Bayesian variable selection and modelling for metastatic breast cancer data

Sarini Sarini[1] James McGree[2] Kerrie Mengersen[3]

(Received 18 December 2013; revised 15 May 2014)

Abstract

A Bayesian model selection procedure is applied to data on 90 women with metastatic breast cancer. Protein covariates are measured on nucleus, cytoplasm, membrane, and stroma of primary breast carcinoma and lymph node metastasis tissue. Multiple imputation is performed to deal with missing data. Zellner's g-prior is used in the Bayesian variable selection procedure. The model space is reduced using posterior variable inclusion probabilities, and then posterior model probabilities are used to derive a candidate set of models. Bayesian model averaging is employed to robustly estimate survival time, and the goodness of fit of the derived model assessed by the correlation between estimated and observed survival times. The results show evidence of proteins having different rules in different parts of the tissue cell with respect to patient

survival. Therefore, a recommendation is given on which part of the cell to observe certain proteins for prognosis. The models obtained are robust toward censoring and showed correlations between the observed and the predicted data between 0.7 and 0.84.

Contents

1 Introduction

Breast cancer is the most prevalent type of cancer in women worldwide [2] and is the second most common cause of death by cancer [5]. Identification of important factors affecting the patient's condition might help in determining treatments that reduce mortality. Clinical evidence suggests that a patient's disease outcome could be inferred from the molecular phenotype of the disease [16, 17], raising the possibility of clinically useful biomarker tests. Moreover, as indicated by Adams et al. [1], age is a potential influential factor in breast cancer, so age at the time of tumour retrieval is also considered.

This study aims to find a model that captures and explains the relationships between observed characteristics. However, poor methods have been applied in modelling the prognostic factors of cancer [13]. In regression modelling, with time to event as the corresponding response, Cox models are favourable as they are flexible in the distributional assumption regarding the survival experience [10]. However, an underlying proportional hazard assumption is often ignored, which can lead to incorrect conclusions.

A parametric multivariate regression based on the Weibull distribution is proposed to identify the prognostic factors. This identification of influential prognostic factors is an important step in deriving a model that provides accurate predictions of patient survival. Bayesian variable selection procedures are used [11, 12, 15, 20].

It is possible to have several reasonable models for one data set. Robust predictions are expected to be obtained by weighting the predictions from each reasonable model by the relative evidence for each model. This approach is termed Bayesian model averaging [3, 6, 7, 8, 9, 18] and is employed in this article as a prognostic tool.

This study is undertaken in order to identify the significant protein biomarkers related to the survival of metastatic breast cancer patients. Moreover, it also examines whether protein biomarkers measured in two tissue types (that is, metastasis and primary) have similar patterns in predicting the survival of the patient. More specifically, since identical measures were undertaken in four cell locations (namely nucleus, cytoplasm, membrane and stroma) in both tissue types, it is interesting to evaluate whether identical measures are needed in all four cell locations. To our knowledge, this is the first study that conducts thorough and detailed analysis on protein biomarkers measured in different cell locations in two tissue types.

2 Methods

2.1 Subjects

Secondary data from McCosker [14], containing measurements on 90 patients with metastasis breast cancer are used in the analysis. At the metastasis stage, the cancer has already spread to other body parts distinct from the original site of cancer formation. In this study, two tissues are examined, primary breast carcinoma (termed primary, which is in the breast where the cancer cells began) and lymph node metastasis (termed LNMet, which is in parts of the body other than the breast where the cancer cells have spread). Tissue contained in tissue microarray cores were collected from the formalin-fixed paraffin-embedded archival carcinoma specimens. In each tissue, measurements of protein covariates were conducted in four cell locations (termed local): nucleus, cytoplasm, membrane, and stroma. The event of interest was time until patient's death, measured from the time of tumour retrieval.

2.2 Pre-processing procedure

Among 19 protein covariates presented by McCosker [14], five are missing more than 60 percent of data, and thus were excluded from the analysis. The exclusion avoids biased imputed values [4]. The protein covariates included are: β_1 integrin, claudin-1, oestrogen, fibronectin, human epidermal growth factor receptor 2 (HER2), insulin-like growth factor type II receptor (IGF-IIR), mitogen activated protein kinase (MAPK), phosphorylated AKT (p-AKT), phosphorylated mitogen-activated protein kinase (pMAPK), progesterone, enhancer-of-split and hairy-related protein 2 (SHARP-2), stratifin, total-AKT/ protein kinase B 1 (Total-AKT1) and vitronectin. In addition to the protein covariates, as indicated by Adams et al. [1], age is a potential influence in breast cancer, so age at the time of tumour retrieval is included in the study.

