Modeling the Probability of Second Cancer in Controlled Clinical Trials

Kao-Tai Tsai
Celgene Corporation, NJ & JPHCOPH, Georgia Southern University, GA

Abstract

Due to the advancement of medical technologies and cancer care, the long-term survival of cancer patients have been substantially increased. For some patients, increases in survival have been offset by the long-term late effects of cancer and its treatment. One of the most life-threatening sequelae is the diagnosis of a new malignant cancer. The number of patients with multiple primary cancers is growing with second new cancer now representing approximately 16% cancers reported to the SEER Program of NCI. Second cancers reflect not only the late effects of therapy but also the influence of shared etiologic factors, genetic susceptibility, environmental exposures, cancer drug exposure, and older age, etc. In this research, we attempt to outline a framework to model the patient-level probability of the occurrence of new cancer malignancies using demographics, efficacy, and safety data which are commonly collected in clinical trials.

Keywords: Drug Safety Data, Second Cancer, Longitudinal Repeated Measures, Control Clinical Trials.

1 Introduction

Due to the advancement of medical technologies and cancer care, the long-term survival of cancer patients have been substantially increased. For example, as of January 2008, there were approximately 11.9 million cancer survivors in the US (about 4% of the population) and the 5-year relative survival rate for all cancers combined has increased steadily to reach 66.1% for patients diagnosed from 1999 to 2006. For some patients, increases in survival have been offset by the long-term late effects of cancer and its treatment. One of the most life-threatening sequelae is the diagnosis of a new malignant cancer. The number of patients with multiple primary cancers is growing with second new cancer now representing approximately 16% cancers reported to the SEER Program of NCI.

Second cancers reflect not only the late effects of therapy but also the influence of shared etiologic factors (in particular, tobacco and excessive alcohol intake), genetic susceptibility, environmental exposures, host effects, and combinations of factors, including gene-environment interactions. Risks for selected second new cancers are also modified by age at exposure and attained age.

During the course of a clinical trial, huge amount of data are collected. This usually include patient demographics data, efficacy and safety data of the intervention, in addition to certain degree of follow-up data. It would be a great advantage for the researcher to look into these data and identify the risk factors which can possible lead to the second new cancers. In this article, we attempt to outline a framework to model the patient-level probability of the occurrence of new cancer malignancies using demographics, efficacy and safety data commonly collected in clinical trials during the active treatment or follow-up periods.

2 The Difficulties in Second Cancer Research

Majority of the journal articles about clinical trials are based on short-term follow-up with small to moderate sample sizes which makes it inadequate to capture the incidence of second cancer due to the medium to long latency. In addition, most of the cancer trials involve combinations of multiple drugs. The patient crossover from one treatment to the other treatment when one drug does not show efficacy makes an inadequate control group for fair comparisons. On the other hand, most of the report of systematic reviews are retrospective with mixture of patient populations. Furthermore, under-reporting of incidence at the patient-level detail data by clinicians makes the tracking of new cancers more difficult.

3 Current Knowledge of Second Cancer

Cancer survivors constitute 3.5% of the US population, but second cancer among this high-risk group now account for 16% of all cancer incidence. *(Travis, L., et al., JNCI, 2006).* Based on the NCI SEER database, compared to the general population, cancer survivors have a 14% increased risk of developing a new cancer *(Curtis, FD., SEER, 2006).*

Drug Effects: Ironically, the medicines or medical procedures to treatment current diseases can also possibly cause new cancers later. For example, besides other reasons, radiotherapy may have a potential role in development of second cancer following a current malignant cancer *(Featherstone, C., et al., Cancer, 2005)*

Disease Effects: In a Swedish study, based on 5652 patients with multiple myeloma (MM) precursor disease (MGUS) with IgG/IgA, an 8-fold increased risk of developing MDS/AML was observed *(Mailankody, S., Blood, 2011).* These observations support a role for disease related factors in MDS/AML following MM *(Mailankody, S., Blood, 2011).* Therefore, a better understanding of underlying molecular mechanisms across MM subgroups and risk of second malignancy will form the basis for target therapies to minimize the risk of second malignancies.

Genetic Effects: It has been estimated that genetic variations can account for up to 95% of variability in drug disposition and effects. Polymorphisms in genes encoding drug-metabolizing enzymes, DNA repair pathways, drug transporters and targets may also contribute to an individual's susceptibility for subsequent malignancies as well *(Evans, WE., et al., NEJM, 2003).* Non-genetic factors which can modulate treatment effects include age, race, organ function, concomitant therapy, drug interactions, and disease itself.

Environmental Effects: Studies indicate that exposure to ionizing radiation at younger ages and higher doses increases the risk of developing cancer diseases *(Landgren, O., et al., Blood, 2009).* Chronic antigen stimulation from prior autoimmune, infectious, inflammatory, allergic disorders and immune dysregulation may play a role in pathogenesis of som specific kind of cancer diseases *(Kristinsson, SY., et al., JCO, 2011).* In addition, socioeconomic status has been shown to influence cancer survival, suggesting that lifestyle factors in these disorders are of importance.

