WEB APPENDIX

Case-centered logistic regression can be a useful way to fit multiplicative intensity models (see reference A1), including stratified Cox models, with familiar software. In this 3-part appendix, we show its relation to stratified Cox regression.

Part 1 shows that it maximizes the same likelihood as stratified Cox regression.

Part 2 shows that it yields the same parameter estimates, and they have the same interpretation.

Part 3 illustrates how it is done in SAS. Several models are fitted to simulated data.

Part 1 – Equivalence of likelihood functions

Given a particular calendar day t, define the following:

$$R(t) = \begin{cases} 1 \text{ if patient was at risk of death on day } t \\ 0 \text{ if patient had died before day } t \end{cases}$$

$$D(t) = \begin{cases} 1 \text{ if patient died on day } t \\ 0 \text{ if patient did not die on day } t \end{cases}$$

$$E(t) = \begin{cases} 1 \text{ if patient had been vaccinated before day } t \\ 0 \text{ if patient had not been vaccinated before day } t \end{cases}$$

Let *X* be a discrete covariate that is to be adjusted for in this analysis. We are interested in the impact of *E*(*t*) on *D*(*t*), within strata defined by *X*. A stratified Cox regression would then model the hazard of death as

$$P(D(t)=1 \mid E(t)=e, X=x, R(t)=1) = \lambda_{0x}(t)e^{\alpha_0 e+\alpha_1 ex}, \qquad (1)$$

where $\lambda_{0x}(t)$ is an unspecified baseline hazard among subjects with $X = x$. The relative hazard associated with vaccination for a subject with $X = x$ is given by

$$RH(t \mid X=x) = \frac{P(D(t)=1 \mid E(t)=1, X=x, R(t)=1)}{P(D(t)=1 \mid E(t)=0, X=x, R(t)=1)} = e^{\alpha_0+\alpha_1 x}.$$

The case-centered analysis fits a logistic regression

$$\log\frac{P(E(t)=1 \mid D(t)=1, X=x, R(t)=1)}{1-P(E(t)=1 \mid D(t)=1, X=x, R(t)=1)} = \log\frac{P(E(t)=1 \mid X=x, R(t)=1)}{1-P(E(t)=1 \mid X=x, R(t)=1)} + \beta_0 + \beta_1 x \quad (2)$$

to a data set containing only decedents that models the log odds of vaccination among patients who died at t and had a covariate value $X = x$ as a function of x and an offset that gives the log odds of vaccination among all patients with $X = X$ who were at risk of death at t.

Let $t_1 < t_2 < \ldots < t_K$ denote the K distinct, ordered event times. Let $e_{l,k}$ denote the vaccination status of patient l on day t_k. Let D_k denote the set of indices for the patients who died at t_k; let $\mathsf{R}_{k,x}$ denote the set of indices for the patients with $X = x$ who were at risk of dying at t_k. Let $n_{k,e,x}$ denote the number of patients at risk of death at $t = t_k$ with $E = e$ and $X = x$. Then the Breslow partial likelihood function for the stratified Cox regression (model 1) is given by

$$L(\alpha_0) = \prod_{k=1}^{K} \prod_{l \in \mathsf{D}_k} \frac{e^{\alpha_0 e_{l,k} + \alpha_1 e_{l,k} x_l}}{\sum_{m \in \mathsf{R}_{k,x_l}} e^{\alpha_0 e_{m,k} + \alpha_1 e_{m,k} x_l}} \tag{3}$$

$$= \prod_{k=1}^{K} \prod_{l \in \mathsf{D}_k} \frac{e^{\alpha_0 e_{l,k} + \alpha_1 e_{l,k} x_l}}{n_{k,0,x_l} + n_{k,1,x_l} e^{\alpha_0 + \alpha_1 x_l}} \tag{4}$$

