

The future of clinical trials: prognostic scores for the development of new cancer medications

Hugo Daniel Paes Loureiro Santos Ambrósio

Vollständiger Abdruck der von der TUM School of Life Sciences der Technischen Universität München zur Erlangung eines Doktors der Naturwissenschaften (Dr. rer. nat.) genehmigten Dissertation.

Vorsitz: Prof. Dr. Dmitrij Frischmann

Prüfende der Dissertation: 1. Prof. Dr. Dr. Fabian Theis

2. Prof. Dr. Conceição Amado

Die Dissertation wurde am 11.03.2024 bei der Technischen Universität München eingereicht und durch die TUM School of Life Sciences am 26.07.2024 angenommen.

To my father, in loving memory.

Abstract

The safety and efficacy of new drug candidates are assessed in drug development. Drug development is a long and complex process comprising a sequence of clinical trials. As new drug candidates advance to the later stages of drug development, large cohorts are enrolled to identify differences in overall survival (OS), the gold-standard measure of efficacy in solid tumors. However, most new oncology drugs fail during the drug development stage as they do not demonstrate superior OS than the standard of care. The drug development process is fraught with several challenges that could prevent potentially efficacious drugs from demonstrating efficacy. For instance, patient drop-out due to adverse events unrelated to the medication, difficulties during patient enrollment, and cross-over between treatment arms can hinder the ability of new medications to show efficacy. Additionally, inappropriate decision-making in early clinical trials can lead inefficacious drugs to progress in drug development. Several approaches have been proposed to assist in clinical trials and support with these problems. Techniques such as patient enrichment can supplement clinical trials with patients with phenotypes that make them more likely to demonstrate drug effect. Still, as an ever increasing number of biomarkers are tested, new computational tools will be required to identify the populations of interest. Additionally, surrogate endpoints assist clinical trial teams by estimating the efficacy in early clinical trial phases or in early interim analyses. The current surrogate endpoints, though, have shown lower correlation with OS for new medication types, such as cancer immunotherapy. In this thesis, I investigate prognostic scores as tools to tackle these problems in clinical trials and to assist decision-making during drug development. Firstly, I benchmarked several statistical and machine learning survival models to assess which ones were more accurate in predicting the mortality risk. A good prognostic score that can reliably forecast the hazard could be used to enrich clinical trials with patients who are less likely to die soon after treatment starts. Next, I investigated the use of prognostic scores to match historical patients into external controls. Using data from historical clinical trials or real-world data to construct the control arms would expedite clinical trials by decreasing the number of patients enrolled. Lastly, I presented the risk trend framework, which predicts the OS results of clinical trials with prognostic scores and mortality values at interim analyses. Clinical trial teams could leverage the early OS estimates to perform data-driven decisions on the development of the drug. The presented analyses provide multiple applications of prognostic scores to drug development. These approaches can assist throughout the trial execution, from enrollment, to interim analyses and the creation of external cohorts. The innovative applications of prognostic scores described in this thesis pave the way for developing more prognostic score solutions for drug development.

Kurzzusammenfassung

In der Arzneimittelforschung werden die Sicherheit und Wirksamkeit neuer Arzneimittelkandidaten bewertet. Die Entwicklung von Arzneimitteln ist ein langer und komplexer Prozess, der u.a. eine Aneinanderreihung von klinischen Studien verschiedener Phasen umfasst. Wenn neue Arzneimittelkandidaten in die späteren Phasen der Arzneimittelentwicklung eintreten, werden große Studien durchgeführt, um Unterschiede im Overall Survival (OS), dem Goldstandard für die Wirksamkeit bei soliden Tumoren, zu ermitteln. Die meisten neuen Krebsmedikamente scheitern jedoch bereits in der Entwicklungsphase, ohne eine bessere Überlebenszeit als die Standardtherapie aufzuweisen.

Der Prozess der Arzneimittelentwicklung ist mit zahlreichen Herausforderungen verbunden, die verhindern könnten, dass potenziell wirksame Arzneimittel ihre Wirksamkeit unter Beweis stellen. So können z. B. Patientenabbrüche aufgrund von unerwünschten Ereignissen, die nichts mit dem Medikament zu tun haben, Schwierigkeiten bei der Patientenrekrutierung und Cross-over zwischen den Behandlungsarmen die Feststellung der Wirksamkeit neuer Medikamente beeinträchtigen. Darüber hinaus können unangemessene Entscheidungen in frühen klinischen Studien dazu führen, dass unwirksame Medikamente in die nächste Phase der Entwicklung kommen.

In dieser Arbeit untersuche ich prognostische Scores als Instrumente zur Lösung dieser Probleme klinischer Studien und zur Unterstützung der Entscheidungsfindung während der Arzneimittelentwicklung. Zunächst habe ich verschiedene statistische Überlebenszeitmodelle und Modelle des maschinellen Lernens miteinander verglichen, um festzustellen, welche Modelle das Sterberisiko besser vorhersagen können. Ein guter prognostischer Score, der das Risiko zuverlässig vorhersagen kann, kann dazu verwendet werden, klinische Studien mit Patienten anzureichern, bei denen die Wahrscheinlichkeit, dass sie bereits kurz nach Beginn der Behandlung versterben, geringer ist. Als Nächstes untersuchte ich die Verwendung von Prognosescores für das Matching von historischen Patienten als externe Kontrollen. Die Verwendung von Daten aus historischen klinischen Studien oder aus der realen Behandlungspraxis zur Bildung von Kontrollgruppen würde klinische Studien beschleunigen, indem die Zahl der tatsächlich in die klinische Studie einzuschließenden Patienten verringert wird. Schließlich habe ich das „Risk trend Framework“ vorgestellt, das die OS-Ergebnisse klinischer Studien aus dem Verlauf der prognostischen Scores und frühen Mortalitätswerten bei Interimsanalysen vorhersagt. Klinische Studienteams könnten die frühen OS-Schätzungen nutzen, um datengesteuerte Entscheidungen über die Weiterführung der Entwicklung des Medikaments zu treffen.

Die vorgestellten Analysen zeigen mehrere Anwendungsmöglichkeiten für prognostische Scores in der Arzneimittelentwicklung. Diese Ansätze können während der gesamten Studiendurchführung hilfreich

sein, von der Rekrutierung bis hin zu Interimsanalysen und der Erstellung externer Kontrollkohorten. Die in dieser Arbeit beschriebenen innovativen Anwendungen von prognostischen Scores ebnen darüber hinaus den Weg für die Entwicklung weiterer prognostischer Score-Lösungen für die Arzneimittelentwicklung.

Acknowledgments

My journey during my PhD studies was supported by many friends, colleagues, and supervisors. Firstly, I thank Anna Bauer-Mehren and the team at Roche for the opportunity to perform my PhD research in her group at Roche. A special thanks for the great support, ideas, brainstorming sessions, and encouragement. Moreover, I thank Fabian Theis and the group at Helmholtz Munich for enabling my research.

I thank Dmitrij Frishman, Fabian Theis, and Conceição Amado for examining my PhD thesis. Additionally, I would like to thank Fabian Theis, Carsten Marr and Anna Bauer-Mehren for their support and feedback during my thesis advisory committee meetings.

I thank my supervisors for their support. Specifically, Anna Bauer-Mehren, Tim Becker, Janick Weberpals, Carlos Talavera-López, and Narges Ahmidi, which supported me throughout this journey and provided invaluable feedback.

I would like to acknowledge the pREDi Data Science IV group from Roche Diagnostics GmbH in Penzberg for their support, and feedback on my work. In particular, I would like to acknowledge Franziska Braun and Fabian Schmich for their great support and feedback. Additionally, my fellow PhD students at Roche, Ali Boushehri, Nikita Makarov, and Maria Bordukova for all the great conversations and brainstorming sessions.

Additionally, I would also like to thank my two groups at Helmholtz Munich, firstly the group led by Narges Ahmidi, specifically, Theresa Wirth, Octavia Ciora, Elisabeth Pachel, Henrik von Kleist and Alireza Zamanian; and second the group led by Carlos Talavera-López, for the great companionship and conversations.

I thank Julia Schlehe and Mara Kieke for the great work that they do in the MUDS graduate school. MUDS was an excellent program that provided me with a community of bright individuals to discuss ideas, and many opportunities to learn about topics outside of my research area. Moreover, I would like to thank the HELENA graduate school for the good structure and support that they offered me.

I thank Theresa M. Kolben, Astrid Kiermaier, Dominik Rüttinger, Andreas Roller, and Meike Schneider for the excellent teamwork and motivation during the projects that we worked together on.

Lastly, but in no way with lower importance, I would also like to acknowledge my friends and family that accompanied me through my PhD and broadly through my life. À minha família, quero deixar um agradecimento especial pelo suporte imensurável e incondicional. Obrigado, António, Maria de Fátima e Cátia.

List of contributed publications

Peer-reviewed full-length research papers

- I. Artificial Intelligence for Prognostic Scores in Oncology: A Benchmarking Study (Loureiro, Becker, Bauer-Mehren, et al. 2021)
 - [H. Loureiro](#), T. Becker, A. Bauer-Mehren, N. Ahmidi*, J. Weberpals* (* contributed equally)
 - Frontiers in Artificial Intelligence Vol. 4 (9), 2021
 - doi: 10.3389/frai.2021.625573
- II. Matching by OS prognostic score to construct external controls in lung cancer clinical trials (Loureiro, Roller, et al. 2023)
 - [H. Loureiro](#), A. Roller, M. Schneider, C. Talavera-López, T. Becker*, A. Bauer-Mehren* (* contributed equally)
 - Clinical Pharmacology & Therapeutics
 - doi: 10.1002/cpt.3109
- III. Correlation between early trends of a prognostic biomarker and overall survival in non-small-cell lung cancer clinical trials (Loureiro, Kolben, et al. 2023)
 - [H. Loureiro](#), T. M. Kolben, A. Kiermaier, D. Rüttinger, N. Ahmidi, T. Becker*, A. Bauer-Mehren* (* contributed equally)
 - JCO Clinical Cancer Informatics
 - doi: 10.1200/CCI.23.00062

Conference Presentations

- IV. Improving Predictive Ability of Survival Models: Comparison of Multiple State of the Art Models
 - [H. Loureiro](#), T. Becker, A. Bauer-Mehren, N. Ahmidi, J. Weberpals
 - International Conference on Pharmacoepidemiology 2020, online due to pandemic
 - The abstract is available in Pharmacoepidemiology and Drug Safety, Vol. 29, Oct. 2020, Pages 35–36.
 - doi: 10.1002/pds.5114

Conference Posters

- V. Towards OS approximation in early clinical trials using prognostic scores
 - [H. Loureiro](#), T. Becker, N. Ahmidi, A. Bauer-Mehren
 - International Conference on Pharmacoepidemiology 2021, online due to pandemic

- The abstract is available in *Pharmacoepidemiology and Drug Safety*, Vol. 30, Aug. 2021, Pages 366–367.
- doi: 10.1002/pds.5305

VI. A longitudinal early-indicator of overall survival based on prognostic scores

- H. Loureiro, T. Becker, A. Bauer-Mehren
- International Conference on Pharmacoepidemiology 2022, Copenhagen, Denmark
- The abstract is available in *Pharmacoepidemiology and Drug Safety*, Vol. 31, Sep. 2022, Page 559.
- doi: 10.1002/pds.5518.

Glossary

Roman letters

A	Segment value (joint models)
b	Random effects (linear mixed effects)
c	Arbitrary function
d	Number of individuals for which the event occurred
D	Random effects covariance matrix (linear mixed models)
F	Arbitrary function
g	Neural network output
h	Hazard function
I	Identity matrix
h_0	Baseline hazard function
H	Cumulative hazard function
H_0	Null hypothesis (hypothesis testing)
H_1	Alternative hypothesis (hypothesis testing)
ℓ	Log likelihood function
\mathcal{L}	Likelihood function
L_1	L one norm (lasso regularization)
L_2	L two norm (ridge regression)
n	Number of individuals at risk
N	Number of events counting process
R	Patient at risk indicator
S	Survivor function
t	Time
T	Time-to-event
x	Explanatory variable/covariate

X	Explanatory variable vector/covariate vector
y	Response variable / longitudinal variable
W	Random effect design matrix (linear mixed effects / joint models)
Y	Response variable vector / longitudinal variable vector
Z	Fixed effect design matrix (linear mixed effects / joint models)

Greek letters

α	Control of the type of shrinkage
β	Fixed effects (Linear mixed effects)
γ	Regression coefficients
δ	Censoring function
ε	Error term
ζ	Coefficient of the Weibull distribution
θ	Linear regression coefficients (risk trend framework)
κ	Spline coefficients (joint models)
λ	Coefficient of the exponential distribution
ξ	Control of the shrinkage strength
ρ	Model weights (gradient boosting)

Abbreviations

ADAM	Adaptive moment estimation
advNSCLC	Advanced non-small-cell lung cancer
AE	Autoencoder
AUC	Area under the curve
DS	DeepSurv
eControls	External controls
GB	Gradient boosting
HR	Hazard ratio
JM	Joint Models for longitudinal and survival data
LME	Linear mixed models

NN	Neural network
NSCLC	Non-small-cell lung cancer
ORR	Overall response rate
OS	Overall survival
PFS	Progression-free survival
ROC	Receiver operating curve
ROPRO	Real world prognostic score
RSF	Random survival forest
RWD	Real-world data
SELU	Scaled exponential linear units
SL	Super learner

List of Figures

Figure 1.1 Example of the sequence of clinical trials in oncology. The patient numbers are only for illustration purposes and do not represent the variability encountered in clinical development. The figure was based on information from the National Cancer Institute (National Cancer Institute 2023).	1
Figure 1.2 Examples of routine blood work biomarkers measured in cancer care.....	2
Figure 1.3 Illustration of the interactions of the cancer patient with the clinic. In each interaction the prognostic score value for the patient is calculated. Values such as routine bloodwork collected at that timepoint alongside baseline biomarkers can be included in the prognostic score.	3
Figure 1.4 Diagram of matching historical patients in a clinical trial.	6
Figure 1.4 Illustration of the “average” prognostic score curves for two different cohorts.	7
Figure 2.1 Illustration of a Survivor function estimated with the Kaplan Meier estimator.....	12
Figure 2.2 Two survivor curves and the corresponding log-rank test p-value.	13
Figure 2.3 Diagram of a simple neural network.	19
Figure 2.4 Example of the biomarker stair function assumed by the Extended Cox model. The biomarker measurements were obtained every 20 days for 12 months.	24
Figure 3.1 Diagram of the analysis. This figure is Figure 1 of the original manuscript by Loureiro et al. (Loureiro, Becker, Bauer-Mehren, et al. 2021).	33
Figure 3.2 Violin plot of the Harrell C-index values for the FH in-sample test and OAK test datasets. The values in the violin plot were obtained by Bootstrap. This figure is based on Figure 6 of the original publication by Loureiro et al. (Loureiro, Becker, Bauer-Mehren, et al. 2021).	35
Figure 3.3 Prediction error of the OS HR for the phase III clinical trials. This figure is based on Figure 3 of the original publication by Loureiro et al. (Loureiro, Roller, et al. 2023).	37
Figure 3.4 Diagram of the steps taken to estimate the OS HR of a new clinical trial with the Risk Trend Framework. This figure is based on Figure 1 of Loureiro et al. (Loureiro, Kolben, et al. 2023).....	40
Figure 3.5 Prediction error of the final OS HR using the ROPRO JM model coefficients at the three- and 6-months interim analysis. Based on Figure 4 of the original publication by Loureiro et al. (Loureiro, Kolben, et al. 2023).	42

Contents

Abstract	I
Kurzzusammenfassung	II
Acknowledgments	IV
List of contributed publications	V
Glossary	VII
List of Figures	X
Contents	XI
1 Introduction	1
1.1 Biomarkers collected in cancer clinical trials.....	2
1.2 Use of prognostic scores in drug development.....	3
1.3 Patient enrichment in drug development.....	3
1.4 Historical data use in drug development	5
1.5 Efficacy estimation in early drug development.....	7
1.6 Aim of the thesis.....	8
2 Background	10
2.1 Survival Analysis	10
2.1.1 Estimation of the Survivor function - Kaplan Meier estimator	11
2.1.2 Calculate difference between survivor functions	12
2.1.3 Estimation of the hazard function.....	13
2.2 Risk / prognostic score models.....	14
2.2.1 Cox model	15
2.2.2 Regularized Cox model	16
2.2.3 Tree-based survival models.....	17
2.2.4 Gradient boosting models applied to risk / prognostic score models	18
2.2.5 Deep learning applied to risk / prognostic score models.....	19
2.2.6 Autoencoder applied to survival analysis.....	20
2.2.7 Super Learner applied to survival analysis.....	21
2.2.8 Goodness of fit of survival models.....	23
2.3 Survival analysis with time-varying covariates.....	23
2.3.1 Extended Cox model	23
2.3.2 Joint models for longitudinal and time-to-event data.....	25
2.4 Endpoints in oncology clinical trials	26
2.4.1 Overall survival	27
2.4.2 Progression-free survival.....	27
2.4.3 Objective response rate.....	28
2.5 Propensity scores.....	28

2.5.1	Propensity scores to build external controls	29
2.6	Non-small-cell lung cancer.....	29
2.6.1	Treatment of NSCLC	29
2.7	Historical and real-world data	30
2.7.1	Real-world dataset.....	30
2.7.2	Roche clinical trials	31
3	Publication summaries.....	32
3.1	First work “Artificial Intelligence for Prognostic Scores in Oncology: A Benchmarking Study”	32
3.2	Second work “Matching by OS prognostic score to construct external controls in lung cancer clinical trials”	36
3.3	Third work “Correlation between early trends of a prognostic biomarker and overall survival in non-small-cell lung cancer clinical trials”	39
4	General discussion.....	43
4.1	Summary	43
4.2	Outlook.....	44
	Bibliography	48
	Appendix	68
	Full length manuscript I - Artificial Intelligence for Prognostic Scores in Oncology: A Benchmarking Study	68
	Full length manuscript II - Matching by OS prognostic score to construct external controls in lung cancer clinical trials.....	85
	Full length manuscript III - Correlation between early trends of a prognostic biomarker and overall survival in non-small-cell lung cancer clinical trials.....	95

1 Introduction

This publication-based thesis focuses on drug development. Specifically, on using statistical and artificial intelligence models to accelerate and enhance the decision-making processes.

Over the last few decades, pharmaceutical drugs have become increasingly important in modern medicine (Goldman, Joyce, and Zheng 2007). The increase in the use of pharmaceutical drugs is due to several factors. First, the number of new molecular entities, i.e., pharmaceutical drugs approved for commercialization, has dramatically increased (Kinch et al. 2014). In parallel, the quality of these drugs, measured in their efficacy and safety, has also increased. Additionally, new types of medications, such as monoclonal antibodies, mRNA vaccines, and gene therapy, increase the number of diseases that can be addressed. However, they also increase the complexity of the available medications.

New drug candidates need to pass through a complex drug development process to assess their safety and efficacy. Drug development is composed of preclinical and clinical development stages. In the preclinical stage, the efficacy and safety of the candidate drug are examined in a laboratory setting in organoids or animals (U. S. Food and Drug Administration 2019a). The next stage, clinical development, comprises multiple sequential clinical trials (U. S. Food and Drug Administration 2019b). In a classical setting outlined in Figure 1.1, the new medication candidate is tested in humans from phase I to phase III clinical trials (U. S. Food and Drug Administration 2019b; Taylor 2015). The number of patients increases from the early phases to the later ones, allowing clinical researchers to estimate efficacy and safety accurately. Finally, if the new drug candidate is proven safe and efficacious, it is evaluated for approval by the regulatory agencies.

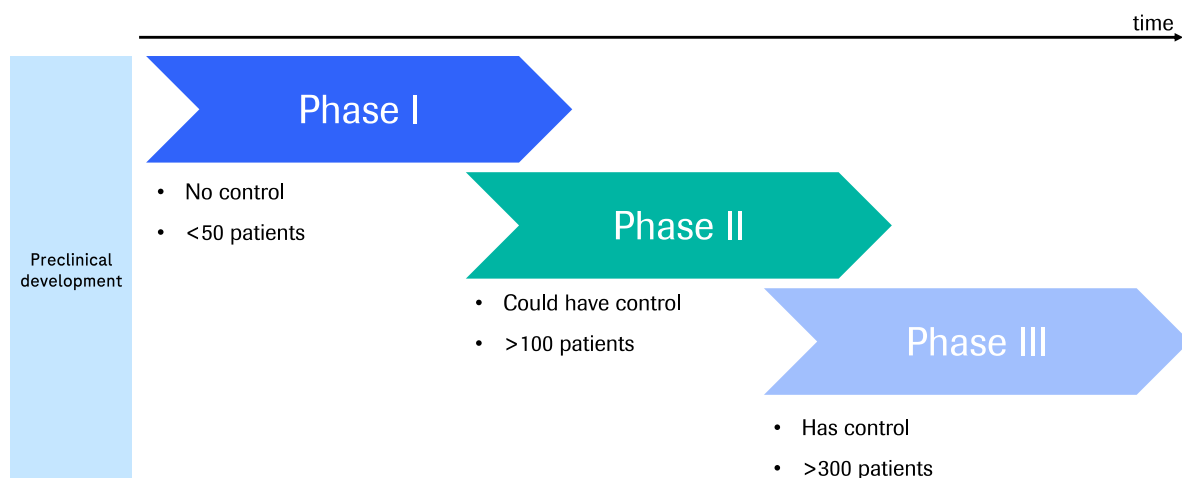


Figure 1.1 Example of the sequence of clinical trials in oncology. The patient numbers are only for illustration purposes and do not represent the variability encountered in clinical development. The figure was based on information from the National Cancer Institute (National Cancer Institute 2023).

The drug development process is lengthy and expensive. Drug development, with its multiple clinical trials, is long (average of 7.9 years) (Kaitin 2010; Taylor 2015), and costly (on average, up to \$2.8 billion for the development of antineoplastic drugs) (Wouters, McKee, and Luyten 2020; Morgan et al. 2011). Considering also the high attrition rate of oncology drugs (Wong, Siah, and Lo 2019), this represents a high expenditure for pharmaceutical companies and a significant time consumption. Therefore, to continue to provide the best treatments to patients, it is imperative to perform concise and data-driven analyses of the likelihood of success of new drugs at early stages (Rubin and Gilliland 2012).

1.1 Biomarkers collected in cancer clinical trials

Patients undergoing cancer treatment (e.g. chemotherapy) have several biomarkers analyzed. At baseline, patients might have cancer specific markers tested, such as EGFR expression in non-small-cell lung cancer or estrogen/progesterone/HER2 expression in breast cancer, to assess the availability of targeted treatments (S.-Y. M. Liu et al. 2023; Heinemann et al. 2013). Additionally, patients will undergo routine blood tests before and during treatment (H. West et al. 2019). The blood tests assess the patient's overall health and, whether it is safe to continue administering the medication (Warr et al. 2013). In general, the blood work panels (see an overview in Figure 1.2) will include blood cell counts, blood chemistry markers, alongside other blood biomarkers that assess the working of specific organs or systems (Warr et al. 2013). The biomarker values are associated with the prognosis of the cancer patients (Becker et al. 2020). Still, the large amount of biomarker values (can be more than 30) generated during treatment make a comprehensive view of the patient difficult. Hence, the use of methods such as prognostic scores can help to create a comprehensive understanding of the state of health of the patients with cancer (Becker et al. 2020).