Behavior Effects: The commonly proposed behavioral risk factors (e.g. tobacco, alcohol and diet) and obesity are among the various types of behavioral effects that can potentially affect the occurrence of cancers *(Kyle, RA., et al., Clin Haematology, 2007)*.

4 Data Sources from Clinical Trial Database

There is usually a huge amount of data collected in clinical trials. These data include the demographics data, such as age, gender, disease stage, performance score, baseline cytogenetic status, treatments received, etc. The efficacy data such as disease progression, time-to-progression, overall survival, censor status, etc. And the safety data including lab test results, medical history, concomitant medicine, etc. It is at a great advantage to utilize these data for the research of event of interest.

As in the general data analysis, not all the data will be important, therefore, it is crucial to select the relevant factors or covariates for analysis. Since the lab data were collected in most of the clinical visits with a longitudinal data type, the selection needs to take into account of this repeated measure nature. A conventional approach is to use the mixed effect model to select the significant lab test parameters. In addition, graphical tools are also important to supplement the inferential findings. The specification of the mixed effect model and the parameter estimations can be found in many monographs. Specifically, for the purpose in the data analysis, one can denote the lab data, at time points $\{t_{ij} : j = 1, \cdots, n_i\}$, for subject i by

$$y_{ij} = \{y_i(t_{ij}), j = 1, \cdots, n_i\}.$$

One can estimate the effect of lab test parameters by assuming a linear mixed effect model for the repeated data:

$$\begin{aligned} y_i(t) &= m_i(t) + \epsilon_i(t) \\ &= x_i'(t)\beta + z_i'(t)b_i + \epsilon_i(t), \quad \epsilon_i(t) \sim N(0, \sigma^2); \end{aligned} \tag{1}$$

where β is the vector of the unknown fixed effect parameters, b_i is a vector of random effect parameters and assume $b_i \sim N(0, \Sigma)$, $x_i(t)$ is the design matrix for the fixed effect, $z_i(t)$ is the design matrix for the random effect, and $\epsilon_i(t)$ is the measurement error independent of b_i.

5 Medical History and Concomitant Medicines

To analyze safety data, it is critically important to understand the patient's medical history as some of these medical history may affect the safety events during the study period. For cancer disease, especially for multiple myeloma, the following medical history were found to be important in studies investigated previously: aortic valve stenosis, basal cell carcinoma, benign prostatic hyperplasia, bronchitis chronic, cardiac failure, cholecystitis, chronic obstructive pulmonary disease, epistaxis, herpes zoster, hypoaesthesia, infection, menopause, oedema peripheral, oesophagitis, periarthritis, polyneuropathy, prostate cancer, sciatica, upper respiratory tract infection, vitamin B12 deficiency, vitamin D deficiency. These medical history items were selected using a simple χ^2 test of significance in the relationship to the occurrence of second cancer.

Similarly, concomitant medications can also affect the treatment efficacy and new cancer occurrences. The following concomitant medicine were found to be significant factors in the analysis of cancer occurrence: A06AD -lLaxatives, C09DA - angiotensin II antagonists and diuretics, C10AB - lipid modifying agents, plain, D06BB - antivirals, J01DB, J01FF - antibacterials, N05CD - benzodiazepine derivatives, N05CF - cyclopyrrolones, R03AK - adrenergics and other anti-asthmatics. Graphical method such as biplot [2] can be quite informative to reveal their relative relation to the second cancers.

6 Efficacy Data Consideration

Since the drug efficacy and safety are generally intertwined, to have a more complete data analysis, one is recommended to also include efficacy data information in the model. Since disease progression is usually among one of the key efficacy variables, it can be quite informative to incorporate this factor into the overall model. The hazard of disease progression was estimated using Muller & Wang [5].

Specifically, let (T_i^*, C_i, T_i) = (failure time, censor time, observation time) for the i-th subject $(i = 1, \cdots, n)$, the hazard estimators can then be obtained by smoothing the increments of the Nelson-Aalen estimator

$$\Lambda_n(t) = \sum_{i=1}^{n} \delta_{[i]} I_{(X_{(i)} \leq t)}/(n - i + 1). \tag{2}$$

of the cumulative hazard function $\Lambda(t)$, where $\delta_{[i]}$ is the censoring indicator of $X_{(i)}$. Using the kernel method, one can estimate the kernel hazard rate function by

$$\Lambda_n(t) = \sum_{i=1}^{n} \{\delta_{[i]}/(n - i + 1)\}(1/h)K((t - X_{(i)})/h). \tag{3}$$

7 Integrated Analysis of Efficacy and Safety Data

With all the important variables selected from various data sets of the clinical trials, one can proceed to estimate the probability of second cancer using these variables in an overall model. For example, one can use the following mixed effect model (in R-codes) to perform the needed estimate of the possible effect of each covariate to the occurrence of second cancer:

```
glmer(SPM ~ lymp + nuet + age + pscore + pfscumhaz + trt +
        sex + stage + creat + sigmedhist + conmed + (1|trt),
        data=yy2, na.action=na.omit, family=binomial)
```

In addition to the parametric modeling, one cal also use the model-based recursive partitioning [1] [3] to further subset the data and find the various subgroups of subjects that seem to have more homogeneous characteristics and can be better classified of their respective probabilities of having second cancer.