In the included protein covariates, there is still some missing data, and multiple imputation is applied to cope with this issue. Of the 90 patients measured, 73 have measurements in LNMet tissue, and 84 have measurements in primary tissue, and in each tissue measurements were taken in all four cell locations. Thus, the data are segmented into eight data subsets based on the tissue types and cell locations.

Prior to model building, the Weibull distributional assumption for the survival time was examined. This was done by plotting $\log[-\log S(t)]$ versus $\log t$, where $S(t)$ is the survival probability obtained using the Kaplan–Meier estimation, and t is the survival time. A straight line plot suggested that the survival time fulfilled the Weibull distribution, and thus the Weibull model is suitable.

Covariates were further selected based on the posterior probability inclusion in the model. A reduced model space was constructed based on the selected covariates and then the resulting models were selected for prediction based on the posterior model probability. The goodness of fit was measured by the correlation between the estimated and observed survival time.

In the modelling process, the first step is to perform Bayesian variable selection, in order to identify significant covariates regarding patient survival time. Model averaging is then used to form robust prediction, where each reasonable model was weighted by its posterior model probability. Finally, this derived model is compared to a 'fixed pooled' model consisting of all covariates that have inclusion posterior probability $\geqslant 0.2$.

3 An overview of statistical method

The set of models considered is denoted by M. Each model is represented by $\gamma \in \{0,1\}^p$, where p is the total number of covariates and $\gamma = (\gamma_1, \ldots, \gamma_p)$ is the indicator of which covariates are included in the model for $j = 1, 2, \ldots, p$.

Using this notation, a linear predictor is $\eta = \sum_j \gamma_j X_j \beta_j$, where X_j is the vector or matrix corresponding to the regression parameter β_j of the jth covariate.

Considering a particular model $m \in M$, $\boldsymbol{\beta}_m = (\beta_{m_1}, \beta_{m_2}, \ldots, \beta_{m_p})$ denotes the vector of regression parameters for the model. According to Bayes theorem, the conjugate distribution given some evidence x is $f(\boldsymbol{\beta}_m \mid x) \propto f(x \mid \boldsymbol{\beta}_m) f(\boldsymbol{\beta}_m)$, and the choice of prior distribution $f(\boldsymbol{\beta}_m)$ is important. We consider Zellner's g-prior [19], assuming a multivariate normal distribution for $\boldsymbol{\beta}_m$. It is defined by specifying $V_m = \left(X_m^T X_m\right)^{-1}$ in the conjugate prior distribution,

$$f\left(\boldsymbol{\beta}_m \mid \sigma^2, m\right) \sim N\left(\mu_{\boldsymbol{\beta}_m}, c^2 V_m \sigma^2\right), \tag{1}$$

with mean $\mu_{\boldsymbol{\beta}_m}$ and residual variance σ^2. In the case of unavailability of prior information, $\boldsymbol{\beta}_m$ is assumed centred around 0, and setting $c^2 = n$ [19] where n is the sample size gives the prior

$$f(\boldsymbol{\beta}_m \mid m) \sim N\left(0, n V_m \sigma^2\right). \tag{2}$$

As for the residual variance, the prior is $f(\sigma^2) \sim \text{IG}(a, b)$, where IG is the inverse Gamma distribution with shape and rate parameters a and b, respectively.

For every indicator γ, the model indicator is assigned to

$$m(\gamma) = \sum_{j=k}^{p} \gamma_j 2^{j-k},$$

where $k = 1$ if a constant term is included in all models under consideration and $k = 0$ otherwise. Given data y, the posterior probability of inclusion for X_j is estimated via

$$f\left(\gamma_j = 1 \mid y\right) = \sum_{\gamma_{\setminus j} \in \{0,1\}^{p-1}} f\left(\gamma_j = 1, \gamma_{\setminus j} \mid y\right), \tag{3}$$

and the posterior model probability is estimated via

$$\hat{f}(m \mid y) = \frac{1}{T - B} \sum_{t=B+1}^{T} I\left(m^{(t)} = m\right), \tag{4}$$

where T and B are the total and burn-in iterations, respectively, and I is the indicator whether the model at iteration t, $m^{(t)}$, equals model m.