Note that at time t_k, the offset term in the logistic regression (model 2) for decedents with $X = x$ is given by $\log(n_{k,1,x} / n_{k,0,x})$. The likelihood function for this model is therefore given by

$$L(\beta_0) = \prod_{k=1}^{K} \prod_{l \in \mathsf{D}_k} \left[\frac{e^{\log(n_{k,1,x_l}/n_{k,0,x_l}) + \beta_0 + \beta_1 x_l}}{1 + e^{\log(n_{k,1,x_l}/n_{k,0,x_l}) + \beta_0 + \beta_1 x_l}} \right]^{e_{l,k}} \times \left[\frac{1}{1 + e^{\log(n_{k,1,x_l}/n_{k,0,x_l}) + \beta_0 + \beta_1 x_l}} \right]^{1-e_{l,k}} \tag{5}$$

$$= \prod_{k=1}^{K} \prod_{l \in \mathsf{D}_k} \left[\frac{n_{k,1,x_l} / n_{k,0,x_l} e^{\beta_0 + \beta_1 x_l}}{1 + n_{k,1,x_l} / n_{k,0,x_l} e^{\beta_0 + \beta_1 x_l}} \right]^{e_{l,k}} \times \left[\frac{1}{1 + n_{k,1,x_l} / n_{k,0,x_l} e^{\beta_0 + \beta_1 x_l}} \right]^{1-e_{l,k}} \tag{6}$$

$$= \prod_{k=1}^{K} \prod_{l \in \mathsf{D}_k} \left[\frac{n_{k,1,x_l} e^{\beta_0 + \beta_1 x_l}}{n_{k,0,x_l} + n_{k,1,x_l} e^{\beta_0 + \beta_1 x_l}} \right]^{e_{l,k}} \times \left[\frac{n_{k,0,x_l}}{n_{k,0,x_l} + n_{k,1,x_l} e^{\beta_0 + \beta_1 x_l}} \right]^{1-e_{l,k}} \tag{7}$$

$$= \prod_{k=1}^{K} \prod_{l \in \mathsf{D}_k} \frac{e^{\beta_0 + \beta_1 x_l e_{l,k}} n_{k,1,x_l}^{e_{l,k}} n_{k,0,x_l}^{1-e_{l,k}}}{n_{k,0,x_l} + n_{k,1,x_l} e^{\beta_0 + \beta_1 x_l}} \tag{8}$$

$$= \left[\prod_{k=1}^{K} \prod_{l \in \mathsf{D}_k} \frac{e^{\beta_0 e_{l,k} + \beta_1 e_{l,k} x_l}}{n_{k,0,x_l} + n_{k,1,x_l} e^{\beta_0 + \beta_1 x_l}} \right] \left[\prod_{k=1}^{K} \prod_{l \in \mathsf{D}_k} n_{k,1,x_l}^{e_{l,k}} n_{k,0,x_l}^{1-e_{l,k}} \right] \tag{9}$$

Note that the first factor in equation 9 is equivalent to equation 4 and that the second factor in equation 9 does not depend on (β_0, β_1). The point estimate and inference obtained by maximum-likelihood estimation of (β_0, β_1) in model 2 are hence identical to those obtained by maximum-likelihood estimation of (α_0, α_1) in 1.

Part 2 – Equivalence of coefficient interpretations

Another way to derive the interpretation of the coefficients β_0 and β_1 in model 2 is based on the following algebra:

$$
\begin{aligned}
e^{\beta_0+\beta_1 x} &= \frac{\dfrac{P(E(t)=1 \mid D(t)=1, X=x, R(t)=1)}{1-P(E(t)=1 \mid D(t)=1, X=x, R(t)=1)}}{\dfrac{P(E(t)=1 \mid X=x, R(t)=1)}{1-P(E(t)=1 \mid X=x, R(t)=1)}} \\
&= \frac{\dfrac{P(E(t)=1 \mid D(t)=1, X=x, R(t)=1)}{P(E(t)=0 \mid D(t)=1, X=x, R(t)=1)}}{\dfrac{P(E(t)=1 \mid X=x, R(t)=1)}{P(E(t)=0 \mid X=x, R(t)=1)}} \\
&= \frac{P(E(t)=1 \| D(t)=1, X=x, R(t)=1)P(E(t)=0 \mid X=x, R(t)=1)}{P(E(t)=0 \| D(t)=1, X=x, R(t)=1)P(E(t)=1 \mid X=x, R(t)=1)} \\
&= \frac{\dfrac{P(D(t)=1 \| E(t)=1, X=x, R(t)=1)P(E(t)=1 \mid X=x, R(t)=1)}{P(D(t)=1 \mid X=x,\ R(t)=1)} P(E(t)=0 \mid X=x,\ R(t)=1)}{\dfrac{P(D(t)=1 \mid E(t)=0, X=x, R(t)=1)P(E(t)=0 \mid X=x, R(t)=1)}{P(D(t)=1 \mid X=x,\ R(t)=1)} P(E(t)=1 \mid X=x,\ R(t)=1)} \\
&= \frac{P(D(t)=1 \| E(t)=1, X=x, R(t)=1)}{P(D(t)=1 \| E(t)=0, X=x, R(t)=1)} \\
&= RH(t \mid X=x),
\end{aligned}
$$

