unspecified distribution,
where b is a vector of random effects with $b \sim G(0, \sum(\theta)$ and θ being the tuning parameter. The random effect is taken as a random intercept at the CSD level.It acts multiplicatively on the baseline hazard-increasing or decreasing the average hazard for ECEBC members belonging to the same census subdivision j in the same way but allowing overall risk to differ across different census subdivisions,
where Z is a design matrix equalling 1 if ECEBC member i belongs to census subdivision j,
where β is a vector of fixed effects
where X includes both member-specific covariates(membership status at last renewal,dummy variable for bursaries and a dummy variable for whether a member is covered by a local branch or not[14] and census subdivision-level covariates(number of schools, licensed facilities and Strongstart facilities in a census subdivision).Since ECEBC is interested in knowing the effect of its local branches, local branch coverage is considered to be the covariate of interest.

[11]Here survival refers to the difference in time between a member's join date and last renewal date at ECEBC or the period of time an individual remains subscribed as an ECEBC member before terminating membership.Since I only have data for join and last renewal dates, I take last renewal date to be a proxy for a member leaving ECEBC.

[12]There is lack of guidance in existing literature on what kind of distribution could be chosen for the baseline hazard so as to enable parametric estimation.

[13]School district regions are chosen because there are a small enough number of school district regions to allow for visual comparison of Kaplan-Meier survival curves.(See data appendix,table 4 for list of school districts and CSDs under each school district region

[14]Although members are automatically considered covered by a local branch if they fall under the branch's catchment area, actual number of members covered or affiliated would be lower.Lack of data from ECEBC prevents me from accurately categorizing local branch members and non-members.
Local branch coverage is a member-specific variable instead of a CSD level variable because a branch can span several CSDs and even be in two non-adjacent areas(for more details on branch coverage see data appendix) and a branch might also cover just one area of a CSD and not others. For example,ECEBC members from 150 Mile House from Cariboo electoral area F are covered by the Williams Lake branch while other areas such as Likely and Horsefly in Cariboo F are not covered by a local branch.

Estimation for the shared frailty model is done through maximum likelihood with joint maximization over β and θ.(see appendix A.3 for the partial likelihood of the shared frailty model with lognormal frailty).

For the covariates above, membership status, local branch coverage, number of Strongstart centres as well as licensed facilities are considered to be time-fixed covariates while ECEBC bursary and number of schools are both time varying covariates. Due to lack of data from ECEBC as to when members are joining a local branch, I have assumed that a member is part of the local branch in their CSD of residence from the moment they join ECEBC to the time they drop out of ECEBC.Therefore, I am overestimating the number of members who are also local branch members at each point in time.Due to lack of data, I have assumed that membership status remains unchanged for members in ECEBC although in reality that covariate should also be time-varying (a student member can switch to being a full-time member who in turn can switch to being an associate member).
For membership status, I have chosen the associate member type as the reference group as associate members occupy the senior most position in the member-type hierarchy. Membership status, bursary and local branch coverage are included as having branch membership and a bursary as well as having certain membership status relative to another might be associated with lower risk of membership termination. Schools, Strongstart facilities and ministry-licensed childcare facilities are all potential areas where early-childhood educators are to be employed. However, I have no ECEBC data that effectively links ECE's to areas of employment through member-specific employment information. However, greater numbers of these potential hubs of ECE employment would ideally be associated with lower risk of membership termination.

In addition to the shared frailty model, I also use a random coefficients model which allows excess risk to be shared by ECEBC members within a CSD in the same way as before but which also allows the effect of the covariate of interest to vary across the different census subdivisions. I consider a random coefficient for branch coverage and a random effect at the CSD level.Estimation is done by maximum likelihood. In this case, the hazard will be given by:

$$\lambda_{ij} = \lambda_0 e^{b_{j0}+(\beta_1+b_{j1})x_{ij1}+\sum_{m=2}^{p}\beta_m x_{ijm}} \tag{2}$$

$$\begin{pmatrix} b_{j0} \\ b_{j1} \end{pmatrix} \sim N\left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \sum_j \equiv \begin{pmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{01} & \sigma_1^2 \end{pmatrix} \right\}$$

$where \mathrm{b}_{j0}$ is the CSD-specific random effect that appears as a random intercept for each CSD j,
where $(\beta_1 + j1)$ is the coefficient of the covariate of interest(branch coverage of an ECEBC member) which is allowed to vary across census subdivisions,
where x_{ijm} are the other $(p-1)$ covariates whose coefficients are not allowed to vary across CSDs(see appendix A.4 for marginal log-likelihood of a random coefficients model).

A shared frailty model however does not solve for omitted variable bias problem as it only allows me to estimate the random effect that is uncorrelated with the covariates. If a fixed effects model was chosen in its place so that θ_j is fixed in CSD j, then observed and unobserved heterogeneity will be controlled for, omitted variable bias will be solved for but several problems will arise. If a fixed effect is considered per CSD, then we will obtain inconsistent estimates for all parameters as number of parameters increase with increase in observations.(Rodriguez, 2018)Assuming we follow the fixed effects model discussed in Kalbfleisch *et.al*(1980),covariates constant within a CSD will drop out of the partial likelihood, incidence of no events in a particular CSD will mean that the specific cluster will not contribute to the partial likelihood(which in turn will mean that for a particular CSD it is plausible to have a fixed effect of 0 which I do not have any way of testing for but which for a random effects model is inconceivable) and CSD-level fixed effects will mean I can only estimate member-level covariates(see appendix A.5 for the partial likelihood of the fixed effects model).Thus, CSD-level covariates will enter both the numerator and denominator of the partial likelihood and cancel out.However a shared frailty model allows all levels of covariates to be estimated,gives an estimate of excess risk and gives an estimate of the strength of these effects.

Members' cities of residence and addresses were cross-checked with respective postal codes from the ECEBC database using Canada Post's Search Bar and Google Maps.The same method was used to check location of childcare facilities and to match schools from school district database to their respective census subdivisions. Members and facilities were matched to relevant census subdivisions using Statistics Canada's Focus on Geography series 2016 and to relevant school districts using school district maps from BC Ministry of Education. [15]
To generate time varying covariates, I split survival time every time a covariate's value changes and generating a person-time count process expands the original data set from 3957 observations to 4622 observations.However, this does not imply re-

[15]In special cases,such as school districts 61(Victoria), 62(Sooke),63(Saanich), 22(Vernon) and 23(Central Okanagan),where areas were difficult to map directly using the above mentioned sources, members were first sorted into HELP (Human Early Learning Partnership) neighborhoods in order to sort them under the correct school districts and ultimately within the correct census municipality boundaries.

peated events for a member or clustering at the individual level as one event from each uncensored ECEBC member enters the partial likelihood. One problem however would be the choice of the distribution for the random effect; many distributional choices are available but the most frequently used distributions are gamma and Gaussian distributions. Hsu. *et al* (2007) found less than 10% bias even when the underlying distribution differed significantly from a gamma distribution.In multi-centre, random control trials, Glidden and Vittinghoff(2004)found a gamma frailty model to be a good fit for the data and found minimal change even with the violation of the assumption of gamma distributed frailty.I found no specific direction in existing literature on distributional choice and use Gaussian distribution in the model(due to the advantages mentioned above) and include a gamma frailty model in the appendix.

Section III-Descriptive statistics

Figures 1 and 2 in appendix B give the number of ECEBC members joining and terminating membership each year at ECEBC.We see that the number of members joining has increased sharply from 2007 onwards[16] but has been declining ever since 2009.In contrast, the number of members renewing has not fluctuated substantially from 2008 onwards. Members joining before 2007 indicate only the longest surviving members from that period and hence could add bias to estimation. As a result, the time period chosen for analysis is between 1st January, 2007(for join date) and 31st May, 2018(for renewal date).

For regional variation, I look at survival time of different member types(appendix B).In terms of absolute numbers, all areas have more than double the number of full-time members than associate or student members and termination rate is the highest for full-time members in comparison to student members. Also in terms of overall termination rate,Kootenays has the highest termination rate and Fraser Valley has the lowest. I also look at percentage of renewals in ECEBC local branches (list of branches, areas covered and renewal data specific to ECEBC local branches are in tables 1,2,3 and 4 of data appendix of the paper). Looking at number of renewals(data appendix, table-2), we see that Caledonia branch has the highest percentage of last renewals although Vancouver branch in absolute numbers has the highest number of renewals at 636 renewals. A sizeable bulk(371 renewals) of these renewals are from the Vancouver CSD (table 3 of data appendix). From these figures, Bulkley Valley, Caledonia, North Shore, Nanaimo, Victoria and Vancouver branches appear to be 'high risk' branches with membership termination rate well over 75%.

Section IV-Kaplan-Meier estimation

Kaplan-Meier survival curves for membership type and for different school district regions are in the next two pages. For membership survival, full-time members show relatively better survival probability at each point in time in comparison to associate and student members-39.5% of full-time members and half of both student and associate members renewed their membership statuses for the last time within the first 400 days of joining ECEBC. Survival probability also takes longer to drop to half for full-time members- roughly within two years of joining ECEBC. Therefore, despite higher termination rate across all BC regions from 2007 to May, 2018, full-time members are still longer surviving ECEBC members than student or associate members.The relatively better survival outcomes for full-time members in contrast to poorer outcomes for student members might be because of student members switching to full-time membership; therefore the poorer survival outcomes for student members only reflect the outcomes for those who held student status throughout their membership period and not those who switched membership(I do not have data for changing statuses as only the last status a member has held is recorded during termination and is entered into the database).Thus, I am overestimating the time a member spends under one status in ECEBC; if a member in question has not terminated membership and is superficially seen to have only one status(whereas they might have had multiple statuses within this time period in reality), the number at risk before event time would be larger for both full-time members and associate members(if full-time members are also switching to associate membership). Therefore, I will obtain better survival probabilities for both these categories than is actually the case.

