Joint Analysis of Independent Datasets: Application to Genetic Effects of Breast Cancer Survival

I. Akushevich[1], A.Yashkin[1], B.Durgin[1], J. Kravchenko[2], K.Arbeev[1], A.I.Yashin[1]

[1]Biodemography of Aging Research Unit, Social Science Research Institute, Duke University, Durham, NC
[2]Department of Surgery, Duke University Medical Center, Durham, NC

Abstract.

No observational study can collect all of the data necessary to fully model a given pathological process in a diverse population. This leads to the existence of complementary datasets varying in size and level of detail that are not directly linkable (e.g. survey data, genetic data, administrative databases, disease registries) but represent important aspects of the same pathological process. In this paper we develop an approach for joint analyses of such data. We relate genetic markers to stage-specific survival after breast cancer diagnoses using i) HRS-Medicare data with hundreds of cases, extensive genetic measurements, but without stage or other cancer characteristics, and ii) SEER-Medicare with millions of cases, detailed cancer characteristics, but without genetic measurements. Since the same underlying model generates both datasets, the likelihood function is expressed using the same set of model parameters for both datasets. The approach is illustrated by simulation studies and application to real data.

Introduction

Important Information which represents different aspects of a process under study is often available in different independent and not directly linkable datasets. Typically methods of statistical matching[1,2]—a series of statistical methods whose objective is the integration of two (or more) data sources (usually samples) referred to the same target population—are used in such situations. The objective of these approaches is to study relationship among variables not jointly observed in a single data source[1]. As a result the variables measured in only one dataset (and therefore needing to be filled in the other one) are completely explained by variable(s) commonly measured in both datasets. In this paper we analyze the effect of genetics on stage-specific survival from a cancer diagnosis. In this setting, the outcome variable (time to death) is the only variable common to both datasets and there are no other common variables that would essentially define dynamics of the process. Specifically we use two datasets that are not directly linkable in any way: i) SEER-Medicare data and ii) HRS-Medicare data. Genetic factors are measured in HRS-Medicare only, while stage, tumor size, and other cancer diagnosis characteristics are only measured in the SEER-Medicare. There are no essential predictors commonly measured in both datasets. Therefore standard methods of statistical matching are not applicable and we need a new approach to solve this task.

The idea of the approach is as follows. If variable(s) is not measured in a specific dataset, but is measured in another dataset, and the distribution of the variable(s) is known or assumed, then the likelihood function can be written as the product of two dataset-specific likelihoods in which the model of missing data is created by analytic averaging using the known (or assumed) distribution. Dataset-specific likelihoods are different but expressed by the same set of parameters. The likelihood function of both datasets is the product of two terms that although different (because of different averaging for each dataset) are still expressed using the same set of model parameters (because of the same true model) that are subject to estimation through maximum likelihood. The methodological challenge is to perform analytical averaging over these missing data for the respective parts of the likelihood function. Such averaging can be performed without the additional assumptions that are necessary for the current approaches of statistical matching and allow for the analyses of outcomes not currently possible due to lack of simultaneously measured key variables.

Model description.

Consider a normally distributed time-invariant predictor g_i (e.g., polygenetic risk score based on biomarker information) and stage at initial breast cancer (BC) diagnosis. The polygenetic risk score is known in HRS but not in SEER-Medicare and conversely for stage. We assume that survival time has a Weibull distribution. This is empirically justified (Figure 1) and used in analyses of cancer survival in the literature [4-7]. If all data on both stage and genetic risk score were measured, the likelihood in standard survival analysis would be:

$$L = \prod_i \prod_j \left(p_j f_N(g_i) \mu_j(t_i, g_i)^{\delta_i} S_j(t_i, g_i) \right)^{\delta_{ij}},$$

where i runs over individuals, j runs over stages at initial diagnosis, t_i denotes survival time, g_i a time-invariant covariate (e.g., polygenetic risk score), δ_i a censoring indicator, and δ_{ij} is the indicator that cancer was initially diagnosed at stage j. The terms p_j and $f_N(g_i)$ stand for the probability to have stage j at diagnosis and the density of the normal distribution for the polygenetic risk score, respectively. Stage-specific hazards and survival functions in the Weibull model for survival time and quadratic hazard for the genetic score are

$$\mu_j(t_i, g_i) = \gamma t_i^{\gamma-1}(\mu_{j0}(1+\beta_j(g_i-g_0)^2)),$$
$$S_j(t_i, g_i) = \exp\left(-t_i^{\gamma}(\mu_{j0}(1+\beta(g_i-g_0)^2))\right),$$

