Analysis of sarcoma and non-sarcoma clinical data with statistical methods and machine learning techniques

Kantidakis, G.

Summary

This thesis sprang from an interdisciplinary collaboration between the European Organisation for Research and Treatment of Cancer (EORTC), the Mathematical Institute of Leiden University, and the Leiden University Medical Center (LUMC) Department of Medical Oncology. Research was split into two parts. In **Part** I (**Chapters** 2, 3, 4), statistical analyses were performed for the EORTC - Soft Tissue and Bone Sarcoma Group (STBSG), whereas in **Part** II (**Chapters** 5, 6, 7, 8) the potential of survival prediction models with machine learning (ML) techniques was compared with traditional statistical models (SM) for sarcoma and non-sarcoma clinical data.

Part I: Clinical trials in soft-tissue sarcomas

This part provided modern efficacy thresholds to design new phase II clinical trials for common histotypes of locally advanced or metastatic soft-tissue sarcoma (STS) patients. The prognostic significance of bone metastasis in STS was investigated to identify high-risk patient populations.

In 2002, Van Glabbeke *et al.* published on behalf of the EORTC - STBSG a pooled analysis to estimate progression-free rates at 3 and 6 months for various groups of STS patients who participated in phase II trials of the EORTC. These historical values have been widely used (> 420 citations) to design new studies for all STS or for specific histology subgroups (in first-line treatment). We performed an extensive in-house literature search to identify all phase II or subsequent clinical trials of advanced or metastatic STS (2003 to 2018), thus documenting the current landscape. Because of the substantial heterogeneity among clinical trials, it was decided to focus first on leiomyosarcoma (LMS) - the most commonly occurring STS subtype in the papers of our literature review. In **Chapter** 2, a random-effects meta-analysis was performed to provide new benchmarks for designing phase II studies of advanced or metastatic LMS patients separately for first-line or pre-treated population. The primary endpoints of interest were progression-free survival rates (PFSRs) at 3 and 6 months, which are nowadays preferred and more frequently reported than progression-free rates (censoring non-disease-related death). When estimates could not be derived from publications, first authors and/or sponsors were contacted. The ESMO Magnitude of Clinical Benefit Scale (MCBS) was used to guide the treatment effect to target in future trials. Information was obtained on 7 first-line and 16 pre-treated trials for 1500 LMS patients. Under the alternative that the true benefit amounts to a hazard ratio of 0.65, a 6-month PFSR $\geq$ 70% can be considered to suggest drug activity in first-line. For pre-treated population, a 3-month PFSR $\geq$ 62% or 6-month PFSR $\geq$ 44% would suggest drug activity. Specific results were also provided for uterine LMS.

In **Chapter** 3, a second meta-analysis was performed for advanced or metastatic liposarcoma (LPS) or synovial sarcoma (SS) - the second and third most common histotypes in our literature review. Study endpoints were PFSRs at 3 and 6 months. The choice of the therapeutic benefit to target in future trials was guided again by the ESMO MCBS. Information was acquired for 1030 LPS patients (25 trials; 7 first-line, 17 pre-treated, 1 both) and 348 SS patients (13 trials; 3 first-line, 10 pre-treated). New benchmarks were proposed to design future histology-specific phase II trials. Minimum values to target in first-line at 3 and 6 months were 79% and 69% for LPS, 82% and 69% for SS. For pre-treated patients, recommended PFSRs at 3 and 6 months suggesting drug activity were 63%

and 44% for LPS, 60% and 41% for SS. Our findings here and in the previous chapter indicate that there is a need to raise the bar of thresholds for the commonest STS types in future histology-tailored phase II trials in order to achieve higher success rates in new prospective confirmatory phase III trials.