However, for student members to be full-time members, they are also required to complete either a four year degree in early childcare or attain 500 hours of work experience. So unless an overwhelming majority of student members join ECEBC towards the end of their degree or remain members long enough to have the option of being a full-time member, the actual number switching over to full-time membership at each point in time is likely to be smaller than expected. Also since associate members hold more senior positions and/or are ECE assistants, the number of full-time members switching over to associate membership is also likely to be small(unless an overwhelming majority of full-time members at each point in time are assistants or have been in early childcare for ten years or more). However, if statuses are changing and multiple outcomes

[16]ECEBC switched from paper-based to electronic record keeping and hence not all records of members joining and renewing are available between 1969-2006

are possible, then KM-estimation would be unsuitable as it is only able to take one survival outcome.

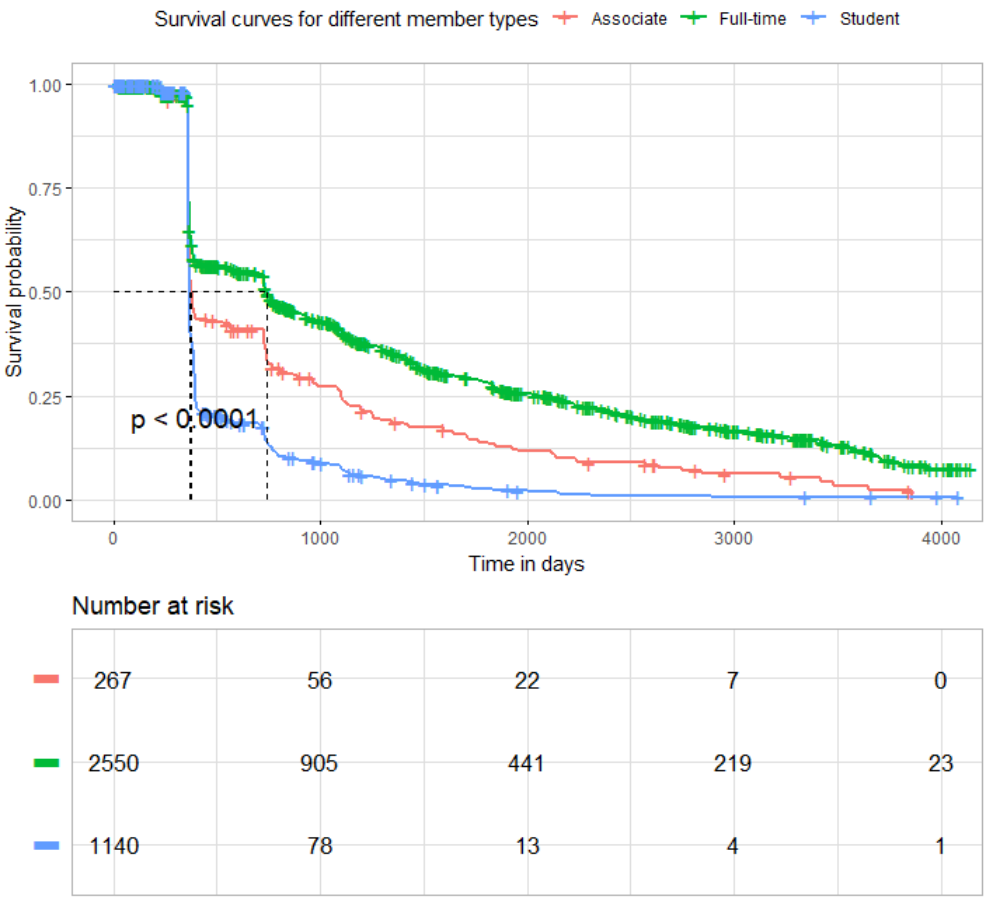


Figure 1 hows Kaplan-Meier survival curves for ECEBC members while figure 2 shows survival curves with members sorted under British Columbia school district regions. Number of members is 3957 and time period is 1st January, 2007 to 31st May, 2018. Number at risk refers to members who are "alive" just before an event-the last time of renewal by an ECEBC member.P-value is from the log-rank test checking if the survival curves for each member type are different.

In the case of survival across different British Columbia regions, I look at survival by school district regions[17] which are a small enough category for visual representation. I find that from joining to last renewal at ECEBC, members in Northwest,Vancouver Island and Kootenays do better relative to members in Fraser Valley and in Thompson Country school district regions.These results again show that higher termination rate does not mean poorer survival outcomes since despite the higher termination rate, members in the Kootenays are longer surviving members than members in Fraser Valley.Around the 400th day, all school district regions show a large drop in membership survival probability with the greatest drop being for Fraser Valley-62.5% of its members have terminated their ECEBC membership. In terms of median survival, we see that 50% of members of the better performing school district regions-Northwest, Vancouver Island and Kootenays have terminated membership between 600 and 725 days of joining ECEBC. However, Northwest's advantage relative to other two better performing regions diminishes gradually and by 2500 days(close to 7 years), survival probability for the three better performing school regions is somewhat similar. Metro/Coast and Greater Victoria despite having a larger membership pool are among the lower performing regions. However their survival probabilities are still better than than that of Fraser Valley and Thompson Country where over half of the members terminate membership around the 400th day of joining ECEBC.
I also test the equality of survival curves for both member types and school district regions (tables-1(a) and 1(b) of appendix-C for member types and tables 2(a) and 2(b) of appendix-C for school district regions)through the log-rank and Peto-Peto-Prentice tests. I find a significant χ^2 statistic for all the tests suggesting difference in survival curves.

[17]School district regions also allow all ECEBC members to be included. If, instead just branches were considered, then a large proportion of members would be excluded as ECEBC does not have branches spanning most BC CSDs.

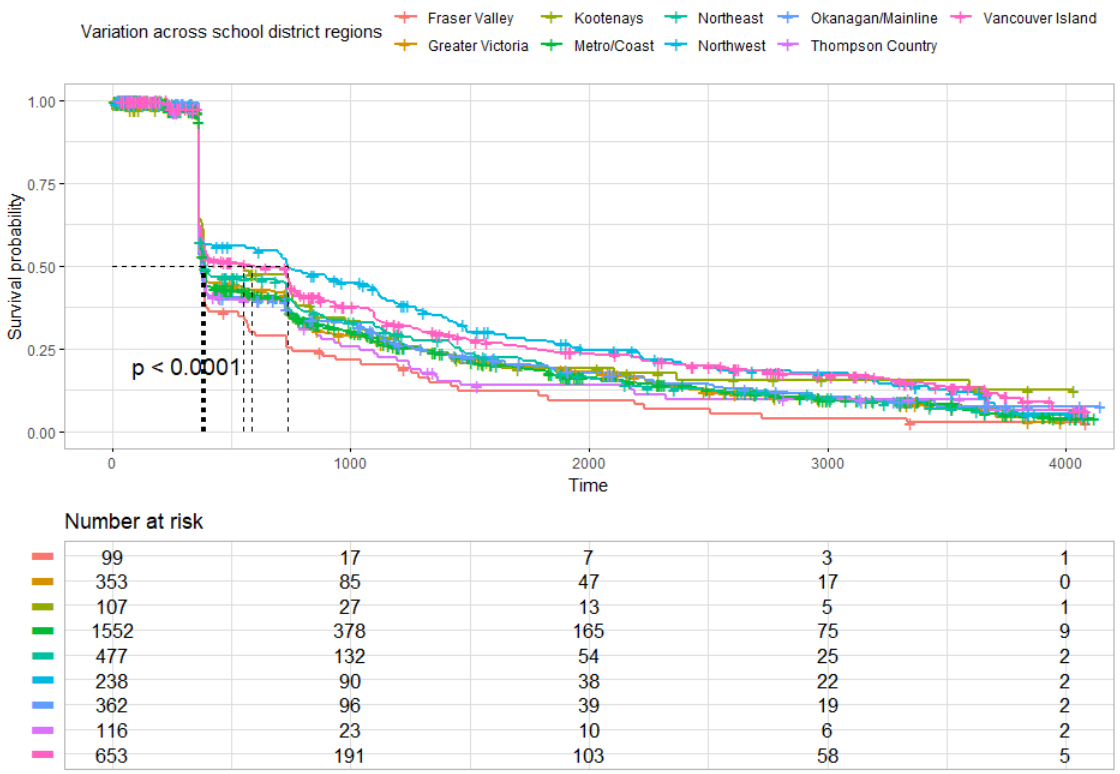


Figure 2 gives variation in survival probabilities for different BC school district regions. Number of observations is 3957 while time period of observation is between 1st January, 2007 and 31st May, 2018.Number at risk is the number of ECEBC members in a school district region who are alive just before an event.

Section V-Extended Cox models

The table in page 8 gives the results from extended Cox models[18]. Column I gives estimates from an extended Cox model with time varying covariates without any regional effects or clustering. Column II gives results of an extended Cox model with standard errors clustered at the CSD-level. Column III gives estimates from an extended Cox model with shared frailty at the CSD level and column IV gives estimates of a random coefficients model that allows for local branch coverage to have a heterogeneous effect across CSDs.

