Biometrics 000, 000–000 DOI: 000
000 0000

# Web-based supplementary materials for "Statistical methods for analysis of radiation effects with tumor and dose location-specific information with application to the WECARE study of asynchronous contralateral breast cancer"

**Bryan Langholz, Duncan C. Thomas, Marilyn Stovall, Susan A. Smith, John D. Boice, Jr,**

**Roy E. Shore, Leslie Bernstein, Charles F. Lynch, Xinbo Zhang,**

**The WECARE Study Group, and Jonine Bernstein**

## Web Appendix A: Formal derivation of the precise location likelihoods

In this section, we provide a probabilistic framework based on a counting process representation of the location-specific outcome and dose data along with the associated intensity processes. These are then extended to accommodate case-control sampling from the failure-time defined risk sets, appropriate to most of the studies that collected location-specific radiation dose information. The CCML, CL, and CCAL likelihoods can be derived as partial likelihoods, conditioning on different summary counting processes. Under realistic conditions, standard techniques assure that, within the context of the modes for location or/and covariate effects, estimating the same dose effects parameters and that the likelihoods all have "standard likelihood properties." As before, we focus our discussion on the WECARE study but the techniques apply generally.

*Subject- and location-specific counting processes and intensities*

We first represent the basic structure of the WECARE cohort with second breast cancer location and location-specific dose data as standard multivariate counting process data, e.g., (Arjas, 1989; Andersen et al., 1993). Let $N_{i,l}(t)$ the counting process that "counts" failure occurrences for subject $i$, location $l$ up to time $t$. A key and very reasonable "continuity assumption" is that CBCs, both across subjects, and across locations within subject, do not occur simultaneously. With $\mathcal{F}$ the filtration based on censoring, covariate, stratum, dose, and counting process information, the continuity assumption is equivalent to 1) the $\mathcal{F}$-intensity associated with $N_{i,l}(t)$ is $\text{pr}[dN_{i,l}(t) = 1|\mathcal{F}_{t^-}] = \lambda_{i,l}(t;\mathcal{F})dt$ where $\lambda_{i,l}(t;\mathcal{F})$ is a non-negative piecewise continuous function of information in $\mathcal{F}_{t^-}$ and 2) $\text{pr}[dN_{i,l}(t) = dN_{j,m}(t) = 1|\mathcal{F}_{t^-}] = o(dt)$ for $i, l \neq j, m$. For the WECARE study, the cohort members are unilateral breast cancer women ascertained by one of the participating cancer registries during the "enrollment period." Nine counting processes are associated with each subject, one for each the of the nine locations.

Given the specification for cohort data, extension to nested case-control data is accommodated by incorporation of the case-control set selection into the counting processes (Borgan et al., 1995; Langholz and Goldstein, 1996, e.g.,). Let $N_{i,l,\mathbf{r}}(t)$ be the counting process for subject $i$, location $l$ as before and let $\mathbf{r}$ be the indices for subjects in a possible case-control set sampled from the cohort risk set if $i$ were to fail at time $t$. Then the intensity process associated with $N_{i,l,\mathbf{r}}(t)$, under mild assumptions (i.e., that failure is independent of previous case-control sampling), is given by $\lambda_{i,l}(t;\mathcal{F})\ \pi_t(\mathbf{r}|i)$ where $\pi_t(\mathbf{r}|i)$ is the probability of picking $\mathbf{r}$ as the case-control set given that $i$ is the case (Borgan et al., 1995). For the WECARE study, $\mathbf{r}$ represents the possible counter-matched sets and $\pi_t(\mathbf{r}|i)$ are the counter-matching probabilities (Langholz and Borgan, 1995).

The precise location likelihoods are partial likelihoods, in the same sense as Cox 1972 (Cox, 1972; Andersen et al., 1993). Let $N_{l,\mathbf{r}} = \sum_i N_{i,l,\mathbf{r}}$, $N_i = \sum_{l,\mathbf{r}} N_{i,l,\mathbf{r}}$, and $N_{\mathbf{r}} = \sum_{i,l} N_{i,l,\mathbf{r}}$. The probabilities that $dN_{i,l,\mathbf{r}}(t) = 1$ conditional on $dN_{l,\mathbf{r}}(t) = 1$, $dN_i(t) = 1$, and $dN_{\mathbf{r}}(t) = 1$ each have the form

$$\frac{\lambda_{i,l}(t;\mathcal{F})\ \pi_t(\mathbf{r}|i)}{\sum_{j,m\in\mathcal{C}} \lambda_{j,m}(t;\mathcal{F})\ \pi_t(\mathbf{r}|j)} \tag{1}$$

where $\mathcal{C}$ is the subject, location set appropriate to the conditioning event. Replacing $\lambda_{i,l}(t;\mathcal{F})$ by the appropriate rate models given in Section 3, yields the precise location CCML, CL, and CCAL likelihoods given in Table 1. Just as in standard event time data, even though the likelihood contributions are not independent, the partial likelihood that is the product of the conditional probabilities has the usual likelihood properties. General conditions for consistency and asymptotic normality of partial likelihood estimators from nested case-control studies apply (Borgan et al., 1995).