Upon obtaining several reasonable models, the predicted values are calculated for each model, and then averaged using

$$\hat{y} = \sum_{l=1}^{L} \frac{\hat{f}_l(m \mid y)}{\sum_{l=1}^{L} \hat{f}_l(m \mid y)}, \tag{5}$$

where L is the number of models considered, $\hat{f}_l$ and $\hat{y}_l$ are the posterior probability and the predicted value based on model l, respectively.

Data processing was undertaken using WinBUGS14.[1] For the normal distribution, the prior parameter mean is set to zero and the variance is set to $\sigma^2 = 100$. As for the inverse gamma distribution, the shape and rate parameters are $a, b = 0.01$. In each process, the total number of iterations was $T = 110,000$ including $B = 10,000$ iterations as burn-in.

4 Results

In the primary tissue, a high posterior probability inclusion for age is only obtained in the nucleus, with the probability estimated to be 1.0. This indicates that in the nucleus, a relationship between the age of a patient at the time of tumour retrieval and protein covariates is detected. In contrast, oestrogen's posterior probability inclusion was estimated to be 1.0 in all primary tissue, except the nucleus. This implies that in the presence of other protein covariates, oestrogen's role is significant in explaining patient's survival, but in the nucleus, this role is diminished. The role of IGF-IIR and p-AKT was also only significant in the nucleus.

Looking at other covariates, HER2 and progesterone play a significant role in all primary tissue locations except in the cytoplasm. However, claudin-1's

[1] http://www2.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml

role is only observable in the cytoplasm. Interestingly, in other locations of the primary tissue many protein covariates have an important role, but in the cytoplasm only oestrogen and β_1 integrin seem to be important covariates. Another protein covariate, stratifin, is important only in the stroma with a posterior probability inclusion of 0.67.

In the metastasis tissue, oestrogen and MAPK show consistent importance in all locations. SHARP2 is also important in all locations except the nucleus, but in the primary tissue it is important in the nucleus and the stroma. HER2 in metastasis was only significant in the stroma. As for the relationship between age and protein covariates, the metastasis tissue is more sensitive as the relationship is detectable in the nucleus and the cytoplasm.

The model posterior probabilities are more concentrated in the primary tissue than in the metastasis tissue, indicated by distinctively higher probabilities in the primary tissue in each cell location. This suggests that in the primary tissue, some sets of covariates are more distinctive than in the metastasis tissue. On the other hand, in the metastasis tissue several reasonable models provide similar values of the posterior probability, suggesting that different sets of covariates provide equivalent information regarding patients' survival. In order to get a more detailed picture of the covariates' role, an example of output with significant β coefficients (at 0.05 significance level) obtained from the first model in each location is given in Table 1.

According to Table 1, although all important protein covariates have a negative effect on the patient's survival, different locations provide different information. In more detail, β_1 integrin, claudin-1, oestrogen and HER2 are important factors in both primary and metastasis tissue. The importance of age in the primary nucleus suggests that the protein covariates are related to age, and this relationship is clearly observed in the nucleus of the primary tissue. On the other hand, a 'unique' feature of LNMet tissue is that the effects of pMAPK, SHARP2, and Total-AKT1 are clearly observed. This implies that, in the metastasis tissue, the number of important covariates is more than that in the primary tissue. However, if age is also considered, then measurements

Table 1: The selected influential variables (estimated coefficients).

Cell location	Significant variables (estimated β)
Primary nucleus	HER2 (−.18), age (−.93)
Primary cytoplasm	claudin-1 (−.55), oestrogen (−.67)
Primary membrane	oestrogen (−.39), HER2 (−.21)
Primary stroma	β_1 integrin (−.48), oestrogen (−.91)
LNMet nucleus	claudin-1 (−.22), oestrogen (−.25), vitronectin (−.22)
LNMet cytoplasm	β_1 integrin (−.06), claudin-1 (−.25)
LNMet membrane	claudin-1 (−.19), oestrogen (−.16),Total-AKT1 (−.15), age(−.05)
LNMet stroma	HER2 (−.60)

Table 2: Correlation coefficients between the observed and the estimated survival times for the selected models (ρ_{1S} and ρ_{2S}) and for the 'fixed pooled' model (ρ_{1F} and ρ_{2F}).