For membership status, the results accord with the pattern obtained from KM-estimation. Relative to associate members, the risk of membership termination is lower for full time members-it is 0.66 times that of associate members holding all other covariates constant. For student members,there is a 1.6 fold increase in risk of membership termination relative to associate members. On the other hand, being a local branch member means risk of membership termination is between 0.86-0.88 times that of non-local branch members for the different model specifications.However, the number of members who are local branch members is overstated as discussed in Section II(footnote 14) and so the effect might actually go in either directions depending on which members are truly active in the local branches.[19]For reasons discussed in section-IV, the estimates for full-time members and student members could be different although it is unlikely that the $\widehat{\beta}$ coefficients would change signs as a small number of members would probably be switching membership.All the other covariates have insignificant association with risk of termination.
All the signs are as expected except for number of schools and ECEBC bursary both of which yield a positive estimate; having more schools in a CSD and having an ECEBC bursary are associated with higher risk of termination. Relative to non-bursary holders, the risk of membership termination is higher for bursary-holding members(11.4%-16.5% higher for different model specifications) while an additional elementary school in a CSD is associated with an increase in risk of membership termination(0.3%-0.5% higher risk for the different specifications). Firstly,a very small proportion of members receive ECEBC bursaries(108 out of 3957 members). Majority of student members do not have ECEBC bursaries(roughly 5% of

[18]I run a simple Cox regression with time-fixed covariates to check if they violate the proportional hazards assumption(appendix D.1 for Cox regression and test for proportional hazards assumption). I find that the time-fixed covariates either satisfy the proportional hazards assumption or the scaled Schoenfeld residuals plots show variation relative to estimate that is small enough to be ignored

[19]A more suitable variable based on exact matches could not be constructed because ECEBC has no record of members who are directly tied to their local branches

student members and 1.96% of full-time members hold ECEBC bursaries). It is possible that an ECEBC bursary does not act as a significant incentive in attracting and retaining members (since very low number of student or full-time members apply to begin with. Having ECEBC membership status is also not a precondition for eligibility as evidenced by majority of the recipients being non-members). Therefore it is plausible that other member characteristics of bursary-holding members might meaningfully explain higher risk of termination among them.In the case of Strongstart facilities, one more Strongstart centre in a CSD is associated with 0.3%-0.8% reduction in risk of membership termination across the different specifications. For ministry-licensed facilities, the reduction in risk is between 0.02%-0.04% for the different models.The overall effect of schools translates to 9.8% difference in risk between first and third quartiles which is higher than the effect of Strongstart centres or licensed facilities. However, higher it is enntirely possible that ECEBC members are not working in elementary schools.Absence of data for different ECEBC members' professions prevents me from making a more judicious choice about including number of elementary schools as a covariate.

Columns II, III and IV are not comparable as the model with clustering gives population-averaged relative risk with all other covariates constant whereas for the shared frailty case, the relative risk is within clusters. However, we see some similarities between estimates of columns II and III.The reason could be because of low within-cluster correlation. Therefore, I also run a shared frailty model with gamma frailty at the CSD level(see appendix E.1)to find the CSD-level random effect in order to compute τ statistic for within-cluster correlation[20]. It turns out to be quite low at 0.017, thus explaining the similarity in estimates.Larger standard errors of model II relative to I point to correlation between CSDs but provides no means of classifying CSDs in terms of excess risk of termination or calculating strength of CSD effects(as only random samples from 3957 observations are considered and hazards are averaged out for model II). Models III and IV compare relative risk within CSDs and are comparable to model I(no clustering). Model III also allows me to find which CSDs are more prone to excess risk of membership termination. However, the estimates are slightly different under model III-the protective effect of being a local branch member relative to a non-branch member in the same CSD with all other covariates fixed is slightly smaller with cluster-specific random effects than in model I with no clustering. For Strongstart facilities, the protective effect is much higher when considering the correlation of member survival outcomes in each CSD. However, the risk of termination is greater if one holds an ECEBC bursary although the effect is still insignificant. Taking into account CSD-level random effects has also slightly increased standard errors.
The value for standard deviation between centres is modest at 0.155. The χ^2 statistic for log partial likelihood after integrating out the random effect on 8 degrees of freedom is 454.67. For model I, the χ^2 statistic for log-partial likelihood is 435.6 on 7 degrees of freedom. The critical value for χ^2 distribution for 7 degrees of freedom is 14.06. Since the difference is larger than the critical value, it is significant. Conducting a likelihood ratio test to choose between models I and III (appendix D.2) also gives a very low p-value suggesting that the shared frailty model is a better model than model I.The frailty model however does not give a suitable confidence interval for the variance of the random effect but a valid confidence interval can be obtained from a 95% profile likelihood(see the page 9 for a plot of a range of values for standard deviation of the random effect against twice the log-likelihood). Plotting the likelihood function over fifty points, I see that the graph is not symmetric about the MLE estimate obtained for the random effect and so regular Wald-confidence interval will not be suitable. Thus, by considering the profile likelihood, I obtain a small confidence interval of (0.1,0.21) for the random effect.
For the random effect, I can exponentiate the standard deviation of the random effect to observe distribution of unobserved risk- thus, a CSD one standard deviation above the mean will have 16.9% higher risk of membership termination which is moderately high. Quartilewise, CSDs in the bottom quartile have 9.9% lower risk than CSDs at the median whereas CSDs in the top quartile have 11% higher risk suggesting a modest overall effect of unobserved CSD-level random effect.

With random effects model, I can also compare how number of membership terminations per CSD compared to average frailty within a CSD. For most CSDs, the excess risk is anywhere up to 10% higher or lower than average but having large number of membership terminations might not automatically mean highly "frail" regions; some CSDs despite having same or even greater number of membership terminations than CSDs in the 10% range have much higher or much lower excess risk than average. If we look back to the descriptive statistics section(excluded in this sample), then we can see that Vancouver has the highest number of renewals(data appendix,table-3)while other areas such as Surrey, North Vancouver and Victoria have much lower membership terminations. Such statistics might lead to the conclusion that Vancouver is a much higher risk region than for example a region such as North Vancouver. However, upon considering the shared frailty model, having higher renewals might not automatically mean higher average frailty(see figure in the next page which gives a plot of excess CSD-level risk against renewals per CSD with a ranking of top 5 CSDs with higher than and lower than average excess risk or average frailty). For example, for North Vancouver city, excess risk is about 30% higher than average which is substantially high while for a CSD like Chetwynd, excess risk is about 19% lower than average. CSDs with larger renewals than usual have lower excess risk because they also tend to have a larger membership pool while CSDs with smaller membership pools have estimated risk shrinking closer to the mean.

[20] τ statistic for within-cluster correlation$=\frac{\theta}{\theta+2}$ where θ is variance of the random effect

	I		II		III		IV	
	$\hat{\beta}$	$\exp(\hat{\beta})$	$\hat{\beta}$	$\exp(\hat{\beta})$	$\hat{\beta}$	$\exp(\hat{\beta})$	$\hat{\beta}$	$\exp(\hat{\beta})$
Fulltime member	-0.411^{***}	0.663	-0.411^{***}	0.663	-0.411^{***}	0.663	-0.414^{***}	0.661
	(0.072)		(0.077)		(0.072)		(0.072)	
Student member	0.473^{***}	1.605	0.473^{***}	1.605	0.467^{***}	1.595	0.461^{***}	1.586
	(0.075)		(0.066)		(0.076)		(0.076)	
Bursary	0.108	1.114	0.108	1.114	0.143	1.153	0.153	1.165
	(0.334)		(0.641)		(0.336)		(0.335)	
Branch coverage	-0.125^{***}	0.883	-0.125^{**}	0.883	-0.150^{***}	0.861	-0.134^{***}	0.874
	(0.040)		(0.057)		(0.054)		(0.053)	
School	0.003	1.003	0.003	1.003	0.005	1.005	0.005	1.005
	(0.002)		(0.003)		(0.003)		(0.003)	
Licensed facility	-0.0004	0.9996	-0.0004	0.9996	-0.0003	0.9998	-0.0006	0.9994
	(0.0006)		(0.00093)		(0.0008)		(0.0008)	
Strongstart	-0.0032	0.997	-0.0032	0.997	-0.008	0.992	-0.008	0.992
	(0.005)		(0.003)		(0.008)		(0.006)	
Random effect(CSD):								
Variance					0.0239		0.00423	
Random effect(branch):								
Variance:							0.0349	
Correlation:							-0.08469	
Observations	4622		4622		4622		4622	
Events	3084		3084		3084		3084	
LRT	435.6^{***}(df=7)		435.6^{***}(df=7)					
Log-likelihood	$-22{,}616.290$		$-22{,}616.290$					
Wald	470.3^{***}(df=7)		296.6^{***}(df=7)					
Score	495.1^{***}(df=7)		495.1^{***}(df=7)					
Integrated loglik					-22606.75		-22605.34	

Table 1: Standard errors are in parentheses.'*','**' and '***' indicate significance levels of 90,95 and 99 percent respectively. Column I gives estimates of extended Cox model with time varying covariates assuming independence of observations. Column II gives estimates of an extended Cox model with standard errors clustered at the CSD level. Column III gives estimates of extended Cox model with shared frailty at the CSD level with random effect following a Gaussian distribution. Column IV is a random coefficients model that allows heterogeneous effect of having a branch across CSDs. Total number of individual observations is 3957.Time period of observation is from 1st, January 1st,2007 to 31st,May 2018. Generating a person-time count process for time varying covariates has split survival time into smaller chunks to generate 4622 observations from the original 3957.