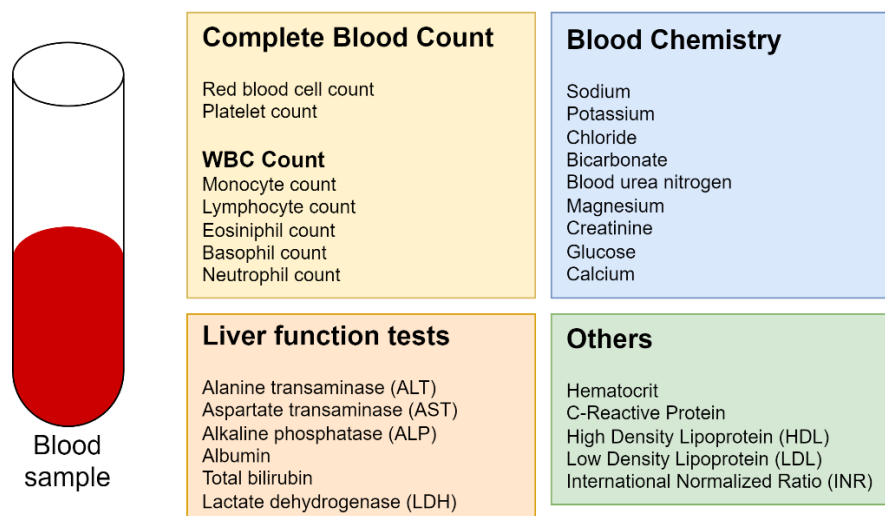


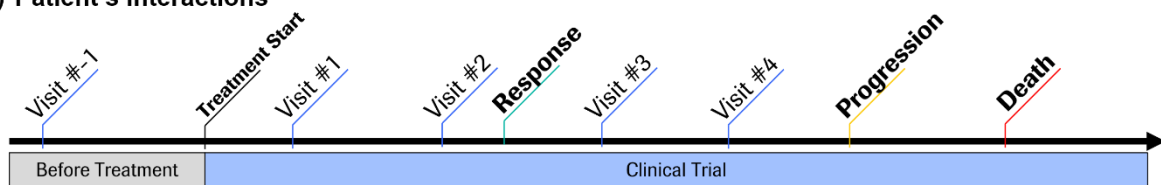
Figure 1.2 Examples of routine blood work biomarkers measured in cancer care.

1.2 Use of prognostic scores in drug development

Prognostic scores, described in detail in subsection 2.2, are statistical models that estimate the prognosis of the patients based on information about their disease and well-being (see Figure 1.3). Prognostic scores are used extensively in medical practice. In oncology, for example, there are multiple prognostic scores available. First, tumor staging, a type of prognostic score, is applied extensively in clinical practice (Telloni 2017; McMillan 2013). Second, the Glasgow prognostic score, based on inflammation (Proctor et al. 2013; McMillan 2013); or the more recent and more comprehensive ROPRO (Becker et al. 2020; 2023). Additionally, outside of oncology, other prognostic scores were developed for severe burns victims (Sheppard et al. 2011), and heart failure (Ferrero et al. 2015).

Given the extent of the use of prognostic scores and their ability to collapse multiple patient characteristics into one score, prognostic scores can constitute a valuable tool in drug development. They are already used in clinical trials (Siegfried, Senn, and Hothorn 2023). In this publication-based thesis, I tackle three different uses of prognostic scores to boost drug development. Firstly, I explored whether more complex, machine-learning prognostic scores could improve the predictive power. Next, I studied whether prognostic scores are a proper tool to define external control arms based on historical datasets. Finally, I investigated whether the longitudinal trend of prognostic scores could predict the efficacy results of clinical trials.

1) Patient's Interactions



2) Transform Vitals/Labs into Prognostic score values

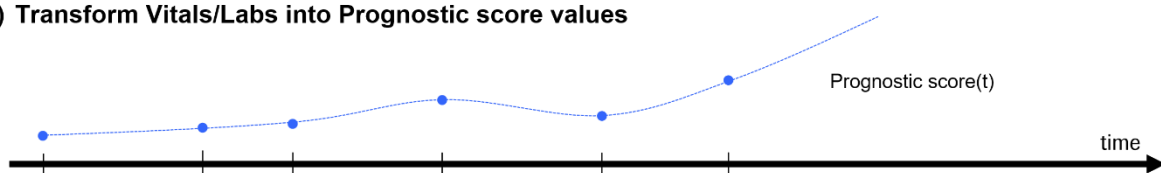


Figure 1.3 Illustration of the interactions of the cancer patient with the clinic. In each interaction the prognostic score value for the patient is calculated. Values such as routine bloodwork collected at that timepoint alongside baseline biomarkers can be included in the prognostic score.

1.3 Patient enrichment in drug development

The primary factor associated with clinical trial failure is the inability to demonstrate efficacy (Fogel 2018). In oncology clinical trials, the primary and gold standard measure of efficacy is overall survival (OS) (F. Fiteni et al. 2014). OS is the time from randomization, or the start of treatment, until the patient's death (F. Fiteni et al. 2014). Several factors associated with the clinical trial design can hinder

a proper estimation of OS. Hence, an improper trial design can lead to potentially efficacious drugs not demonstrating efficacy in clinical trials.

There are several requirements for a reliable estimation of OS. Specifically, OS is highly dependent on the number of patients and on the duration of follow-up of a clinical trial (Mushti, Mulkey, and Sridhara 2018; Zhuang, Xiu, and Elsayed 2009). Additional factors, such as the crossover between study arms (Jönsson et al. 2014) or which treatments are prescribed to patients after their drop-out, can also affect OS (Korn, Freidlin, and Abrams 2011). To minimize these effects, clinical trials should refocus on populations expected to benefit the most from the medication (Freidlin and Korn 2014). Approaches such as patient enrichment can reduce the number of experienced adverse events while also increasing the number of events of interest, mitigating the aforementioned issues with OS (U.S. Food and Drug Administration 2019).

Patient enrichment consists of selecting a patient population where the detection of the drug effect is more likely (U.S. Food and Drug Administration 2019; Temple 2010; Freidlin and Korn 2014). Strategies to “enrich” the cohort fall into three categories: practical, predictive, and prognostic (Temple 2010). Firstly, in practical enrichment, patients are selected to reduce the cohort’s overall variability. Specifically, patients can be excluded if they have highly fluctuating measurements or comorbidities that could contribute to their drop-out from the trial (Temple 2010). Second, in predictive enrichment, the clinical trial team attempts to enroll patients who are more likely to respond to the treatment of interest. A higher number of responders to the treatment increases the likelihood of detecting a treatment benefit. Lastly, in prognostic enrichment, the cohort is augmented with patients for which the occurrence of the event of interest, which the medication should prevent, is higher. These three strategies target different statistical aspects of the clinical trial. Specifically, the practical, predictive, and prognostic enrichment approaches attempt to increase the study power, the absolute effect difference, and the effect size, respectively (U.S. Food and Drug Administration 2019).

Prognostic scores are a clear candidate method to perform prognostic enrichment since they can aggregate complex patient and disease information into a score. Historically, prognostic scores in oncology have been simple models composed of only a few (less than 10) variables (International Non-Hodgkin’s Lymphoma Prognostic Factors Project 1993; Ko et al. 2015; Kinoshita et al. 2013; Carsten Nieder and Astrid Dalhaug 2010). Recently, though, ROPRO, which is composed of 27 covariates was introduced (Becker et al. 2020; 2023). The higher number of covariates grants a higher discriminatory performance to ROPRO than other simpler prognostic scores. ROPRO and other advanced prognostic scores facilitate data-driven approaches to prognostic enrichment by providing a comprehensive view of the health status of the patients.

Although the ROPRO significantly increases the number of variables included in prognostic scores, it is based on the Cox proportional hazards model (D. R. Cox 1972), which is a linear model. Due to its simplicity, the Cox model does not consider nonlinearities or interactions between covariates unless

these interactions are specified in the model (Loureiro, Becker, Bauer-Mehren, et al. 2021). Other machine learning models applied to survival data, such as random survival forests (Ishwaran et al. 2008), gradient boosting (Ridgeway 1999), and DeepSurv (Katzman et al. 2018), also estimate the patient's risk from a set of covariates but do not suffer the same limitations of the Cox model (Loureiro, Becker, Bauer-Mehren, et al. 2021). Therefore, it is possible that these more advanced models could extract additional information from complex patient/disease information. The extra information could better predict the patient's risk and improve prognostic enrichment.

First research question:

Can more complex survival models build more performant prognostic scores?

1.4 Historical data use in drug development

Issues with the enrollment of patients in clinical trials are the major hindrance to drug development (Desai 2020). Over 80% of clinical trials fail to finish the enrollment process on time (Desai 2020), leading to either the termination of the trial or its extension to meet the enrollment goal. Enrolling the number of patients initially specified is imperative to guarantee that the study has the necessary statistical power to obtain a reliable estimate of efficacy. Unger et al. list the lack of access to cancer clinics, the lack of suitable clinical trials for the specific subtype of cancer patients, or stringent inclusion and exclusion criteria as some of the causes of the enrollment issues (Unger et al. 2016). Additionally, patients might fear experimentation or be concerned that randomization into a treatment arm might not lead to the best possible treatment for their disease (Unger et al. 2016).

The efficacy of a new treatment is measured in comparison with the standard of care (Schmidli et al. 2020). Hence, in phase III clinical trials, the cohort is usually split into two or more arms of treatment, where one of the arms is prescribed the standard of care. Given the high number of clinical trials performed yearly, and the slowly changing standard of care for some cancer types, many patients in the control group are prescribed the same treatment in a clinical trial as they would be in a regular clinical setting (Viele et al. 2014). Therefore, historical data, from previous clinical trials or real-world data (RWD) (Mishra-Kalyani et al. 2022), might contain a sizeable number of patients prescribed the standard of care. Hence, to reduce the number of patients that need to be enrolled, one possibility would be to augment the control cohort with historical patients (Viele et al. 2014), or to create an external control (eControl), i.e., to fully replace the control cohort with historical data (Carrigan et al. 2020; K. Tan et al. 2022; Weberpals et al. 2021). Although these techniques could lead to the enrollment of fewer patients, and, hence, assist with the issues in drug development, the resulting trial would no longer be randomized. Without randomization, there could be confounding factors between the cohorts that bias the study results. The bias due to the non-balanced augmented control could display as a worse prognosis for one of the groups in the trial. The worse prognosis could lead to a higher mortality rate of one group that is unrelated to the efficacy of the medication, leading to incorrect clinical trial readouts.

In observational studies, there are tools to mitigate the effect of the lack of randomization. One of the most used methods, propensity scores, balances a set of variables in both the treatment and control cohorts. Which variables to balance with the propensity scores has been an active topic of study. Several authors have determined that all prognostic variables (related to the outcome) should be included in the propensity score (Brookhart et al. 2006; Westreich et al. 2011). Hence, variables such cancer-specific biomarker expression (e.g. EGFR, ALK, PD-1 expression), lifestyle choices (weight, smoking history) or blood-work biomarkers (such as blood cell counts, or biomarkers related to kidney or liver function) should be included in the propensity score.

Controlling as many variables as possible in propensity score might sound reasonable. Still, including many variables in propensity scores can lead to an increase of the variance of the propensity scores (Caliendo and Kopeinig 2008). The increased variance can lead to lower overlap between the scores and fewer matches, leading to a more limited eControl cohort. This effect can be particularly problematic in cases where not only the effect estimates but also the statistical significance of the result is an outcome of the analysis. In summary, fewer patients lead to a lower statistical power of the analysis, which could make obtaining a significant p-value impossible.

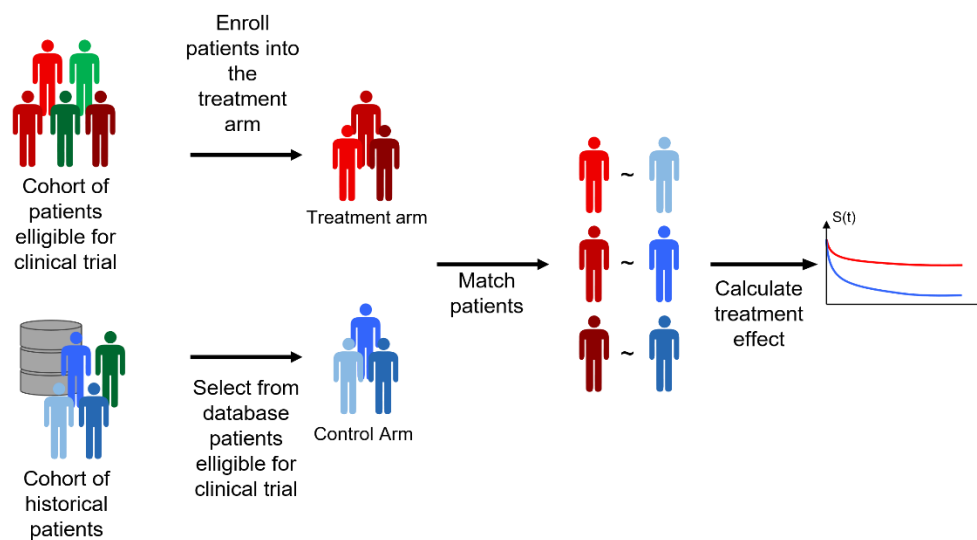


Figure 1.4 Diagram of matching historical patients in a clinical trial.

Instead of propensity scores, prognostic scores can also balance the cohorts (Stuart, Lee, and Leacy 2013). In a simulation study, Stuart et al. showed that prognostic scores can adequately balance the cohorts and reduce bias (Stuart, Lee, and Leacy 2013). Still, this approach was never used to create eControls in oncology. The use of prognostic scores in matching eControls could allow for the inclusion of a wide range of variables without excessively increasing the variability. Still, it might lead to a partial biomarker-level matching of the variables between the cohorts.

Second research question:

Can prognostic scores be used to match external controls?

1.5 Efficacy estimation in early drug development

In drug development, a large trial cohort size and a long follow-up are required to obtain a reliable estimate of OS (Mushti, Mulkey, and Sridhara 2018; Zhuang, Xiu, and Elsayed 2009). Therefore, the estimation of OS is hindered in early clinical trial phases due to the low number of patients, and in interim analyses of phase III clinical trials due to limited follow-up (Loureiro, Kolben, et al. 2023). To assist with decision-making in the early stages of clinical trials of solid tumors, surrogate endpoints, such as progression-free survival or overall response rate are used to obtain an early estimate of efficacy (Savina et al. 2018; Frédéric Fiteni, Westeel, and Bonnetain 2017).

The surrogate endpoints are commonly used in solid tumors as surrogate endpoints for OS (Savina et al. 2018; Frédéric Fiteni, Westeel, and Bonnetain 2017). The surrogate endpoints analyze the difference between the number of progressions and responses, respectively. In solid tumors, progression and response events are determined based on changes in tumor size obtained by imaging techniques (Villaruz and Socinski 2013; Sullivan, Schwartz, and Zhao 2013). Progression and response are events that generally occur earlier than death, hence their use instead of OS (Fallowfield and Fleissig 2012). Although these surrogate endpoints have shown a high correlation with OS in classic antineoplastic treatments (Mauguen et al. 2013; Savina et al. 2018), their association with OS is weaker for new types of treatment, such as immunotherapy (Mushti, Mulkey, and Sridhara 2018; Ye et al. 2020). Additionally, these endpoints also depend on the subjectivity of the readers when measuring the lesions (Sullivan, Schwartz, and Zhao 2013).

The "average" prognostic score curve for each cohort:

$$\text{Prognostic}_{\text{drug}=1}(t) \begin{cases} \text{Prognostic}_{\text{Patient } 1}(t) \\ \text{Prognostic}_{\text{Patient } 2}(t) \\ \dots \\ \text{Prognostic}_{\text{Patient } n}(t) \end{cases}$$

$$\text{Prognostic}_{\text{drug}=2}(t) \begin{cases} \text{Prognostic}_{\text{Patient } n+1}(t) \\ \text{Prognostic}_{\text{Patient } n+2}(t) \\ \dots \\ \text{Prognostic}_{\text{Patient } n+m}(t) \end{cases}$$

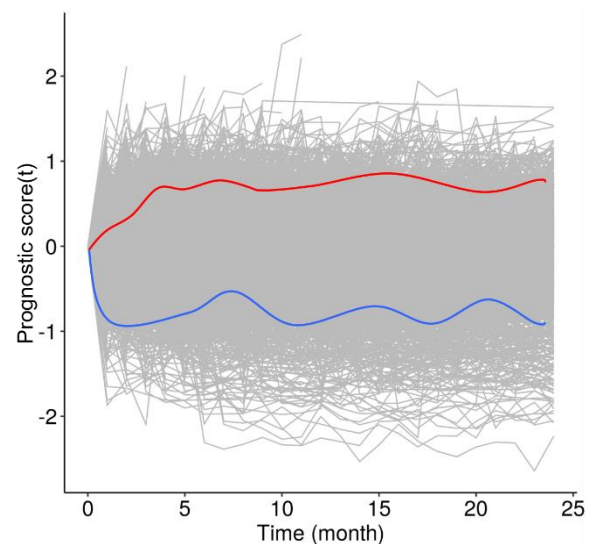


Figure 1.5 Illustration of the "average" prognostic score curves for two different cohorts.

An alternative to the aforementioned imaging-based techniques are prognostic scores, which are also associated with survival time. Specifically, the longitudinal evolution of prognostic scores could be an innovative tool to track the evolution of the cohort over time. Specifically, in this new method, the variation of the prognostic score value is computed between the baseline (symbolizing the risk before the start of treatment) and during the treatment (when the treatment influences the risk score). Hence,

the variation of the prognostic score value for each patient during treatment should characterize the influence of the drug on the risk of death (Loureiro, Kolben, et al. 2023). The variation of the prognostic scores at the cohort level (illustrated in Figure 1.5), i.e., for each tested medication, should represent the average improvement/deterioration of the cohort after the start of the medication. Hence, in a traditional clinical trial setting with an experimental and a control arm, the average improvement of the prognostic score should characterize the survival difference between the drugs (Loureiro, Kolben, et al. 2023).

Third research question:

Can the longitudinal values of prognostic scores characterize the efficacy results of a clinical trial?

1.6 Aim of the thesis

Prognostic scores are methods that combine a set of biomarkers into a risk value, which is correlated to the survival time of the patient. Although prognostic scores are models that are routinely used in survival analysis, there are additional uses of prognostic scores in clinical trials that remain unexplored. Therefore, in this publication-based thesis, I investigated multiple methods of using prognostic scores to accelerate drug development by analyzing the three research questions below.

First research question: Can more complex survival models build more performant prognostic scores?

Second research question: Can prognostic scores be used to match external controls?

Third research question: Can the longitudinal values of prognostic scores characterize the efficacy results of a clinical trial?

Firstly, although there have been advances in machine learning in the last few decades, the most used survival models are still simple regression models (Klatte, Rossi, and Stewart 2018). The simple regression models could limit the performance of the prognostic scores. For example, in simple regression models, nonlinearities or interactions between covariates need to be specified by the investigator (Loureiro, Becker, Bauer-Mehren, et al. 2021). In contrast, more complex models can infer these nonlinearities from the dataset. Hence, I performed a benchmarking analysis of survival analysis models of different complexities in a large multi-cancer dataset. I presented a preliminary version of this result as a presentation at the ICPE 2020 conference (Loureiro et al. 2020), alongside a full paper in the *Frontiers in Artificial Intelligence* journal (Loureiro, Becker, Bauer-Mehren, et al. 2021), of which I am the first author. The publication described the main characteristics of each survival model. Additionally, I complemented the literature by providing a systematic analysis that included a comprehensive array of models applied to a large dataset. Notably, I introduced two new survival

models based on the autoencoder and on the super learner. I made all my code available so the community could use these models.

Secondly, prognostic scores could be used to match external controls. With the increase of the available historical data, its use in clinical trials is becoming an increasingly important subject. Currently, the most used method to match patients are propensity scores. Still, a limited number of simulation studies have shown that prognostic scores can yield more balanced eControls. Hence, in the second part of this publication-based thesis, I contrast the performance of prognostic scores and propensity scores to create eControls for cancer clinical trials. I performed eControl analyses on 12 recent lung cancer clinical trials. The analysis showcases that it is possible to create eControls using RWD, and that prognostic scores are suitable tools to balance the cohorts. I published the results of this analysis in the *Clinical Pharmacology & Therapeutics* journal in 2023 (Loureiro, Roller, et al. 2023). I am the first author of this manuscript. Hence, the second part of this thesis expands the literature on both eControls and prognostic scores.

Thirdly, the state-of-the-art imaging-based surrogate endpoints have a lower correlation with OS for more recent medications, such as cancer immunotherapy (Frédéric Fiteni, Westeel, and Bonnetain 2017). Alternatively to imaging-based surrogate endpoints, prognostic scores also correlate with the endpoints of interest. Hence, in the final part of my publication-based thesis, I consider the use prognostic score values to estimate efficacy. I presented the main research idea in a poster at ICPE 2021 (Loureiro, Becker, Ahmidi, et al. 2021) and some preliminary results in the poster at ICPE 2022 (Loureiro, Becker, and Bauer-Mehren 2022). Next, I published analysis as a paper in *JCO Clinical Cancer Informatics* in 2023 (Loureiro, Kolben, et al. 2023). I am the first author of the manuscript. I presented the risk trend framework that combines prognostic scores with the Joint Modeling framework (section 2.3.2). The risk trend framework estimates the efficacy results of a clinical trial from interim analysis data. As a significant deviation from other methods in the literature, instead of relying on data from cancer imaging, the risk trend framework considers data from blood biomarkers that are cheap and obtained frequently during cancer treatment. In summary, the risk trend framework provides the clinical trial teams with an alternative tool to analyze efficacy early in clinical development. All the code necessary for the analysis is available to other researchers.

2 Background

In the second chapter, I define the main concepts and the statistical methods used in this thesis. Firstly, I introduce the basics of survival analysis alongside several of the models used to estimate the risk/hazard function (in section 2.1 and 2.2). Second, in chapter 2.3 I introduce a more recent and complex topic in survival analysis, the inclusion of time-dependent biomarkers in the models. Next, I present the main endpoints in cancer clinical trials (chapter 2.4), and propensity scores, a popular method in statistics used to account for possible measured confounding in eControls (chapter 2.5). Lastly, in chapter 2.6 I describe the RWD and clinical trial datasets used in this thesis.

2.1 Survival Analysis

Survival analysis is a field of statistics that studies the time until the occurrence of one or more events of interest (Kleinbaum and Klein 2012; Kalbfleisch and Prentice 2002). In survival analysis, the time-to-event of an individual is defined as a non-negative random variable T . Although T can be defined as a discrete, continuous or mixed random variable, depending on the analysis type (Kalbfleisch and Prentice 2002), in this work, I consider T to be a continuous random variable.