Cell location	ρ_{1S}	ρ_{2S}	ρ_{1F}	ρ_{2F}
Primary nucleus	.70	.63	.71	.65
Primary cytoplasm	.79	.77	.63	.79
Primary membrane	.82	.79	.71	.65
Primary stroma	.79	.77	.69	.62
LNMet nucleus	.82	.81	.79	.71
LNMet cytoplasm	.77	.68	.78	.69
LNMet membrane	.84	.87	.78	.69
LNMet stroma	.82	.82	.78	.69

of the primary tissue nucleus provide a better result.

The goodness of fit for each selected model is quite acceptable. Despite the models consisting of only a small number of selected covariates (as listed in Table 1), the correlations between the observed and predicted survival times are $\rho_{1S} \geqslant 0.70$ when all data is considered. When only non-censored data is considered, the corresponding correlations are $\rho_{2S} > 0.6$.

To assess the performance of the model, a 'fixed pooled' model is fitted to all eight locals. The selected covariates are those with posterior inclusion probability > 0.2, giving eleven protein biomarkers and age in the model (fibronectin, p-AKT, and pMAPK were not selected). The correlations estimated from the fitted 'fixed pooled' model are given in Table 2, where ρ_{1F} is the correlation between the observed and the predicted survival times for all data and ρ_{2F} is the correlation when only non-censored data is considered. Almost all correlations estimated from the 'fixed pooled' model are less than those estimated from the selected models with covariates given in Table 1, implying that adding more covariates does not improve the models' predictability.

5 Discussion

To derive a deep insight into factors contributing to metastasis breast cancer patients' survival times, a multivariate Weibull regression model was proposed. The covariates were specific protein measures required for insulin-like growth factor and extracellular matrix induced signalling events [14] as well as age at tumour retrieval. This is the first study, to our knowledge, that conducts thorough and detailed analysis based on different locations in two tissues, primary and metastasis, and presents the role of protein covariates in various locations.

With regard to the method and reporting of the data analysis as discussed by Mallett et al. [13], we examined the survival time distribution to ensure that the required distributional assumption for modelling was met by the data. We also divided the data into eight data sets based on cell locations and tissue types to allow us to identify the different influences of different covariates in different locations. We employed Zellner's g-prior to take into account potential multicollinearity that might exist among covariates. Since there was no external data and the size of data at hand was too small to split for validation purposes, model validation was done by fitting a 'fixed pooled' model to the same dataset. Information from the few selected covariates were

robust and better in predictability. This was indicated by higher correlations of the observed and the predicted survival times for almost all models across cell locations and tissue types compared to those estimated from the 'fixed pooled' model.

It is possible that interactions exist between the protein covariates, as well as other non-linear contributions to survival time. Further study such as non-linear variable selection and identification of important interactions using a Bayesian approach is worth further consideration.

Findings from this study can help clinicians better design breast cancer examinations. The identification of the influences on specific tissues could improve protein-based treatment for breast cancer patients. This might be examined further regarding factors that should be considered prior to cancer spreading, and possibly to design new treatments.

Acknowledgements The authors thank Susanna Cramb from the Cancer Council Queensland for discussion on breast cancer staging and statistics. The present work also benefited from the input of Nicole White, an associate researcher at QUT, who provided valuable comments.

References

[1] J. Adams, P. J. Carder, S. Downey, M. A. Forbes, K. MacLennan, V. Allgar, S. Kaufman, S. Hallam, R. Bicknell, J. J. Walker, F. Cairnduff, P. J. Selby, T. J. Perren, M. Lansdown, and R. E. Banks. Vascular endothelial growth factor (VEGF) in breast cancer: comparison of plasma, serum, and tissue VEGF and microvessel density and effects of tamoxifen. *Cancer Res.*, 60(11):2898–2905, 2000. http://cancerres.aacrjournals.org/content/60/11/2898. C169, C171

[2] F. Bray, J.-S. Ren, E. Masuyer, and J. Ferlay. Global estimates of cancer prevalence for 27 sites in the adult population in 2008. *Int. J. Cancer*, 132(5):1133–1145, 2013. doi:10.1002/ijc.27711. C169

[3] M. Clyde and E. I. George. Model uncertainty. *Stat. Sci.*, 19(1):81–94, 2004. http://projecteuclid.org/download/pdfview_1/euclid.ss/1089808274. C170