In **Chapter** 4, we investigated whether, and if so, to what extent, skeletal metastases at presentation affect the outcome of patients with advanced or metastatic STS. Selected patients participated in five clinical trials of EORTC - STBSG. Individuals were included if they started treatment with an active drug and had advanced/metastatic STS. The endpoints of interest were overall survival (OS) and progression-free survival (PFS). Univariate and multivariate pooled analyses (after correcting for 12 covariates) were employed with Kaplan–Meier and Cox regression to model the impact of bone metastasis at presentation per treatment line stratified by study. For the subset of patients with bone metastasis, the impact of another metastatic organ site (among liver, lymph node, lung, soft-tissue, or other) at diagnosis was explored with multivariate Cox regression models. 565 out of 1034 (54.6%) patients received first-line systemic treatment for locally advanced or metastatic disease. Bone metastases were present in 140 patients (77 first-line, 63 second-line or higher). The unadjusted difference in OS/PFS with or without bone metastasis was statistically significant only for first-line patients. For OS, the adjusted hazard ratios for bone metastasis presence were 1.33 (95%-CI: 0.99–1.78) and 1.11 (95%-CI: 0.81–1.52) for first-line/second-line or higher treated patients, respectively. Likewise, the adjusted hazard ratios for PFS were 1.31 (95%-CI: 1.00–1.73) and 1.07 (95%-CI: 0.80–1.43). Hence, the adjusted effects were not statistically significant, despite a trend for first-line patients. Subgroup analyses indicated bone and lymph node metastasis as the most detrimental combination for OS and bone and lung metastasis for PFS. Since skeletal metastases at study entry cannot be ascertained as a significant risk factor (per line of treatment), stratification is not justified in randomised studies with these patients.

Part II: Statistical models versus machine learning to predict survival for sarcoma and non-sarcoma clinical data

In this part of the thesis, the predictive performance of existing and novel ML methods was compared with traditional SM for real-life clinical data (small/medium or large sample sizes, low- or high-dimensional settings) with time-to-event endpoints.

Nowadays, a growing interest can be observed in applying ML for clinical prediction by the medical community. Over the years, several algorithms have been developed and adapted to right censored data. Neural networks have been repeatedly employed to build clinical prediction models in healthcare. Even so, despite their non-negligible use, a comprehensive review on survival neural networks (SNNs) using prognostic factors is missing. In **Chapter** 5, we presented the first ever attempt at a structured overview of SNNs with prognostic factors for clinical prediction. Our aim was to provide a broad understanding of the literature (1st January 1990 - 31st August 2021, global search in PubMed). Relevant manuscripts were classified as methodological/technical (novel methodology or new theoretical model; 13 studies) or applications (11 studies). We discussed how SNNs are employed in the medical field for prediction and detailed how researchers have tried to adapt a classification method to right censored survival data. There are two methodological trends: either time is added as part of the input features and a single output node is specified, or multiple output nodes are defined for each time interval. This work was supplemented with a critical appraisal of model aspects that should be designed and reported more carefully. We identified key characteristics of prediction models (i.e., number of patients/predictors, evaluation measures, calibration), and compared SNNs' predictive performance to the Cox proportional hazards model. The median sample size was 920 patients, and the median number of predictors was 7. Major findings included poor reporting (e.g., regarding missing data, hyperparameters), as well as inaccurate model development/validation. Calibration was neglected in more than half of the studies. Cox models were not developed to their full potential, and claims for the performance of SNNs were exaggerated. Light was shed on the current state of art of SNNs in medicine with prognostic factors. Limitations were discussed, and future directions were proposed for researchers who seek to develop existing methodology.

There is an open discussion about the value of ML versus SM within clinical and healthcare practice. ML techniques might be an attractive choice for modelling complex data (large sample size, high-dimensional setting). In **Chapter** 6, three ML techniques: a) random survival forests (RSF), and b-c) two methodological extensions of the partial logistic artificial neural network (PLANN) with one and two hidden layers were tested to large retrospective data of 62294 patients from the United States provided by the Scientific Registry for Transplant Recipients. A total of 97 predictors were selected, over more than 600, to predict survival since liver transplantation on clinical/statistical grounds. A comparison was performed between these ML techniques and three different Cox models (all variables, backward, LASSO). Emphasis was given on the advantages and pitfalls of each method and on extracting interpretability from the ML methods. Well-established predictive measures were employed from the survival field (C-index, Brier score and Integrated Brier Score) and the strongest prognostic factors were identified for each model. Clinical endpoint was overall graft-survival defined as the time between transplantation and the date of graft-failure or death. The RSF showed slightly better predictive performance than Cox models based on the C-index. Neural networks showed better performance than both Cox models and RSF based on the Integrated Brier Score at 10 years. From the three ML techniques, PLANN extended with one hidden layer predicted survival probabilities the most accurately being as calibrated as the Cox model with all variables. The RSF and the PLANN extended with two hidden layers were less calibrated on test data. Regarding interpretability, the Cox model with all variables and the PLANNs identified *re-transplantation* as the strongest predictor and *donor age*, *diabetes*, and *life support* as relatively strong predictors. According to RSF, the most prognostic variable was *donor age*, followed by *re-transplantation*, *life support* and *serology status of Chronic hepatitis C virus*. All in all, it was shown that ML techniques can be a useful tool for both prediction and interpretation in this survival context.