Survival analysis, despite its name, it is not restricted to death events. Other events like progression, response, and cure, among others can also be modelled with the same techniques. Recurrent events (i.e., that happen multiple times) can also be modelled.

The distribution of the time-to-event T can be characterized using multiple methods, of which, the most common are the survivor and hazard functions. The survivor function usually defined as $S(t)$, is defined as the probability that an individual will only suffer the event after a specific time t

$$S(t) = P(T > t). \quad (2.1)$$

Additionally, the hazard function $h(t)$ represents the instantaneous risk of the patient suffering the event at each time t

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}. \quad (2.2)$$

The simple expression relates the survivor and hazard functions

$$S(t) = \exp[-H(t)] = \exp\left[-\int_0^t h(u) du\right], \quad (2.3)$$

where $H(t)$ is the cumulative hazard function, and symbolizes the accumulated risk until time t (Cleves, Gould, and Marchenko 2016).

In practical terms, to determine the time to the event of interest, it is necessary to define which point is considered the baseline, or starting time ($t = 0$) (Clark et al. 2003). The definition of an unambiguous starting time in a study is imperative to avoid biases, such as immortal time bias (Suisa 2007; Weberpals et al. 2017). In oncology clinical trials, one type of prospective analysis, a common start date is the randomization date (Oba et al. 2013). Whereas, in retrospective analyses using historical datasets, it is common to define the baseline at the start of treatment.

In longitudinal studies, due to patient dropout, loss of follow-up or any other known or unknown reason, the time of event of a patient might not be observed (Turkson, Ayiah-Mensah, and Nimoh 2021). Hence, the time-to-event of these patients is partially missing since the event had still not occurred until the last captured follow-up date. This phenomenon is called as right censoring (Kalbfleisch and Prentice 2002; Turkson, Ayiah-Mensah, and Nimoh 2021; K.-M. Leung, Elashoff, and Afifi 1997). The censoring indicator $\delta(t)$ is a random variable that represents whether a patient's event was observed

$$\delta(t) = \begin{cases} 1, & \text{if time-to-event is known} \\ 0, & \text{otherwise} \end{cases} \quad (2.4)$$

Naturally, the cause of the censoring event can be associated to the time-to-event. The right censoring mechanism is independent of the time-to-event if for whichever subset of patients, the patients that were censored at time t are representative of the remaining patients (that were not censored) (Kleinbaum and Klein 2012; Turkson, Ayiah-Mensah, and Nimoh 2021). Conversely, a censoring mechanism is non-independent of the time-to-event if censoring is more likely when the risk of event of the patients is higher. Kalbfleisch and Prentice, present as an example of non-independent censoring a case where a patient would be withdrawn from the study for being in imminent danger of death (Kalbfleisch and Prentice 2002). In clinical trials, right censoring can occur for a variety of reasons. For example, patients may move away from the clinic where they are followed for reasons completely independent of the disease or the risk of the event (independent/non-informative censoring).

Additionally, patients may withdraw from a clinical trial due to adverse events and lose contact with the study team (dependent/informative censoring) (Wilson et al. 2021). The study can also end before all possible events; hence, some patients would be lost to follow-up (Singh and Mukhopadhyay 2011). The works by Kalbfleisch and Prentice or Kleinbaum and Klein contain more information on censoring, including left and interval censoring (Kalbfleisch and Prentice 2002; Kleinbaum and Klein 2012).

2.1.1 Estimation of the Survivor function - Kaplan Meier estimator

The Kaplan-Meier estimator is a non-parametric estimator of the survivor function (Kaplan and Meier 1958; Kalbfleisch and Prentice 2002, chap. 1). It assumes that the censoring mechanism is independent of the time of the event (Kaplan and Meier 1958; Overgaard and Hansen 2021; K.-M. Leung, Elashoff, and Afifi 1997). The Kaplan Meier estimator defines the estimate of the survivor function as

$$\hat{S}(t) = \prod_{j:T_j \leq t} \left(1 - \frac{n_j}{d_j}\right), \quad (2.5)$$

where n_j , and d_j represent, respectively, the number of individuals at risk at time t_j , and the number of individuals that had the event at time t_j . Therefore, the estimated survivor function is a stepwise function that starts at $\hat{S}(t = 0) = 1$ at the start of the study (no patients have suffered the event) and decreases at each time-point when an event is observed. With a large enough sample size, the estimate $\hat{S}(t)$ converges to the true survivor function of the studied population (Kaplan and Meier 1958; Kalbfleisch and Prentice 2002).

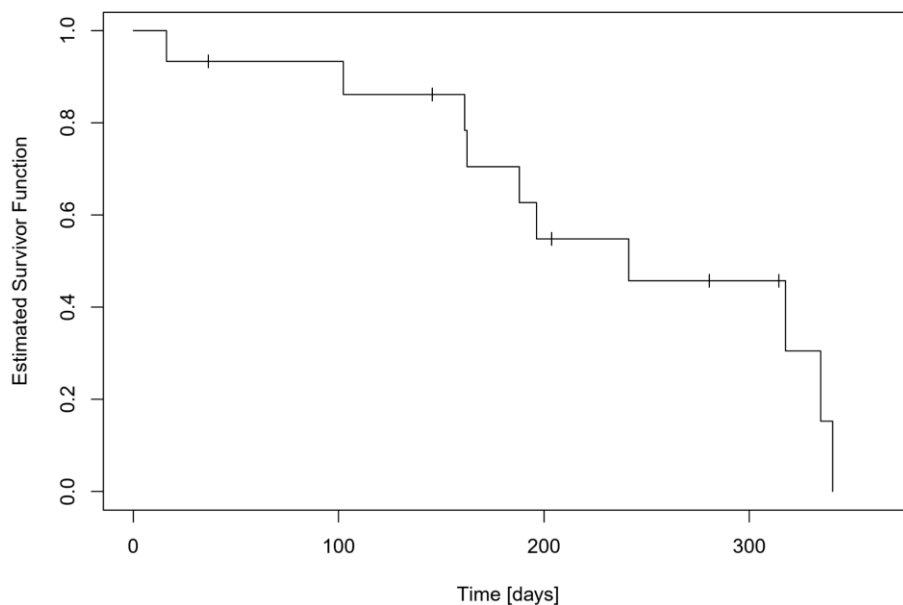


Figure 2.1 Illustration of a Survivor function estimated with the Kaplan Meier estimator.

Figure 2.1 displays an example survivor function. In this example, the dataset comprised fifteen patients who were followed for 365 days. As expected, at $t = 0$ no patients had experienced the event, hence $\hat{S}(t = 0) = 1$. The survivor function demonstrates the stepwise nature of the Kaplan-Meier estimator. Each “step” corresponds to the event of one or more patients. The “+” symbols represent right censoring events. In this example, at the end of the study $t = 365$, all patients had suffered the event or were censored, hence $\hat{S}(t = 365) = 0$.

2.1.2 Calculate difference between survivor functions

Analyzing the difference between two survivor functions is a common application of survival analysis. For example, in a clinical trial, the difference between the treatment and control survivor functions is analyzed as an outcome. The log-rank test is one of the methods used to define the difference between

the survivor functions. The log-rank test is a nonparametric hypothesis test that compares the estimated number of events in each group with the overall number of events (Peto and Peto 1972; Kalbfleisch and Prentice 2002, chap. 1.5). When comparing two survival curves, the null hypothesis of the log-rank test is that both survival functions are equal

$$H_0: S_1(t) = S_2(t) \quad \text{vs} \quad H_1: S_1(t) \neq S_2(t). \quad (2.6)$$

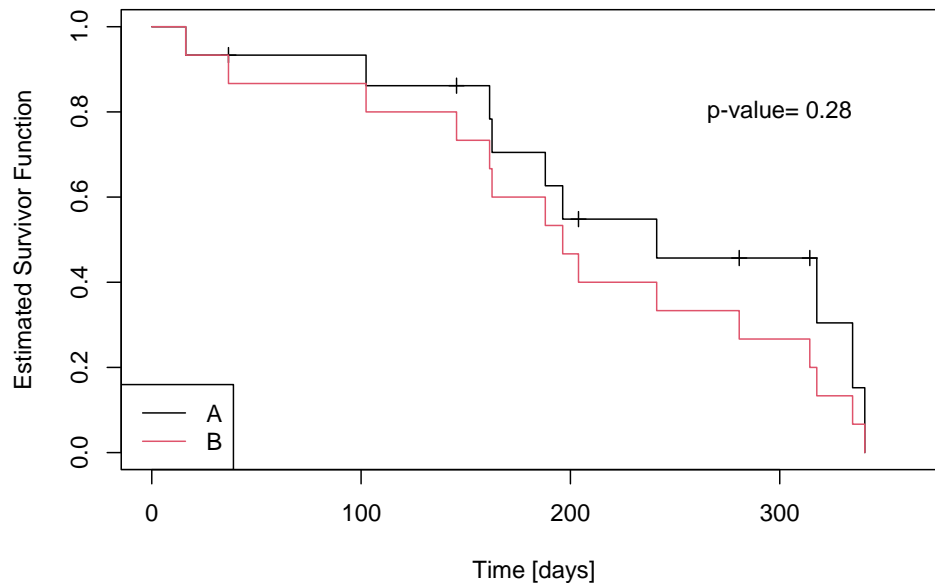


Figure 2.2 Two survivor curves and the corresponding log-rank test p-value.

Figure 2.2 compares two survival curves (A and B). Although the patients in curve B appear to have a lower survival time based on the figure, according to the log-rank test, the difference between the survival curves is not statistically significant.

2.1.3 Estimation of the hazard function

Several parametric distributions have been proposed to estimate the hazard function (C. Cox et al. 2007; C. Cox 2008). For example, the hazard function for the exponential distribution is a constant value

$$h_{\text{exponential}}(t) = \lambda > 0. \quad (2.7)$$

Whereas the survival function for the exponential distribution (following the relationship in equation 2.3) is the exponential function

$$S_{\text{exponential}}(t) = e^{-\lambda t}. \quad (2.8)$$

For the exponential distribution, since the hazard does not depend on the time t , the probability of the event at any interval is also independent of the time since the start of the analysis.

The assumption that the hazard is invariant over time might be too strict in many scenarios (Afify and Mohamed 2020). Hence, other more complex distributions, such as the Weibull distribution allow for a more flexible hazard function that can vary over time (Kalbfleisch and Prentice 2002). Specifically, the hazard function following the Weibull distribution

$$h_{\text{Weibull}}(t) = \lambda \zeta (\lambda t)^{\zeta-1}, \quad (2.9)$$

is specified by the λ , and $\zeta > 0$ parameters. The extra parameter ζ controls the slope of the hazard function. For $\zeta > 0$ the hazard function is monotonically increasing, while for $\zeta < 0$ the hazard function is monotonically decreasing. Additionally, for $\zeta = 0$, the Weibull distribution is reduced to the constant exponential distribution hazard function. The survivor function following the Weibull distribution is

$$S_{\text{Weibull}}(t) = \exp[-(\lambda t)^\zeta]. \quad (2.10)$$

Although the Weibull distribution allows the hazard function to vary over time, the variation must be monotonic. There are other distributions, such as the log-normal distribution, gamma, or F distribution that allow for the hazard function to be non-monotonic (Kalbfleisch and Prentice 2002).

The previous hazard functions assumed a homogeneous patient population, this is, that there were no patient characteristics that influenced the event time T . Usually, though, there are covariates, such as disease characteristics, patient biomarkers or environmental variables that increase the risk of the event. The previously introduced models can be extended to incorporate these covariates into the hazard functions. Considering a vector of covariates available for each of the patients $X_i = (x_{i,1}, x_{i,2}, \dots)$, the extended hazard functions to account for these effects for the exponential distribution is

$$h_{\text{exponential}}(t|X_i) = \lambda c(\gamma^\top X_i), \quad (2.11)$$

while for the Weibull distribution the hazard function assumes the form

$$h_{\text{Weibull}}(t|X_i) = \zeta (\lambda t)^{\zeta-1} c(\gamma^\top X_i). \quad (2.12)$$

In the previous expressions, γ represents regression coefficients, and $c(\cdot)$ represents a function dependent on $\gamma^\top X_i$, such as a simple linear function or an exponential function. In the aforementioned models, though, the covariates have the same effect regardless of the time

2.2 Risk / prognostic score models

Prognostic scores are models that estimate an event's risk at the baseline (i.e., the prognosis) given a set of covariates X . Although the models from section 2.1.3 can be used as prognostic scores, their parametric nature makes it necessary to assume the distribution of the survival times. The Cox proportional hazards model is frequently used to avoid specifying the distribution.

2.2.1 Cox model

The Cox proportional hazards model (or in short, Cox model) (D. R. Cox 1972) is a semi-parametric survival model that specifies the hazard function as

$$h_i(t|X_i) = h_0(t) \exp(\gamma^\top X_i), \quad (2.13)$$

where $h_0(t)$ is the baseline hazard function. The Cox model describes, thus, the hazard function as the product of two terms, a baseline hazard function $h_0(t)$ that is time-dependent but unspecified and a second term, $\exp(\gamma^\top X_i)$, that is constant over time and depends on the covariates X_i . The Cox model assumes proportional hazards, i.e., the comparison (hazard ratio) between two patients i and j

$$\frac{h_i(t|X_i)}{h_j(t|X_j)} = \frac{h_0(t) \exp(\gamma^\top X_i)}{h_0(t) \exp(\gamma^\top X_j)} = \exp[\gamma^\top (X_i - X_j)], \quad (2.14)$$

is constant over time.

The Cox model has other properties, specifically, when all explainable variables are 0, the hazard function $h(t; X)$ collapses to the baseline hazard function $h_0(t)$. This property is particularly useful when considering categorical variables are split into multiple sub-variables with the one-hot encoding method.

The semiparametric nature of the Cox model, where the baseline hazard function is not specified, allows for the estimation of γ even in cases where an appropriate distribution for T is not known (Kleinbaum and Klein 2012). The estimation of the coefficients γ is performed with the partial likelihood estimator. For each patient, the partial likelihood has the form

$$\mathcal{L}_i(\gamma) = \frac{h(t_i|X_i)}{\sum_{j:T_j \geq t_i} h(t_i|X_j)} = \frac{h_0(t) \exp(\gamma^\top X_i)}{\sum_{j:T_j \geq t_i} h_0(t) \exp(\gamma^\top X_j)} = \frac{\exp(\gamma^\top X_i)}{\sum_{j:T_j \geq t_i} \exp(\gamma^\top X_j)}. \quad (2.15)$$

The partial likelihood iterates over all the event times. The rationale is that patients whose time-to-event is lower should have higher hazard, while patients that suffer the event later in the study should have lower hazard values. The complete partial likelihood corresponds to the product of the partial likelihoods for all patients

$$p\mathcal{L}(\gamma) = \prod_{i=1}^N \delta_i \cdot \mathcal{L}_i(\gamma). \quad (2.16)$$

For simplicity, a log transform is applied to the log-likelihood

$$p\ell(\gamma) = \sum_{i=1}^N \delta_i \left(\gamma^\top X_i - \log \sum_{j:T_j \geq t_i} \exp(\gamma^\top X_j) \right). \quad (2.17)$$

Additionally, by determining the partial derivative of the log partial likelihood, we can estimate the best possible vector γ

$$p\ell'(\gamma) = \sum_{i=1}^N \delta_i \left(X_i - \sum_{j:t_j \geq t_i} \frac{\exp(\gamma^\top X_j) X_j}{\exp(\gamma^\top X_j)} \right). \quad (2.18)$$

The Cox model can be extended to allow the baseline hazard function to take different shapes for different subsets of patients. The stratified version of the Cox model, can be rewritten as

$$h_j(t|X) = h_{0j}(t) \exp(\gamma^\top X), j = 1, \dots, r, \quad (2.19)$$

where r is the number of different strata. The stratified version can be useful if any covariate does not appear to follow the proportional hazards assumption. In those cases, the patient population can be divided into r strata, denoting the different values of the variable. Still, in all strata, the value of the γ covariates are maintained.

The Cox model is a highly adaptable and extensively used model in survival analysis. Though, its simplicity is criticized, notably that the covariates γ are not directly associated with the time-to-event T . Additionally, as a linear model, the Cox model cannot model nonlinear effects, such as higher order terms or interaction terms, unless specified. Some of the models introduced below attempt to tackle this limitation.

2.2.2 Regularized Cox model

The regularized Cox model (Tibshirani 1997; Simon et al. 2011) is an extension of the Cox model that incorporates regularization. Regularization is a process used in machine learning to prevent the model coefficients from increasing exponentially. Regularization can be performed by including in the likelihood/loss function a term based on the value of the model weights γ . Additionally, regularization is a useful technique in ill-posed machine learning problems, e.g., when there are more covariates than observations, and can prevent overfitting. In lay terms, regularization attempts to create a simpler model, retaining a good performance (Zou et al. 2019).

The regularized Cox model keeps the same expression (equation 2.15) but adds a second term to the log partial likelihood function

$$\ell^{\text{regularized}}(\gamma) = \sum_{i:\delta_i=1} \left(\gamma^\top X_i - \log \sum_{j:t_j \geq t_i} \exp(\gamma^\top X_j) \right) + \xi \left(\alpha \|\gamma\|_1 + \frac{1}{2} (1 - \alpha) \gamma^\top \gamma \right), \quad (2.20)$$

that is responsible for the regularization. The α term controls the type of regularization performed, while ξ controls its strength. Ridge regression ($\alpha = 0$), also referred to as L_2 regression tends shrink the coefficients γ by penalizing large coefficient values ($\gamma^\top \gamma$ term) (Hastie, Tibshirani, and Friedman 2009), although it does not completely set them to 0 (completely removing the contribution of certain covariates),

$$\ell^{\text{ridge}}(\gamma) = \sum_{i:\delta_i=1} \left(\gamma^\top X_i - \log \sum_{j:t_j \geq t_i} \exp(\gamma^\top X_j) \right) + \xi \left(\frac{1}{2} \gamma^\top \gamma \right). \quad (2.21)$$

The Lasso ($\alpha = 1$) performs L_1 regularization. L_1 regularization also imposes a penalty on the coefficients, but it tends to shrink the coefficients to 0, effectively performing variable selection (Hastie, Tibshirani, and Friedman 2009; Loureiro, Becker, Bauer-Mehren, et al. 2021)

$$\ell^{\text{lasso}}(\gamma) = \sum_{i:\delta_i=1} \left(\gamma^\top X_i - \log \sum_{j:t_j \geq t_i} \exp(\gamma^\top X_j) \right) + \xi \|\gamma\|_1. \quad (2.22)$$

Values of α between 0 and 1 correspond to the elastic net, which combines of both methods (lasso and ridge regression).

2.2.3 Tree-based survival models

Tree-based methods are routinely used in medicine due to their simplicity and interpretability (Banerjee et al. 2019; Podgorelec et al. 2002). Tree-based models partition the covariate space sequentially into subsets, which are assigned a value. A simple tree contains all observations at its root node, then the observations are sequentially split in a binary way according to each of their covariates X . At each level of the tree the fitting process identifies the covariate x that leads to the optimal split. The tree is grown by adding more splits until a stopping criterion is met. The fitting of the tree usually results in a long and overfit tree. Therefore, after the fitting, the tree usually undergoes a pruning process to determine a simpler tree with a good predictive performance (Imad Bou-Hamad, Denis Larocque, and Hatem Ben-Ameur 2011).

Depending on the type of the outcome of interest, multiple algorithms to determine the optimal splits (Hastie, Tibshirani, and Friedman 2009). In classification, for example, one popular method is to minimize the information entropy. For survival outcomes, a common approach is to split the tree by the difference between the survival functions of each resulting node with the log-rank statistic (chapter 2.1.2) (Segal 1988; Ishwaran et al. 2008). Although multiple other splitting algorithms have been proposed (Imad Bou-Hamad, Denis Larocque, and Hatem Ben-Ameur 2011; Ishwaran et al. 2008).

The terminal nodes of the trees contain the estimate that the tree attributes to that set of covariates X . Specifically, for trees that estimate survival times, the output of a terminal node h is

$$H_{\text{Tree}}(t|X) = \hat{H}_{\text{Node}}(t) = \sum_{j|t_j \leq t} \frac{d_{j,\text{Node}}}{n_{j,\text{Node}}}, \quad \text{if } X \in \text{Node}, \quad (2.23)$$

which is the Nelson-Aalen estimator (Nelson 1969) of the cumulative hazard function. All samples that fall into a terminal node in survival trees have the same cumulative hazard function (Ishwaran et al. 2008).

Random forests are an ensemble learning method that relies on many tree-based estimators to inform the response. Each tree is built on a subset of the database in a random forest. Specifically, a new dataset is created with bootstrap for each of the B trees of a random forest. Additionally, each of the bootstrapped datasets contain only a subset of the covariates of the original dataset to reduce the overall variance of the random forest (Hastie, Tibshirani, and Friedman 2009). With each bootstrapped dataset, a tree is built following the algorithm described above but without pruning. All of the B trees contribute to the final output of the random forest (Hastie, Tibshirani, and Friedman 2009).

Random survival forests (RSF) (Ishwaran et al. 2008) are an extension of random forests that support the analysis of right-censored data. The development of random survival forests was fueled by the need for more complex models capable of dealing with nonlinear effects in the covariates (Ishwaran et al. 2008). RSF are non-parametric models that do not follow the same Cox model assumptions, such as the proportional hazards assumption (Ishwaran et al. 2008). Additionally, RSFs are fundamentally different from the aforementioned models as they estimate the cumulative hazard function

$$H_{\text{RSF}}(t|X) = \frac{1}{B} \sum_{b=1}^B H_b(t|X), \quad (2.24)$$

instead of the hazard function. Similarly, the survival function estimated by the RSF is defined by

$$S_{\text{RSF}}(t|X) = \frac{1}{B} \sum_{b=1}^B S_b(t|X). \quad (2.25)$$

2.2.4 Gradient boosting models applied to risk / prognostic score models

Gradient boosting (GB) is another method based on ensemble learning (Friedman 2001). The predicted values result from the contribution of several weak learners (Y. Chen et al. 2013). Weak learners are added iteratively to the model to minimize the cost function. GB was initially introduced as a regression model, but has since been extended to survival analysis (Ridgeway 1999). GB can be used to estimate the hazard function, which is defined by

$$h_{\text{GB},i}(t|X_i) = F_{\text{GB}}(X_i) = \sum_{b=1}^B \rho_b f_{\text{GB},b}(X_i), \quad (2.26)$$

where $f_{\text{GB},b}(X)$ corresponds to the weak learners introduced at each iteration b , and ρ_b to model weights. The gradient boosting algorithm presented by Ridgeway (Ridgeway 1999), uses the cox partial hazard to fit the model

$$p\ell_{\text{GB}}(\gamma) = - \sum_{i=1}^N \delta_i \left(F_{\text{GB}}(X_i) - \log \sum_{j:T_j \geq T_i} \exp(F_{\text{GB}}(X_j)) \right). \quad (2.27)$$

2.2.5 Deep learning applied to risk / prognostic score models

Deep learning is a popular branch of machine learning that uses large neural networks with different architectures to model highly complex data relationships. Deep learning has proven especially useful in image (L. Jiao and J. Zhao 2019; Qian et al. 2013; Weiss et al. 2022) and natural language processing (D. W. Otter, J. R. Medina, and J. K. Kalita 2021; Wu et al. 2020; Maslej-Krešňáková et al. 2020). In both imaging and natural language processing, deep learning has outperformed most of the classical machine learning algorithms (Gehrmann et al. 2018; Magnini, Lavelli, and Magnolini 2020).