Profile likelihood and confidence interval for random effect

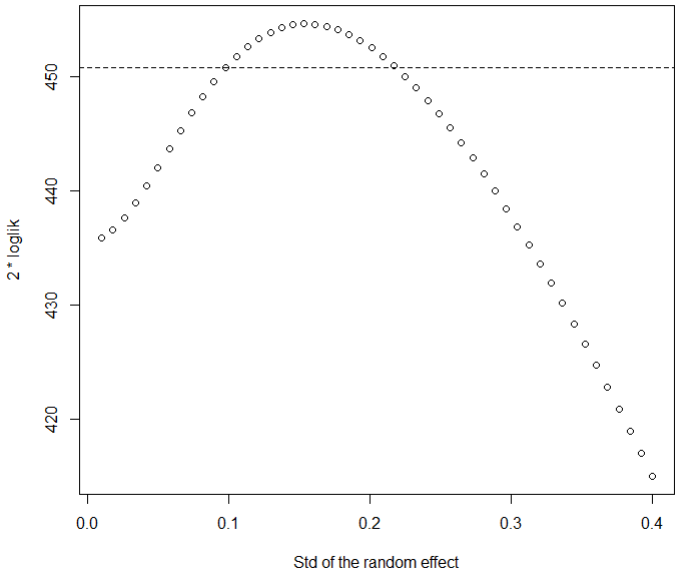


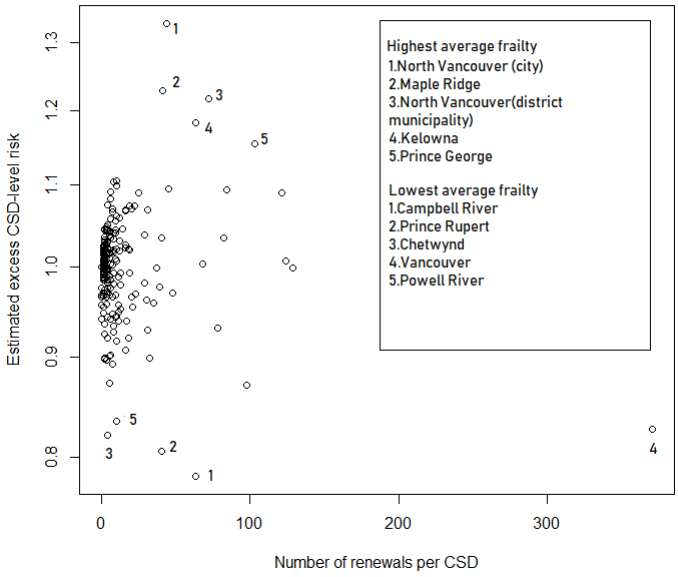
Figure above gives profile likelihood of the random effect of extended Cox model III and the resulting confidence interval using 50 points.

For model IV, there is negative correlation between random effect at the CSD level as well as the random slope for local branch coverage; members in CSDs with higher average frailty would experience a diminished protective effect of being part of local ECEBC branches although the correlation is small. The variance for local branches is also not high suggesting not enough heterogeneity of local branch coverage across CSDs.Checking model fit(see Analysis of Deviance Table in appendix D.2), I find that there is no significant variation in local branch coverage effect across CSDs and thus model III fits the data better than model IV.

I also conduct some robustness checks(appendix E) using a gamma shared frailty model(appendix E.1) as well as a piecewise exponential model with and without mixed effects(appendix E.2). I find the variance of the random effect to be slightly larger in the gamma frailty case and the variation is found to be highly significant. CSDs with frailty at the first quartile have 12.2% lower risk than CSDs with median frailty whereas CSDs at the third quartile have 13.2% higher risk than CSDs with frailty at the median.These results are similar to what I obtained under model III.Moreover, local branch coverage as well as membership types give similar estimates as in model III and they have significant association with risk of membership termination. Other estimates are also of similar magnitude.

In the case of piecewise exponential model with and without mixed effects, I observe similar patterns- lower risk of membership termination for full-time members relative to associate members as well as lower risk of membership termination for local branch members as opposed to non-branch members.For full-time members, the risk of termination is 0.65-0.68 times that of associate members while for student members, the risk of termination is 1.42 times higher. Being a local branch member is associated with lower risk of termination than members not covered by any local branches-the risk of termination is 0.87-0.88 times that of those not covered by a branch.All the estimates have same signs as before except for bursaries which has a negative sign(although as before, it is insignificant).

Estimated excess CSD-level risk against number of renewals per CSD



Section VI-Conclusion

This paper analyses the effect of member-specific and CSD-level factors on the risk of membership termination at Early Childhood Educators of BC. Being a full-time ECEBC member relative to being an associate member as well as being a local ECEBC branch member relative to being a non-local branch member are significantly associated with lower risk of membership termination. In addition, through a shared frailty model, I also found CSDs with much higher and lower excess risk than average.

The results are robust to consideration of a piecewise exponential model with or without mixed effects as well as consideration of a different distribution for the random effect for the shared frailty case. Future work can focus on increasing the number of levels in the hierarchical data so as to see if ECEBC membership has any association with school district-level or regional district-level variables. In case of higher level modelling, more variables such as low income measure for families with children under 6 years of age can be incorporated which are available at much higher geographical units such as census agglomerations and regional districts. A variable such as low income measure is a likely channel through which survival time in ECEBC is affected. This, is because early childhood educators are likely to be earning below the living wage in various BC regions and also because childcare costs are high in BC. So demand for childacre might be smaller in areas with a higher proportion of low-income families. If CSDs are chosen to be nested within school districts or local health areas, then population of children under 6 years of age could also be considered as such population data is not available at the CSD-level.Another possible extension could include focusing on census tracts which could act as a suitable proxy for neighborhoods in CSDs in order to focus on neighborhood differences in income levels, childcare costs and distribution of children within a CSD in order to capture the heterogeneity that exists within CSDs themselves in affecting ECE retention.Data sources such as PRIZM5 which segments Canadian neighborhoods using common demographic traits could be used to capture heterogeneity within CSDs so as to point to new areas where ECEBC could help expand member activity, particularly regarding the establishment of any new,local branches. Moreover, such variables can provide a more thorough picture of why ECEBC members are dropping out in a CSD by explaining any factors external to ECEBC that could be responsible for low membership.

References

Atkinson, E.,Crowson, C. & Therneau, T.(2018).*Using Time Dependent Covariates and Time Dependent Coefficients in the Cox Model*. URL:https://cran.r-project.org/web/packages/survival/vignettes/timedep.pdf

Austin,P.C.(2017).*A Tutorial on Multilevel Survival Analysis:Methods,Models and Applications."International Statistical Review,85*(2),185-203

Cox,D.R. (1972).*Regression Models and Life Tables." Journal of the Royal Statistical Society B, 34*(2), 187–220.

Fox J. & Weisberg S. (2011). *Appendix: An R Companion to Applied Regression.* Sage, Thousand Oaks.

Guo,G.& Rodríguez,G. (1992).*Estimating a Multivariate Proportional Hazards Model for Clustered Data Using the EM Algorithm, with an Application to Child Survival in Guatemala. Journal of the American Statistical Association,87*:969-976

Grambsch,P.M. & T.M. Therneau.(2000).*Modelling Survival Data:Extending the Cox Model*:Springer-Verlag, New York.

Ha,I.D.,Sylvester,R.,Legrand,C.& MacKenzie,G.(2011). *Frailty Modelling for Survival Data from Multi-Centre Clinical Trial.Statistics in Medicine, 30*(17), 2144–2159

Helmers, C.& H.Rogers. (2008).*Innovation and Survival of New Firms across British Regions.*Department of Economics, Discussion Paper Series.University of Oxford.

Hsu.L.,Gorfine,M. Malone,K.(2007).*On robustness of marginal regression coefficient estimates and hazard functions in multivariate survival analysis of family data when the frailty distribution is mis-specified.Statistics in Medicine,26*(25):4657-4678 https://doi.org/10.1002/sim.2870

Kalbfleisch J.D. & Prentice R.L. (2002).*The Statistical Analysis of Failure Time Data.* New York:John Wiley & Sons, .

Kaplan,E.L.& P.Meier(1958).*Nonparametric estimation from Incomplete Observations. Journal of the American Statistical Association. 53*(282):457-481

Nakata,Y. & X. Zhang (2012).*A survival analysis of patent examination requests by Japanese electrical and electronic manufacturers, Economics of Innovation and New Technology, 21*(1):31-54

Pebley Stupp.(1987).*Reproductive Patterns and Child Mortality in Guatemala. Demography,24*:43-60.

Rodriguez, G. (2007). *Lecture Notes on Generalized Linear Models.* URL: http://data.princeton.edu/wws509/notes/

Rodriguez, G. (2018).*Lecture Notes on Survival Analysis.* URL: http://data.princeton.edu/wws509/notes/

Rondeau,V.,Mazroui,Y.& Gonzale,J.R.(2012).*frailtypack:An R Package for the Analysis of Correlated Survival Data with Frailty Models using Penalized Likelihood Estimation or Parametric Estimation. Journal of Statistical Software,47*(4).

Sastry,N.(1997).*Family-Level Clustering of Childhood Mortality Risk in Northeast Brazil.Population Studies,51*(3),245-261.

Therneau,T.(2018).*Mixed effects Cox model.*

TABLE 1

Branch	CSDs covered
Bulkley Valley	Smithers,Hazelton,Telkwa
Caledonia	Prince George
Campbell River	Campbell River,Campbell River 11,Strathcona C
Central Okanagan	Kelowna, West Kelowna, Lake Country,Peachland
Comox	Comox,Courtenay Comox Valley A(Royston, Fanny Bay) Comox Valley C(Merville)
Dawson Creek	Dawson Creek
Kamloops	Kamloops,Barriere,Chase, Cache Creek,Merritt
Nanaimo	Nanaimo,Nanaimo River, Nanaimo Town 1, Qualicum Beach, Parksville, Nanaimo F
North Okanagan	Vernon,Armstrong, Lumby,Coldstream
North Shore	North Vancouver city, West Vancouver, North Vancouver district municipality
Port Alberni	Port Alberni
Prince Rupert	Prince Rupert
Quesnel	Quesnel
Southern Gulf Islands	Saltspring Island, Pender Island, Mayne Island
Terrace	Terrace, Kitimat, New Aiyansh, Gitwangak 1
Vancouver	Vancouver,Surrey, Burnaby
Victoria	Victoria, Sidney, North Saanich, Central Saanich,Esquimalt,Langford
West Coast	Tofino, Ucluelet
West Kootenay	Nelson,Trail, Nakusp, Salmo, Grand Forks, Central Kootenay H(Winlaw, South Slocan), Montrose,Rossland
Williams Lake	100 Mile House,Cariboo F(150 Mile House)

Table 1 above gives the names of ECEBC local branches as well as the CSDs covered by each local branch.