where the fourth equality makes use of Bayes' rule. The coefficient β_0 in this logistic regression model 2 thus has exactly the same interpretation as the coefficient α_0 in the Cox model 1: It gives the log relative hazard of death comparing vaccinated patients with $X=0$ to nonvaccinated patients with $X=0$. Likewise, β_1 in the logistic regression model (model 2) has the same interpretation as α_1 in the Cox regression (model 1): It gives the log of the ratio of this relative hazard at $X=x+1$ to the same relative hazard at $X=x$.

Part 3 – Illustration with Simulated Data in SAS

The SAS program included below (SAS Institute Inc., Cary, North Carolina) illustrates the equivalence of case-centered logistic regression with Cox regression. We simulate a cohort of *N* people who risk death at *T* time points. We specify the proportion of the cohort that is randomly vaccinated, and we specify a relative risk by which vaccination multiplies the risk of death at each time point.

We fit 3 models and show that they yield identical estimates of the relative risk:

Model 1: Cox regression fitted to a typical cohort data set with 1 record per person (using the approximate likelihood of Breslow (reference A2)).
Model 2: Case-centered logistic regression fitted to a data set with 1 record per death.
Model 3: Case-centered logistic regression fitted to a smaller data set with 1 record per risk-set.

These 3 models are fitted to 3 data sets named Cohort, Deaths, and Risksets, respectively.

The program can be copied and run as is, or modified to try alternative specifications. It produces a summary table of results like the one included below (Appendix Table).

References for Appendix

A1. Anderson PK, Gill RD. Cox's regression model for counting processes: large sample study. *Ann Stat*. 1982;10(4):1100–1120.