The simplest neural networks are built of perceptrons, base units that sum the inputs and apply a nonlinear transformation to them. A perceptron with an input X generates the output

$$f_{\text{Perceptron}}(X) = c(\gamma_0 + \gamma^T X) \quad (2.28)$$

where $c(\cdot)$ corresponds to a user specified function (e.g., a sigmoid), γ_0 and γ correspond to the coefficients that multiply the bias term and the input vector, respectively.

Fully connected feed-forward neural networks (NN) consist of multiple layers of multilayer perceptrons (Hastie, Tibshirani, and Friedman 2009). The first layer (input layer) receives the covariate vector X , followed by at least one hidden layer and an output layer that generates the neural network's output $f_{\text{NN}}(X)$. Figure 2.3 represents a simple neural network.

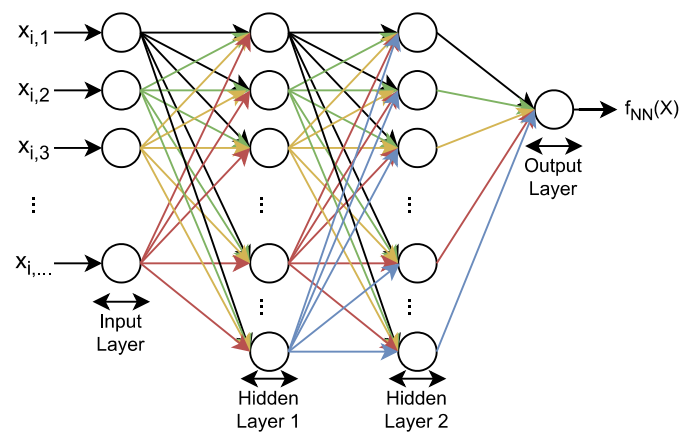


Figure 2.3 Diagram of a simple neural network.

Given the nonlinear nature of the perceptrons, as more layers are added, the neural network can reproduce more complex relationships between the covariates X and output Y . Neural networks, if given enough data and processing capacity can approximate any given function (Funahashi 1989; Kon and Plaskota 2000). The flexibility of neural networks to approximate any function makes them ideal candidates to model the complex relationships of biomedical data.

Neural networks (Faraggi and Simon 1995) have been applied to survival problems since they were introduced by Faraggi and Simon. Faraggi and Simon proposed to replace the linear term in the Cox model hazard function (Equation 2.13) with the output of a neural network $g(\cdot)$,

$$h_{\text{NN}}(t|X_i) = h_0(t) \exp[g(X_i, \gamma)]. \quad (2.29)$$

Replacing the linear term with the output of a neural network should allow for more complex relationships between the covariates X and outcome (T, δ) to be modelled. For the neural network hazard function, the partial likelihood is defined as

$$\mathcal{L}_{i,\text{NN}}(\gamma) = \frac{h_{\text{NN}}(t_i|X_i)}{\sum_{j:t_j \geq t_i} h_{\text{NN}}(t_i|X_j)} = \frac{\exp[g(X_i, \gamma)]}{\sum_{j:t_j \geq t_i} \exp[g(X_j, \gamma)]}, \quad (2.30)$$

which is very similar to the original Cox partial likelihood (Equation 2.15). Faraggi and Simon considered the logistic function

$$f(X) = \frac{1}{1 + \exp(-X)}, \quad (2.31)$$

as the nonlinear function used in the perceptrons and determined the first and second derivatives of the partial likelihood (Faraggi and Simon 1995). With both derivatives, they used the Newton-Raphson algorithm to estimate the coefficients of the network.

Faraggi and Simon considered a neural network with only one hidden layer, which could be a limiting network architecture to model complex interactions (Faraggi and Simon 1995). Katzman et al. expanded upon their original work (Faraggi and Simon 1995) by generalizing it to any given feed-forward neural network (Katzman et al. 2018). The DeepSurv (DS) method, introduced by Katzman et al. is implemented in a contemporary deep learning framework (Bergstra et al. 2010; Bastien et al. 2012). Additionally, the DS model supports other more recent deep learning techniques, such as the scaled exponential linear units (commonly referred to as SELU) (Klambauer et al. 2017), adaptive moment estimation (or ADAM) (Kingma and Ba 2014), dropout layers (Srivastava et al. 2014) and regularization of the weights of the network (Lewkowycz and Gur-Ari 2020).

The DS is also fit using the log partial hazard function

$$\ell_{\text{DS}}(\gamma) = \sum_{i:\delta_i=1} \left(g(X_i, \gamma) - \log \sum_{j:t_j \geq t_i} \exp(g(X_j, \gamma)) \right) + \xi \|\gamma\|_2^2. \quad (2.32)$$

Although it can incorporate an L_2 regularization term and uses gradient descent to estimate the parameters of the network.

2.2.6 Autoencoder applied to survival analysis

An autoencoder is a type of feedforward neural network whose objective is to reproduce the input at the output of the network (Goodfellow, Bengio, and Courville 2016; Weberpals et al. 2021). Structurally, the hidden layers of the autoencoder are composed of two parts, an encoder, and a decoder. The encoder collapses the input vector X into a vector $Z_{\text{Bottleneck}}$ in lower dimension. The decoder attempts to do the opposite, to reconstruct the vector X from its lower dimensional representation

$Z_{\text{Bottleneck}}$. The output of the middle layer $Z_{\text{Bottleneck}}$ that connects the encoder and decoder is named the bottleneck layer.

One of the uses for the autoencoder is to perform non-linear dimensionality reduction. I introduced a new survival analysis model that incorporated the autoencoder predictions in the Cox model (Loureiro, Becker, Bauer-Mehren, et al. 2021). Specifically, the covariate vector is processed by the autoencoder and the bottleneck values $Z_{\text{Bottleneck}}$ are then used in a Cox model with hazard function

$$h_{\text{AE},i}(t|Z_{\text{Bottleneck},i}) = h_0(t) \exp(\gamma^\top Z_{\text{Bottleneck},i}). \quad (2.33)$$

2.2.7 Super Learner applied to survival analysis

The Super Learner (SL) is a machine learning method that diverges from the aforementioned methods. Specifically, instead of learning the relationship between the covariates X and the outcome of interest Y , the SL combines the outputs of multiple individual models to compose an output that is based on all the previous learners (van der Laan, Polley, and Hubbard 2007). By combining multiple different learner algorithms, the SL attempts to combine the strengths of multiple models.

The SL technique has been used in a variety of biomedical studies. Ehwerhemuepha et al. showed that the Super Learner could obtain a better prediction performance of severe Covid-19 than the individual models (Ehwerhemuepha et al. 2021). Zhu et al. used the super learner alongside other models to estimate the treatment effect of different medications (Zhu and Gallego 2020). In a different field, propensity score prediction both Wyss et al. and Ju et al. used the SL to estimate the propensity scores (Wyss et al. 2018; Ju et al. 2019). Propensity scores are described in chapter 2.5.

The SL relies heavily on cross-validation (van der Laan, Polley, and Hubbard 2007). Cross-validation is a technique used in machine learning that builds upon the idea of splitting the learning data $(X_{\text{learning}}, Y_{\text{learning}})$ into a training $(X_{\text{train}}, Y_{\text{train}})$ and a validation set $(X_{\text{validation}}, Y_{\text{validation}})$ (Arlot and Celisse 2010). In cross-validation, the learning dataset $(X_{\text{learning}}, Y_{\text{learning}})$ is divided into two or more training and validation sets (Arlot and Celisse 2010; Browne 2000). In an analysis that uses cross-validation with V different splits, the selected model is trained in each of the training sets $(X_{\text{train},v}, Y_{\text{train},v})$, $v = 1, \dots, V$. Afterwards, the performance of each of the models is evaluated with the corresponding cross-validation validation dataset $(X_{\text{validation},v}, Y_{\text{validation},v})$, $v = 1, \dots, V$. Cross-validation in prediction problems is used to assess the predictive effectiveness of the model (Browne 2000). Cross-validation enables a more efficient use of the dataset, which may lead to better average performance of the models, and additionally, to reduce the risk of obtaining a model with significantly worse performance (Schaffer 1993).

In the SL method, the dataset is split into V different splits. The selected models that compose the SL are fit on each of the V training sets $(X_{\text{train},v}, Y_{\text{train},v})$, $v = 1, \dots, V$. Next, the outputs of each of the V models on the validation dataset are collected $\hat{Y}_{\text{model},v}$. The full output \hat{Y}_{models} matrix is composed of

one column for each model and a row for each observation in the learning dataset X_{learning} . The contribution of each of the models γ_{SL} to the SL output is then calculated by determining $E(Y, \hat{Y}_{\text{models}}) = c(\hat{Y}_{\text{models}} | \gamma_{\text{SL}})$. To calculate the output of the model for new data, the super learner framework fits models that use the totality of the learning dataset. Then it balances the output of the models using the obtained γ_{SL} .

To estimate the SL parameters γ_{SL} , van der Laan et al. considered using least squares to minimize the prediction error (van der Laan, Polley, and Hubbard 2007)

$$\text{Cost}_{\text{CV}}(\gamma_{\text{SL}}) = \sum_{i=1}^N \left(Y_i - c(\hat{Y}_{\text{models},i} | \gamma_{\text{SL}}) \right)^2, \quad (2.34)$$

while LeDell et al. suggested optimizing the area under the receiver operating characteristic (LeDell, van der Laan, and Peterson 2016) (AUC-ROC) (Fawcett 2006). The structure of $c(\hat{Y}_{\text{models}} | \gamma_{\text{SL}})$ can be defined by the user, a simple example is $c(\hat{Y}_{\text{models}} | \gamma_{\text{SL}}) = \gamma_{\text{SL}} \cdot \hat{Y}_{\text{models}}$ when the response variable Y is numeric.

Although the SL was introduced for regression and classification problems, it has been extended to survival analysis (Polley et al. 2011; Golmakani and Polley 2020). The first extension, suggested by Polley et al., was to convert the right-censored dataset (X, T, δ) into a counting process dataset that contained the number of events of interest, and the censoring events for each time-point t (Polley et al. 2011). The counting process approach differs from the survival analysis algorithms introduced in this section. Hence, a super learner following this approach could not use the aforementioned algorithms. Conversely, Golmakani et al. proposed two additional algorithms to merge the model estimates (Golmakani and Polley 2020). These algorithms are based on the Cox partial likelihood (Equation 2.15). The algorithms assume that all models estimate the same function, the hazard function $h(t)$. Hence, models such as the random survival forests introduced in chapter 2.2.3 could not be incorporated into the Super Learner, as they estimate the cumulative hazard function $H(t)$.

Instead of using the aforementioned approaches, Loureiro et al. proposed an extension of the approach by LeDell et al. (LeDell, van der Laan, and Peterson 2016), where the Concordance Index (in short, C-index, described below) (Frank E. Harrell Jr et al. 1982) is used (Loureiro, Becker, Bauer-Mehren, et al. 2021). Specifically, the approach by Loureiro et al. considers that $c(\hat{Y}_{\text{models}} | \gamma_{\text{SL}}) = \gamma_{\text{SL}}^T \hat{Y}_{\text{models}}$, and estimates the SL coefficients with

$$\hat{\gamma}_{\text{SL}} = \underset{\gamma_{\text{SL}}}{\text{argmax}} \text{C-index}(\gamma_{\text{SL}}^T \hat{Y}_{\text{models}}). \quad (2.35)$$

Since the C-index is a non-linear function, the maximization is performed using the L-BFGS-B algorithm (Byrd et al. 1995).

2.2.8 Goodness of fit of survival models

The Harrell C-index (or C-statistic) is a goodness of fit measure for prognostic models (Frank E. Harrell Jr et al. 1982). It is a generalization of the area under the receiver operating characteristic (AUC-ROC). Specifically, the C-index is a rank correlation measure between the risk values of the patients and their time-to-event (Frank E. Harrell Jr et al. 1982). The C-index is defined between -1 and 1. C-index values of 0 and 1 represent, respectively, no correlation, and perfect correlation between the risk values and time-to-event.

The Harrell C-index depends on the study specific censoring distribution (Uno et al. 2011). Another measure is the Uno C-index (Uno et al. 2011) that does not suffer from this limitation, and retains the same interpretation as the Harrell C-index, i.e., and is defined between -1 and 1.

2.3 Survival analysis with time-varying covariates

Most of the aforementioned models (e.g., the Cox model) consider that the covariates are constant over time. Hence, only one measurement of these variables is considered in the whole model, e.g., the value before the start of the longitudinal study. Additionally, the variable will have the same effect on the hazard $h_i(t)$ at any time-point t . Although some variables, such as age at baseline, ethnicity, or sex are constant during the study, other covariates, such as blood biomarkers, will change over time. In studies where the effect of the time-varying variable on the hazard is of interest, the aforementioned models will be inappropriate. In this chapter, I briefly introduce the extended Cox model, a simple extension of the Cox model, alongside the joint models for survival and longitudinal data (in short, JM), representing a more sophisticated approach. Both models incorporate repeated measurements into the hazard function.

2.3.1 Extended Cox model

The Cox model, introduced in chapter 2.2.1, was extended by Andersen and Gill to incorporate longitudinal covariates $Y_i(t)$ (Andersen and Gill 1982). The extended Cox model uses counting process notation to generalize the Cox model. The new model depends on two new counting processes $N_i(t)$ and $R_i(t)$. $N_i(t)$ represents the number of events that have happened to patient i at time t , while $R_i(t)$ indicates whether the subject i is at risk $R_i(t) = 1$ at time t , or otherwise $R_i(t) = 0$ (Rizopoulos 2012). The hazard function of the extended Cox model is

$$h_i(t|X_i(t)) = h_0(t)R_i(t) \exp[\gamma^T Y_i(t)]. \quad (2.36)$$

One important consideration of the extended Cox model is that it does not assume that the hazard ratio is constant over time like the Cox model (Equation 2.14).

Similarly, to the Cox model, the parameters γ of the extended Cox model are also estimated with the partial log-likelihood function

$$\ell(\gamma) = \sum_{i=1}^N \int_0^{\infty} \left\{ R_i(t) \exp\{\gamma^\top Y_i(t)\} - \log \left[\sum_{j:T_j \geq T_i} R_j(t) \exp\{\gamma^\top Y_j(t)\} \right] \right\} dN_i(t), \quad (2.37)$$

which uses the counting process integral notation (Rizopoulos 2012).

Although the extended Cox model can incorporate time-varying covariates, it takes several assumptions about these variables. Specifically, the extended Cox model assumes that the values of the time-dependent variables $Y_i(t)$ are known at every time-point t during follow-up. Hence, for variables that are only measured at each visit (e.g., blood biomarkers), the extended Cox model assumes that $Y_i(t)$ can be represented by a stair function with jumps at each measurement (Figure 2.4). After the measurement, the value of the biomarker $Y_i(t)$ stays constant until the next measurement. The extended Cox model performs a “last observation carried forward” approach to deal with unknown values.

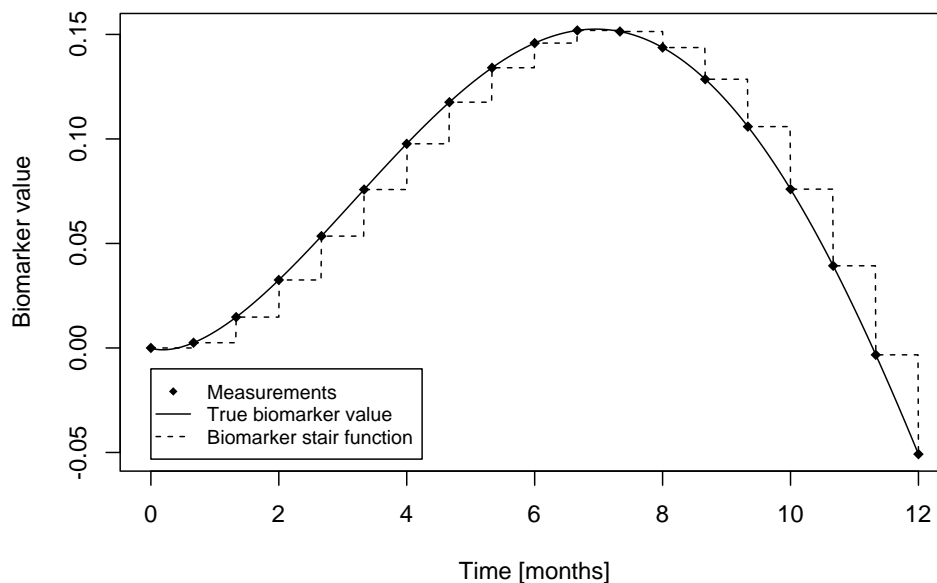


Figure 2.4 Example of the biomarker stair function assumed by the Extended Cox model. The biomarker measurements were obtained every 20 days for 12 months.

Additionally, the extended Cox model can only model exogenous variables. Exogenous variables are, briefly, time-varying covariates associated with the rate of failures over time. Still, the evolution of these exogenous variables after any t does not depend on the occurrence of failures up to the current time t (Rizopoulos 2012; Kalbfleisch and Prentice 2002). Some examples of exogenous variables are the pollution levels at a certain location, or seasonal patterns, such as the sunlight hours or amount of rainfall. These variables could be related with the event of interest, though; the occurrence of the event (e.g., death of a patient) will not influence future values of these variables.

Other biomarkers, such as those collected from the patient are considered endogenous variables. Blood biomarkers cannot be measurable after the event (e.g., death) happened, hence, it does not fit the definition of an exogenous variable. Another aspect of endogenous variables is that they are usually measured with error. The extended Cox model does not model the possible measurement error; hence, it might produce biased results when endogenous variables are used in the model.

2.3.2 Joint models for longitudinal and time-to-event data

The assumptions and limitations of the extended Cox model led to the creation of the joint models for longitudinal and time-to-event data (or, in short JM). JM is composed of two separate sub-models, a survival, and a longitudinal model, that are joint. The survival sub-model is based on the Cox model, and the longitudinal sub-model is based on the linear mixed-effects model.

Survival sub-model

The hazard function of the survival sub-model is similar to the extended cox model

$$h_{JM,i}(t|y_i(t), X_i) = h_0(t) \exp[\gamma^\top X_i + \alpha y_i(t)], \quad (2.38)$$

as it also incorporates a time-varying function $y_i(t)$, although it no longer depends on the counting process notation of the extended Cox model. Another difference of the survival sub-model of JM when compared to the Cox and extended Cox models is that the baseline hazard function $h_0(\cdot)$ must be specified. In the JM framework, not specifying the baseline hazard function can lead to underestimating of the standard errors of the model parameter estimates (Rizopoulos 2012, chap. 4; Hsieh, Tseng, and Wang 2006). The baseline hazard function chosen can be based on a parametric distribution, e.g., follow the exponential or Weibull distributions, described in chapter 2.1.3. Instead, the baseline hazard function can follow other more flexible structures. For example, it can be defined by a series of step-functions

$$h_0(t) = \sum_{j=1}^J A_j I(v_{j-1} < t \leq v_j), \quad (2.39)$$

where the time scale is split into J segments, whose time-value is represented by v_j . A_j denotes the segment value between times v_{j-1} and v_j . Additionally, the baseline hazard function can also be defined by more complex functions, such as a sum of linear splines or B-splines

$$\log h_0(t) = \kappa_0 + \sum_{j=1}^J \kappa_j B_j(t), \quad (2.40)$$

where κ represent the spline coefficients for each of the J splines. For both models, as J is increased, the flexibility of the baseline hazard function increases, but the extra flexibility might also lead to overfitting or issues in the regression of the model (Rizopoulos 2012).

Longitudinal sub-model

The longitudinal sub-model is based on the linear mixed effects method. Linear mixed effects (in short, LME) are an extension of linear models focused on the analysis of grouped data (Pinheiro and Bates 2000, chap. 1). LME models combine fixed effects, which are parameters common to the whole study population, alongside random effects that are patient/group specific and make the model more flexible than simple linear models. The LME model is formulated as

$$Y_i(t) = Z_i(t)\beta_{\text{fixed}} + W_i(t)b_i + \varepsilon_i(t), \quad (2.41)$$

where $Z_i(t)$ and $W_i(t)$ are design matrices of the fixed and random effects, respectively. β_{fixed} are the fixed-effects coefficients, and b_i are the random-effects regression coefficients. The random effects $b_i \sim \mathcal{N}(0, D)$ are normally distributed, with mean zero and a covariance matrix D , and the measurements are assumed to have been collected with error $\varepsilon_i \sim \mathcal{N}(0, \sigma^2 I)$ that is normally distributed with variance σ^2 .

Full model

The full joint model connects the survival and longitudinal sub-models. Specifically, the full JM is defined by

$$\begin{cases} h_i(t|y_i(t), X_i) = h_0(t) \exp[\gamma^\top X_i + \alpha y_i(t)] \\ y_i(t) = z_i(t)\beta_{\text{fixed}} + w_i(t)b_i + \varepsilon_i(t) \end{cases} \quad (2.42)$$

In the JM implementation I adopted (Rizopoulos 2012; 2010), only one longitudinal variable was modelled, hence the use of lowercase $x_i(t)$, $z_i(t)$, and $y_i(t)$.

The JM coefficients can be estimated in a two-step approach or a combined, joint likelihood. The two-step approach computes the linear mixed effects model first. Then, the longitudinal sub-model and its coefficients are used in the survival model estimation. Although this method is computationally simpler, it produces biased results. Conversely, the joint likelihood approach estimates the coefficients of both sub-models at once. The joint likelihood approach is an intricate computational problem prone to convergence issues. Therefore, the JM package (Rizopoulos 2010) uses a complex framework comprising the Markov chain Monte Carlo and regular quasi-Newton to fit the JM.

2.4 Endpoints in oncology clinical trials

The primary goal of clinical trials is to demonstrate that a new medication is superior (or at least non-inferior) in effectiveness and/or safety versus the current standard of care (Driscoll and Rixe 2009). In a classical setting, a new drug is tested in clinical trials. The first clinical trial, a phase I trial, focuses on safety and determining an appropriate medication dose. Next, in a phase II study, the biological effect of the drug is determined. In oncology, the biological effect is usually determined by the ability of the drug to treat the tumor (e.g., lead to tumor shrinkage). Lastly, in phase III clinical trials, the drug

is compared to the standard of care. In the phase III trial, the drug is evaluated according to different endpoints that assess its performance, the drug should prove to be at least as potent as the standard of care. All clinical trials phases assess the safety of the drug. As more patients are included in the clinical trials, more rare events might be detected (Mahipal and Nguyen 2014; MD Anderson Cancer Center, n.d.).

In oncology clinical trials, several efficacy endpoints are considered (F. Fiteni et al. 2014). The endpoints analyze different aspects of the disease, such as the death of the patients, the development of the tumor, or the length of the treatment (U.S. Food and Drug Administration 2018). These endpoints usually result of time-to-event data with right censoring. Hence, methods from survival analysis are used to model them. Usually, the survivor functions of each trial are calculated, then the main characteristics of the distribution can be compared with the standard of care, or statistical tests such as the log-rank are used. Additionally, the influence of some covariates on the survivor function might be analyzed.

