
Universiteit
Leiden
The Netherlands

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

Kantidakis, G.

Citation
Kantidakis, G. (2022, November 23). *Analysis of sarcoma and non-sarcoma clinical data with statistical methods and machine learning techniques*. Retrieved from https://hdl.handle.net/1887/3486743

Version:	Publisher's Version
License:	Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden
Downloaded from:	https://hdl.handle.net/1887/3486743

Note: To cite this publication please use the final published version (if applicable).

Nederlandse samenvatting

Dit proefschrift is voortgekomen uit een interdisciplinaire samenwerking tussen de European Organization for Research and Treatment of Cancer (EORTC), het Mathematisch Instituut van de Universiteit Leiden en de afdeling Medische Oncologie van het Leids Universitair Medisch Centrum (LUMC). Het onderzoek werd opgesplitst in twee delen. **Deel** I (**Hoofdstukken** 2, 3, 4) beschrijven statistische analyses uitgevoerd voor de EORTC - Soft Tissue and Bone Sarcoma Group (STBSG), in **Deel** II (**Hoofdstukken** 5, 6, 7, 8) werd het potentieel van overlevingsvoorspellingsmodellen met machine learning (ML) technieken vergeleken met traditionele statistische modellen (SM) voor sarcoom en niet-sarcoom klinische data.

Deel I: Klinische proeven bij wekedelensarcomen

Dit deel verschafte moderne drempelwaarden om nieuwe klinische fase II studies te ontwerpen voor de werkzaamheid van nieuwe behandelingen voor veelvoorkomende histotypes van lokaal gevorderde of gemetastaseerde wekedelensarcoom (STS)-patiënten. De prognostische betekenis van botmetastasen bij STS werd onderzocht om patiëntenpopulaties met een hoog risico te identificeren.

In 2002 publiceerde Van Glabbeke *et al.* namens de EORTC - STBSG een gepoolde analyse om de progressievrije percentages patiënten op 3 en 6 maanden te schatten voor verschillende groepen STS-patiënten die deelnamen aan fase II-onderzoeken van de EORTC. Deze historische waarden zijn op grote schaal gebruikt (> 420 citaties) om nieuwe studies te ontwerpen voor alle STS of voor specifieke histologische subgroepen (in eerstelijnsbehandeling). We hebben een uitgebreid literatuuronderzoek uitgevoerd om alle fase II of daaropvolgende klinische onderzoeken van geavanceerde of gemetastaseerde STS (2003 tot 2018) te identificeren, en zo het huidige landschap te beschrijven. Vanwege de aanzienlijke heterogeniteit tussen klinische onderzoeken werd besloten om eerst te focussen op leiomyosarcoom (LMS) - het meest voorkomende STS-subtype in de artikelen van ons literatuuronderzoek. In **Hoofdstuk** 2 werd een meta-analyse met een random-effecten model uitgevoerd om nieuwe drempelwaarden te bepalen voor het opzetten van fase II-onderzoeken van patiënten met gevorderd of gemetastaseerd LMS, afzonderlijk voor eerstelijns- of voorbehandelde populatie. De primaire eindpunten van belang waren progressievrije overlevingspercentages (PFSR's) na 3 en 6 maanden, die tegenwoordig de voorkeur hebben en vaker worden gerapporteerd dan progressievrije percentages (waarbij niet-ziektegerelateerde sterfte gecensureerd wordt). Wanneer schattingen niet konden worden afgeleid uit publicaties, is contact opgenomen met eerste auteurs en/of sponsors. De ESMO Magnitude of Clinical Benefit Scale (MCBS) werd gebruikt om een indicatie te krijgen van het behandelingseffect dat voorzien moet worden in toekomstige onderzoeken. Er werd informatie verkregen over 7 eerstelijns en 16 voorbehandelde onderzoeken bij 1500 LMS-patiënten. Onder het alternatief dat het werkelijke voordeel een hazard ratio van 0.65 bedraagt, kan een PFSR op 6 maanden van $\geq$ 70% worden overwogen om te zoeken naar een actieve behandeling in de eerste lijn. Voor een voorbehandelde populatie zou een PFSR van 3 maanden $\geq$ 62% of een PFSR van 6 maanden $\geq$ 44% wijzen op een werkzame behandeling. Specifieke resultaten werden ook verstrekt voor LMS van de baarmoeder.