[4] A. R. T. Donders, G. J. M. G. van der Heijden, T. Stijnen, and K. G. M. Moons. Review: a gentle introduction to imputation of missing values. *J. Clin. Epidemiol.*, 59(10):1087–1091, 2006. doi:10.1016/j.jclinepi.2006.01.014. C171

[5] J. Ferlay, I. Soerjomataram, D. F. M. Ervik, F. Bray, R. Dikshit, S. Elser, C. Mathers, M. Rebelo, and D. M. Parkin. GLOBOCAN 2012: Estimated cancer incidence, mortality and prevalence worldwide in 2012. International Agency for Research on Cancer, World Health Organization, 2014. http://globocan.iarc.fr. C169

[6] E. I. George and R. E. McCulloch. Variable selection via Gibbs sampling. *J. Am. Stat. Assoc.*, 88(423):881–889, 1993. doi:10.1080/01621459.1993.10476353. C170

[7] E. I. George and R. E. McCulloch. Approaches for Bayesian variable selection. *Stat. Sinica*, 7(2):339–373, 1997. http://www3.stat.sinica.edu.tw/statistica/j7n2/j7n26/j7n26.htm. C170

[8] J. Geweke. *Variable selection and model comparison in regression*, volume 5 of *Bayesian statistics*, pages 609–620. Oxford University Press, 1996. C170

[9] J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky. Bayesian model averaging: a tutorial. *Stat. Sci.*, 14(4):382–401, 1999. http://www.jstor.org/stable/2676803. C170

[10] D. G. Kleinbaum and M. Klein. *Survival Analysis: A self-learning text*. Statistics for Biology and Health. New York, Springer-Verlag, 2011. doi:10.1007/0-387-29150-4. C170

[11] E. E. Leamer. Regression selection strategies and revealed priors. *J. Am. Stat. Assoc.*, 73(363):580–587, 1978. doi:10.1080/01621459.1978.10480058. C170

[12] E. E. Leamer. *Specification searches: Ad hoc inference with nonexperimental data*. Wiley New York, 1978. C170

[13] S. Mallett, P. Royston, R. Waters, S. Dutton, and D. G. Altman. Reporting performance of prognostic models in cancer: a review. *BMC Med.*, 8(1):21, 2010. doi:10.1186/1741-7015-8-21. C170, C177

[14] H. C. McCosker. *Prognostic significance of IGF and ECM induced signalling proteins in breast cancer patients*. PhD thesis, School of Biomedical Sciences, QUT, 2012. http://eprints.qut.edu.au/53580/. C171, C177

[15] T. J. Mitchell and J. J. Beauchamp. Bayesian variable selection in linear regression. *J. Am. Stat. Assoc.*, 83(404):1023–1032, 1988. doi:10.1080/01621459.1988.10478694. C170

[16] K. Pantel and R. H. Brakenhoff. Dissecting the metastatic cascade. *Nat. Rev. Cancer*, 4(6):448–456, 2004. doi:10.1038/nrc1370. C169

[17] R. Radpour, Z. Barekati, C. Kohler, W. Holzgreve, and X. Y. Zhong. New trends in molecular biomarker discovery for breast cancer. *Genet. Test. Mol. Bioma.*, 13(5):565–571, 2009. doi:10.1089/gtmb.2009.0060. C169

[18] A. E. Raftery, D. Madigan, and J. A. Hoeting. Bayesian model averaging for linear regression models. *J. Am. Stat. Assoc.*, 92(437):179–191, 1997. doi:10.1080/01621459.1997.10473615. C170

[19] A. Zellner. *On assessing prior distributions and Bayesian regression analysis with g-prior distributions*, volume 6 of *Bayesian inference and decision techniques: Essays in Honor of Bruno De Finetti*, pages 233–243. North-Holland, Amsterdam, The Netherlands, 1986. C173

[20] A. Zellner. *An introduction to Bayesian inference in econometrics*. New York: John Wiley & Sons, 1996. C170

Author addresses

1. **Sarini Sarini**, School of Mathematical Sciences, QUT, Brisbane 4000, Australia; and Mathematics Department, University of Indonesia, Indonesia.
   mailto:sarini@student.qut.edu.au
2. **James McGree**, School of Mathematical Sciences, QUT, Brisbane 4000, Australia.
3. **Kerrie Mengersen**, School of Mathematical Sciences, QUT, Brisbane 4000, Australia.