2.4.1 Overall survival

Overall survival (OS) corresponds to the percentage of patients alive at a certain instant in time (Driscoll and Rixe 2009). OS is usually the primary endpoint in phase III oncology clinical trials (Cheema and Burkes 2013).

Considering death as the event of interest is advantageous, since it is unambiguous and independent of investigator bias (U.S. Food and Drug Administration 2018). Still, since OS considers death by any cause, it could be influenced by other causes unrelated to the disease of interest (Kutikov et al. 2010). Additionally, it might be affected by the crossover of patients from the control to the treatment arm (Jönsson et al. 2014), or from the chosen subsequent treatments after the patients drop out of the trial (Korn, Freidlin, and Abrams 2011).

2.4.2 Progression-free survival

Progression-free survival (PFS) is a tumor-burden biomarker that measures the time from randomization until disease progression (U.S. Food and Drug Administration 2018). Progression can be defined as the increase in tumor growth, the appearance of new metastasis or the death of the patient (Villaruz and Socinski 2013).

PFS presents several benefits versus OS. Firstly, progression events typically occur before death, hence, PFS might take less time to mature than OS (Fleming, Rothmann, and Lu 2009). Additionally, the subsequent treatment does not confound PFS as it does OS (U.S. Food and Drug Administration 2018). Still, there are disadvantages in using PFS. The definition of progression can be subjective and depend on investigator assessment (Fallowfield and Fleissig 2012; Villaruz and Socinski 2013). It might also not be possible to pinpoint the date of progression accurately as scans will only be carried out weeks

apart (U.S. Food and Drug Administration 2018). Finally, PFS is not always correlated with survival (A. Tan et al. 2017; Shameer et al. 2021).

2.4.3 Objective response rate

Objective response rate (sometimes referred to as overall response rate, ORR) measures the percentage of patients for which the tumor reduced by a predefined amount within a certain period (U.S. Food and Drug Administration 2018; Aykan and Özatlı 2020). ORR is an endpoint that measures the direct antitumor effect of the drug.

ORR offers several benefits versus OS. Firstly, ORR can be measured earlier and with less patients than OS (U.S. Food and Drug Administration 2018). Still, there are also several disadvantages between ORR and OS. Specifically, the investigator outlines the definition of response and may vary between clinical trials. Additionally, ORR might not always correlate with survival (Mushti, Mulkey, and Sridhara 2018; Aykan and Özatlı 2020).

2.5 Propensity scores

In a late stage active comparator clinical trial, the trial cohort is randomized into at least a treatment and a control arm (Yoshida, Solomon, and Kim 2015). The randomization process ensures no bias or confounding between the arms of treatment. Since the patients are randomized into the treatment and control arms, the baseline variables are independent of the chosen treatment by design (S. G. West et al. 2014; Incerti et al. 2023). Hence, all baseline covariates of the patients are balanced between the cohorts (Williamson and Forbes 2014). Conversely, in analyses containing historical data, such as eControls, there can be intrinsic differences between the cohorts that influence the outcomes. In an optimal setting, all covariates related to the outcome should be balanced, so that there is no difference between the cohorts (Austin, Grootendorst, and Anderson 2007). Still, this might not be attainable, because the historical dataset might only include some combinations of observed covariates. Instead, another approach is to use methods such as propensity scores to balance the cohorts.

Propensity scores model the probability of the prescription of a treatment conditioned on the baseline values of the patient's covariates

$$P(\text{Treatment}_i|X_i) = f_{\text{Propensity Score}}(X_i). \quad (2.43)$$

The propensity score values of each patient can balance the cohorts to avoid biases, such as confounding. Balancing on the propensity score will balance the joint contribution of each variable to the treatment prescription (Glynn, Schneeweiss, and Stürmer 2006; Webster-Clark et al. 2021).

As the treatment indicator is usually a binary random variable, the propensity score can be estimated using any statistical method appropriate for classification problems. The most commonly used model

is the logistic regression (Williamson and Forbes 2014; Glynn, Schneeweiss, and Stürmer 2006), although other more complex models have also been studied (Weberpals et al. 2021).

The propensity score values can balance the cohorts with different methods, e.g., by matching or stratification. In the propensity score matching process, each patient in the treatment cohort is matched to one or more patients in the RWD control cohort with the closest propensity scores. To avoid large differences between matched patients, it is common to define a “caliper” value, which restricts the maximum propensity score difference allowed for a match (Austin 2011).

2.5.1 Propensity scores to build external controls

eControls are artificial control arms created using RWD or historical data from clinical trials. An eControl is constructed with a historical cohort with patients eligible to be enrolled in the clinical trial is necessary (Carrigan et al. 2020; Xiaomeng Wang et al. 2023). The patients in the historical cohort need to pass all the eligibility criteria considered in the clinical trial. Additionally, the patients must have been prescribed the control medication of interest (Xiaomeng Wang et al. 2023). Then, the propensity score can be calculated for each patient of the treatment and historical cohorts. Lastly, the propensity score values are used to match or stratify patients from the control cohort to the treated patients (Schmidli et al. 2020).

2.6 Non-small-cell lung cancer

Lung cancer is the second most common type of cancer worldwide. Lung cancer accounts for 11.4% of all cancer diagnoses and 18% of all cancer related deaths (Sung et al. 2021). Non-small-cell lung cancer (NSCLC) accounts for over 80% of all lung cancer cases (Ganti et al. 2021). The two main subtypes of NSCLC are adenocarcinoma, and squamous-cell carcinoma. Adenocarcinoma is the most common type of NSCLC and develops from glandular cells of the lung, which secrete mucus. It is more common in the outer part of the lung. Conversely, squamous-cell carcinoma occurs mostly the squamous cells of the larger bronchi of the lung. The main cause of both types of NSCLC is smoking.

2.6.1 Treatment of NSCLC

Most NSCLC cases are diagnosed at later stages (referred to as advanced NSCLC, advNSCLC), when the tumor has already metastasized to other parts of the body (Ganti et al. 2021). The treatment of NSCLC at later stages is composed mostly of chemotherapy, radiation or a combination of both (Ganti et al. 2021). Chemotherapy targeting NSCLC has evolved dramatically in the last decades. While previously NSCLC was treated with regular cytotoxic chemotherapy. New medications have been introduced that target individual genetic changes in the tumors (e.g. to ALK, EGFR or KRAS). Additionally, cancer immunotherapy drastically increased the median survival for patients with advanced NSCLC for which the tumor does not exhibit the aforementioned mutations.

In clinical trials that focus on advNSCLC, the main efficacy endpoint is OS. Still, both PFS and ORR are analyzed. For cancer immunotherapy, it has been observed that both PFS and ORR are not as correlated with OS as for regular chemotherapy. There are several hypotheses about the cause of this discrepancy. For example, the different mode of action of cancer immunotherapies can lead to tumor flare reaction. Tumor flare reaction is a side effect of cancer immunotherapy, also known as pseudoprogression, results from the infiltration of T-cells in the tumor site and causing an inflammation and apparent increase of the tumor burden or appearance of new tumor lesions (Taleb B 2019). Tumor flare reaction might occur before noticeable antitumour effect of the drug, resulting from the infiltration of the T-cells in the tumor sites. The tumor flare reaction was considered as disease progression according to the RECIST guidelines and led to treatment discontinuation. The discontinuation might prove counterproductive because it might lead to the patient not reaching treatment benefit of the immunotherapy. PFS might be inflated for cancer immunotherapies since they rely on the RECIST guidelines to define progression. Hence, in comparisons with other types of chemotherapy that do not lead to tumor flare reactions, the immunotherapies might show an increase of progression that does not lead to lower survival rates.

2.7 Historical and real-world data

With the continuous development and implementation of information technology systems in the clinical setting, available clinical data is expanding rapidly (F. Liu and Panagiotakos 2022; Booth, Karim, and Mackillop 2019). Data is available from both previously run clinical trials and RWD. RWD is usually defined as routinely collected clinical data that was obtained outside of the controlled setting of a clinical trial (Makady et al. 2017; McDonald et al. 2016). Oncology is a clinical area that is particularly well covered by RWD (Booth, Karim, and Mackillop 2019). The available datasets can describe the treatment and its effect longitudinally, as well as information such as the prescribed and administered medications, demographics, disease characteristics and longitudinal biomarker values that characterize the patient comprehensively.

Still, although RWD contains such high-quality patient-level data, RWD is not expected to substitute clinical trials. Still, there is an active push to determine additional ways to use RWD to complement and accelerate clinical development (Becker et al. 2020; Loureiro, Becker, Bauer-Mehren, et al. 2021; Yap et al. 2022; Ton et al. 2022; Xiaomeng Wang et al. 2023).

2.7.1 Real-world dataset

The Flatiron Health (FH) database is a longitudinal database, which focuses on oncology, and contains information such as the primary cancer, prescribed treatments, blood work and other biomarker values, alongside outcome information such as the date-of-death or progression. The patient-level information was de-identified and curated via technology-enabled abstraction (Ma et al. 2020; Birnbaum et al.

2020). Additionally, the raw data was augmented with oncologist-defined rule-based lines of therapy. The de-identified data originated from approximately 280 cancer clinics (~800 sites of care) across the United States. Most of the patient data originated in community oncology settings. The ratio between community/academic patients may vary depending on the study cohort. The de-identified data was subject to obligations to prevent re-identification and protect patient confidentiality. Data from the FH database was used extensively across this thesis. I specify which cohorts and the number of patients are included in chapter 3.

2.7.2 Roche clinical trials

F. Hoffmann-La Roche AG (in short, Roche) is a multinational pharmaceutical company. Roche has developed several approved oncology medications (e.g., Erlotinib, Bevacizumab or Atezolizumab) and has sponsored multiple clinical trials in oncology for these medications. To complement and validate the results from RWD, I used data from several Roche clinical trials. Specifically, in this thesis I used data from 11 clinical trials in advanced non-small-cell lung cancer (advNSCLC). The trials cover both approved and non-approved medications across phase II and III. The chosen trials focused on the first line of therapy, although two trials considered patients after the failure of cytotoxic chemotherapy.

3 Publication summaries

This chapter summarizes the three publications of this publication-based thesis that addressed the research questions described in chapter 1.

3.1 First work “Artificial Intelligence for Prognostic Scores in Oncology: A Benchmarking Study”

by H. Loureiro, T. Becker, A. Bauer-Mehren, N. Ahmidi*, J. Weberpals* (* contributed equally)

The article “Artificial Intelligence for Prognostic Scores in Oncology: A Benchmarking Study” was published in 2021 in the *Frontiers in Artificial Intelligence* journal (Loureiro, Becker, Bauer-Mehren, et al. 2021). I am the main author of this publication.

In this manuscript, I benchmarked the ROPRO with a comprehensive list of complex prognostic score models. DeepSurv (based on deep learning) had a higher performance than ROPRO for the RWD dataset, still the higher performance was not generalizable to a clinical trial dataset.

Research problem

In the last few decades, there have been significant developments in machine learning. Still, most state-of-the-art prognostic scores are based on simple statistical models, such as the Cox model or logistic regression (Loureiro, Becker, Bauer-Mehren, et al. 2021; Arkenau et al. 2009; International Non-Hodgkin’s Lymphoma Prognostic Factors Project 1993; Ko et al. 2015; Kinoshita et al. 2013), and consider a low number of variables (usually less than 10).

Neither the Cox nor logistic regression models can incorporate nonlinearities of the data, such as variable interaction or higher order terms, unless specified in the model structure (F. E. Harrell, Lee, and Mark 1996). The nonlinear effects in the complex clinical data might contain additional prognostic information. Hence, the current prognostic scores might be limited by using simpler methods. In essence, this publication addresses the first research question introduced in chapter 1.3:

First research question:

Can more complex survival models build more performant prognostic scores?

Approach

Firstly, I gathered a comprehensive list of statistical and machine learning-based survival analysis models. The list included the regularized Cox model, random survival forest (RSF), gradient boosting (GB), a deep learning-based model known as DeepSurv (DS), a novel autoencoder-based model (AE), and the super learner (SL). All the models were introduced in chapter 2.2. I setup a benchmarking study with these models to analyze their performance versus the ROPRO (Figure 3.1).

The ROPRO model, which is based on the cox model, was considered the baseline in the benchmark, and had been trained on 27 variables (Becker et al. 2020; 2023). Since I benchmarked machine-learning models, considering only 27 variables could be a limitation. Therefore, to the set of 27 variables, I included two additional sets, considering 44 and 88 covariates.

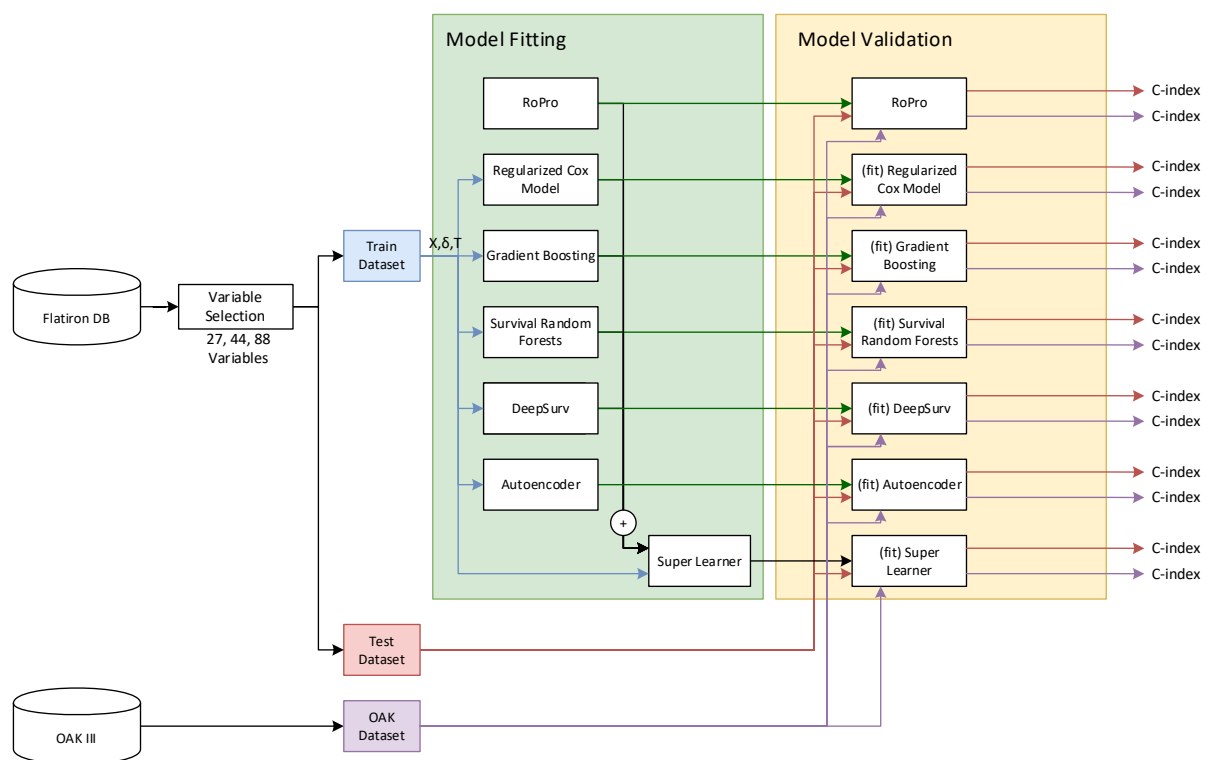


Figure 3.1 Diagram of the analysis. This figure is Figure 1 of the original manuscript by Loureiro et al. (Loureiro, Becker, Bauer-Mehren, et al. 2021).

To train and validate the models, I extracted a dataset from the FH dataset comprised of information on 136,719 patients with cancer across 18 different primary tumors. I divided the FH dataset into train (90%, 121,644 patients) and in-sample test (10%, or 15,075) sets. Additionally, to test the models in a non-RWD setting, I extracted an out-of-sample test dataset based on the OAK clinical trial (Rittmeyer et al. 2017). The OAK dataset comprises 1,187 patients with advanced non-small-cell lung cancer (advNSCLC). For both datasets, I considered only biomarkers at baseline, i.e., obtained before either the start of treatment (for FH), or the randomization date (for OAK).

I fit all survival models on the FH train dataset. Then, I predicted the risk values for the in-sample and out-sample test datasets. I used the Harrell and Uno's C-indexes to evaluate the performance of the

models. Additionally, I used bootstrap to obtain confidence intervals of both C-index, so that I could compare the performance of the models. The comparison of the C-index confidence intervals is equivalent to a hypothesis test of difference of the means.

Results

Firstly, for the 27 and 44 variable in-sample test sets, the performance of the GB, RSF, DS and SL models was significantly higher than ROPRO (Figure 3.2). For the 88 variable in-sample test dataset, all models, except AE obtained a significantly higher Harrell and Uno C-index than ROPRO. Still, none of the more complex models had significantly higher C-index values than ROPRO for the OAK (out-of-sample) datasets. The more complex model with the highest C-index for the OAK datasets was SL, although it was not a significant increase.

Conclusion

The benchmarking analysis contrasted the performance of multiple machine-learning models, against the Cox model (ROPRO). Overall, the more complex models did not outperform the simple Cox model. Additionally, adding more covariates to the training dataset did not significantly increase the performance. These results could signal the absence of nonlinearities in the considered datasets. As a hypothesis, the machine-learning models might perform better in complex types of information, such as images, genomic information, or additional disease information since simple models like the Cox model would not be capable of modeling these datatypes.

Individual contributions

Janick Weberpals, Tim Becker, and I defined the initial idea of the benchmarking study. I performed the literature review to identify suitable survival analysis models to be considered. Janick Weberpals contributed to this search by suggesting the autoencoder and super learner models. I extracted the FH and OAK datasets with Tim Becker and homogenized them into a standardized form. I developed all the modeling and prediction code used in the benchmarking study. This included Python and R code and shell scripts that automated the benchmarking.

Additionally, I modified several of the software packages used in this analysis. All the modifications were published alongside the manuscript as open source. Finally, I wrote the first draft of the manuscript and all authors contributed to its editing.

I am the main author of this publication.

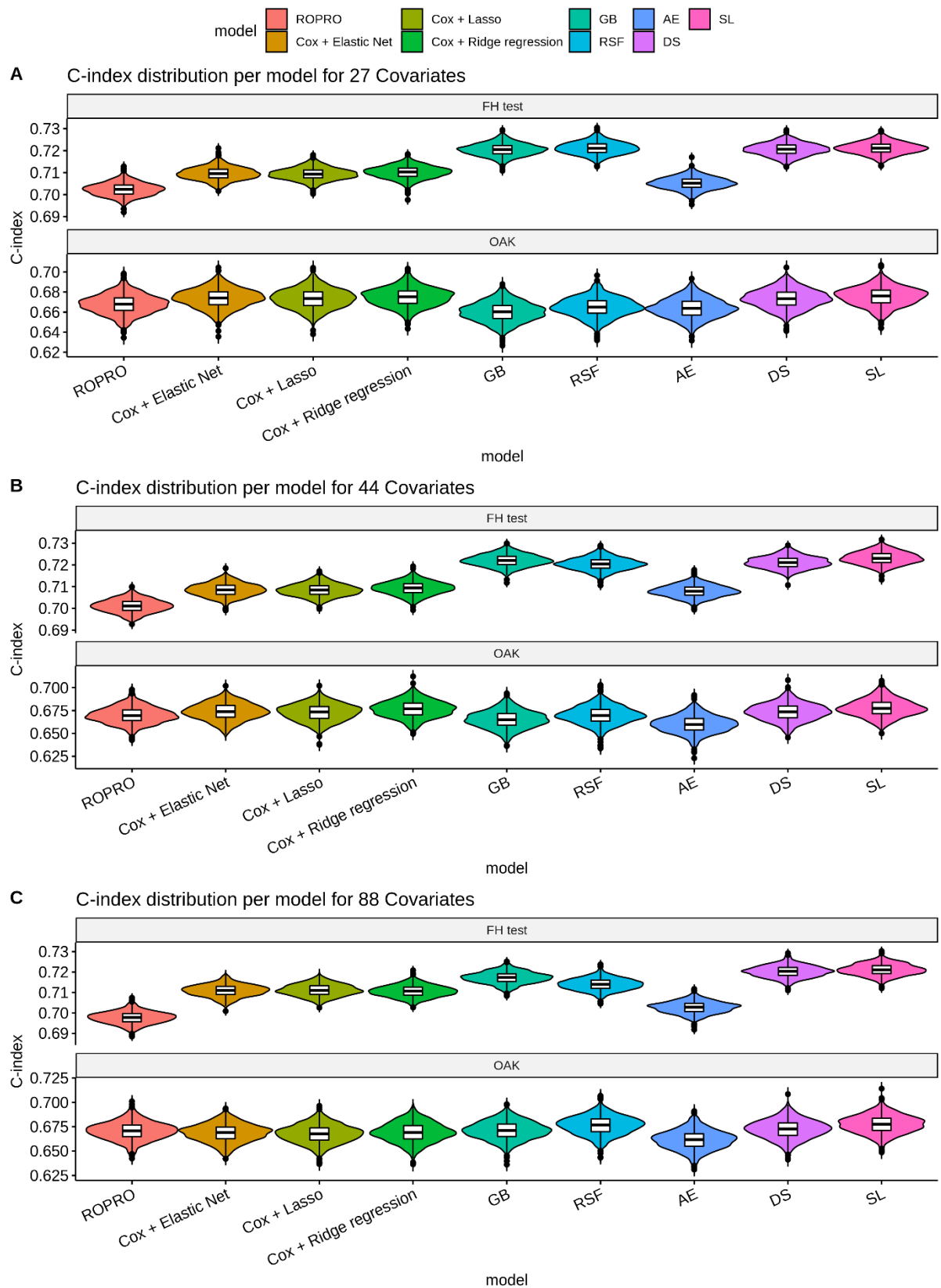


Figure 3.2 Violin plot of the Harrell C-index values for the FH in-sample test and OAK test datasets. The values in the violin plot were obtained by Bootstrap. This figure is based on Figure 6 of the original publication by Loureiro et al. (Loureiro, Becker, Bauer-Mehren, et al. 2021).

3.2 Second work “Matching by OS prognostic score to construct external controls in lung cancer clinical trials”

by H. Loureiro, A. Roller, M. Schneider, C. Talavera-López, T. Becker*, A. Bauer-Mehren* (* contributed equally)

The second work included in this thesis “Matching by OS prognostic score to construct external controls in lung cancer clinical trials”, was published in the *Clinical Pharmacology & Therapeutics* journal in 2023 (Loureiro, Roller, et al. 2023). I am the main author of this publication.

In the second publication, I compared external control arms matched with prognostic scores and with propensity scores. The external controls constructed with prognostic scores obtained higher accuracy in predicting OS.