In **Hoofdstuk** 3 werd een tweede meta-analyse uitgevoerd voor gevorderd of gemetastaseerd liposarcoom (LPS) of

synoviosarcoom (SS) - het tweede en derde meest voorkomende histotype in ons literatuuronderzoek. Onderzoekseindpunten waren PFSR's na 3 en 6 maanden. De keuze van het therapeutische voordeel dat in toekomstige onderzoeken zou moeten worden nagestreefd, werd opnieuw geleid door de ESMO MCBS. Informatie werd verkregen voor 1030 LPS-patiënten (25 onderzoeken; 7 eerstelijns, 17 voorbehandeld, 1 beide) en 348 SS-patiënten (13 onderzoeken; 3 eerstelijns, 10 voorbehandeld). Er werden opnieuw drempelwaarden voorgesteld voor toekomstige histologie-specifieke fase II-studies. Streefwaarden in de eerste lijn na 3 en 6 maanden waren 79% en 69% voor LPS, 82% en 69% voor SS. Voor voorbehandelde patiënten waren de streefwaarden voor PFSR's na 3 en 6 maanden 63% en 44% voor LPS, 60% en 41% voor SS. Onze bevindingen hier en in het vorige hoofdstuk geven aan dat er behoefte is aan een hogere drempel voor de meest voorkomende STS-types in toekomstige, op histologie afgestemde fase II-studies om hogere succespercentages te bereiken in nieuwe prospectieve bevestigende fase III-studies.

In **Hoofdstuk** 4 hebben we onderzocht of, en zo ja, in welke mate botmetastasen bij start van de behandeling de prognose van patiënten met gevorderde of gemetastaseerde STS beïnvloeden. Geselecteerde patiënten namen deel aan vijf klinische studies van EORTC - STBSG. Individuen werden geïncludeerd als ze begonnen met de behandeling met een actief medicijn en gevorderd/gemetastaseerd STS hadden. De eindpunten voor dit onderzoek waren algehele overleving (OS) en progressievrije overleving (PFS). Univariate en multivariate gepoolde analyses (na correctie voor 12 covariaten) werden gebruikt met Kaplan-Meier en Cox-regressie om de impact van botmetastase bij presentatie per behandelingslijn (eerstelijns of later) te modelleren, gestratificeerd per studie. Voor de groep van patiënten met botmetastase werd de impact van de aanwezigheid van andere uitzaaiingen (onder andere in de lever, lymfeklieren, long, zacht weefsel of andere) op het moment van de diagnose onderzocht aan de hand van multivariate Cox-regressiemodellen. 565 van de 1034 (54.6%) patiënten kregen eerstelijns systemische behandeling voor lokaal gevorderde of gemetastaseerde ziekte. Botmetastasen waren aanwezig bij 140 patiënten (77 eerstelijns, 63 tweedelijns of later). Het niet-gecorrigeerde verschil in OS/PFS met of zonder botmetastase was alleen statistisch significant voor eerstelijnspatiënten. Voor OS waren de aangepaste hazard ratios voor de aanwezigheid van botmetastasen 1.33 (95%-BI: 0.99-1.78) en 1.11 (95%-BI: 0.81-1.52) voor eerstelijns/tweedelijns of later behandelde patiënten, respectievelijk. De gecorrigeerde hazard ratios voor PFS waren 1.31 (95%-BI: 1.00–1.73) en 1.07 (95%-BI: 0.80–1.43). De gecorrigeerde effecten waren dus niet statistisch significant, ondanks een trend voor eerstelijnspatiënten. Subgroepanalyses wezen op bot- en lymfekliermetastase als de meest schadelijke combinatie voor OS en bot- en longmetastase voor PFS. Aangezien skeletmetastasen bij aanvang van het onderzoek niet als significante risicofactor (per behandelingslijn) kunnen worden geïdentificeerd, is stratificatie in gerandomiseerde onderzoeken met deze patiënten niet aangewezen.

Deel II: Statistische modellen versus machine learning om overleving te voorspellen aan de hand van sarcoom en niet-sarcoom klinische gegevens

In dit deel van het proefschrift werden de voorspellende prestaties van bestaande en nieuwe ML-methoden vergeleken met traditionele SM voor de analyse van real-world time-to-event data (van kleine/middelgrote of grote steekproefomvang, met laag- of hoogdimensionale gegevens).

Tegenwoordig is er een groeiende interesse van de medische gemeenschap in toepassingen van ML voor klinische voorspelling. In de loop der jaren zijn er verschillende algoritmen ontwikkeld en aangepast aan rechtsgecensureerde data. Neurale netwerken zijn herhaaldelijk gebruikt om klinische voorspellingsmodellen in de gezondheidszorg te bouwen. Ondanks hun niet te verwaarlozen gebruik, ontbreekt er een uitgebreide beoordeling van overlevingsneurale netwerken (SNN's) op basis van prognostische factoren. In **Hoofdstuk** 5 presenteerden we de allereerste poging tot een gestructureerd overzicht van SNN's met prognostische factoren voor klinische voorspelling. Ons doel was om een breed overzicht van de literatuur te bieden (1 januari 1990 - 31 augustus 2021, global search in PubMed). Relevante manuscripten werden geclassificeerd als methodologisch/technisch (nieuwe methodologie of nieuw theoretisch model; 13 studies) of toepassingen (11 studies). We bespraken hoe SNN's in de medische wereld worden gebruikt voor voorspellingen en beschreven hoe onderzoekers hebben geprobeerd een classificatiemethode aan te passen aan rechtsgecensureerde overlevingsgegevens. Er zijn twee methodologis-