TABLE-2

Branch	% of CSD resident members renewing	number of members covered
Bulkley Valley	83.3%	36
Caledonia	85.8%	120
Campbell River	61.1%	113
Central Okanagan	77.1%	105
Comox	58.4%	101
Dawson Creek	64.3%	28
Kamloops	76.9%	104
Nanaimo	84.2%	183
North Okanagan	70.07%	155
North Shore	81.2%	154
Port Alberni	75.8%	33
Prince Rupert	61.5%	65
Quesnel	53.3%	15
Southern Gulf Islands	52.9%	17
Terrace	75.8%	62
Vancouver	78.6%	809
Victoria	80.7%	171
West Coast	73.3%	15
West Kootenay	64%	50
Williams Lake	66.7%	12

Table 2 above gives number of members in each local branch as well as the percentage of members renewing in each branch.

TABLE-3

CSDs	renewals(highest)	renewals(lowest)
Vancouver	371	
Surrey	129	
Nanaimo	124	
Burnaby	121	
Prince George	103	
Alberni-Clayoquot B,C,D, Ahahswinis 1,Ashcroft,Cache Creek, Creston,Gitanmaax 1, Kimberley, Lytton, East Kootenay C, Merritt, Metchosin		1
Alberni-Clayoqout F, Fraser Lake,Kaslo, Montrose, Mount Waddington C		2

Table 3 above gives a broad snapshot of number of renewals in BC CSDs.

5 Southeast Kootenay	Cranbrook,Fernie,Sparwood, East Kootenay C(Fort Steele)	
6 Rocky Mountain	Kimberley,Invermere,Golden	
8 Kootenay Lake	Nelson,Creston,Salmo,Kaslo,Slocan,	Central Kootenay A(G
10 Arrow Lakes	Nakusp,Central Kootenay K(Fauquier), Silverton	
20 Kootenay-Columbia	Castlegar,Montrose, Rossland, Trail Fruitvale, Kootenay Boundary B(Genelle), Central Kootenay J(Robson)	
51 Boundary	Grand Forks	
19 Revelstoke	Revelstoke	
22 Vernon	Vernon, Lumby,Coldstream, Okanagan 1, North Okanagan C	
23 Central Okanagan	Kelowna,West Kelowna, Central Okanagan,Tsinstikeptum 9,Peachland	
53 Okanagan Similkameen	Okanagan-Similkameen D(Okanagan Falls), Oliver	
58 Nicola-Similkameen	Merritt,Princeton	
67 Okanagan Skaha	Penticton,Summerland,Penticton 1, Okanagan-Similkameen D(Kaleden)	
83 North Okanagan-Shuswap	Armstrong, Salmon Arm, Sicamous, Enderby, Columbia-Shuswap C(Blind Bay,Tappen, Sorrento)	
27 Cariboo-Chilcotin	Cariboo-D,E(Alkali Lake), F(Big Lake Ranch, 150 Mile House),G(108 Mile Ranch), H(Canim Lake, Forest Grove), Williams Lake,One Hundred Mile House	
28 Quesnel	Cariboo A, B, C, I, Quesnel	
57 Prince George	Fraser-Fort George A, C, D,F, Mackenzie,McBride,Prince George	
59 Peace River South	Chetwynd,Dawson Creek,Peace River E, Pouce Coupe,Tumbler Ridge	
60 Peace River North	Fort St.John, Hudson's Hope	
81 Fort Nelson	Northern Rockies	
91 Nechako Lakes	Bulkley-Nechako D,Vanderhoof, Fort St.James, Fraser Lake	
49 Central Coast	Central Coast A, Central Coast C, Central Coast E	
33 Chilliwack	Chilliwack,Fraser Valley H,Tzeachten 13	
34 Abbotsford	Abbotsford	
75 Mission	Mission	
78 Fraser-Cascade	Agassiz,Fraser Valley A,Fraser Valley B, Hope	

Census subdivisions listed are the CSDs that match the ECEBC membership database.Areas in parentheses are specific areas in electoral areas that are in the ECEBC database(members do not come from all areas within an electoral area)
School district regions are: Kootenays(SD 5-SD 51),Okanagan/Mainline(SD 19-SD 83),Northwest(SD 27-SD 49),Fraser Valley(SD 33-SD 78)

SD Name	CSDs covered
73 Kamloops/Thompson	Kamloops,Thompson-Nicola J, Thompson-Nicola L, Cache Creek, Barriere, Chase
74 Gold Trail	Ashcroft, Lytton, Lillooet

School district region is: Fraser Valley(SD 73- SD 74). School district 93-Conseil scolaire francophone has schools across British Columbia and so does not constitute a separate school district region

APPENDIX-A.1 Kaplan-Meier estimation

If $t_{(1)}$, $t_{(2)}$,$t_{(3)}$.....$t_{(n)}$ are all distinct ordered events (events being last renewals for ECEBC members) such that $t_{(1)}$<$t_{(2)}$<$t_{(3)}$<.....<$t_{(n)}$, then the Kaplan-Meier estimate of the survival function will be:

$$\widehat{S}(t) = \prod_{i:t_i<t} \left(1 - \frac{d_i}{n_i}\right) \tag{3}$$

where d_i= number of renewals at time t_i
n_i =number of members at risk or number of members "alive" (that is those who still hold ECEBC membership) just before time t_i
1-d_i/n_i is the conditional probability of being "alive" at time t_i (where d_i/n_i being the conditional probability of renewing for the last time at time t_i)

A.2-Extended Cox model-time varying covariates

If $t_{(1)}$, $t_{(2)}$, $t_{(3)}$,.....,$t_{(n)}$ are distinct, ordered event or renewal times for ECEBC members such that, $t_{(1)}$<$t_{(2)}$< $t_{(3)}$< .....<$t_{(n)}$,

$$\lambda_i = \lambda_0 e^{X_i\beta} \tag{4}$$

where i=1,2,.....,n indicates any random ECEBC member
X_i(or $X_i(t)$) is the i'th row of the matrix X(or $X(t)$ for time-varying covariates) for ECEBC member i;X(or $X(t)$) is an (nxp) matrix of time-fixed(or time-varying) covariates
where λ_i is the hazard of each ECEBC member i such that it is the product of the baseline hazard λ_0 and the risk score, r_i(r_i being $e^{X_i\beta}$ for time-fixed covariates and $e^{X_i(t)\beta}$) for time-varying covariates.
The hazard ratio for any two ECEBC members i and k would be:

$$\frac{\lambda_i}{\lambda_k} = \frac{\lambda_0 e^{X_i\beta}}{\lambda_0 e^{X_k\beta}} \tag{5}$$

Since the baseline hazard cancels out, the hazard ratio is simply the ratio of risk scores for each member- therefore the hazard ratio would be a constant over time making the Cox model time invariant-this is the proportional hazards assumption of the Cox model. Assuming no ties (no simultaneous last renewals so only one unique last renewal or event at each event time) the partial likelihood[21] would just be:

$$Partial\ likelihood = \prod_{i=1}^{n} \prod_{t\geqslant 0} \left\{ \frac{Y_i(t) r_i(\beta,t)}{\sum_j Y_j(t) r_j(\beta,t)} \right\}^{dN_i(t)} \tag{6}$$

where $r_i(\beta,t)$=$e^{X_i(t)\beta}$,
where member j in the denominator represents any random ECEBC member at risk of membership termination at event time, t_i. Log-partial likelihood is:

$$Log\ partial\ likelihood = \sum_{i=1}^{n} \int_0^{\infty} \left\{ Y_i(t) X_i(t)\beta - log \sum_j Y_j(t) r_j(\beta,t) \right\} dN_i(t) \tag{7}$$

Differentiating the log partial likelihood, I obtain the score vector $U(\beta)$ which is:

$$U(\beta) = \sum_{i=1}^{n} \int_0^{\infty} [X_i(s) - \overline{x}(\beta,s)] dN_i(s) \tag{8}$$

where,

$$\overline{x}(\beta,s) = \frac{\sum Y_i(s) r_i(s) X_i(s)}{\sum Y_i(s) r_i(s)} \tag{9}$$

Estimates of $\widehat{\beta}$ are obtained by setting $U(\widehat{\beta}) = 0$ For handling simultaneous renewals at each distinct event time, the Efron method has been used as it is more accurate than the more frequently used Breslow method.

[21]This partial likelihood is applicable for both time-fixed and time varying covariates. For time varying covariates, $r_i(\beta,t)$=$e^{X_i(t_j)}$ where $X_i(t_j)$ gives the time-varying covariate's value at event time t_j

APPENDIX-A.3-Extended Cox model with shared frailty

(Source:Therneau,T.(2018)) The log partial likelihood will be:

$$log[(\beta, b)] = \sum_{i=1}^{n} \int_0^{\infty} \left[Y_i(t) \left(X_i(t)\beta + Z_i(t)b \right) - log \left(\sum_j Y_j(t) e^{X_i(t)\beta + Z_i(t)b} \right) \right] \tag{10}$$

Then we have a penalized Cox model with a penalty function $p(b) = \sum_j b_j$ The random effect is integrated out to obtain the integrated log-likelihood which is:

$$IPL(\beta, \theta) = \frac{1}{(2\pi)^{q/2} |\sum(\theta)|^{1/2}} \int PPL(\beta, b) e^{-b' \sum^{-1}(\theta) b/2} db \tag{11}$$

where q is the number of random effects
A maximum likelihood estimate is obtained by joint maximization over β and θ.