Research problem

External control (eControl) arms are cohorts created exclusively with historical data. Since the control arm does not originate from randomization, there can be biases and confounding that affect the analysis results. Prognostic scores have been suggested to balance the treatment and external control cohorts and account for bias and confounding. Specifically, the prognostic score values can match or stratify the patients of the treatment and control arms, reducing the differences between the cohorts. Still, prognostic scores have never been used to create external control in oncology. The second publication focused on the second research question, introduced in chapter 1.4:

Second research question:

Can prognostic scores be used to match external controls?

Approach

Firstly, I obtained a list of 11 recent advNSCLC Roche clinical trials whose population could be reproduced with the FH dataset. All trials that contained more than two comparison arms were split into individual treatment-control arms, yielding 16 experimental-control comparisons. Afterwards, I extracted data from 46,595 patients with advNSCLC from the FH database to create external controls. I applied the same inclusion and exclusion criteria to the FH dataset for each clinical trial. Hence, for each trial, I generated a list of patients from FH who would have been eligible to the trial.

Next, to characterize the performance of prognostic scores to match eControls, I considered three different matching algorithms. The first was ROPRO which was the prognostic score used in this analysis. The second algorithm was a propensity score composed of the 27 covariates from ROPRO, for simplicity, I refer to this model as ROPROvars. The last model was a propensity score composed of the five most prognostic variables from ROPRO (or simply, 5Vars). The 5Vars model was included to understand the influence of the number of variables in the results.

The OS HR error was the performance metric. I used the Cox model to obtain the OS HR for each of the eControls. To collapse the results, I used the median to calculate the average error and bootstrap to calculate the confidence intervals.

Results

Overall, the ROPRO (prognostic score model) obtained the lowest OS HR error (MAD [bootstrap CI] 0.072 [0.036, 0.185]), followed by the 5Vars model, and lastly the ROPROvars model (MAD [bootstrap CI] 0.087 [0.054, 0.383]). There were small errors (less than 0.05) for many of the clinical trials analyzed. Additionally, when only phase III trials were considered, the error further decreased for all models (Figure 3.3).

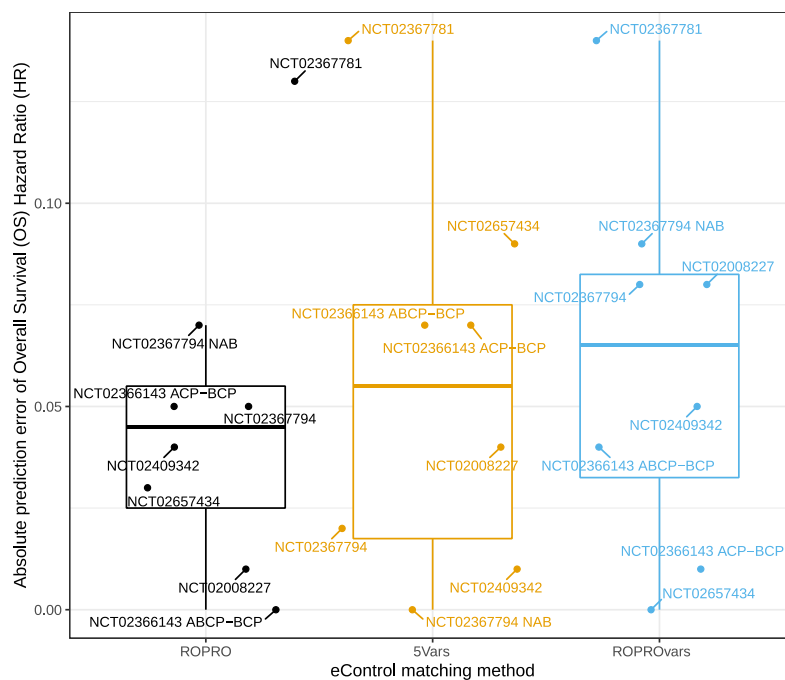


Figure 3.3 Prediction error of the OS HR for the phase III clinical trials. This figure is based on Figure 3 of the original publication by Loureiro et al. (Loureiro, Roller, et al. 2023).

Conclusion

All considered approaches could adequately reproduce the OS HR of most clinical trials. Still, the ROPRO obtained the lowest prediction error of the three models. The prediction error was especially low for phase III clinical trials, suggesting that the prognostic scores could be a good method to construct external controls in late drug development.

Individual contributions

Tim Becker, Anna Bauer-Mehren, and I conceived the hypothesis of using prognostic scores in external controls. Andreas Roller, Meike Schneider and Carlos Talavera-López supervised the medical aspects of the project. I performed the literature search for suitable clinical trials to be included in the analysis.

I extracted the clinical trial and RWD datasets. Additionally, I wrote all the analysis code, and performed the analysis. I wrote the draft of the manuscript, and all remaining authors edited the manuscript significantly.

I am the main author of this publication.

3.3 Third work “Correlation between early trends of a prognostic biomarker and overall survival in non-small-cell lung cancer clinical trials”

by [H. Loureiro](#), T. M. Kolben, A. Kiermaier, D. Rüttinger, N. Ahmidi, T. Becker*, A. Bauer-Mehren* (* contributed equally)

The article “Correlation between early trends of a prognostic biomarker and overall survival in non-small-cell lung cancer clinical trials” was published in 2023 in the JCO Clinical Cancer Informatics journal (Loureiro, Kolben, et al. 2023). I am the main author of this publication.

In the last publication of this publication-based thesis, I presented the “risk-trend framework”. The new framework estimates the OS HR of clinical trials from preliminary biomarker information from interim analyses.

Research problem

Most oncology drugs do not reach approval because they do not show an improvement in efficacy. Efficacy is measured with OS, an endpoint that requires a large population size and long follow-up to demonstrate improvements. To obtain an estimate of OS at interim analyses of phase III trials, the clinical trial teams use PFS, trial characteristics and additional information from the previous phases. Still, information from blood work and other biomarkers are not used frequently, although it is prognostic. In this analysis, I investigated if prognostic scores that combine so much information about a patient can be used to estimate OS. The final publication included in this publication-based thesis addressed the third research question:

Third research question:

Can the longitudinal values of prognostic scores characterize the efficacy results of a clinical trial?

Approach

With the FH dataset, I recreated 12 recent lung cancer clinical trials. The recreated clinical trials covered various medications, including regular cytotoxic chemotherapy, targeted treatments, and immunotherapy. Each of the recreated clinical trials consist of one treatment and one control arms. I created a new method, the risk trend framework, that models the prognostic score values for each of the clinical trials with a JM

$$\begin{cases} h_i(t) &= h_0(t) \exp[\gamma \cdot \text{Treatment Arm}_i + \alpha \cdot \text{risk}_{\text{trend},i}(t)] \\ \text{risk}_{\text{trend},i}(t) &= \beta_0 + (b_1 + \beta_1) \cdot t + \beta_2 t \cdot \text{Treatment Arm}_i + \varepsilon(t) \end{cases} \quad (3.1)$$

where the γ and β_2 coefficients model the difference between the treatments. The γ coefficient models the direct influence of the treatment on the hazard function. While the β_2 coefficient models a slope that is treatment dependent. A value of $\beta_2 \neq 0$, symbolizes a difference in the trends of the prognostic scores.

I considered three and six months to represent interim analyses. As a first analysis, I analyzed the correlation between the γ and β_2 coefficients and the OS HR at the interim analyses. Next, I fit a simple linear regression on the true OS HR. The regression had the form

$$\log(\text{OS HR}_j) = \theta_0 + \theta_1\beta_{2,j} + \theta_2\gamma_j + \epsilon, j \in \{1, \dots, 12\} \quad (3.2)$$

where θ are the linear regression coefficients, and j is a trial iterator. To assess the performance of estimating the OS HR with this method, I performed a leave one out analysis. In each iteration, I fit the linear model with 11 of the trials and predicted the OS value of the trial left out of the regression. Figure 3.4 summarizes the risk trend framework.

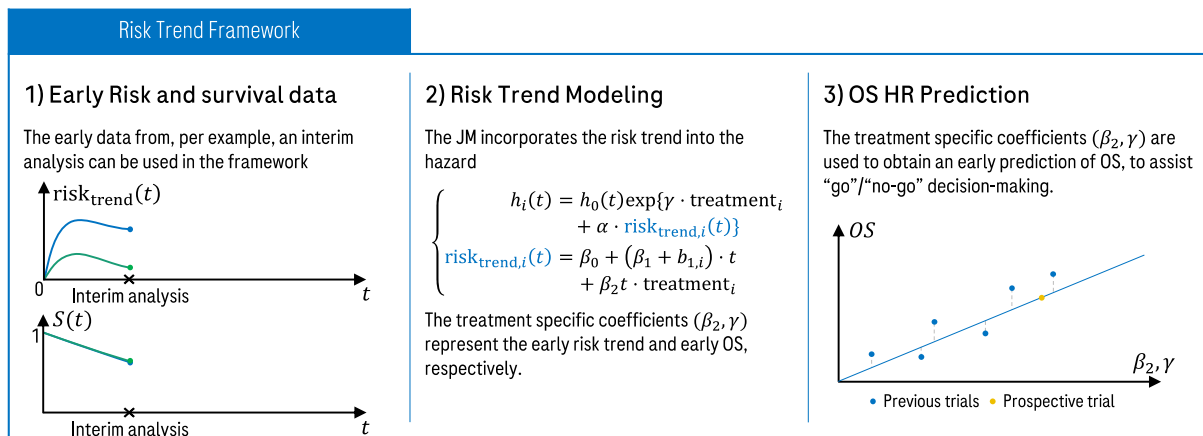


Figure 3.4 Diagram of the steps taken to estimate the OS HR of a new clinical trial with the Risk Trend Framework. This figure is based on Figure 1 of Loureiro et al. (Loureiro, Kolben, et al. 2023).

Results

The sign of the risk trend (β_2 coefficient) was concordant with the final OS effect for 11 out of 12 clinical trials. The only exception was PRONOUNCE, where the Carboplatin+Pemetrexed treatment arm had a significantly higher risk trend $\beta_2 > 0$, although there was no difference in OS.

The JM coefficients (β_2, γ) were highly correlated with the final OS HR at the three months (ROPRO JM adjusted R^2 values [bootstrap CI]: 0.88 [0.62, 0.98]), and at six months interim analyses (ROPRO JM adjusted R^2 values [bootstrap CI]: 0.85 [0.52, 0.98]). The DeepROPRO coefficients were also highly correlation with the OS HR. Additionally, in a sensitivity analysis with $\gamma = 0$, there was also high correlation between the β_2 coefficient and the final OS for both three months (ROPRO adjusted R^2 [bootstrap CI]: 0.85 [0.53, 0.97]), and six months (ROPRO adjusted R^2 [bootstrap CI]: 0.86 [0.46, 0.98]).

Lastly, the JM coefficients predicted the final OS HR with a low error at three months (RMSE [bootstrap CI] for ROPRO JM: 0.11 [0.08, 0.14], Figure 3.5), and six months (RMSE [bootstrap CI] for ROPRO: 0.12 [0.07, 0.16]).

Conclusion

The JM coefficients at early time points were correlated with the final OS HR. Additionally, these coefficients could be used to predict the final OS with a low error (usually lower than 0.1). Given the low prediction error obtained in this analysis, the risk trend framework could be used in two scenarios. Firstly, it could inform futility analyses at interim analyses of phase III studies. Additionally, it could be used before at an earlier phase to obtain an estimate of what the final OS HR would be at a later phase. The last use was not considered in this work and is only an outlook.

The correlation between the JM coefficients and final OS was higher than in other analyses (adjusted R^2 of 0.88). Shameer et al. (Shameer et al. 2021) reported a R^2 of 0.23 for all considered trials, and R^2 of 0.86 for PD-1/PD-L1 inhibitor studies. Still, my analysis considered a much lower number of studies.

Although the correlation results and the low prediction error highlight the interest of the risk trend framework, the analysis has some limitations. Firstly, I did not formally show surrogacy of the risk trend with OS, using the predefined guidelines. Next, the analysis focused exclusively on lung cancer, more research needs to be performed to understand the efficacy of this framework for other cancer types. Since this was an introductory analysis, I only performed the analysis in on RWD. The framework needs to be proven to also work in clinical trials.

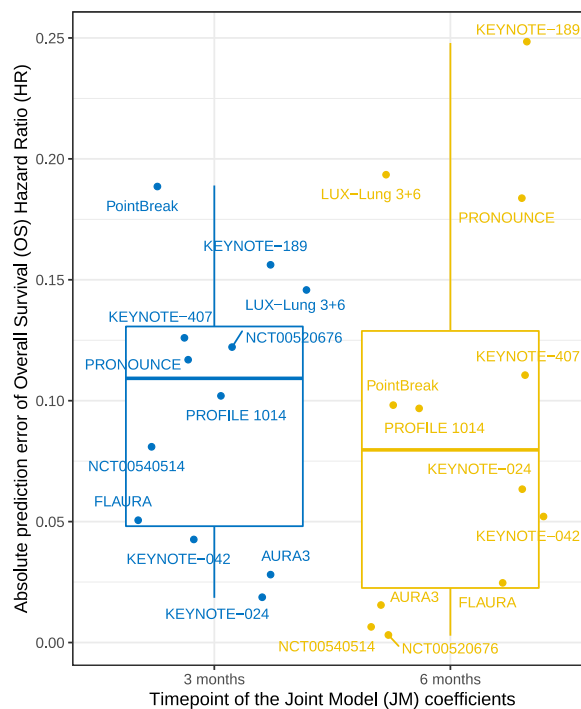


Figure 3.5 Prediction error of the final OS HR using the ROPRO JM model coefficients at the three- and 6-months interim analysis. Based on Figure 4 of the original publication by Loureiro et al. (Loureiro, Kolben, et al. 2023).

Individual contributions

Tim Becker, Anna Bauer-Mehren, and I devised using the longitudinal prognostic scores to define OS. Tim Becker and Anna Bauer-Mehren supervised the whole project, Dominik Rüttinger and Theresa Kolben supervised the medical aspects, and Narges Ahmidi supervised the machine learning aspects of the analysis. Tim Becker and I defined the JM strategy to model the prognostic score values over time. I extracted the FH dataset, performed the emulation of the clinical trials, and wrote all the code used in the analysis. I wrote the manuscript draft, and all other authors edited the manuscript.

I am the main author of this publication.

4 General discussion

Several prognostic covariates are routinely collected during treatment in oncology clinical practice. The prognostic variables are frequently interpreted individually to characterize specific aspects of the disease (e.g., progression) or the impact of the medication (e.g., adverse events). Conversely, prognostic scores collapse many prognostic variables into a single score that reflects the risk of an event, such as death. The risk value provides a comprehensive overview of the patient's health status. Hence, its benefit is twofold. In clinical practice, doctors obtain a comprehensive measure of risk and can focus their attention on patients that are at a greater hazard. Additionally, in research settings, the prognostic scores enable the analysis of large cohorts of patients based on their complete risk values. In this publication-based thesis, I addressed each of three research questions below with a publication.

First research question: Can more complex survival models build more performant prognostic scores?

Second research question: Can prognostic scores be used to match external controls?

Third research question: Can the longitudinal values of prognostic scores characterize the efficacy results of a clinical trial?

4.1 Summary

In this thesis, I explored multiple applications of prognostic scores in drug development. Firstly, I extensively benchmarked several prognostic score models (Loureiro, Becker, Bauer-Mehren, et al. 2021). I included several state-of-the-art methods and introduced two new models. I trained the models with a large RWD database. Additionally, I analyzed the generalizability of the results obtained in RWD with data from a recent advNSCLC clinical trial. I described each model summarily and analyzed their benefits and disadvantages. The results of the analysis enable researchers to perform a data-driven choice of which models they should consider for their survival analysis problem. I used the results of the first analysis to inform my following studies.

Next, as another use of prognostic scores in drug development, I analyzed whether prognostic scores can be used to construct external control arms. Overall, the external controls matched with prognostic scores obtained OS hazard ratios that were more similar to the original trials. The external controls matched with prognostic scores obtained similar OS HR values to the external control arms of the considered phase III studies.

Finally, I presented the risk trend framework (Loureiro, Kolben, et al. 2023). The risk trend framework uses the longitudinal prognostic scores and early mortality values to predict the OS results of clinical

trials. My analysis provides a first look at this methodology in RWD. The risk trend framework could enable an early estimation of the OS hazard ratio at interim analyses, providing clinical trial teams with data-driven insights.

The applications of prognostic scores developed in this thesis can be directly applied in drug development. Additionally, they can work as catalysts for further developments of prognostic scores in drug development. Although randomized clinical trials and endpoints such as OS are the standard methods recognized by regulatory bodies, new tools and endpoints are increasingly warranted for decision-making and complementing regulatory applications.

4.2 Outlook

Prognostic scores with rich data types

The increasing digitalization of clinical practice is leading to the capture of more data in electronic health records (EHR) and other healthcare databases (Pisaniello and Dixon 2020). The growing capture of clinical data is increasing the available anonymized patient-level data that is available to researchers. As an example, datasets such as Flatiron Health contain data from hundreds of thousands of patients, with patient-level information on treatment, biomarker results, outcome information, among others (Ma et al. 2020; Birnbaum et al. 2020). Additionally, some datasets such as the UK biobank, also contain some basic genomic information, e.g., a selection of single-nucleotide polymorphisms (SNPs) (Sudlow et al. 2015). Still, in EHRs, other rich data types such as images or complex genomic information (e.g., sequencing) are usually unavailable or available exclusively for a subset of patients (Dagenais et al. 2022). I expect that as more rich data types become more widely available; they will further increase the performance of prognostic scores.

Specifically, in the case of imaging in prognostic scores, most prognostic scores do not use the direct imaging information. Instead, the models incorporate features extracted from the images, such as progression or other measures (Elias K. Mai et al. 2015; Lun et al. 2020; Lindegaard et al. 2020; Rhee et al. 2022; Sato et al. 2021). Some analyzes suggest that the direct use of the images on the model could increase its performance (Kyono, Gilbert, and van der Schaar 2019). Providing the images to the model could lead to more information being extracted, increasing the performance (Abler et al. 2023; K. H. Leung et al. 2021). Additionally, it could also provide a way to remove the subjectivity in interpretation of the results between different scans and patients (Sullivan, Schwartz, and Zhao 2013).

Comprehensive tool to combine all types of patients data

The third publication included in this thesis (Loureiro, Kolben, et al. 2023) introduced the risk trend framework, which estimates OS at an early stage with preliminary mortality and the biomarker data. The risk trend framework estimates OS with a different approach from the other techniques available

in literature. Specifically, Shameer et al. created a web tool that predicts the final OS effect estimated from early PFS estimates (Shameer et al. 2022). The tool considers additional information such as the medication's mode of action to provide an estimate specific to the molecule type. Seo et al. presented an advanced convolutional neural network approach that integrated information about the drug and its target (Seo et al. 2021). The convolutional neural network considered the chemical features of the molecule, the genotype-tissue expression and other biological features of the target (Seo et al. 2021). Additionally, Hegge et al. created a model to predict the probability of success of a new medication using a Bayesian model that incorporates a comprehensive set of information from the trial, the number of patients, as well as the prior study outcomes (Hegge et al. 2020). These approaches tackled efficacy estimation at an early stage with multiple different modeling techniques and distinct data types.

The aforementioned early efficacy estimators consider different models and types of data. An estimator combining all the data types could possibly further increase the performance. Additionally, these models could be incorporated into the tools already available to visualize clinical trial results. Moreover, these tools were trained usually with open results or results from one sponsor alone. A comprehensive approach including several industry and academic partners could potentially build a much more powerful and comprehensive tool.

Risk scores for adverse events

In this thesis, I focused exclusively on efficacy (specifically, on mortality). Still, clinical trials analyze not only efficacy but also the safety profiles of drugs. The survival analysis framework used in this thesis can also create adverse events risk scores (Kondalsamy-Chennakesavan et al. 2009). The adverse event risk models could be applied in the same situations as the efficacy prognostic scores. Specifically, the adverse event risk scores could be used as a measure to enrich clinical trials, for example, to include less patients for which adverse events are more likely. Furthermore, the models could estimate the number of adverse events for each clinical trial arm, in an approach similar to the risk trend framework. In essence, the adverse event risk models could complement inclusion/exclusion criteria of clinical trials.

Several authors have analyzed adverse events as the event of interest in survival analysis (Kondalsamy-Chennakesavan et al. 2009; Osterman et al. 2022). Still, the datasets used in most the analyses were of moderate size (most below 5000 patients). Hence, larger datasets such as those in RWD, could be used to create models with higher performance and generalizability.

Adoption of RWD

RWD has become a great asset in health care decision making, constituting many studies in the past decade (Booth, Karim, and Mackillop 2019). One type of analysis that can be performed with RWD is to verify the generalizability of clinical trial results on the general population (Blonde et al. 2018). This type of analysis is possible since all patients are available in RWD. Conversely, in clinical trials, only

patients who are eligible for enrollment are available. Still, high quality RWD is usually only available for regions such as north America or some European countries (Hiramatsu et al. 2021; Wharton et al. 2023; Xuan Wang et al. 2019). Hence, the data might only be representative of some populations. Though, some programs aim to increase the gathering, availability and linking of RWD across countries, e.g. in Asia (Kc et al. 2023). Hence, in the future as more data from underrepresented regions becomes available, it should allow for better patient representation. Additionally, common data models such as the Observational Medical Outcomes Partnership (OMOP) (Garza et al. 2016) should simplify database analyses by standardizing the data structures.

Another challenge in using RWD are the several types of bias that can arise from an improper analysis (Levenson et al. 2023; Blonde et al. 2018). To avoid bias, researchers should consider how the data was collected, should perform several safe checks to test the quality of the data, and should test the assumptions that were taken during the analysis with sensitivity analyses. Several methods, including software packages, have been introduced recently to facilitate the analysis of RWD (e.g. how to interpret missing data patterns) (Sondhi et al. 2023). Moreover, there is a growing body of knowledge on the appropriate methodology to use in the analysis of RWD. There are also several educational offers, including degrees and workshops. In the future, the analysis of RWD should undergo a standardization that will increase its quality.

Deep learning developments and their use in drug development

Several studies have analyzed the performance of advanced machine learning based models to construct prognostic models. Chen et al. presented multiple applications where machine learning models performed better than simpler linear regression models (D. Chen et al. 2019). Although they noted that the more complex models needed to be more interpretable. Conversely, Christodoulou et al. showed in a systematic review that there was no performance benefit between machine learning models and logistic regression (Christodoulou et al. 2019). Therefore, there is no consensus in literature about whether machine learning models could improve the performance of prognostic scores. The first publication included in this thesis (Loureiro, Becker, Bauer-Mehren, et al. 2021) showed only minor performance benefits when using deep learning models in tabular (from EHR) data.