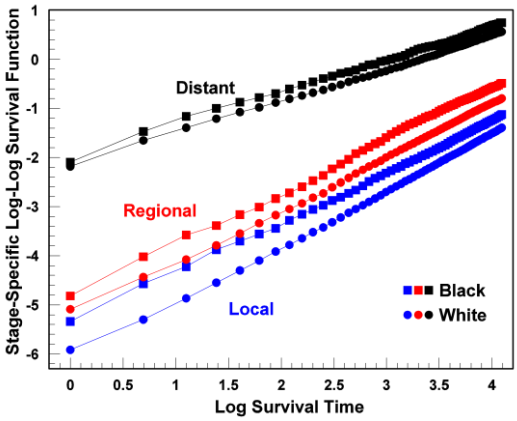

Figure 1. Stage and race-specific survival functions (in the $\log(-(\log(S(t)))$ vs. $\log(t)$ format) for individuals diagnosed in 1995-2007

where γ and μ_{j0} are Weibull parameters (the parameter γ could be considered stage-independent, this is empirically justified at least for local and regional stages (Figure 1); however this assumption is not critical and can be relaxed in model extensions), g_0 is the norm with respect to the covariate, i.e., the value of the covariate when survival is highest, and β is the effect of a deviation of the genetic risk score from the norm. In the simplest formulation, g_0 and β are stage-independent.

In HRS, stage is unknown. Thus, the survival function is averaged over all stages and the hazard function is then calculated through the derivative of averaged survival:

$$S(t_i, g_i) = \sum_j p_j S_j(t_i, g_i),$$
$$\mu(t_i, g_i) = \sum_j p_j \mu_j(t_i, g_i) S_j(t_i, g_i) / S(t_i, g_i).$$

The likelihood for the HRS data is

$$L_{HRS} = \prod_i \mu(t_i, g_i)^{\delta_i} S(t_i, g_i) f_N(g_i).$$

The likelihood for the SEER-Medicare data is obtained by averaging over g_i (through integration of the survival function $S_j(t_i, g_i)$ with respect to g_i with the density of normal distribution):

$$L_{SEER} = \prod_i \prod_j \left(p_j \mu_j(t_i)^{\delta_i} S_j(t_i)\right)^{\delta_{ij}},$$

where

$$S_j(t_i) = \sqrt{d_j}\exp\left(-\mu_{j0} t_i^{\gamma}(1+d_j\beta(g_0-g_m)^2)\right),$$
$$\mu_j(t_i) = \mu_{j0}\gamma t_i^{\gamma-1}\left(1+d_j\beta\sigma^2+d_j^2\beta(g_0-g_m)^2\right),$$

with $d_j = (1+2\mu_{j0}\beta\sigma^2 t_i^{\gamma})^{-1}$. Here g_m and σ^2 are the mean and variance of the normal distribution of g_i in the population. Model parameters are estimated from a pooled dataset using the likelihood

$$L = L_{SEER} \cdot L_{HRS}.$$

Simulation Studies

We first apply the most straightforward approach to parameter estimation, maximizing the likelihood $L = L_{SEER} \cdot L_{HRS}$ while evaluating all parameters. Since this main approach can have deficiencies, several alternative approaches can be defined. For example, the second term in $d_j = (1+2\mu_{j0}\beta\sigma^2 t_i^{\gamma})^{-1}$ could be small (i.e., $2\mu_{j0}\beta\sigma^2 t_i^{\gamma} \ll 1$), resulting in the identifiability of only the parameters μ_{j0} and γ for each stage-specific subsample. Furthermore, the number of individuals in SEER-Medicare is much larger than that in the HRS, so uncertainty in estimates of $\log(L_{SEER})$ could be comparable with $\log(L_{HRS})$ that worsen the estimates of