APPENDIX-A.4-Extended Cox model with mixed effects-random coefficients model

(Source:Rondeau *et.al*(2008);Ha *et.al*(2011))
The marginal log-likelihood for the random coefficients model is:

$$l(\Phi) = ln \prod_{i=1}^{G} \int_R \int_R \left\{ \prod_{j=1}^{n_i} \lambda(Y_{ij}|u_i, w_i)^{\delta_{ij}} S(Y_{ij}|u_i, w_i) \right\} f(u_i, w_i) du_i dw_i \tag{12}$$

where Φ=$(\lambda_0(.), \beta, \sigma_0^2, \sigma_1^2, \rho)$

APPENDIX-A.5-FIXED EFFECTS MODEL FOR HETEROGENEITY

Source:Kalbfleisch and Prentice(1980) as cited in Rodriguez(2018)

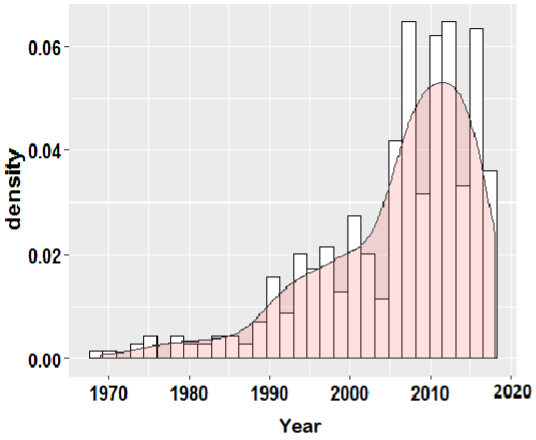
$$\lambda_{ij}\left(t_{ij}\right) = \theta_j \lambda_0\left(t_{ij}\right) e^{x_{ij}^{'}\beta} \quad (1) \tag{13}$$

If θ_j is fixed effect for cluster j, hazard for ECEBC member i is given by the above equation. θ_j enters as a fixed value for each cluster.

$$L = \prod_{i=1}^{mi} \frac{\theta_j \lambda_0(t_{ij}) e^{x_{ij}^{'}\beta}}{\sum_{k \in R_{ij}} \theta_j \lambda_0(t_{ij}) e^{x_{jk}^{'}\beta}} \tag{14}$$

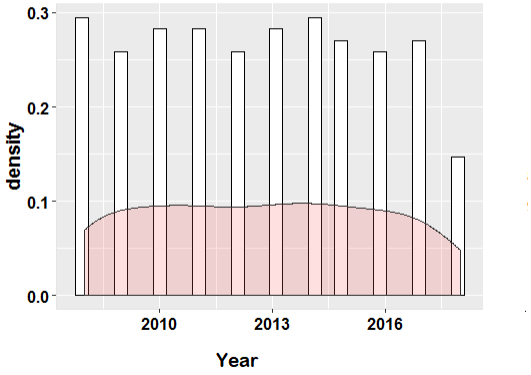
We see that both the fixed effects and the baseline hazard cancel out.If we have covariates that are constant for a cluster such as a time-fixed covariate,it will also cancel out.

Appendix B-Members joining and renewing



Year	No.joining	Year	No.joining
1969	1	1997	15
1970	1	1998	30
1974	2	1999	24
1976	3	2000	21
1978	1	2001	24
1979	2	2002	38
1980	3	2003	28
1982	2	2004	40
1983	3	2005	55
1984	1	2006	91
1985	2	2007	402
1986	1	2008	566
1987	3	2009	401
1988	3	2010	333
1989	6	2011	306
1990	7	2012	279
1991	13	2013	245
1992	18	2014	254
1993	11	2015	312
1994	16	2016	373
1995	15	2017	323
1996	28	2018	209

Figure-1-Members joining



Year	Number of renewals
2008	302
2009	425
2010	324
2011	298
2012	292
2013	314
2014	260
2015	233
2016	284
2017	449
2018	388

Figure-2-Members renewing

School district region	Member type			Termination rate			Overall termination rate
	Full-time	Associate	Student	Full-time	Associate	Student	
Northwest	159	22	57	32.7%	13.6%	8.77%	25.2%
Vancouver Island	411	48	194	31.9%	27.1%	18.0%	27.4%
Kootenays	77	10	20	29.9%	85.4%	20%	28.0%
Northeast	269	27	181	23.8%	7.41%	18.8%	20.96%
Metro/Coast	1094	96	362	23.3%	18.8%	10.2%	19.97%
Okanagan/Mainline	206	32	124	28.6%	12.5%	20.2%	24.3%
Greater Victoria	198	20	135	25.3%	15%	8.15%	18.1%%
Thompson Country	66	8	42	24.2%	25%	14.3%	20.7%
Fraser Valley	70	4	25	20%	-	16%	18.2%

The table above gives the number of member types in each school district region as well as their respective termination rate by member type. Termination rate is percentage of members who have terminated membership at ECEBC. Termination rate=$(1 - \frac{Members\ of\ type\ alive\ in\ 2018}{Member\ of\ type\ in\ 2007})$

APPENDIX-C-Kaplan-Meier estimation

	N	Observed	Expected	$(O-E)^2/E$	$(O-E)^2/V$
Full-time member	2550	1886	2330	84.76	432.40
Student member	1140	979	567	299.38	466.16
Associate	267	219	187	5.64	6.86
χ^2 statistic=496*** (df=2)					

Table 1(a) above gives log-rank test for equality of survival curves.Number of observations is 3957.Time period considered is between 1st January, 2007 and 31st May, 2018.

	N	Observed	Expected	$(O-E)^2/E$	$(O-E)^2/V$
Full-time member	2550	1084	1346	51.0	276.8
Student member	1140	689	440	140.8	287.9
Associate	267	136	122	1.4	2.4
χ^2 statistic=306*** (df=2)					

Table 1(b)above gives results of Peto-Peto-Prentice test for equality of survival curves.Number of observations is 3957.Time period considered is between 1st January, 2007 and 31st May, 2018.

Region	N	Observed	Expected	$(O-E)^2/E$	$(O-E)^2/V$
Fraser Valley	99	81	63.3	4.9372	5.8015
Greater Victoria	353	289	277.2	0.5039	0.6273
Kootenays	107	77	84.5	0.6654	0.7714
Metro/Coast	1552	1242	1170.2	4.409	8.0298
Northeast	477	377	372.3	0.0585	0.0749
Northwest	238	178	216.2	6.7589	8.1015
Okanagan/Mainline	362	274	263.4	0.4233	0.5244
Thompson Country	116	92	84.1	0.7510	0.8835
Vancouver Island	653	474	552.8	11.2249	15.3431

χ^2=33.6***(df=8)

Table 2(a)above gives the results of a log-rank test testing equality of survival curves for 9 BC school district regions.

Region	N	Observed	Expected	$(O-E)^2/E$	$(O-E)^2/V$
Fraser Valley	99	49.2	42.6	1.038282	1.68
Greater Victoria	353	174.4	174.7	0.000391	0.000699
Kootenays	107	46.6	52.4	0.642068	1.08
Metro/Coast	1552	784	736.5	3.064465	8.12
Northeast	477	234	230.4	0.057162	0.106
Northwest	238	103.6	125.4	3.803888	6.82
Okanagan/Mainline	362	178.1	165.8	0.920565	1.63
Thompson Country	116	57.6	55	0.118423	0.195
Vancouver Island	653	280.7	325.5	6.162093	12.3

χ^2 statistic=26.1***(df=8)

Table 2(b)above gives the results of a Peto-Peto-Prentice test testing equality of survival curves for 9 BC school district regions.

D.1-Cox regression with time-fixed covariates

	$\widehat{\beta}$
Full-time member	-0.407^{***}
	(0.072)
Student member	0.459^{***}
	(0.075)
Strongstart	-0.002
	(0.004)
Branch	-0.139^{*}
	(0.040)
Licensed facilities	0.004
	(0.0002)
Observations	3,957
Events	3084
R^2	0.100
Max. Possible R^2	1.000
Log Likelihood	$-22{,}811.530$
Wald Test	446.740^{***} (df = 5)
LR Test	415.330^{***} (df = 5)
Score (Logrank) Test	469.052^{***} (df = 5)
Note:	*p<0.1; **p<0.05; ***p<0.01

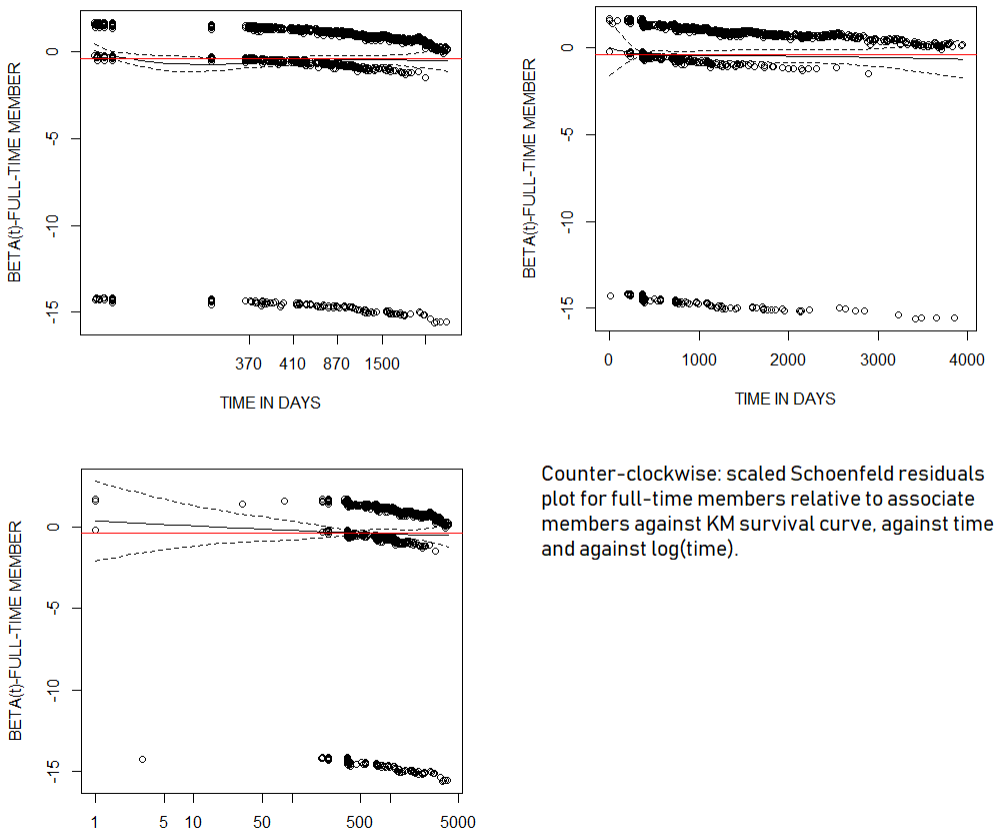
Table 2: Time period of observation is between 1st January, 2007 and 31st May, 2018

