Computational Intelligence in Clinical Oncology: Lessons Learned from an Analysis of a Clinical Study

B. Haibe-Kains¹,², C. Desmedt², S. Loi³, M. Delorenzi⁴, C. Sotiriou², and G. Bontempi¹

¹ Machine Learning Group, Université Libre de Bruxelles, Brussels, Belgium
² Functional Genomics Unit, Institut Jules Bordet, Brussels, Belgium
³ Peter MacCallum Cancer Center, East Melbourne, Victoria, Australia
⁴ Bioinformatics Core Facility, Institut Suisse de Recherche Expérimentale sur le Cancer, Lausanne, Switzerland

Summary. In this chapter, we present a retrospective clinical study where the adoption of computational intelligence approaches for performing knowledge extraction from gene expression data enabled an improved oncological clinical analysis. This study focuses on a survival analysis of estrogen receptor (ER) positive breast cancer patients treated with tamoxifen. The chapter describes each step of the gene expression data analysis procedure, from the quality control of data to the final validation going through normalization, feature transformation, feature selection, and model building. Each section proposes a set of guidelines and motivates the specific choice made for this particular study. Finally, the main guidelines that emerged from this study are the use of simple and effective techniques rather than complex non-linear models, the use of interpretable methods and the use of scalable computational solutions able to deal with multiplatform and multisource data.

10.1 Introduction

Recent advances in biomedical measurement technologies, such as gene expression profiling, expose clinicians to an exponential increase of complex data. Major challenges on the computational side arise from the huge dimensionality of the data, the relatively low number of samples, the high redundancy of input variables, the heterogeneity of the data sources and the high level of noise. This is why traditional clinical analysis needs the help of computational intelligence approaches to manage the complexity of the analysis task, without being overwhelmed by the massive amount of data or mislead by spurious patterns [1].

In this chapter, we present a retrospective clinical study in which the adoption of computational intelligence approach for performing knowledge extraction from gene expression data enabled an improved oncological clinical analysis. We focus on a survival analysis of estrogen receptor (ER) positive breast cancer (BC) patients treated with tamoxifen, a well-known treatment in BC therapy. The patients included in the study are heterogeneous with respect to their clinical
behavior and response to tamoxifen therapy. It is known that current biomarkers (e.g. expression levels of ER) give little insight into tumor biology and potential response to treatment. Indeed 30-40% of women with ER-positive disease develop distant metastases and die despite tamoxifen treatment, illustrating the urgent clinical need for new biomarkers that can predict which women with ER-positive BC are at high risk of relapse despite the use of tamoxifen. These patients would benefit from new therapies such as the aromatase inhibitors [2, 3].

Gene expression profiling of tumors appears to be a promising new strategy for predicting clinical outcome in BC patients. According to recent studies [4, 5, 6, 7] the heterogeneity of clinical response can be correlated with different molecular "portraits". Additionally, gene classifiers have been developed that can distinguish subgroups of patients with different prognoses or responses to therapies [8]. Due to the pressing clinical need, several other investigators have also developed gene classifiers for ER-positive BC patients treated with tamoxifen [9, 10, 11, 12]. After an initial period of enthusiasm about the potentiality of computational techniques, problematic issues about the design of the analysis and the performance assessment of gene classifiers were raised [13, 14]. Moreover, skepticism about the robustness of the approach arose in the medical community because of the small overlap of genes in signatures (i.e. set of highly discriminating genes) derived from different clinical studies on similar cohorts of BC patients [15]. In this chapter, we address these open questions and describe the entire computational procedure leading to the development of a model which predicts the risk of recurrence in tamoxifen treated patients. Thereby we also investigate the potential of this model to reveal new biological processes associated with the clinical outcome of these patients.

The chapter is organized as a set of sections, each serving as illustration of a single step of the clinical study (entirely represented in Figure 10.1). Each section introduces the issue, refers to existing reviews on the topic and discusses the most relevant methodological aspects. A subsection called Tamoxifen Study details the methodological choices made during the study, the motivations behind them and the related results. The section terminates with a subsection called Lessons Learned where we summarize the main lessons to be retained by the reader.

In particular, Section 10.1 introduces the clinical problem and emphasizes the need of computational intelligence techniques. Section 10.2 illustrates the properties of gene expression data collected in large clinical studies. Section 10.3 serves as an introduction to survival data and the related statistical methods. Sections 10.4 and 10.5 discuss the preprocessing of gene expression data, in particular quality controls and normalization. Sections 10.6 and 10.7 emphasize the importance of dimension reduction and the issue of stability in feature selection. Section 10.8 details the model building procedure. Sections 10.9 and 10.10 summarizes the accuracy assessment of the model in a cross-validation framework and an independent validation set respectively. Section 10.11 gives biological interpretation of the model. Conclusions are drawn in Section 10.12.