parameters contributed only in $\log(L_{HRS})$. Since different parts of the likelihood could depend only on a part of parameters (L_{HRS} and stage-specific parts of L_{SEER}), we have opportunities to define and test alternative (potentially more effective) approaches to maximize the joint likelihood. We expect that the final approach should include all or several of the following components: i) estimation of g_m and σ^2 from the empirical distribution in HRS, ii) estimation of stage-at-diagnosis frequencies p_j from the empiric distribution in SEER-Medicare, iii) maximizing stage-specific parts of L_{SEER} to estimate μ_{j0} and compare estimates of γ among different stages, iv) estimate β and g_0 from L_{HRS}. If estimates of γ are significantly different we will extent the model with stage-specific γ.

We evaluate the methodological framework developed above by a series of simulation studies. In the simulation studies we assume the correctness of the model within a range of parameters and then reconstruct their true values using the model estimation procedures based on the likelihood function $L = L_{SEER} \cdot L_{HRS}$. We assessed the accuracy of these parameter estimates using the t-test to compare the estimated parameter means over 100 simulations with the true parameter values based on the data. The simulation studies allow us to identify areas in the parameter space with good and bad identifiability. If areas with bad identifiability are found, we repeat simulation studies replacing certain parameters by their estimates obtained based on clinical expertise. We also evaluated the sensitivity of estimated parameters for alternative assumptions for the construction of our model.

Technically, we simulate a dataset setting the true parameters as shown in the first line of Table 1. Then the data are simulated in five steps. First, the polygenic risk score g_i is simulated normally with mean g_m and variance σ^2. Second, stage at diagnosis ($s = 1,2,3$) is simulated using an uniformly distributed random number r_1, i.e., $s = 1,2,3$ if $p_2 + p_3 < r_1 \le 1$, $p_3 \le r_1 \le p_2 + p_3$, and $0 \le r_1 \le p_3$ respectively; I_s is the indicator of simulated stage. Third, time to death τ_i is calculated as $\left(-\log(r_2)/(\mu_{0s}(1+b(g_i-g_0)^2))\right)^{1/\gamma}$, where r_2 is uniformly distributed random number. Fourth, death indicator d_i is 1 if $\tau_i < 30$ and 0 if $\tau_i \ge 30$; $\tau_i = \min(\tau_i, 30)$. We simulated 100K patients. We additionally assume that N_g =1%, 10% and 30% of patients do not have stage data, but have g_i. In this case we denote $I_0 = 1$ and $I_s = 0$. The log likelihood in this notation is:

$$\log L = \sum_i \left\{ \log\left[\sum_s I_s p_s f_{is} \exp\left(-\mu_{0s}\tau_i^\gamma (1+\beta f_{is}^2 (g_m - g_0)^2)\right) + I_0 f_N(g_i; m, \sigma^2) \sum_s p_s \exp\left(-\mu_{0s}\tau_i^\gamma (1+\beta(g_i-g_0)^2)\right)\right] + \right.$$

$$\left. d_i \log\left[\sum_s I_s \mu_{0s}\gamma\tau_i^{\gamma-1}\left(1+\beta\sigma^2 f_{is}^2 + \beta(g_m-g_0)^2 f_{is}^4\right) + I_0 \frac{\sum_s p_s \gamma\mu_{0s}\tau_i^{\gamma-1}\left(1+\beta(g_i-p_0)^2 \exp\left(-\mu_{0s}\tau_i^\gamma(1+\beta(g_i-g_0)^2)\right)\right)}{\sum_s p_s \exp\left(-\mu_{0s}\tau_i^\gamma(1+\beta(g_i-g_0)^2)\right)} \right] \right.$$

where $f_{is} = (1+2b\mu_{0s}\sigma^2\tau_i^\gamma)^{-1/2}$, $p_1 = 1 - p_2 - p_3$, and $f_N(g_i; m, \sigma^2)$ is the density of the normal distribution with mean m and variance σ^2 taken for g_i.