A2. Breslow NE. Covariance analysis of censored survival data. *Biometrics*. 1974;30(1):89–99.

```
/* SAS PROGRAM. Set the parameters for the simulation in the line below.*/

%macro sim (N=10000, T=10, X=0.5, risk=0.05, relrisk=0.5);
    **************************************************************************************;
    * where       N = number of people                                                   ;
    *             T = number of timepoints                                               ;
    *             X = proportion of people exposed (vaccinated before the 1st timepoint)  ;
    *          risk = risk of death in each riskset in the unexposed (unvaccinated)      ;
    *       relrisk = relative risk                                                      ;
    **************************************************************************************;
data COHORT            (keep = time person death vax)
     expanded_RISKSETS (keep = riskset person death vax);
  N = &N; T = &T; prob_vax = &X; Risk=&risk; RelRisk=&relrisk;

  do person = 1 to N;
    death = 0;                                *person starts alive;
    vax = ranbin(0,1,prob_vax);               *person gets vaccinated or not;
    if vax = 0 then prob_dth = risk;          *if unvaccinated: person gets baseline risk;
    if vax = 1 then prob_dth = risk * relrisk; *if vaccinated: risk is multiplied by relrisk;

   do timepoint = 1 to T while (death=0);     *person gets into every riskset until death;
      death = ranbin(0,1,prob_dth);           *person dies or not;
      riskset = timepoint;                    *a riskset is identified for each timepoint;
      output expanded_RISKSETS;               *risksets are output here, and condensed below;
      end;
   time = timepoint - 1;               *Set time for Cox regression back to the last timepoint;
   output COHORT;                      *Cox regression model is fit to this COHORT dataset;
   end;

 proc sort data=expanded_risksets; by riskset person;

/* make a dataset with one record per riskset, for case-centered analysis by model 3 */
data RISKSETS  (keep=riskset vax_dths tot_dths odds logodds);
   set expanded_RISKSETS;
   by riskset;
   retain vax_dths tot_dths vax_n unvax_n 0;
  if first.riskset then do;
    vax_dths=0; tot_dths=0; vax_n=0; unvax_n=0;
    end;
  if death=1      then tot_dths + 1;
  if death=1 and vax=1 then vax_dths + 1;
  if vax = 1      then vax_n + 1;
  if vax = 0      then unvax_n + 1;

 if last.riskset and (vax_n>0) and (unvax_n>0) then do;
   odds = vax_n / unvax_n;
   logodds = log(odds);
   output RISKSETS;
   end;

/* make a dataset with one record per death, for case-centered analysis by model 2 */
Data DEATHS (keep = Case_exposed logodds);
   Set RISKSETS;
   Do death = 1 to tot_dths;
    If death <= vax_dths then case_exposed = 1;
    Else case_exposed = 0;
    output DEATHS; end;
```

```
/* Model 1: Cox regression */
   ods output phreg.parameterestimates=parms1;
 proc PHREG data=COHORT nosummary;
 model time*death(0)= vax / ties=BRESLOW convergeparm=0.000001;
   title 'Model 1: Cox regression on cohort dataset';
   data parms1; set parms1; length model $40.;
   model = 'Cox regression';
   keep model estimate stderr;

/* Model 2: Case-centered logistic regression with 1 record per death */
   ods output logistic.parameterestimates=parms2;
 proc LOGISTIC descending data = DEATHS;
 model case_exposed = / offset=logodds converge=0.000001;
 title "Model 2: Case-centered logistic regression using the dataset with 1 record per death";
   data parms2; set parms2; if _n_=1;
   model = "Case-centered, 1 record per death";
   keep model estimate stderr;

/* Model 3: Case-centered logistic regression with 1 record per riskset */
   ods output logistic.parameterestimates=parms3;
 proc LOGISTIC descending data= RISKSETS;
 model vax_dths / tot_dths = / offset=logodds converge=0.000001;
 title 'Model 3: Case-centered logistic regression using the dataset with 1 record per riskset';
   data parms3; set parms3; if _n_=1;
        model = 'Case-centered, 1 record per riskset';
        keep model estimate stderr;

/* Merge results from the three models */
 data results;
   set parms1 parms2 parms3;
   N = &N;T=&T;relrisk=&relrisk;risk=&risk;prob_vax=&X;
   rr = exp(estimate);
   spread = 1.96*stderr;
   lower = exp(estimate - spread);
   upper = exp(estimate + spread);
   covered='No ';                          *Yes, if the 95% CI covers the true relative risk;
   if lower < relrisk < upper then covered='Yes';

/* Label and print results */
  proc print data=results split='*';
   id model; var N T prob_vax risk relrisk rr lower upper covered;
   format N comma9. rr lower upper 7.4 risk prob_vax relrisk 5.2;
   label model = 'Type of*Model'
      prob_vax = 'Probability*of*vaccination'
      risk     = 'Risk of*death*at each*time point*in unvaxed'
      relrisk  = 'True*underlying*relative*risk'
      rr       = "Estimate*  of  *relative*risk"
      T        = 'Number of* time *points'
      N        = 'Number*  of  *people'
      lower    = "Lower limit*of*95% CI"
      upper    = "Upper limit*of*95% CI"
      covered  = "Is true*rel. risk*inside the*95% CI?";
 title  'Table. Cox regression compared with two versions of case-centered logistic regression';
 title2 'Summary of three models fit to the same simulated data'; run;
%mend sim;
%sim  (N=10000, T=10, X=0.5, risk=0.05, relrisk=0.5);
```

Appendix Table. Summary of 3 models fitted to the same simulated data (Cox regression compared with 2 versions of case-centered logistic regression)

Type of Model	Number of people	Number of time points	Probability of vaccination	Risk of Death at each time point in unvaxed	True underlying Relative Risk
Cox regression	10,000	10	0.50	0.05	0.50
Case-centered, 1 record per death	10,000	10	0.50	0.05	0.50
Case-centered, 1 record per riskset	10,000	10	0.50	0.05	0.50

Type of Model	Estimate of relative risk	Lower limit of 95% Conf Int	Upper limit of 95% Conf Int	Is the true RR inside the 95% CI?
Cox regression	0.4825	0.4483	0.5194	Yes
Case-centered, 1 record per death	0.4825	0.4483	0.5194	Yes
Case-centered, 1 record per riskset	0.4825	0.4483	0.5194	Yes