Biomedical Signal Processing and Control 33 (2017) 169–177

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Change-point of multiple biomarkers in women with ovarian cancer

Inés P. Mariño[a,b,*,1], Oleg Blyuss[b,1], Andy Ryan[b], Aleksandra Gentry-Maharaj[b], John F. Timms[b], Anne Dawnay[c], Jatinderpal Kalsi[b], Ian Jacobs[b,d], Usha Menon[b,2], Alexey Zaikin[b,2]

[a] Departamento de Biología y Geología, Física y Química Inorgánica, Universidad Rey Juan Carlos, Móstoles, Madrid 28933, Spain
[b] Department of Women's Cancer, Institute for Women's Health, University College London, London WC1E 6BT, United Kingdom
[c] Clinical Biochemistry, University College London Hospitals, London W1T 4EU, United Kingdom
[d] University of New South Wales, Sydney, NSW 2052, Australia

ARTICLE INFO

Article history:
Received 21 August 2016
Received in revised form 21 October 2016
Accepted 14 November 2016

Keywords:
Ovarian cancer
Gibbs sampling
Change-point detection
Bayesian estimation

ABSTRACT

To date several algorithms for longitudinal analysis of ovarian cancer biomarkers have been proposed in the literature. An issue of specific interest is to determine whether the baseline level of a biomarker changes significantly at some time instant (change-point) prior to the clinical diagnosis of cancer. Such change-points in the serum biomarker Cancer Antigen 125 (CA125) time series data have been used in ovarian cancer screening, resulting in earlier detection with a sensitivity of 85% in the most recent trial, the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS, number ISRCTN22488978; NCT00058032). Here we propose to apply a hierarchical Bayesian change-point model to jointly study the features of time series from multiple biomarkers. For this model we have analytically derived the conditional probability distribution of every unknown parameter, thus enabling the design of efficient Markov Chain Monte Carlo methods for their estimation. We have applied these methods to the estimation of change-points in time series data of multiple biomarkers, including CA125 and others, using data from a nested case-control study of women diagnosed with ovarian cancer in UKCTOCS. In this way we assess whether any of these additional biomarkers can play a role in change-point detection and, therefore, aid in the diagnosis of the disease in patients for whom the CA125 time series does not display a change-point. We have also investigated whether the change-points for different biomarkers occur at similar times for the same patient. The main conclusion of our study is that the combined analysis of a group of specific biomarkers may possibly improve the detection of change-points in time series data (compared to the analysis of CA125 alone) which, in turn, are relevant for the early diagnosis of ovarian cancer.

1. Introduction

Cancer of the ovary is the fifth most common cause of cancer-related deaths among women, with more than 140,000 deaths worldwide each year. Most cases occur in post-menopausal women (75%), with an incidence of 40 per 100,000 per year in women aged over 50. Diagnosis of ovarian cancer at Stage I, when the tumour is confined to the ovary, results in a 5-year survival of 90% [1]. However, the 5-year survival decreases sharply when cancer diagnosis occurs at later stages such as Stage III (20%) and Stage IV (3%). This suggests that the development of new approaches for longitudinal multi-marker analysis that result in earlier detection of ovarian cancer may significantly impact on mortality [2–4].

One of the most successful methods of detection of ovarian cancer in a screening context to date is the statistical inference technique for the longitudinal analysis of ovarian cancer biomarkers developed by Skates et al. [5–7], where the main assumption is the existence of a change-point in the serum CA125 time series as the tumour develops. In particular, the level of CA125 is assumed to remain approximately

* Corresponding author.
E-mail address: ines.perez@urjc.es (I.P. Mariño).
[1] These authors contributed equally to this work.
[2] These authors also contributed equally to this work.

http://dx.doi.org/10.1016/j.bspc.2016.11.015
1746-8094/

constant until, at some time instant, it begins to increase significantly. The latter time point is referred to as a change-point. The algorithm proposed in [5] is based on a hierarchical Bayesian model that includes the change-point as one of the random parameters to be estimated.

In this paper, we jointly analyse time series data from multiple biomarkers to determine whether the level of these markers changes significantly and coherently at specific time instants. The detection of such a change-point may contribute to the earlier diagnosis of the disease. Although the serum CA125 is the most useful biomarker in the screening of ovarian cancer, multiple serum biomarkers have been reported to be associated with the development of ovarian cancer and to possibly improve the performance of CA125 when used in combination [8–17]. The biomarker that has received more attention is the Human Epididymis Protein 4 (HE4), which has been used in the ROMA (Risk of Ovarian Malignancy Algorithm) to discriminate ovarian cancer from benign diseases [8,18] as well as in different panels for the purpose of early detection [9–11]. In a study within the Prostate Lung Colorectal and Ovarian (PLCO) cancer screening trial [19], HE4 was the second best marker after CA125 with a sensitivity of 73% (95% confidence interval 0.60–0.86) compared to 86% (95% confidence interval 0.76–0.97) for CA125 [12,20]. Another serum biomarker glycodelin has also shown promising performance in the detection of ovarian cancer [13,14,21]. Other markers that appear to be promising when used in multi-marker panels include matrix metalloproteinase-7 (MMP7) [13,20,22], cytokeratin 19 fragment (CYFRA 21-1) [15,20] and mesothelin (MSLN) [11,16].