We also consider estimate the models based on only HRS or SEER-Medicare data. In the case SEER-Medicare data respective likelihood L_{SEER} is the product of stage-specific likelihoods. In the model for one stage only four parameters can be identified: μ_{0s}, γ, $\beta\sigma^2$, and $\beta(g_m - g_0)^2$. There is no any reduction in the number of model parameters in the model of HRS data.

Table 2 represents the results of simulation studies. We notice that the combined model reproduces the initial parameter values more faithfully than either the stage-only model or the genetic-only model. When we use only genetic data, the optimization routine is unable to settle upon consistent boundaries for the stages. Thus we see that parameter estimates for each stage vary wildly from simulation to simulation. On the other hand, when we use only the data that contains information on the stage of BC but not genetic information, the model cannot determine the effect of the genetic data on time to survival. The term $\beta(m-g_0)^2$ becomes inflated, and as a result the estimates for the stage parameters are depressed. Only when using enough data from both sources are the parameters estimated accurately.

Analysis of real data.
This task is designed to estimate polygenic influence on BC survival using SEER-Medicare and HRS data. Four genetic risk scores were constructed and used in the analysis based on subsets of SNPs in candidate

genes: i) whose association with breast cancer survival was shown in earlier GWAS (Specific SNPs will be extracted from NHGRI-EBI Catalog of published GWAS, https://www.ebi.ac.uk/gwas), ii) found in our preliminary GWAS of all stages BC survival, (top genes that have been associated with breast cancer survival in our preliminary GWAS belong to pathways regulating TP53 and TP63 mediated apoptosis, RB1-related senescence and ERBB2 dependent metastasis); iii) from key pathways involved in cancer invasion and metastasis (such as epithelial to mesenchymal transition, adherence junctions and matrix metalloproteinases pathways[8-11], and iv) involved in response to cancer treatment (e.g., cytochromes, ERBB and AKT pathway genes[12-14]. The scores are evaluated as linear predictors of related SNPs on BC survival in the Cox model. Then we will estimate the joint model to evaluate the effect of the genetic risk score on stage specific survival. The results of application of the method to analysis of real data will be shown at PAA 2019.

Discussion

The approach for analyzing disparate and non-mergeable data sets and the substantive results on breast cancer survival disparities improve knowledge on the breast cancer disparities and will provide other researchers new opportunities for joint analyses of data drawn from different non-linkable sources.

The methodology developed in this paper enable pooled analysis of disparate (in terms of design and content) but complementary (measuring factors relevant to the same process) datasets. Previously, analysis of information collected in different datasets (and often with different designs) was not possible. Using the tools generated by this study, researchers will be able to improve the body of knowledge on effects of risk factors and/or chosen treatment strategies given the presence of multiple inter-related confounders using available observational data drawn from different non-linkable sources. In addition, to the mathematical aspects, a practical application of the new methodologies to the case of BC will be provided to document sources of disparities in BC survival due to genetic and social factors. Estimates of the effects on BC outcomes (in the proposed study BC survival) and their racial disparities are new and were not possible beyond the methodology to be developed in our study.