It is important to examine the resulting outputs from these canned programming tools to ensure the plausibility of the findings in either biological or clinical senses. Furthermore, one can also construct the confidence Intervals of the estimates to better understand the stability of the estimates.

8 Assessing Model Adequacy

As a general practice, one needs to assess the goodness-of-fit after the model is fitted. Following the approach of Landwehr, et al. [4] to create pseudo-repeated measures, the following steps were taken to assess the adequacy of the mode: (1) partition the N observations into K non-overlapping clusters with N_k observations in each; (2) form the $N \times K$ matrix $\mathbf{Z} = (z_{ik})_{N\times K}$ with $z_{ik} = 1$, if ith observation is in the kth cluster and 0 otherwise; (3) compute the local estimate $\hat{p}_l$ by fitting

$$\text{logit}(p_l) = \mathbf{Z}\gamma + \mathbf{X}\beta + \epsilon;$$

(4) use the model fitted above to compute the local deviance contribution of each observation, $d(\hat{p}_{l,jk}, y_{jk})$, and sum the deviances within each group giving D_{lk}; (5) compute running estimates of approximate pure error

$$D_L(t) = \sum_{k=1}^{t} D_{lk} / \sum_{k=1}^{t} (N_k - 1).$$

The values $D_L(t)$ represent the local mean deviance calculated from the tightest t clusters; (6) plot $y(t) = D_L(t)$ against its degrees of freedom $x(t) = \sum_{k=1}^{t}(N_k - 1)$, for $t = 1, \cdots, k$ and superimpose the plot above with the line of global mean deviance, $Y = D_L(K)$, and observe its position relative to the points plotted above. Substantial deviation indicates systematic lack-of-fit.

Depending on the adequacy of the model, one may further extend the linear model to generalized additive models such as

$$\mathbf{y} = g(\mu^{\mathbf{b}}) + \epsilon, \tag{4}$$

with

$$g(\mu^{\mathbf{b}}) = \mathbf{X}\beta + \sum_{j=1}^{k} f_j(\mathbf{x_{ji1}}, \cdots, \mathbf{x_{ji_j}}) + \mathbf{Z}\mathbf{b}, \tag{5}$$

and

$$\mathbf{b} \sim N(0, \psi_\theta), \quad \epsilon \sim N(0, \mathbf{\Lambda}\sigma^2), \tag{6}$$

where $g(\cdot)$ is link function, $\mu^{\mathbf{b}} \equiv \mathbf{E}(\mathbf{y}|\mathbf{b})$, $\mathbf{y}|\mathbf{b} \sim$ exponential family, and f_j is a smooth/non-smooth function of columns of $\mathbf{X}$. The error term ϵ may have several variance components to take care of the possible random effects from multiple factors in the data.

9 Summary

Second cancer issue posts a unique challenge in medical research. The advancement of medical technologies and medications prolongs human life and, simultaneously, increases the possibility for people to acquire diseases which related to either the older age or the exposure of cancer-causing substances and life-styles. Research on this issue based on controlled clinical trial data usually is insufficient due to the medium to long latency of majority of second cancers and insufficient duration of patient follow-up after finishing the active controlled treatment. Aggregated results from epidemiological research cannot really meaningfully pinpoint the underline causes of the second cancers, not even mention the prediction of disease at patient-level.

In this article, we outline the steps to estimate the probabilities of the occurrence of second tumor malignancies based on both efficacy and safety data of clinical trials. With carefully collected data, we had shown sufficiently good prediction of the occurrence of second cancer for the patients in the database, which is not shown here due to the data confidentiality.

Prior to statistical modeling, we utilize well designed graphical methods to help understand the intricacies and inter-relationship of data. Several statistical methods were used to analyze the data and their performances were compared with the method proposed above. We found that careful examinations of data, results, and discussions with subject experts are essential to avoid the pitfalls of some canned programs and to obtain results which are clinically sensible. Even though we have achieved reasonable success, lots of unknowns still need to be researched and further results will be reported in the later publications.

References

[1] Breiman, L. (2001), Random Forests, Machine Learning v45(1), p5-32.

[2] Gabriel, K. R. (1971) The biplot graphical display of matrices with application to principal component analysis. Biometrika v58, p453-467.

[3] Hothorn, T., Hornik, K., and Zeileis, A. (2006). Unbiased Recursive Partitioning: A Conditional Inference Framework. Journal of Computational and Graphical Statistics, v15(3), p651- 674.

[4] Landwehr, J.M., Pregibon, D., and Shoemaker, A.C. (1984) Graphical Methods for Assessing Logistic Regression Models, JASA, v79, p61-71.

[5] Mueller, H.G. and Wang, J.L. (1994) Hazard Rates Estimation Under Random Censoring with Varying Kernels and Bandwidths, Biometrics v50, p61-76.