Schoenfeld Residuals

	I			II			III		
	rho	chisq	p	rho	chisq	p	rho	chisq	p
Full-time member	−0.0342	3.608	0.0575	−0.01298	0.52094	0.470	−0.01627	0.8193	0.365
Student member	0.0351	3.840	0.05	−0.00756	0.17084	0.673	0.01382	0.5946	0.441
Strongstart	0.0156	0.724	0.395	−0.0015	0.00667	0.935	−0.00479	0.0681	0.794
Branch	−0.0161	0.788	0.375	−0.00234	0.01665	0.897	0.00423	0.0545	0.815
Licensed facility	−0.0136	0.530	0.467	0.0022	0.0138	0.906	0.00496	0.0701	0.791
Global	NA	49.183	0	NA	0.65468	0.985	NA	8.8408	0.116

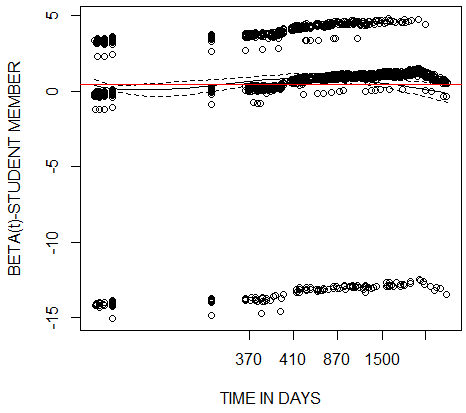
Number of observations is 3957.Time period considered is 1st January, 2007 to 31st May, 2018.Rho,chisq and p values values are for Schoenfeld residuals transformed against left continuous Kaplan-Meier survival curve(I), time (II) and log(time) (III).

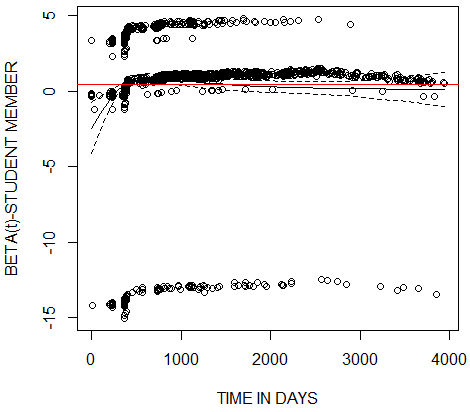
We see that for all three transformations, the p values are high for most covariates.Only the global value for the transformation against the KM plot gives a low p value. I also cross-check by plotting scaled Schoenfeld residuals against left continuous Kaplan Meier survival curves, against time and against log(time). The red line in the following Schoenfeld residuals plots(plots are in pg.35-39) are the $\widehat{\beta}$ values from the original Cox regression. For full-time members, strongstart centres and licensed facilities, we see minimal variation in $\widehat{\beta}(t)$ relative to $\widehat{\beta}$ despite the smaller p values for full-time members for

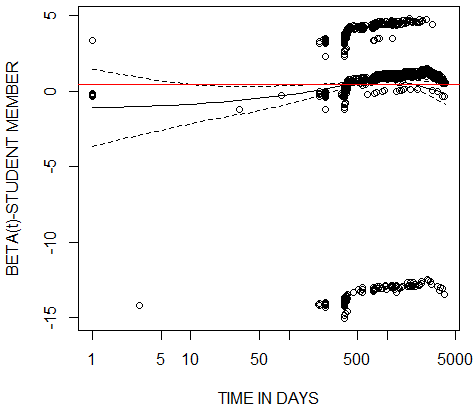


Counter-clockwise: scaled Schoenfeld residuals plot for full-time members relative to associate members against KM survival curve, against time and against log(time).

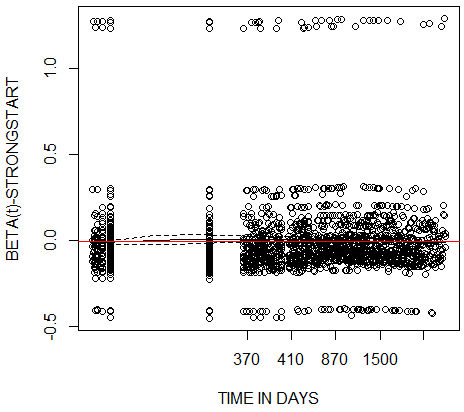
transformation against KM survival curves. For student members we do see some variation in $\widehat{\beta}(t)$ relative to $\widehat{\beta}$ in the initial 450 days of joining ECEBC for transformation against time while we see minimal variation for transformation against KM survival curve. However, the transformation against time does not show a linear trend and hence I use a log-time transformation. For log(time) transformation, there is much smaller variation between $\widehat{\beta}(t)$ and $\widehat{\beta}$ and the difference is relatively constant before diminishing after 450 days. However, since the difference is small, I can ignore this variation in time. For branch coverage, the smoother for the scaled Schoenfeld plot against the KM survival curve shows very little variation whereas for transformation against time and against log(time), there is initially(roughly 50 days from joining ECEBC)larger variation in $\widehat{\beta}(t)$ relative $\widehat{\beta}$. However, it diminishes rapidly. Therefore, I ignore the small variation in $\widehat{\beta}(t)$. I also check for any outliers using deviance residuals. In case of deviance residuals plot against linear predictor(pg.40), we see the plot to be symmetric about zero in the range $(-3.5, +3.5)$ and hence can conclude that there are no outliers. For influential observations(pg.40), one particular observation is seen to have a particularly large dfbeta value relative to $\widehat{\beta}$ coefficient and doing a sensitivity analysis by removing this point did not change the results.

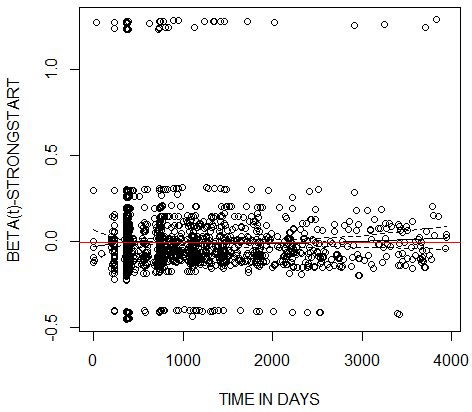


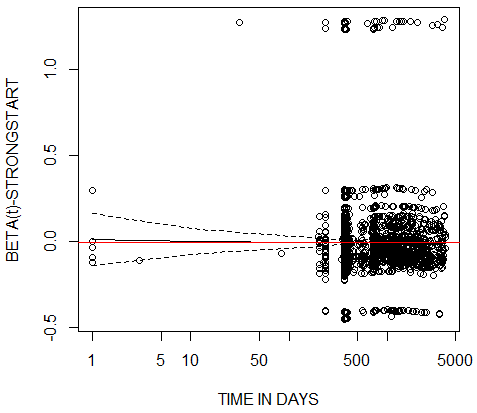




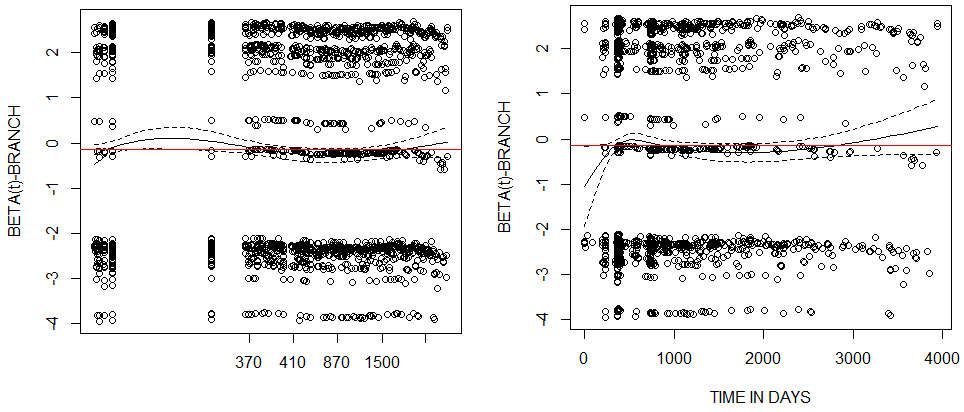
Counter-clockwise: scaled Schoenfeld residual plots for student members against KM survival curves, time and log(time) respectively.