My analysis (Loureiro, Becker, Bauer-Mehren, et al. 2021) did not consider situations where richer information is available. Recently, several models and software packages have been created that apply deep learning to problems using a complex array of data types (Mak and Pichika 2019). For example, the software package “ehrapy” provides an end-to-end exploratory framework to analyze complex EHR databases (Lukas Heumos et al. 2023). Additionally, it includes several built-in models to analyze different types of data within the database. In terms of models, Rajkomar et al. and Tomašev et al. created deep-learning models considering raw EHR data (Rajkomar et al. 2018; Tomašev et al. 2021). Their models predicted several event types such as mortality, length of stay, or patient diagnosis during patient hospitalization, with a higher accuracy than simpler models (logistic regression or gradient

boosting). One of the benefits highlighted in the two analyses was that the models could incorporate “raw” clinical notes without preprocessing. Cheerla and Gevaert exploited the flexible structure of deep learning models to incorporate several types of genomic data (e.g. mRNA expression data and microRNA expression) into a complex survival model (Cheerla and Gevaert 2019). The data in the dataset was very heterogenous. Some patients had multiple kinds of expression data, while others had only one. Finally, LoReTTa is a self-supervised framework developed to facilitate the training of foundational models in multimodal datasets. LoReTTa is capable of modelling datasets of disjoint modalities (Tran et al. 2024), e.g. one dataset composed of images and text, and another dataset with images and audio. The developed deep learning model could accommodate these data variations. The flexibility of these models makes them ideal candidates to jointly analyze data from different databases since they can incorporate heterogeneous data types (Eraslan et al. 2019).

Federated learning in RWD

Due to privacy concerns, it might not be possible to extract patient-level information from certain medical databases. Effectively, many patient-level data is present in data silos that are inaccessible (Rieke et al. 2020). A possible solution to this problem is federated learning. Essentially, federated learning performs the analyses in each data source separately and then combines the results into a final joint analysis (Kairouz et al. 2021). Federated learning could enable us to perform analyses that were impossible before due to the lack of data. It might enable global analyses, while preserving privacy of local data.

In conclusion, there is a rapid increase in the number of clinical data sources available to scientists (Blonde et al. 2018). Moreover, enhanced epidemiological guidelines and methods, coupled with advanced models, enable thorough and robust analysis of these data. Using prognostic scores, historical data and advanced models in drug development offers exciting opportunities for researchers.

Bibliography

- Abler, Daniel, Perrine Courlet, Matthieu Dietz, Roberto Gatta, Pascal Girard, Alain Munafo, Alexandre Wicky, et al. 2023. "Semiautomated Pipeline to Quantify Tumor Evolution From Real-World Positron Emission Tomography/Computed Tomography Imaging." *JCO Clinical Cancer Informatics*, no. 7 (September): e2200126. <https://doi.org/10.1200/CCI.22.00126>.
- Afify, Ahmed Z., and Osama A. Mohamed. 2020. "A New Three-Parameter Exponential Distribution with Variable Shapes for the Hazard Rate: Estimation and Applications." *Mathematics* 8 (1). <https://doi.org/10.3390/math8010135>.
- Andersen, P. K., and R. D. Gill. 1982. "Cox's Regression Model for Counting Processes: A Large Sample Study." *The Annals of Statistics* 10 (4): 1100–1120. <https://doi.org/10.1214/aos/1176345976>.
- Arkenau, Hendrik-Tobias, Jorge Barriuso, David Olmos, Joo Ern Ang, Johann de Bono, Ian Judson, and Stan Kaye. 2009. "Prospective Validation of a Prognostic Score to Improve Patient Selection for Oncology Phase I Trials." *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology* 27 (16): 2692–96. <https://doi.org/10.1200/JCO.2008.19.5081>.
- Arlot, Sylvain, and Alain Celisse. 2010. "A Survey of Cross-Validation Procedures for Model Selection." *Statistics Surveys* 4 (January): 40–79. <https://doi.org/10.1214/09-SS054>.
- Austin, Peter C. 2011. "Optimal Caliper Widths for Propensity-Score Matching When Estimating Differences in Means and Differences in Proportions in Observational Studies." *Pharmaceutical Statistics* 10 (2): 150–61. <https://doi.org/10.1002/pst.433>.
- Austin, Peter C., Paul Grootendorst, and Geoffrey M. Anderson. 2007. "A Comparison of the Ability of Different Propensity Score Models to Balance Measured Variables between Treated and Untreated Subjects: A Monte Carlo Study." *Statistics in Medicine* 26 (4): 734–53. <https://doi.org/10.1002/sim.2580>.
- Aykan, Nuri Faruk, and Tahsin Özatlı. 2020. "Objective Response Rate Assessment in Oncology: Current Situation and Future Expectations." *World Journal of Clinical Oncology* 11 (2): 53–73. <https://doi.org/10.5306/wjco.v11.i2.53>.
- Banerjee, Mousumi, Evan Reynolds, Hedvig B. Andersson, and Brahmajee K. Nallamothu. 2019. "Tree-Based Analysis." *Circulation: Cardiovascular Quality and Outcomes* 12 (5): e004879. <https://doi.org/10.1161/CIRCOUTCOMES.118.004879>.

- Bastien, Frédéric, Pascal Lamblin, Razvan Pascanu, James Bergstra, Ian Goodfellow, Arnaud Bergeron, Nicolas Bouchard, David Warde-Farley, and Yoshua Bengio. 2012. “Theano: New Features and Speed Improvements.” arXiv. <http://arxiv.org/abs/1211.5590>.
- Becker, T., Marc Mailman, Sandy Tan, Ernest Lo, and A. Bauer-Mehren. 2023. “Comparison of Overall Survival Prognostic Power of Contemporary Prognostic Scores in Prevailing Tumor Indications.” *Medical Research Archives* 11 (4). <https://doi.org/10.18103/mra.v11i4.3638>.
- Becker, T., J. Weberpals, A.M. Jegg, W.V. So, A. Fischer, M. Weisser, F. Schmich, D. Rüttinger, and A. Bauer-Mehren. 2020. “An Enhanced Prognostic Score for Overall Survival of Patients with Cancer Derived from a Large Real-World Cohort.” *Annals of Oncology* 31 (11): 1561–68. <https://doi.org/10.1016/j.annonc.2020.07.013>.
- Bergstra, James, Olivier Breuleux, Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, Guillaume Desjardins, Joseph Turian, David Warde-Farley, and Yoshua Bengio. 2010. “Theano: A CPU and GPU Math Expression Compiler.” In *Proceedings of the Python for Scientific Computing Conference (SciPy)*, 4:1–7. Austin, TX.
- Birnbaum, Benjamin, Nathan Nussbaum, Katharina Seidl-Rathkopf, Monica Agrawal, Melissa Estevez, Evan Estola, Joshua Haimson, Lucy He, Peter Larson, and Paul Richardson. 2020. “Model-Assisted Cohort Selection with Bias Analysis for Generating Large-Scale Cohorts from the EHR for Oncology Research.” arXiv. <http://arxiv.org/abs/2001.09765>.
- Blonde, Lawrence, Kamlesh Khunti, Stewart B. Harris, Casey Meizinger, and Neil S. Skolnik. 2018. “Interpretation and Impact of Real-World Clinical Data for the Practicing Clinician.” *Advances in Therapy* 35 (11): 1763–74. <https://doi.org/10.1007/s12325-018-0805-y>.
- Booth, Christopher M., Safiya Karim, and William J. Mackillop. 2019. “Real-World Data: Towards Achieving the Achievable in Cancer Care.” *Nature Reviews Clinical Oncology* 16 (5): 312–25. <https://doi.org/10.1038/s41571-019-0167-7>.
- Brookhart, M. Alan, Sebastian Schneeweiss, Kenneth J. Rothman, Robert J. Glynn, Jerry Avorn, and Til Stürmer. 2006. “Variable Selection for Propensity Score Models.” *American Journal of Epidemiology* 163 (12): 1149–56. <https://doi.org/10.1093/aje/kwj149>.
- Browne, Michael W. 2000. “Cross-Validation Methods.” *Journal of Mathematical Psychology* 44 (1): 108–32. <https://doi.org/10.1006/jmps.1999.1279>.
- Byrd, Richard H., Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. 1995. “A Limited Memory Algorithm for Bound Constrained Optimization.” *SIAM Journal on Scientific Computing* 16 (5): 1190–1208. <https://doi.org/10.1137/0916069>.

- Caliendo, Marco, and Sabine Kopeinig. 2008. "Some Practical Guidance for the Implementation of Propensity Score Matching." *Journal of Economic Surveys* 22 (1): 31–72. <https://doi.org/10.1111/j.1467-6419.2007.00527.x>.
- Carrigan, Gillis, Samuel Whipple, William B. Capra, Michael D. Taylor, Jeffrey S. Brown, Michael Lu, Brandon Arnieri, Ryan Copping, and Kenneth J. Rothman. 2020. "Using Electronic Health Records to Derive Control Arms for Early Phase Single-Arm Lung Cancer Trials: Proof-of-Concept in Randomized Controlled Trials." *Clinical Pharmacology & Therapeutics* 107 (2): 369–77. <https://doi.org/10.1002/cpt.1586>.
- Carsten Nieder and Astrid Dalhaug. 2010. "A New Prognostic Score Derived from Phase I Study Participants with Advanced Solid Tumours Is Also Valid in Patients with Brain Metastasis." *Anticancer Research* 30 (3): 977.
- Cheema, P.K., and R.L. Burkes. 2013. "Overall Survival Should Be the Primary Endpoint in Clinical Trials for Advanced Non-Small-Cell Lung Cancer." *Current Oncology* 20 (2): 150–60. <https://doi.org/10.3747/co.20.1226>.
- Cheerla, Anika, and Olivier Gevaert. 2019. "Deep Learning with Multimodal Representation for Pancancer Prognosis Prediction." *Bioinformatics* 35 (14): i446–54. <https://doi.org/10.1093/bioinformatics/btz342>.
- Chen, David, Sijia Liu, Paul Kingsbury, Sunghwan Sohn, Curtis B. Storlie, Elizabeth B. Habermann, James M. Naessens, David W. Larson, and Hongfang Liu. 2019. "Deep Learning and Alternative Learning Strategies for Retrospective Real-World Clinical Data." *Npj Digital Medicine* 2 (1): 43. <https://doi.org/10.1038/s41746-019-0122-0>.
- Chen, Yifei, Zhenyu Jia, Dan Mercola, and Xiaohui Xie. 2013. "A Gradient Boosting Algorithm for Survival Analysis via Direct Optimization of Concordance Index." *Computational and Mathematical Methods in Medicine* 2013: 1–8. <https://doi.org/10.1155/2013/873595>.
- Christodoulou, Evangelia, Jie Ma, Gary S. Collins, Ewout W. Steyerberg, Jan Y. Verbakel, and Ben Van Calster. 2019. "A Systematic Review Shows No Performance Benefit of Machine Learning over Logistic Regression for Clinical Prediction Models." *Journal of Clinical Epidemiology* 110 (June): 12–22. <https://doi.org/10.1016/j.jclinepi.2019.02.004>.
- Clark, T G, M J Bradburn, S B Love, and D G Altman. 2003. "Survival Analysis Part I: Basic Concepts and First Analyses." *British Journal of Cancer* 89 (2): 232–38. <https://doi.org/10.1038/sj.bjc.6601118>.
- Cleves, Mario Alberto, William Gould, and Yulia V. Marchenko. 2016. *An Introduction to Survival Analysis Using Stata*. Revised third edition. College Station, Texas: Stata Press.

- Cox, Christopher. 2008. "The Generalized F Distribution: An Umbrella for Parametric Survival Analysis." *Statistics in Medicine* 27 (21): 4301–12. <https://doi.org/10.1002/sim.3292>.
- Cox, Christopher, Haitao Chu, Michael F. Schneider, and Alvaro Muñoz. 2007. "Parametric Survival Analysis and Taxonomy of Hazard Functions for the Generalized Gamma Distribution." *Statistics in Medicine* 26 (23): 4352–74. <https://doi.org/10.1002/sim.2836>.
- Cox, D. R. 1972. "Regression Models and Life-Tables." *Journal of the Royal Statistical Society. Series B (Methodological)* 34 (2): 187–220.
- D. W. Otter, J. R. Medina, and J. K. Kalita. 2021. "A Survey of the Usages of Deep Learning for Natural Language Processing." *IEEE Transactions on Neural Networks and Learning Systems* 32 (2): 604–24. <https://doi.org/10.1109/TNNLS.2020.2979670>.
- Dagenais, Simon, Leo Russo, Ann Madsen, Jen Webster, and Lauren Becnel. 2022. "Use of Real-World Evidence to Drive Drug Development Strategy and Inform Clinical Trial Design." *Clinical Pharmacology & Therapeutics* 111 (1): 77–89. <https://doi.org/10.1002/cpt.2480>.
- Desai, Mira. 2020. "Recruitment and Retention of Participants in Clinical Studies: Critical Issues and Challenges." *Perspectives in Clinical Research* 11 (2). https://doi.org/10.4103/picr.PICR_6_20.
- Driscoll, James J., and Oliver Rixe. 2009. "Overall Survival: Still the Gold Standard: Why Overall Survival Remains the Definitive End Point in Cancer Clinical Trials." *The Cancer Journal* 15 (5). <https://doi.org/10.1097/PPO.0b013e3181bdc2e0>.
- Ehwerhemuepha, Louis, Sidy Danioko, Shiva Verma, Rachel Marano, William Feaster, Sharief Taraman, Tatiana Moreno, Jianwei Zheng, Ehsan Yaghmaei, and Anthony Chang. 2021. "A Super Learner Ensemble of 14 Statistical Learning Models for Predicting COVID-19 Severity among Patients with Cardiovascular Conditions." *Intelligence-Based Medicine* 5 (January): 100030. <https://doi.org/10.1016/j.ibmed.2021.100030>.
- Elias K. Mai, Thomas Hielscher, Jost K. Kloth, Maximilian Merz, Sofia Shah, Marc S. Raab, Michaela Hillengass, et al. 2015. "A Magnetic Resonance Imaging-Based Prognostic Scoring System to Predict Outcome in Transplant-Eligible Patients with Multiple Myeloma." *Haematologica* 100 (6): 818–25. <https://doi.org/10.3324/haematol.2015.124115>.
- Eraslan, Gökçen, Žiga Avsec, Julien Gagneur, and Fabian J. Theis. 2019. "Deep Learning: New Computational Modelling Techniques for Genomics." *Nature Reviews Genetics* 20 (7): 389–403. <https://doi.org/10.1038/s41576-019-0122-6>.
- Fallowfield, Lesley J., and Anne Fleissig. 2012. "The Value of Progression-Free Survival to Patients with Advanced-Stage Cancer." *Nature Reviews Clinical Oncology* 9 (1): 41–47. <https://doi.org/10.1038/nrclinonc.2011.156>.

- Faraggi, David, and Richard Simon. 1995. "A Neural Network Model for Survival Data." *Statistics in Medicine* 14 (1): 73–82. <https://doi.org/10.1002/sim.4780140108>.
- Fawcett, Tom. 2006. "An Introduction to ROC Analysis." *ROC Analysis in Pattern Recognition* 27 (8): 861–74. <https://doi.org/10.1016/j.patrec.2005.10.010>.
- Ferrero, Paolo, Attilio Iacovoni, Emilia D'Elia, Muthiah Vaduganathan, Antonello Gavazzi, and Michele Senni. 2015. "Prognostic Scores in Heart Failure — Critical Appraisal and Practical Use." *International Journal of Cardiology* 188 (June): 1–9. <https://doi.org/10.1016/j.ijcard.2015.03.154>.
- Fiteni, F., V. Westeel, X. Pivot, C. Borg, D. Vernerey, and F. Bonnetain. 2014. "Endpoints in Cancer Clinical Trials." *Journal of Visceral Surgery* 151 (1): 17–22. <https://doi.org/10.1016/j.jviscsurg.2013.10.001>.
- Fiteni, Frédéric, Virginie Westeel, and Franck Bonnetain. 2017. "Surrogate Endpoints for Overall Survival in Lung Cancer Trials: A Review." *Expert Review of Anticancer Therapy* 17 (5): 447–54. <https://doi.org/10.1080/14737140.2017.1316196>.
- Fleming, Thomas R., Mark D. Rothmann, and Hong Laura Lu. 2009. "Issues in Using Progression-Free Survival When Evaluating Oncology Products." *Journal of Clinical Oncology* 27 (17): 2874–80. <https://doi.org/10.1200/JCO.2008.20.4107>.
- Fogel, David B. 2018. "Factors Associated with Clinical Trials That Fail and Opportunities for Improving the Likelihood of Success: A Review." *Contemporary Clinical Trials Communications* 11 (August): 156–64. <https://doi.org/10.1016/j.conctc.2018.08.001>.
- Freidlin, Boris, and Edward L. Korn. 2014. "Biomarker Enrichment Strategies: Matching Trial Design to Biomarker Credentials." *Nature Reviews Clinical Oncology* 11 (2): 81–90. <https://doi.org/10.1038/nrclinonc.2013.218>.
- Friedman, Jerome H. 2001. "Greedy Function Approximation: A Gradient Boosting Machine." *The Annals of Statistics* 29 (5): 1189–1232.
- Funahashi, Ken-Ichi. 1989. "On the Approximate Realization of Continuous Mappings by Neural Networks." *Neural Networks* 2 (3): 183–92. [https://doi.org/10.1016/0893-6080\(89\)90003-8](https://doi.org/10.1016/0893-6080(89)90003-8).
- Ganti, Apar Kishor, Alyssa B. Klein, Ion Cotarla, Brian Seal, and Engels Chou. 2021. "Update of Incidence, Prevalence, Survival, and Initial Treatment in Patients With Non–Small Cell Lung Cancer in the US." *JAMA Oncology* 7 (12): 1824–32. <https://doi.org/10.1001/jamaoncol.2021.4932>.
- Garza, Maryam, Guilherme Del Fiol, Jessica Tenenbaum, Anita Walden, and Meredith Nahm Zozus. 2016. "Evaluating Common Data Models for Use with a Longitudinal Community Registry."

- Journal of Biomedical Informatics* 64 (December): 333–41.
<https://doi.org/10.1016/j.jbi.2016.10.016>.
- Gehrmann, Sebastian, Franck Dernoncourt, Yeran Li, Eric T. Carlson, Joy T. Wu, Jonathan Welt, John Foote Jr., et al. 2018. “Comparing Deep Learning and Concept Extraction Based Methods for Patient Phenotyping from Clinical Narratives.” *PLOS ONE* 13 (2): e0192360.
<https://doi.org/10.1371/journal.pone.0192360>.
- Glynn, Robert J., Sebastian Schneeweiss, and Til Stürmer. 2006. “Indications for Propensity Scores and Review of Their Use in Pharmacoepidemiology.” *Basic & Clinical Pharmacology & Toxicology* 98 (3): 253–59. https://doi.org/10.1111/j.1742-7843.2006.pto_293.x.
- Goldman, Dana P., Geoffrey F. Joyce, and Yuhui Zheng. 2007. “Prescription Drug Cost Sharing Associations With Medication and Medical Utilization and Spending and Health.” *JAMA* 298 (1): 61–69. <https://doi.org/10.1001/jama.298.1.61>.
- Golmakani, Marzieh K, and Eric C Polley. 2020. “Super Learner for Survival Data Prediction.” *The International Journal of Biostatistics* 16 (2): 20190065. <https://doi.org/10.1515/ijb-2019-0065>.
- Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press.
<http://www.deeplearningbook.org>.
- Harrell, F. E., K. L. Lee, and D. B. Mark. 1996. “Multivariable Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors.” *Statistics in Medicine* 15 (4): 361–87. [https://doi.org/10.1002/\(SICI\)1097-0258\(19960229\)15:4<361::AID-SIM168>3.0.CO;2-4](https://doi.org/10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4).
- Harrell, Frank E., Jr, Robert M. Califf, David B. Pryor, Kerry L. Lee, and Robert A. Rosati. 1982. “Evaluating the Yield of Medical Tests.” *JAMA* 247 (18): 2543–46.
<https://doi.org/10.1001/jama.1982.03320430047030>.
- Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Second Edition. Springer Series in Statistics. New York: Springer Science & Business Media.
- Hegge, Stephan, Markus Thunecke, Matthias Krings, Léonard Ruedin, Jan Saputra Müller, and Paul von Büнау. 2020. “Predicting Success of Phase III Trials in Oncology.” medRxiv.
<https://doi.org/10.1101/2020.12.15.20248240>.
- Heinemann, V., J.Y. Douillard, M. Ducreux, and M. Peeters. 2013. “Targeted Therapy in Metastatic Colorectal Cancer – An Example of Personalised Medicine in Action.” *Cancer Treatment Reviews* 39 (6): 592–601. <https://doi.org/10.1016/j.ctrv.2012.12.011>.
- Hiramatsu, Katsutoshi, Annabel Barrett, Yasuhiko Miyata, and PhRMA Japan Medical Affairs Committee Working Group 1. 2021. “Current Status, Challenges, and Future Perspectives of

- Real-World Data and Real-World Evidence in Japan.” *Drugs - Real World Outcomes* 8 (4): 459–80. <https://doi.org/10.1007/s40801-021-00266-3>.
- Hsieh, Fushing, Yi-Kuan Tseng, and Jane-Ling Wang. 2006. “Joint Modeling of Survival and Longitudinal Data: Likelihood Approach Revisited.” *Biometrics* 62 (4): 1037–43. <https://doi.org/10.1111/j.1541-0420.2006.00570.x>.
- Imad Bou-Hamad, Denis Larocque, and Hatem Ben-Ameur. 2011. “A Review of Survival Trees.” *Statistics Surveys* 5 (January): 44–71. <https://doi.org/10.1214/09-SS047>.
- Incerti, Devin, Michael T. Bretscher, Ray Lin, and Chris Harbron. 2023. “A Meta-Analytic Framework to Adjust for Bias in External Control Studies.” *Pharmaceutical Statistics* 22 (1): 162–80. <https://doi.org/10.1002/pst.2266>.
- International Non-Hodgkin’s Lymphoma Prognostic Factors Project. 1993. “A Predictive Model for Aggressive Non-Hodgkin’s Lymphoma.” *New England Journal of Medicine* 329 (14): 987–94. <https://doi.org/10.1056/NEJM199309303291402>.
- Ishwaran, Hemant, Udaya B. Kogalur, Eugene H. Blackstone, and Michael S. Lauer. 2008. “Random Survival Forests.” *The Annals of Applied Statistics* 2 (3): 841–60. <https://doi.org/10.1214/08-aoas169>.
- Jönsson, Linus, Rickard Sandin, Mattias Ekman, Joakim Ramsberg, Claudie Charbonneau, Xin Huang, Bengt Jönsson, Milton C. Weinstein, and Michael Drummond. 2014. “Analyzing Overall Survival in Randomized Controlled Trials with Crossover and Implications for Economic Evaluation.” *Value in Health* 17 (6): 707–13. <https://doi.org/10.1016/j.jval.2014.06.006>.
- Ju, Cheng, Mary Combs, Samuel D Lendle, Jessica M Franklin, Richard Wyss, Sebastian Schneeweiss, and Mark J van der Laan. 2019. “Propensity Score Prediction for Electronic Healthcare Databases Using Super Learner and High-Dimensional Propensity Score Methods.” *Journal of Applied Statistics* 46 (12): 2216–36. <https://doi.org/10.1080/02664763.2019.1582614>.
- Kairouz, Peter, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, et al. 2021. “Advances and Open Problems in Federated Learning.” *Foundations and Trends® in Machine Learning* 14 (1–2): 1–210. <https://doi.org/10.1561/22000000083>.
- Kaitin, K. I. 2010. “Deconstructing the Drug Development Process: The New Face of Innovation.” *Clinical Pharmacology & Therapeutics* 87 (3): 356–61. <https://doi.org/10.1038/clpt.2009.293>.
- Kalbfleisch, John D, and Ross L Prentice. 2002. *The Statistical Analysis of Failure Time Data*. Wiley Series in Probability and Statistics. New Jersey: John Wiley & Sons, Inc.