## Web Appendix B: Simulation study of CCML, CCAL, and CL likelihood efficiency

In order to more precisely compare efficiency of the precise location likelihoods in the context of the WECARE study, we used a simulation study in which we randomly assigned subject/location $(D, L)$ within the case-control set based the conditional probability Web equation (1) under model (7)with $\gamma$ and $\delta$ set to values close to those estimated from the data. The ERR model was used for radiation dose response and we performed simulations for different values of $\beta_0$. We then computed the information for estimating $\beta$ as the inverse of the $\beta, \beta$ "corner" of the parameter covariance matrix at the true parameter values for the CCAL, CCML, and CL analyses. Note that for the CCAL the expected information at the true values does not depend on the simulated subject/location so no simulation is needed. Although the variability was quite small, the CCML and CL information depend on the (randomly chosen) location and subject, respectively, so we averaged the CCML and CL $\beta$ information from 20 replications. The relative efficiencies for a range of plausible $\beta_0$ are given in Web Table 1. The CL and CCAL $\beta_0 = 0$ relative efficiencies are consistent with those based on the estimated null information (21% and 111%, respectively). Both the CL and CCAL relative efficiency decrease as $\beta_0$ increases. In particular, the increases over CCML realized from the CCAL are trivial over the range of $\beta_0$ interest.

[Table 1 about here.]

## Web Appendix C: Formal derivation derivation of location-group likelihoods

We define counting processes $N_{i,\mathbf{l}}$ to indicate a CBC for subject $i$ with location-group $\mathbf{l}$, $\mathbf{l} \subset \{1, \ldots, 9\}$. Let $\mathcal{F}^*$ be the "expanded filtration," $\mathcal{F}$ plus the $N_{i,\mathbf{l}}$. As before, let $N_{il}$ be the counting process associated with precise location $l$. Then, the intensity $\lambda_{i,\mathbf{l}}(t; \mathcal{F})$ associated with $N_{i,\mathbf{l}}(t)$ is derived as

$$\begin{aligned} \mathrm{pr}(dN_{i,\mathbf{l}}(t) = 1 | \mathcal{F}^*_{t-}) &= \sum_{l \in \mathbf{l}} \mathrm{pr}(dN_{i,l}(t) = 1 \text{ and } dN_{i,\mathbf{l}}(t) = 1 | \mathcal{F}^*_{t-}) \\ &= \sum_{l \in \mathbf{l}} \mathrm{pr}(dN_{i,l}(t) = 1 | \mathcal{F}_{t-}) \ \mathrm{pr}(\mathbf{L} = \mathbf{l} | dN_{i,l}(t) = 1, \mathcal{F}^*_{t-}) \\ &= \sum_{l \in \mathbf{l}} \lambda_{i,l}(t; \mathcal{F}) \ \rho_{i,l,t}(\mathbf{l}) \ dt \qquad (2) \end{aligned}$$

where $\rho_{i,l,t}(\mathbf{l})$ is the probability of a group location $\mathbf{l}$ given the actual location $l$ and could, in general, depend on subject $i$- and time $t$-specific factors. Extension to case-control sampled data is precisely as for the precise location situation.

Assuming, for instance, the CCAL subject/location rates model (1), the natural induced location-group model based on (2) is $\lambda(t, \mathbf{l}, s, c, \mathbf{z}; \alpha, \eta, \beta) = \sum_{l \in \mathbf{l}} [\alpha_s(t) \ x(c, l; \eta) \ r(z_l; \beta) \ \rho(\mathbf{l}|l)]$ where, for simplicity, we have taken the $\rho$ to depend only on $l$. The induced CCML and CL models are similarly defined. Completely analogous to the precise location situation, likelihood construction is then based on the probabilities that $dN_{i,\mathbf{l}}(t) = 1$ conditional on $dN_{\mathbf{l},\mathbf{r}}(t) = 1$, $dN_i(t) = 1$, and $dN_{\mathbf{r}} = 1$, respectively, replacing $\lambda_{i,\mathbf{l}}(t; \mathcal{F})$ by the appropriate induced rates model.

## References

Andersen, P. K., Borgan, Ø., Gill, R. D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes.* Springer Verlag, New York.

Arjas, E. (1989). Survival models and martingale dynamics (with discussion). *Scand. J. Statist.* **16,** 177–225.

Borgan, Ø., Goldstein, L., and Langholz, B. (1995). Methods for the analysis of sampled cohort data in the Cox proportional hazards model. *Annals of Statistics* **23,** 1749–1778.

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

Langholz, B. and Borgan, Ø. (1995). Counter-matching: A stratified nested case-control sampling method. *Biometrika* **82,** 69–79.

Langholz, B. and Goldstein, L. (1996). Risk set sampling in epidemiologic cohort studies. *Statistical Science* **11,** 35–53.

**Table 1**

*Relative expected information (to CCML) for estimation of $\beta$ over a range of true $\beta_0$ values from CCML, CL, and CCAL analyses from 609 WECARE precise location case-control sets with simulated subject/location outcomes.*

| Analysis | ERR $\beta_0$ | | | |
|---|---|---|---|---|
| method | 0 | 0.25 | 0.50 | 0.75 |
| CCML | 100% | 100% | 100% | 100% |
| CL | 28% | 14% | 9% | 7% |
| CCAL | 109% | 105% | 102% | 102% |