che trends: ofwel wordt tijd toegevoegd als onderdeel van de invoerfuncties en wordt een enkel uitvoerknooppunt gespecificeerd, of worden meerdere uitvoerknooppunten gedefinieerd voor elk tijdsinterval. Dit werk werd aangevuld met een kritische beoordeling van modelaspecten die zorgvuldiger zouden moeten worden ontworpen en gerapporteerd. We identificeerden de belangrijkste kenmerken van voorspellingsmodellen (d.w.z. aantal patiënten/voorspellers, evaluatiemaatregelen, kalibratie) en vergeleken de voorspellende prestaties van SNN's met het Cox-model voor 'proportional hazards'. De mediaan van de steekproefomvangen was 920 patiënten en de mediaan van het aantal voorspellende factoren was 7. De belangrijkste bevindingen waren onder meer slechte rapportering (bijvoorbeeld met betrekking tot ontbrekende gegevens, hyperparameters), evenals onnauwkeurige modelontwikkeling/-validatie. Kalibratie werd in meer dan de helft van de onderzoeken verwaarloosd. Cox-modellen werden niet tot hun volle potentieel ontwikkeld en claims voor de prestaties van SNN's waren overdreven. Er werd licht geworpen op de huidige stand van de techniek van SNN's in de geneeskunde met prognostische factoren. Beperkingen werden besproken en toekomstige richtingen werden voorgesteld voor onderzoekers die bestaande methodologie verder willen ontwikkelen.

Er is een open discussie over de waarde van ML versus SM binnen de klinische en zorgpraktijk. ML-technieken kunnen een aantrekkelijke keuze zijn voor het modelleren van complexe gegevens (grote steekproefomvang, hoogdimensionale setting). In **Hoofdstuk** 6 werden drie ML-technieken: a) random survival forests (RSF), en b-c) twee methodologische uitbreidingen van het partiële logistieke kunstmatige neurale netwerk (PLANN) met één en twee verborgen lagen getest op een grote retrospectieve gegevensset van 62294 patiënten uit de Verenigde Staten, verstrekt door de Scientific Registry for Transplant Recipients. In totaal werden 97 variabelen geselecteerd, uit een total van meer dan 600, om overleving sinds levertransplantatie te voorspellen op klinische/statistische gronden. Er is een vergelijking gemaakt tussen deze ML-technieken en drie verschillende Cox-modellen (volledig model met alle variabelen, achterwaartse selectie, LASSO). De nadruk werd gelegd op de voordelen en valkuilen van elke methode en op de interpreteerbaarheid van de ML-methoden. Er werden goed gefundeerde parameters gebruikt (C-index, Brier-score en Integrated Brier-score) en de sterkste prognostische factoren werden voor elk model geïdentificeerd. Het klinische eindpunt was de transplantaat-overleving, gedefinieerd als de tijd tussen de transplantatie en de datum van het falen van het transplantaat of overlijden van de patiënt. De RSF vertoonde iets betere voorspellende prestaties dan Cox-modellen op basis van de C-index. Neurale netwerken vertoonden betere prestaties dan zowel Cox-modellen als RSF op basis van de Integrated Brier Score na 10 jaar. Van de drie ML-technieken waren de voorspelde overlevingskansen van de PLANN met één verborgen laag het nauwkeurigst, en net zo goed gekalibreerd als het Cox-model met alle variabelen. De RSF en de PLANN uitgebreid met twee verborgen lagen waren minder goed gekalibreerd op de testgegevens. Wat betreft de interpreteerbaarheid, identificeerden het Cox-model met alle variabelen en de PLANNs *hertransplantatie* als de sterkste voorspellende factor en *donorleeftijd*, *diabetes*, en *levensondersteuning* als relatief sterke voorspellers. Volgens RSF was *donorleeftijd* de meest voorspellende variabele, gevolgd door *hertransplantatie*, *levensondersteuning* en *serologiestatus van het chronische hepatitis C-virus*. Al met al werd aangetoond dat ML-technieken een nuttig hulpmiddel kunnen zijn voor zowel voorspelling als interpretatie in deze overlevingscontext.