In order to incorporate this information we assume a hierarchical Bayesian change-point model for different biomarkers in addition to serum CA125. Statistical inference in this model can be carried out using Markov Chain Monte Carlo (MCMC) methods [23]. In particular, we have analytically obtained the full conditional probability distributions for all the unknown parameters in the model, thus enabling the design of an efficient Metropolis-within-Gibbs algorithm [24] for their Bayesian estimation. We apply this technique to the estimation of change-points in time series data, including CA125 and the other biomarkers in patients diagnosed with ovarian cancer and in a control group of healthy individuals. We assess whether any of these additional biomarkers can play a role in ovarian cancer diagnosis by either detecting a change-point in any of the available biomarkers earlier than in CA125 or by detecting a change-point in women in whom the CA125 does not display a change-point. We also investigate whether the change-points for different biomarkers occur at similar time points.

The Bayesian estimation approach advocated in this paper aims at producing a full statistical characterisation of the unknown model parameters, given in the form of their posterior probability distribution conditional on the available data. The proposed Metropolis-within-Gibbs algorithm yields a Monte Carlo approximation of this posterior distribution, which enables the implementation of a variety of estimators for the parameters and provides the means to evaluate their accuracy and reliability as well. While in this paper we keep the change-point estimation process relatively simple, the framework and algorithms described in Sections 2 and 3 lends itself to potentially advantageous extensions.

The rest of this paper is organised as follows. Section 2 is devoted to the description of the dataset and the hierarchical Bayesian model used to represent it. The inference algorithm for the detection and estimation of change-points in biomarker time series is introduced in Section 3. The results obtained for the available dataset are shown and discussed in Section 4. Section 5 is devoted to the conclusions.

2. Data model

2.1. Dataset

In this study we have used a dataset from the multimodal arm of the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS) [25], where women underwent annual screening using the blood tumour marker CA125. HE4, MMP7, CYFRA_21-1, glycodelin and MSLN assays were performed on stored serial samples from a subset of women in the multimodal arm diagnosed with ovarian cancer. The dataset included 179 controls (healthy subjects) and 44 cases (diagnosed patients): 35 invasive epithelial ovarian cancer (iEOC), 3 fallopian tube cancer and 6 peritoneal cancer. Out of these 44 cases, 16 are early stage (FIGO [26] Stages I and II) and 28 are late stage (FIGO Stages III and IV). In terms of histology, there are 27 serous cancers, 2 papillary, 3 endometrioid, 2 clear cell, 3 carcinosarcoma, and 7 not specified cancers. Each control has 4–5 serial samples available (177 controls with 5 samples and 2 controls with 4 samples) and each case has 2–5 serial samples available (24 cases with 5 samples, 10 cases with 3 samples and 10 cases with 2 samples). The range of ages for the healthy subjects (controls) is 50.3–78.8 years and the average age over all these subjects and samples is 63.6 years. The range of ages for the diagnosed patients (cases) is 52.0–77.4 years and the average age over all these patients and samples is 65.5 years.

It should be noted here that all the biomarker measurements have been modified via a logarithmic transformation, as detailed in [5,14], in the form of $Y = \log(Z + 4)$, where Z is the value of a particular marker. For most patients with ovarian cancer prior to disease diagnosis, serum CA125 rises exponentially. This transformation allows us to observe a linear change in time.

2.2. Model

Fig. 1 shows the scheme of the hierarchical Bayesian model for patients diagnosed with ovarian cancer. Let Y_{ij} denote the log-transformed measurement of the biomarker Z (where Z can be any of CA125, HE4, glycodelin, MSLN, MMP7 or CYFRA_21-1) for the ith patient in the study at age t_{ij}, where $j = 1, \ldots, k_i$ represents the ordinal of the observation for patient i (i.e., the first observation, the second observation, and so on), being k_i the total number of measurements for patient i. The values of the biomarkers are collected at time points t_{ij}, which can depend on previous values $Y_{ij'}$, $j' < j$. For this model, an unobserved binary indicator I_i is included to distinguish subjects whose ovarian cancer does ($I_i = 1$), or does not ($I_i = 0$), produce an increased biomarker level. Notice that for patients with $I_i = 0$ there is no way to tell them from healthy individuals by looking at that biomarker alone. Within the proposed hierarchical model, separation of cases that do not produce a change-point from those that do produce it enables a more precise estimation of individual change-points and rates of change. The indicator I_i for each case is assumed to follow, a priori, a Bernoulli distribution with success probability π, where π represents the proportion of cases that produce an increased biomarker level. As in [5], considering the observation of Kabawat et al. [27] that approximately 15% of ovarian cancer cases do not produce an excess of biomarker CA125, a Beta(42.5,7.5) distribution with mean 0.85 and standard deviation 0.05 is adopted as the prior of π. We assume the same prior distribution of π for all the biomarkers studied in this paper.