References

1. D'Orazio M, Di Zio M, Scanu M. *Statistical matching: Theory and practice.* John Wiley & Sons; 2006.
2. Rässler S. *Statistical matching: A frequentist theory, practical applications, and alternative Bayesian approaches.* Vol 168: Springer Science & Business Media; 2012.
3. Warren JL, Klabunde CN, Schrag D, Bach PB, Riley GF. Overview of the SEER-Medicare data: content, research applications, and generalizability to the United States elderly population. *Medical Care.* 2002;40(8):IV-3-IV-18.
4. Nardi A, Schemper M. Comparing Cox and parametric models in clinical studies. *Statistics in medicine.* 2003;22(23):3597-3610.
5. Baghestani A, Moghaddam S, Majd H, Akbari M, Nafissi N, Gohari K. Survival analysis of patients with breast cancer using weibull parametric model. *Asian Pac J Cancer Prev.* 2015;16(18):8567-8571.
6. Carroll KJ. On the use and utility of the Weibull model in the analysis of survival data. *Controlled clinical trials.* 2003;24(6):682-701.
7. Zhu HP, Xia X, Chuan HY, Adnan A, Liu SF, Du YK. Application of Weibull model for survival of patients with gastric cancer. *BMC gastroenterology.* 2011;11(1):1.
8. Krøigård AB, Larsen MJ, Lænkholm A-V, et al. Identification of metastasis driver genes by massive parallel sequencing of successive steps of breast cancer progression. *PloS one.* 2018;13(1):e0189887.
9. Škalamera D, Dahmer-Heath M, Stevenson AJ, et al. Genome-wide gain-of-function screen for genes that induce epithelial-to-mesenchymal transition in breast cancer. *Oncotarget.* 2016;7(38):61000.
10. Moirangthem A, Bondhopadhyay B, Mukherjee M, et al. Simultaneous knockdown of uPA and MMP9 can reduce breast cancer progression by increasing cell-cell adhesion and modulating EMT genes. *Scientific reports.* 2016;6:21903.
11. Slattery ML, John E, Torres-Mejia G, et al. Matrix metalloproteinase genes are associated with breast cancer risk and survival: the Breast Cancer Health Disparities Study. *PloS one.* 2013;8(5):e63165.
12. Jernström S, Hongisto V, Leivonen S-K, et al. Drug-screening and genomic analyses of HER2-positive breast cancer cell lines reveal predictors for treatment response. *Breast Cancer: Targets and Therapy.* 2017;9:185.
13. Toomey S, Madden SF, Furney SJ, et al. The impact of ERBB-family germline single nucleotide polymorphisms on survival response to adjuvant trastuzumab treatment in HER2-positive breast cancer. *Oncotarget.* 2016;7(46):75518.

14. Simonsson M, Veerla S, Markkula A, Rose C, Ingvar C, Jernström H. CYP1A2–a novel genetic marker for early aromatase inhibitor response in the treatment of breast cancer patients. *BMC cancer.* 2016;16(1):256.
15. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *Journal of chronic diseases.* 1987;40(5):373-383.
16. Elixhauser A, Steiner C, Harris DR, Coffey RM. Comorbidity measures for use with administrative data. *Medical care.* 1998;36(1):8-27.
17. Akushevich I, Kravchenko J, Arbeev KG, Ukraintseva SV, Land KC, Yashin AI. Health Effects and Medicare Trajectories: Population-Based Analysis of Morbidity and Mortality Patterns. *Biodemography of Aging*: Springer; 2016:47-93.
18. Akushevich I, Arbeev K, Kravchenko J, Berry M. Causal Effects of Time-Dependent Treatments in Older Patients with Non-Small Cell Lung Cancer. *PloS one.* 2015;10(4):e0121406.

Table 1. Results of simulation studies. The "model" =0-4 corresponds to the cases when only genetic (0), stage-specific (1-3), all (4) data are available for model estimation