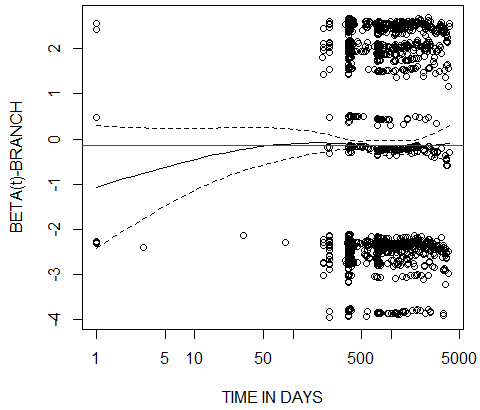




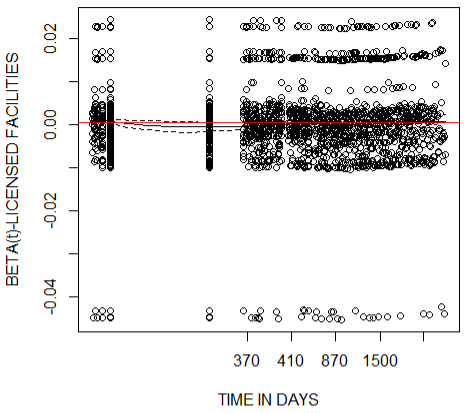


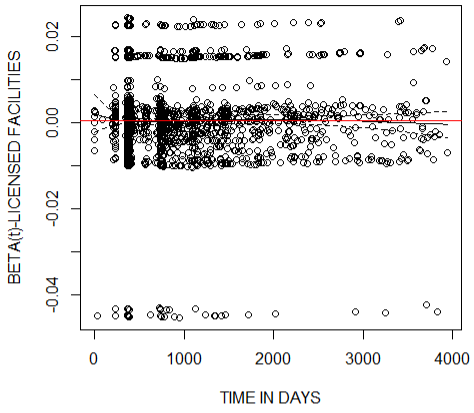
Counter clockwise-Scaled schoenfeld residual plots for Strongstart centres transformed against KM survival curves. time and log(time) respectively.

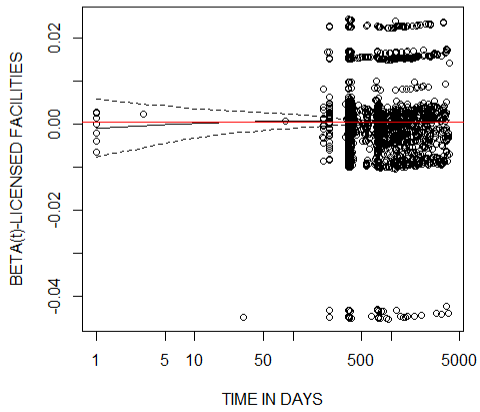




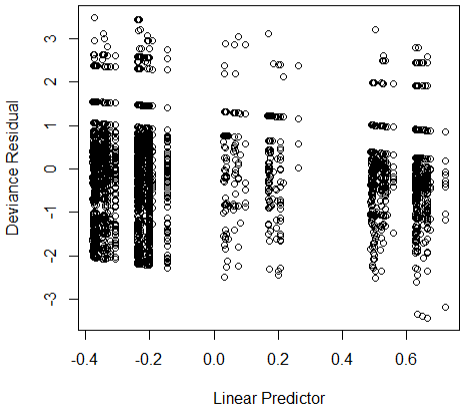
Counter-clockwise: scaled Schoenfeld residuals plot of branch coverage relative to no branch coverage against KM survival curves. time and log(time) respectively.



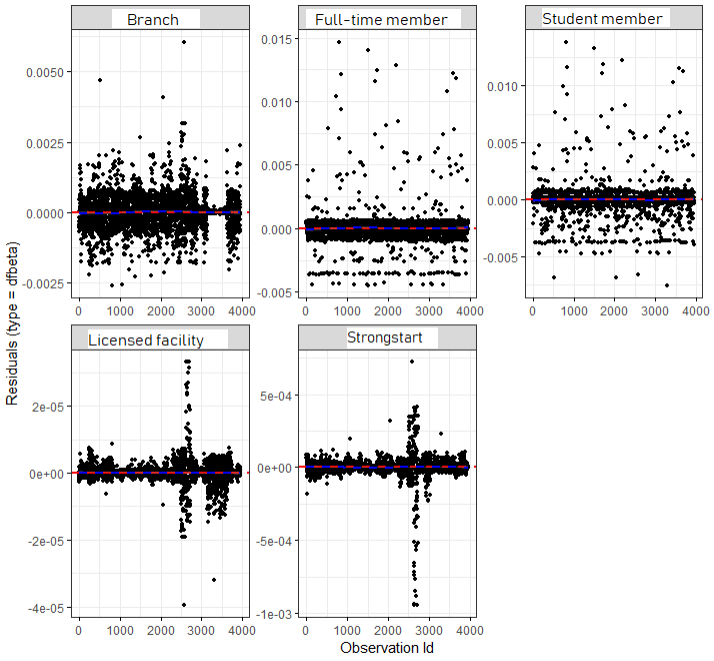




Counter-clockwise: scaled Schoenfeld plots for licensed facilities transformed against KM survival curves, time and log(time) respectively.



Deviance residuals. The above show the deviance residuals plotted against a linear predictor identify any outliers for 3957 observations.



Df beta values have been plotted against 3957 observations for each covariate to see if any particular observation is influential- that is the degree to which it is an outlier and its influence on the covariate had it been removed.

D.2-Likelihood ratio test(Analysis of Deviance Table)

	Cox model	Frailty model	Random coefficients model		
	x	x			
Log likelihood	-22616	-22607			
Df			1		
χ^2 statistic			19.072		
P(>|Chi|)			0***		
		x		x	
Log likelihood		-22607		-22605	
Df					2
χ^2 statistic					2.8278
P(>|Chi|)					0.2432

Table above gives results from two likelihood ratio tests- the top row gives results of a likelihood ratio test for models I and III from the paper whereas the bottom row gives the results of a likelihood ratio test for models III and IV from the paper.

APPENDIX E-ROBUSTNESS CHECK

E.1-Shared frailty model with random effect following gamma distribution

	$\widehat{\beta}$	$\exp(\widehat{\beta})$
Full-time member	-0.411^{***}	0.663
	(0.072)	
Student member	0.466^{***}	1.59
	(0.0757)	
Bursary	0.144	1.15
	(0.337)	
Strongstart	-0.0092	0.991
	(0.003)	
Branch	-0.153^{***}	0.858
	(0.043)	
Licensed facilities	-0.0003	0.997
	(0.001)	
School	0.005	1.005
	(0.003)	
Random effects(CSD):		
Variance	0.035	
Observations	4622	
Events	3084	
Integrated loglik	-22607.7	
LRT test	558^{***}(df=49.31)	

Table 3: Standard errors are in parentheses."*","**","***" indicate significance levels of 90,95 and 99 percent respectively.Table above gives a shared frailty model with gamma frailty at the CSD level. Number of individual observations is 3957.Time period considered is from January 1st,2007 to 31st,May 2018. Generating a person-time count process for time varying covariates has split survival time into smaller chunks to generate 4622 observations from the original 3957.

I use a shared frailty model with random effect following a gamma distribution as a robustness check. The table below gives the estimates for the shared frailty case. As in the main paper, the estimates are remarkably similar to model III with highly significant results for membership type and local branch membership. The gamma frailty model also provides a χ^2 statistic for the frailty term which is highly significant.

E.2-Piecewise exponential model and piecewise exponential model with mixed effects

Exploiting the patterns found in Kaplan-Meier estimates, I split time into the following intervals-(0-400 days),(400-720 days) and (720-4143 days).

I could divide time into even smaller chunks but the problem is that given a modest number of ECEBC members(with most of the members dropping out before the first three years or so of joining ECEBC), each of these smaller intervals will have too few observations and hence fewer events. Each interval receives its own exposure (that is difference between start time and end time for each interval) as well as its own "death" indicator indicating whether an ECEBC member has dropped out of ECEBC or not. With time varying covariates(number of schools and bursaries), I divide up the interval a time varying covariate might fall into, into sub-intervals; each interval receives its own death indicator and exposure and the covariate changes its value at the boundary of the sub-intervals. Both the sub-intervals however will have the same baseline risk as the main interval that has been subdivided. The piecewise exponential model has been proven to be equivalent to a Poisson model(Rodriguez, 2018)where the death indicators are independent Poisson observations with mean=$t_{ij}\lambda_{ij}$ where t_{ij} gives the time period individual i spends in interval j and λ_{ij} gives the hazard of individual i in the j'th interval. In addition to the piecewise exponential model that assumes independence of observations in the same CSD, I also use a piecewise exponential model with a random intercept at the CSD level. As seen below, the estimates are very similar to the values obtained in the extended Cox model.

	$\hat{\beta}$	$\exp(\hat{\beta})$	$\hat{\beta}$	$\exp(\hat{\beta})$
Full-time member	-0.391^{***}	0.676	-0.389^{***}	0.6478
	(0.072)		(0.072)	
Student member	0.355^{***}	1.42	0.349^{***}	1.42
	(0.075)		(0.076)	
Bursary	-0.421	0.66	-0.401	0.77
	(0.334)		(0.335)	
Strongstart	-0.0013	0.999	-0.0039	0.996
	(0.0045)		(0.006)	
Branch	-0.129^{***}	0.88	-0.145^{***}	0.87
	(0.0398)		(0.047)	
Licensed facility	-0.00043	0.9996	-0.00025	0.9998
	(0.00059)		(0.0007)	
School	0.0032	1.003	0.0037	1.004
	(0.0023)		(0.0026)	
Interval1(400,720]	-1.476***	0.229	-1.466^{***}	0.231
	(0.09)		(0.090)	
Interval2(720,4143]	-0.463^{***}	0.629	-0.447^{***}	0.6395
	(0.042)		(0.042)	
Intercept	-6.478^{*}	0.0015	-6.48^{***}	0.0015
	(0.073)		(0.0796)	
Random effect(CSD level):				
Variance:			0.011	
Observations	7772		7772	
Events	3084		3084	

Table 4: Standard errors are in parentheses.'*','**','***' indicate significance levels of 90,95 and 99 percent respectively.Table above gives a piecewise exponential model. Number of individual observations is 3957.Time period considered is from January 1st,2007 to 31st,May 2018. Dividing survival time into intervals has split survival time into smaller chunks to generate 9596 observations from the original 3957.