In de vorige studie leverde onze groep nieuwe methodologische uitbreidingen van het PLANN-model. De PLANN uitbreiding werd ontwikkeld en gevalideerd voor complexe levertransplantatiegegevens. Het is echter niet ongebruikelijk dat een klein aantal patiënten wordt gerekruteerd in klinische studies en een beperkt aantal voorspellende kenmerken verzameld worden, bijvoorbeeld in sarcoomonderzoeken. Toch verwachten clinici dat ML-modellen mogelijk beter presteren dan SM. Daarom lag de focus in **Hoofdstuk** 7 op de vergelijking tussen dergelijke modellen voor niet-complexe klinische gegevensbanken (kleine / middelgrote steekproefomvang, laagdimensionaal) gecomplementeerd met een Monte Carlo-simulatiestudie om een verschillende real-life settingen te kunnen bestuderen. Er werden synthetische gegevens (250 of 1000 patiënten) gegenereerd die sterk leken op vijf prognostische factoren die vooraf waren geselecteerd op basis van een Europese Osteosarcoom Intergroup-studie (MRC BO06/EORTC 80931) waarin het effect van dosis-intensieve chemotherapie werd onderzocht bij patiënten met gelokaliseerd osteosarcoom in de extremiteiten. De voorspellende prestaties van PLANN original en PLANN extended (met één verborgen laag) werden vergeleken met Cox-modellen voor 20, 40, 61 en 80% gecensureerde gegevens. Overlevingstijden werden gegenereerd op basis van een log-normale verdeling. Het eindpunt van belang was de totale overleving, gedefinieerd als de tijd tot overlijden door welke oorzaak dan ook sinds de datum

van de operatie. Modellen werden geëvalueerd op basis van de C-index, Brier-score op 0-5 jaar, geïntegreerde Brier-score (IBS) op 5 jaar en miskalibratie op 2 en 5 jaar (een parameter die meestal verwaarloosd wordt). De ML-modellen waren in staat om een vergelijkbare voorspellende prestatie te bereiken op gesimuleerde gegevens voor de meeste scenario's met betrekking tot de C-index, Brier-score of IBS. De SM waren echter vaak beter gekalibreerd. De prestaties waren robuust in scenario's waarin gecensureerde patiënten werden verwijderd voor het 2^{de} jaar of administratieve censurering na 5 jaar werd uitgevoerd (op trainingsgegevens). Onderzoekers moeten zich bewust zijn van de tijdsintensieve aspecten van het werken met ML-technieken, zoals data preparatie, afstemming van hyperparameters en rekentijd, waardoor ze nadelig zijn ten opzichte van conventionele regressiemodellen in een eenvoudige klinische setting.

In gezondheidsonderzoek zijn verschillende chronische ziekten vatbaar voor concurrerende risico's (CR's). Aanvankelijk werden SM ontwikkeld om de cumulatieve incidentie van een interessante gebeurtenis te schatten in de aanwezigheid van CR's. Dankzij de groeiende interesse in het toepassen van ML voor klinische voorspelling, zijn deze technieken ook uitgebreid naar CR's, maar de literatuur is nog beperkt. In **Hoofdstuk** 8 wilden we prognostische klinische voorspellingsmodellen voor CR's ontwikkelen en valideren met SM- en ML-technieken. Twee SM a) oorzaak-specifieke Cox, b) Fine-Gray-model en drie ML-modellen i) PLANN origineel voor CR's (PLANNCR origineel), ii) een methodologische uitbreiding genaamd PLANNCR uitgebreid, en iii) RSF voor CR's (RSFCR) werden gebruikt. De voorspellende prestaties van alle methoden werden beoordeeld in termen van discriminatie en kalibratie in een andere eenvoudige klinische setting (kleine / middelgrote steekproefomvang, klein aantal factoren). De beschikbare dataset bevat 3826 retrospectief verzamelde gegevens van patiënten met extremiteit STS (eSTS) en negen variabelen van de gepersonaliseerde SARcoma Care (PERSARC) Study Group. Voor zover wij weten, was dit de allereerste studie van deze soort voor eSTS. Het klinische eindpunt was de tijd in jaren tussen operatie en ziekteprogressie (interessante gebeurtenis) of overlijden (concurrerende gebeurtenis). De Brier-score, het gebied onder de curve (AUC) en de miskalibratie van het model werden gebruikt om de voorspellende prestaties na respectievelijk 2, 5 en 10 jaar te evalueren. De resultaten toonden aan dat de ML-modellen een vergelijkbare prestatie kunnen bereiken met de SM op basis van de Brier-score en AUC voor ziekteprogressie en overlijden (95% betrouwbaarheidsintervallen bij 2, 5 en 10 jaar overlapten). Niettemin waren de SM vaak beter gekalibreerd. Over het algemeen zijn ML-technieken minder praktisch omdat ze een aanzienlijke implementatietijd vergen (voorbereiding van de data, afstemming van hyperparameters, rekenintensiteit). Als zodanig moeten deze technieken voor niet-complexe real-life problemen alleen worden toegepast als aanvulling op SM als verkennende instrumenten voor de prestaties van modellen. Meer aandacht voor modelkalibratie is dringend nodig.