N_g	model		μ_{10}	μ_{20}	μ_{30}	β	g_0	γ	p_2	p_3	g_m	σ^2	$\tilde{\beta}$ *	$\beta\sigma^2$
%		Units	$[t]^{-\gamma}$	$[t]^{-\gamma}$	$[t]^{-\gamma}$	$[g]^{-2}$	$[g]$	1	1	1	$[g]$	$[g]$	1	1
		True	0.10	0.12	0.14	0.10	3.00	1.10	0.35	0.10	3.00	0.09	0	0.009
1	0	Mean	0.065	0.107	0.386	0.167	3.164	1.196	0.457	0.094	3.001	0.091		
1	0	StdErr	0.003	0.005	0.018	0.014	0.127	0.009	0.022	0.006	0.001	0.000		
1	0	Ratio	12.544	2.823	-13.940	-4.857	-1.295	-10.704	-4.773	1.125	-0.737	-2.568		
10	0	Mean	0.075	0.114	0.311	0.112	3.040	1.127	0.562	0.078	3.000	0.090		
10	0	StdErr	0.003	0.003	0.017	0.008	0.040	0.003	0.027	0.006	0.000	0.000		
10	0	Ratio	8.074	2.282	-10.117	-1.655	-1.018	-9.627	-7.741	3.375	-0.465	-0.665		
30	0	Mean	0.074	0.110	0.265	0.100	2.982	1.114	0.582	0.095	3.000	0.090		
30	0	StdErr	0.003	0.001	0.016	0.005	0.019	0.002	0.028	0.007	0.000	0.000		
30	0	Ratio	7.861	7.889	-7.678	0.020	0.990	-7.906	-8.412	0.643	-0.545	0.022		
1	1	Mean	0.066					1.103					0.595	0.007
1	1	StdErr	0.001					0.001					0.044	0.001
1	1	Ratio	24.702					-6.271					-13.506	2.271
10	1	Mean	0.077					1.103					0.806	0.013
10	1	StdErr	0.003					0.001					0.129	0.001
10	1	Ratio	7.433					-5.429					-6.256	-2.675
30	1	Mean	0.077					1.103					0.832	0.012
30	1	StdErr	0.003					0.001					0.130	0.001
30	1	Ratio	7.368					-4.700					-6.390	-2.526
1	2	Mean		0.087				1.102					0.424	0.007
1	2	StdErr		0.002				0.001					0.025	0.001
1	2	Ratio		21.163				-3.428					-16.820	2.146
10	2	Mean		0.103				1.101					0.492	0.012
10	2	StdErr		0.003				0.001					0.108	0.001
10	2	Ratio		5.162				-2.158					-4.561	-1.941
30	2	Mean		0.105				1.102					0.399	0.012
30	2	StdErr		0.003				0.001					0.094	0.001
30	2	Ratio		5.227				-2.858					-4.222	-2.215
1	3	Mean			0.100			1.107					0.561	0.011
1	3	StdErr			0.003			0.001					0.060	0.001
1	3	Ratio			14.029			-6.224					-9.287	-1.503
10	3	Mean			0.106			1.107					0.792	0.014
10	3	StdErr			0.004			0.001					0.120	0.001
10	3	Ratio			7.812			-6.157					-6.612	-3.712
30	3	Mean			0.104			1.106					0.873	0.014
30	3	StdErr			0.005			0.001					0.126	0.001
30	3	Ratio			7.952			-5.635					-6.908	-3.676
1	4	Mean	0.095	0.114	0.134	0.159	3.147	1.101	0.350	0.100	3.001	0.091		
1	4	StdErr	0.001	0.001	0.001	0.013	0.114	0.000	0.000	0.000	0.001	0.000		
1	4	Ratio	5.494	5.569	5.429	-4.622	-1.294	-2.778	-0.523	1.257	-0.754	-2.588		
10	4	Mean	0.099	0.119	0.139	0.110	3.039	1.100	0.350	0.100	3.000	0.090		
10	4	StdErr	0.000	0.000	0.000	0.007	0.038	0.000	0.000	0.000	0.000	0.000		
10	4	Ratio	4.300	5.092	3.205	-1.332	-1.022	-0.159	-0.562	1.230	-0.487	-0.656		
30	4	Mean	0.100	0.120	0.140	0.099	2.982	1.100	0.350	0.100	3.000	0.090		
30	4	StdErr	0.000	0.000	0.000	0.005	0.019	0.000	0.000	0.000	0.000	0.000		
30	4	Ratio	1.011	1.830	0.475	0.256	0.969	0.135	0.202	1.101	-0.548	0.032		

* $\tilde{\beta} = \beta(g_m - g_0)^2$