In the previous study, our group provided new methodological extensions of the PLANN model. PLANN extended was developed and validated for complex liver transplantation data. However, it is not uncommon to have a small number of patients recruited in clinical trials and a limited set of predictive features, for instance in sarcoma trials. Even so, there is an expectation by clinicians that ML models may perform better than SM. Therefore, in **Chapter** 7, the focus was on the comparison between such models for non-complex clinical data (small / medium sample size, low dimensional) with a Monte Carlo simulation study to investigate a different real-life setting. Synthetic data (250 or 1000 patients) were generated that closely resembled five prognostic factors preselected based on a European Osteosarcoma Intergroup study (MRC BO06/EORTC 80931) that investigated the effect of dose-intense chemotherapy in patients with localised extremity osteosarcoma. The predictive performance of PLANN original and PLANN extended (with one hidden layer) was compared with Cox models for 20, 40, 61, and 80% censoring. Survival times were generated from a log-normal distribution. The endpoint of interest was overall survival defined as the time to death from any cause since the date of surgery. Models were evaluated in terms of the C-index, Brier score at 0-5 years, integrated Brier score (IBS) at 5 years, and miscalibration at 2 and 5 years (usually neglected). The ML models were able to reach a similar predictive performance on simulated data for most scenarios with respect to the C-index, Brier score, or IBS. However, the SM were frequently better calibrated. Performance was robust in scenarios where censored patients were removed before the 2^{nd} year or administrative censoring at 5 years was performed (on training data). Researchers should be aware of burdensome aspects of ML techniques such as data preprocessing, tuning of hyperparameters, and computational intensity that render them disadvantageous against conventional regression models in a simple clinical setting.

In health research, several chronic diseases are susceptible to competing risks (CRs). Initially, SM were developed to estimate the cumulative incidence of an event of interest in the presence of CRs. As recently there is a growing interest in applying ML for clinical prediction, these techniques have also been extended to CRs but the literature is limited. In **Chapter** 8, we aimed to develop and validate prognostic clinical prediction models for CRs with SM and ML techniques. Two SM a) cause-specific Cox, b) Fine-Gray model and three ML models i) PLANN original for CRs (PLANNCR original), ii) a methodological extension called PLANNCR extended, and iii) RSF for CRs (RSFCR) were employed. The predictive performance of all methods was assessed in terms of discrimination and calibration in another simple clinical setting (small / medium sample size, small number of predictors). The dataset at hand contained 3826 retrospectively collected patients with extremity STS (eSTS) and nine predictors from the PERsonalised SARcoma Care (PERSARC) Study Group. To the best of our knowledge, this was the first ever study of this kind for eSTS. The clinical endpoint was the time in years between surgery and disease

progression (event of interest) or death (competing event). The Brier score, the area under the curve (AUC) and the model's miscalibration were used to evaluate predictive performance at 2, 5, and 10 years, respectively. Results showed that the ML models are able to reach a comparable performance with the SM based on the Brier score and AUC for disease progression and death (95% confidence intervals at 2, 5, and 10 years overlapped). Nevertheless, the SM were frequently better calibrated. Overall, ML techniques are less practical as they require substantial implementation time (data preprocessing, hyperparameter tuning, computational intensity). As such, for non-complex real life data, these techniques should only be applied complementary to SM as exploratory tools of model's performance. More attention to model calibration is urgently needed.