- Kaplan, E. L., and Paul Meier. 1958. "Nonparametric Estimation from Incomplete Observations." *Journal of the American Statistical Association* 53 (282): 457–81. <https://doi.org/10.1080/01621459.1958.10501452>.
- Katzman, Jared L., Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. 2018. "DeepSurv: Personalized Treatment Recommender System Using a Cox Proportional Hazards Deep Neural Network." *BMC Medical Research Methodology* 18 (1): 24. <https://doi.org/10.1186/s12874-018-0482-1>.
- Kc, Sarin, Lydia Wenxin Lin, Diana Beatriz Samson Bayani, Yaroslava Zemlyanska, Amanda Adler, Jeonghoon Ahn, Kelvin Chan, et al. 2023. "What, Where, and How to Collect Real-World Data and Generate Real-World Evidence to Support Drug Reimbursement Decision-Making in Asia: A Reflection Into the Past and A Way Forward." *International Journal of Health Policy and Management* 12 (1): 1–9. <https://doi.org/10.34172/ijhpm.2023.6858>.
- Kinch, Michael S., Austin Haynesworth, Sarah L. Kinch, and Denton Hoyer. 2014. "An Overview of FDA-Approved New Molecular Entities: 1827–2013." *Drug Discovery Today* 19 (8): 1033–39. <https://doi.org/10.1016/j.drudis.2014.03.018>.
- Kingma, Diederik P., and Jimmy Ba. 2014. "Adam: A Method for Stochastic Optimization." arXiv. <https://doi.org/10.48550/arXiv.1412.6980>.
- Kinoshita, Akiyoshi, Hiroshi Onoda, Nami Imai, Akira Iwaku, Mutumi Oishi, Ken Tanaka, Nao Fushiya, et al. 2013. "The Glasgow Prognostic Score, an Inflammation Based Prognostic Score, Predicts Survival in Patients with Hepatocellular Carcinoma." *BMC Cancer* 13 (1): 52. <https://doi.org/10.1186/1471-2407-13-52>.
- Klambauer, Günter, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. 2017. "Self-Normalizing Neural Networks." In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 972–81. NIPS'17. Long Beach, California, USA. <https://doi.org/10.48550/arXiv.1706.02515>.
- Klatte, Tobias, Sabrina H. Rossi, and Grant D. Stewart. 2018. "Prognostic Factors and Prognostic Models for Renal Cell Carcinoma: A Literature Review." *World Journal of Urology* 36 (12): 1943–52. <https://doi.org/10.1007/s00345-018-2309-4>.
- Kleinbaum, David G., and Mitchel Klein. 2012. *Survival Analysis: A Self-Learning Text*. 3rd ed. Statistics for Biology and Health. New York: Springer.
- Ko, Jenny J, Wanling Xie, Nils Kroeger, Jae-lyun Lee, Brian I Rini, Jennifer J Knox, Georg A Bjarnason, et al. 2015. "The International Metastatic Renal Cell Carcinoma Database Consortium Model as a Prognostic Tool in Patients with Metastatic Renal Cell Carcinoma Previously Treated with First-Line Targeted Therapy: A Population-Based Study." *The Lancet Oncology* 16 (3): 293–300. [https://doi.org/10.1016/S1470-2045\(14\)71222-7](https://doi.org/10.1016/S1470-2045(14)71222-7).

- Kon, Mark A., and Leszek Plaskota. 2000. "Complexity of Predictive Neural Networks." In *Proceedings from the Third International Conference on Complex Systems*, 181–91. https://doi.org/10.1007/978-3-540-35866-4_18.
- Kondalsamy-Chennakesavan, Srinivas, Chantal Bouman, Suzanne De Jong, Karen Sanday, Jim Nicklin, Russell Land, and Andreas Obermair. 2009. "Clinical Audit in Gynecological Cancer Surgery: Development of a Risk Scoring System to Predict Adverse Events." *Gynecologic Oncology* 115 (3): 329–33. <https://doi.org/10.1016/j.ygyno.2009.08.004>.
- Korn, Edward L., Boris Freidlin, and Jeffrey S. Abrams. 2011. "Overall Survival As the Outcome for Randomized Clinical Trials With Effective Subsequent Therapies." *Journal of Clinical Oncology* 29 (17): 2439–42. <https://doi.org/10.1200/JCO.2011.34.6056>.
- Kutikov, Alexander, Brian L. Egleston, Yu-Ning Wong, and Robert G. Uzzo. 2010. "Evaluating Overall Survival and Competing Risks of Death in Patients with Localized Renal Cell Carcinoma Using a Comprehensive Nomogram." *Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology* 28 (2): 311–17. <https://doi.org/10.1200/JCO.2009.22.4816>.
- Kyono, Trent, Fiona J. Gilbert, and Mihaela van der Schaar. 2019. "Multi-View Multi-Task Learning for Improving Autonomous Mammogram Diagnosis." In *Proceedings of the 4th Machine Learning for Healthcare Conference*, 106:571–91. Proceedings of Machine Learning Research. PMLR. <https://proceedings.mlr.press/v106/kyono19a.html>.
- L. Jiao and J. Zhao. 2019. "A Survey on the New Generation of Deep Learning in Image Processing." *IEEE Access* 7: 172231–63. <https://doi.org/10.1109/ACCESS.2019.2956508>.
- Laan, Mark J. van der, Eric C. Polley, and Alan E. Hubbard. 2007. "Super Learner." *Statistical Applications in Genetics and Molecular Biology* 6 (1). <https://doi.org/10.2202/1544-6115.1309>.
- LeDell, Erin, Mark J. van der Laan, and Maya Peterson. 2016. "AUC-Maximizing Ensembles through Metalearning." *The International Journal of Biostatistics* 12 (1): 203–18. <https://doi.org/10.1515/ijb-2015-0035>.
- Leung, Kevin H., Steven P. Rowe, Martin G. Pomper, and Yong Du. 2021. "A Three-Stage, Deep Learning, Ensemble Approach for Prognosis in Patients with Parkinson's Disease." *EJNMMI Research* 11 (1): 52. <https://doi.org/10.1186/s13550-021-00795-6>.
- Leung, Kwan-Moon, Robert M. Elashoff, and Abdelmonem A. Afifi. 1997. "CENSORING ISSUES IN SURVIVAL ANALYSIS." *Annual Review of Public Health* 18 (1): 83–104. <https://doi.org/10.1146/annurev.publhealth.18.1.83>.
- Levenson, Mark, Weili He, Jie Chen, Yixin Fang, Douglas Faries, Benjamin A. Goldstein, Martin Ho, et al. 2023. "Biostatistical Considerations When Using RWD and RWE in Clinical Studies for

- Regulatory Purposes: A Landscape Assessment.” *Statistics in Biopharmaceutical Research* 15 (1): 3–13. <https://doi.org/10.1080/19466315.2021.1883473>.
- Lewkowycz, Aitor, and Guy Gur-Ari. 2020. “On the Training Dynamics of Deep Networks with L₂ Regularization.” In *Advances in Neural Information Processing Systems*, edited by H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, 33:4790–99. https://proceedings.neurips.cc/paper_files/paper/2020/file/32fcc8cfe1fa4c77b5c58dafd36d1a98-Paper.pdf.
- Lindegaard, Jacob Christian, Primoz Petric, Anne Marie Lindegaard, Kari Tanderup, and Lars Ulrik Fokdal. 2020. “Evaluation of a New Prognostic Tumor Score in Locally Advanced Cervical Cancer Integrating Clinical Examination and Magnetic Resonance Imaging.” *International Journal of Radiation Oncology*Biophysics* 106 (4): 754–63. <https://doi.org/10.1016/j.ijrobp.2019.11.031>.
- Liu, Fang, and Demosthenes Panagiotakos. 2022. “Real-World Data: A Brief Review of the Methods, Applications, Challenges and Opportunities.” *BMC Medical Research Methodology* 22 (1): 287. <https://doi.org/10.1186/s12874-022-01768-6>.
- Liu, Si-Yang Maggie, Mei-Mei Zheng, Yi Pan, Si-Yang Liu, Yangqiu Li, and Yi-Long Wu. 2023. “Emerging Evidence and Treatment Paradigm of Non-Small Cell Lung Cancer.” *Journal of Hematology & Oncology* 16 (1): 40. <https://doi.org/10.1186/s13045-023-01436-2>.
- Loureiro, Hugo, Tim Becker, Narges Ahmidi, and Anna Bauer-Mehren. 2021. “Towards OS Approximation in Early Clinical Trials Using Prognostic Scores.” In *Pharmacoepidemiology and Drug Safety*, 30:366–67. John Wiley & Sons, Ltd. <https://doi.org/10.1002/pds.5305>.
- Loureiro, Hugo, Tim Becker, and Anna Bauer-Mehren. 2022. “A Longitudinal Early-Indicator of Overall Survival Based on Prognostic Scores.” In *Pharmacoepidemiology and Drug Safety*, 31:559. John Wiley & Sons, Ltd. <https://doi.org/10.1002/pds.5518>.
- Loureiro, Hugo, Tim Becker, Anna Bauer-Mehren, Narges Ahmidi, and Janick Weberpals. 2020. “Improving Predictive Ability of Survival Models: Comparison of Multiple State of the Art Models.” In *Pharmacoepidemiology and Drug Safety*, 29:35–36. John Wiley & Sons, Ltd. <https://doi.org/10.1002/pds.5114>.
- . 2021. “Artificial Intelligence for Prognostic Scores in Oncology: A Benchmarking Study.” *Frontiers in Artificial Intelligence* 4: 9. <https://doi.org/10.3389/frai.2021.625573>.
- Loureiro, Hugo, Theresa M. Kolben, Astrid Kiermaier, Dominik Rüttinger, Narges Ahmidi, Tim Becker, and Anna Bauer-Mehren. 2023. “Correlation Between Early Trends of a Prognostic Biomarker and Overall Survival in Non-Small-Cell Lung Cancer Clinical Trials.” *JCO Clinical Cancer Informatics*, no. 7 (September): e2300062. <https://doi.org/10.1200/CCI.23.00062>.

- Loureiro, Hugo, Andreas Roller, Meike Schneider, Carlos Talavera-López, Tim Becker, and Anna Bauer-Mehren. 2023. "Matching by OS Prognostic Score to Construct External Controls in Lung Cancer Clinical Trials." *Clinical Pharmacology & Therapeutics*, November. <https://doi.org/10.1002/cpt.3109>.
- Lukas Heumos, Philipp Ehmele, Tim Treis, Julius Upmeier zu Belzen, Altana Namsaraeva, Nastassya Horlava, Vladimir A. Shitov, et al. 2023. "Exploratory Electronic Health Record Analysis with Ehrapy." *medRxiv*, January, 2023.12.11.23299816. <https://doi.org/10.1101/2023.12.11.23299816>.
- Lun, Ronda, Vignan Yogendrakumar, Andrew M. Demchuk, Richard I. Aviv, David Rodriguez-Luna, Carlos A. Molina, Yolanda Silva, et al. 2020. "Calculation of Prognostic Scores, Using Delayed Imaging, Outperforms Baseline Assessments in Acute Intracerebral Hemorrhage." *Stroke* 51 (4): 1107–10. <https://doi.org/10.1161/STROKEAHA.119.027119>.
- Ma, Xinran, Lura Long, Sharon Moon, Blythe J.S. Adamson, and Shrujal S. Baxi. 2020. "Comparison of Population Characteristics in Real-World Clinical Oncology Databases in the US: Flatiron Health, SEER, and NPCR." *medRxiv*. <https://doi.org/10.1101/2020.03.16.20037143>.
- Magnini, Bernardo, Alberto Lavelli, and Simone Magnolini. 2020. "Comparing Machine Learning and Deep Learning Approaches on NLP Tasks for the Italian Language." In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 2110–19. Marseille, France: European Language Resources Association. <https://aclanthology.org/2020.lrec-1.259>.
- Mahipal, Amit, and Danny Nguyen. 2014. "Risks and Benefits of Phase 1 Clinical Trial Participation." *Cancer Control* 21 (3): 193–99. <https://doi.org/10.1177/107327481402100303>.
- Mak, Kit-Kay, and Mallikarjuna Rao Pichika. 2019. "Artificial Intelligence in Drug Development: Present Status and Future Prospects." *Drug Discovery Today* 24 (3): 773–80. <https://doi.org/10.1016/j.drudis.2018.11.014>.
- Makady, Amr, Anthonius de Boer, Hans Hillege, Olaf Klungel, and Wim Goettsch. 2017. "What Is Real-World Data? A Review of Definitions Based on Literature and Stakeholder Interviews." *Value in Health* 20 (7): 858–65. <https://doi.org/10.1016/j.jval.2017.03.008>.
- Maslej-Krešňáková, Viera, Martin Sarnovský, Peter Butka, and Kristína Machová. 2020. "Comparison of Deep Learning Models and Various Text Pre-Processing Techniques for the Toxic Comments Classification." *Applied Sciences* 10 (23). <https://doi.org/10.3390/app10238631>.
- Mauguen, Audrey, Jean-Pierre Pignon, Sarah Burdett, Caroline Domerg, David Fisher, Rebecca Paulus, Samithra J Mandrekar, et al. 2013. "Surrogate Endpoints for Overall Survival in Chemotherapy and Radiotherapy Trials in Operable and Locally Advanced Lung Cancer: A Re-Analysis of Meta-Analyses of Individual Patients' Data." *The Lancet Oncology* 14 (7): 619–26. [https://doi.org/10.1016/S1470-2045\(13\)70158-X](https://doi.org/10.1016/S1470-2045(13)70158-X).

- McDonald, Laura, Dimitra Lambrelli, Radek Wasiak, and Sreeram V. Ramagopalan. 2016. "Real-World Data in the United Kingdom: Opportunities and Challenges." *BMC Medicine* 14 (1): 97. <https://doi.org/10.1186/s12916-016-0647-x>.
- McMillan, Donald C. 2013. "The Systemic Inflammation-Based Glasgow Prognostic Score: A Decade of Experience in Patients with Cancer." *Cancer Treatment Reviews* 39 (5): 534–40. <https://doi.org/10.1016/j.ctrv.2012.08.003>.
- MD Anderson Cancer Center. n.d. "Phases of Clinical Trials." MD Anderson Cancer Center. Accessed October 5, 2023. <https://www.mdanderson.org/patients-family/diagnosis-treatment/clinical-trials/phases-of-clinical-trials.html>.
- Mishra-Kalyani, P.S., L. Amiri Kordestani, D.R. Rivera, H. Singh, A. Ibrahim, R.A. DeClaro, Y. Shen, et al. 2022. "External Control Arms in Oncology: Current Use and Future Directions." *Annals of Oncology* 33 (4): 376–83. <https://doi.org/10.1016/j.annonc.2021.12.015>.
- Morgan, Steve, Paul Grootendorst, Joel Lexchin, Colleen Cunningham, and Devon Greyson. 2011. "The Cost of Drug Development: A Systematic Review." *Health Policy* 100 (1): 4–17. <https://doi.org/10.1016/j.healthpol.2010.12.002>.
- Mushti, Sirisha L., Flora Mulkey, and Rajeshwari Sridhara. 2018. "Evaluation of Overall Response Rate and Progression-Free Survival as Potential Surrogate Endpoints for Overall Survival in Immunotherapy Trials." *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research* 24 (10): 2268–75. <https://doi.org/10.1158/1078-0432.CCR-17-1902>.
- National Cancer Institute. 2023. "How Do Clinical Trials Work? - NCI." August 31, 2023. <https://www.cancer.gov/research/participate/clinical-trials/how-trials-work>.
- Nelson, Wayne. 1969. "Hazard Plotting for Incomplete Failure Data." *Journal of Quality Technology* 1 (1): 27–52. <https://doi.org/10.1080/00224065.1969.11980344>.
- Oba, Koji, Xavier Paoletti, Steven Alberts, Yung-Jue Bang, Jacqueline Benedetti, Harry Bleiberg, Paul Catalano, et al. 2013. "Disease-Free Survival as a Surrogate for Overall Survival in Adjuvant Trials of Gastric Cancer: A Meta-Analysis." *JNCI: Journal of the National Cancer Institute* 105 (21): 1600–1607. <https://doi.org/10.1093/jnci/djt270>.
- Osterman, Chelsea K., Hanna K. Sanoff, William A. Wood, Megan Fasold, and Jennifer Elston Lafata. 2022. "Predictive Modeling for Adverse Events and Risk Stratification Programs for People Receiving Cancer Treatment." *JCO Oncology Practice* 18 (2): 127–36. <https://doi.org/10.1200/OP.21.00198>.

- Overgaard, Morten, and Stefan Nygaard Hansen. 2021. "On the Assumption of Independent Right Censoring." *Scandinavian Journal of Statistics* 48 (4): 1234–55. <https://doi.org/10.1111/sjos.12487>.
- Peto, Richard, and Julian Peto. 1972. "Asymptotically Efficient Rank Invariant Test Procedures." *Journal of the Royal Statistical Society. Series A (General)* 135 (2): 185–207. <https://doi.org/10.2307/2344317>.
- Pinheiro, José C., and Douglas M. Bates. 2000. *Mixed-Effects Models in S and S-PLUS*. Statistics and Computing. New York: Springer. <https://doi.org/10.1007/b98882>.
- Pisaniello, Huai Leng, and William Gregory Dixon. 2020. "What Does Digitalization Hold for the Creation of Real-World Evidence?" *Rheumatology* 59 (1): 39–45. <https://doi.org/10.1093/rheumatology/kez068>.
- Podgorelec, Vili, Peter Kokol, Bruno Stiglic, and Ivan Rozman. 2002. "Decision Trees: An Overview and Their Use in Medicine." *Journal of Medical Systems* 26 (5): 445–63. <https://doi.org/10.1023/A:1016409317640>.
- Polley, Eric C., Mark J. van der Laan, Mark J. van der Laan, and Sherri Rose. 2011. "Super Learning for Right-Censored Data." In *Targeted Learning: Causal Inference for Observational and Experimental Data*, 249–58. New York, NY: Springer New York. https://doi.org/10.1007/978-1-4419-9782-1_16.
- Proctor, Michael J., Paul G. Horgan, Dinesh Talwar, Colin D. Fletcher, David S. Morrison, and Donald C. McMillan. 2013. "Optimization of the Systemic Inflammation-Based Glasgow Prognostic Score." *Cancer* 119 (12): 2325–32. <https://doi.org/10.1002/cncr.28018>.
- Qian, Xiaohua, Jiahui Wang, Shuxu Guo, and Qiang Li. 2013. "An Active Contour Model for Medical Image Segmentation with Application to Brain CT Image: An Active Contour Model for Medical Image Segmentation." *Medical Physics* 40 (2): 021911. <https://doi.org/10.1118/1.4774359>.
- Rajkomar, Alvin, Eyal Oren, Kai Chen, Andrew M. Dai, Nissan Hajaj, Michaela Hardt, Peter J. Liu, et al. 2018. "Scalable and Accurate Deep Learning with Electronic Health Records." *Npj Digital Medicine* 1 (1): 18. <https://doi.org/10.1038/s41746-018-0029-1>.
- Rhee, Hyungjin, Sang Hyun Choi, Ji Hoon Park, Eun-Suk Cho, Suk-Keu Yeom, Sumi Park, Kyunghwa Han, Seung Soo Lee, and Mi-Suk Park. 2022. "Preoperative Magnetic Resonance Imaging-Based Prognostic Model for Mass-Forming Intrahepatic Cholangiocarcinoma." *Liver International* 42 (4): 930–41. <https://doi.org/10.1111/liv.15196>.
- Ridgeway, Greg. 1999. "The State of Boosting." *Computing Science and Statistics*, 172–81.

- Rieke, Nicola, Jonny Hancox, Wenqi Li, Fausto Milletari, Holger R. Roth, Shadi Albarqouni, Spyridon Bakas, et al. 2020. "The Future of Digital Health with Federated Learning." *Npj Digital Medicine* 3 (1): 119. <https://doi.org/10.1038/s41746-020-00323-1>.
- Rittmeyer, Achim, Fabrice Barlesi, Daniel Waterkamp, Keunchil Park, Fortunato Ciardiello, Joachim von Pawel, Shirish M Gadgeel, et al. 2017. "Atezolizumab versus Docetaxel in Patients with Previously Treated Non-Small-Cell Lung Cancer (OAK): A Phase 3, Open-Label, Multicentre Randomised Controlled Trial." *The Lancet* 389 (10066): 255–65. [https://doi.org/10.1016/S0140-6736\(16\)32517-X](https://doi.org/10.1016/S0140-6736(16)32517-X).
- Rizopoulos, Dimitris. 2010. "JM: An R Package for the Joint Modelling of Longitudinal and Time-to-Event Data." *Journal of Statistical Software* 35 (9): 1–33. <https://doi.org/10.18637/jss.v035.i09>.
- . 2012. *Joint Models for Longitudinal and Time-to-Event Data: With Applications in R*. Chapman & Hall/CRC Biostatistics Series 6. Boca Raton: CRC Press.
- Rubin, Eric H., and D. Gary Gilliland. 2012. "Drug Development and Clinical Trials—the Path to an Approved Cancer Drug." *Nature Reviews Clinical Oncology* 9 (4): 215–22. <https://doi.org/10.1038/nrclinonc.2012.22>.
- Sato, Masaya, Ryosuke Tateishi, Yutaka Yatomi, and Kazuhiko Koike. 2021. "Artificial Intelligence in the Diagnosis and Management of Hepatocellular Carcinoma." *Journal of Gastroenterology and Hepatology* 36 (3): 551–60. <https://doi.org/10.1111/jgh.15413>.
- Savina, Marion, Sophie Gourgou, Antoine Italiano, Derek Dinart, Virginie Rondeau, Nicolas Penel, Simone Mathoulin-Pelissier, and Carine Bellera. 2018. "Meta-Analyses Evaluating Surrogate Endpoints for Overall Survival in Cancer Randomized Trials: A Critical Review." *Critical Reviews in Oncology/Hematology* 123 (March): 21–41. <https://doi.org/10.1016/j.critrevonc.2017.11.014>.
- Schaffer, Cullen. 1993. "Selecting a Classification Method by Cross-Validation." *Machine Learning* 13 (1): 135–43. <https://doi.org/10.1007/BF00993106>.
- Schmidli, Heinz, Dieter A. Häring, Marius Thomas, Adrian Cassidy, Sebastian Weber, and Frank Bretz. 2020. "Beyond Randomized Clinical Trials: Use of External Controls." *Clinical Pharmacology & Therapeutics* 107 (4): 806–16. <https://doi.org/10.1002/cpt.1723>.
- Segal, Mark Robert. 1988. "Regression Trees for Censored Data." *Biometrics* 44 (1): 35. <https://doi.org/10.2307/2531894>.
- Seo, Sangwoo, Youngmin Kim, Hyo-Jeong Han, Woo Chan Son, Zhen-Yu Hong, Insuk Sohn, Jooyong Shim, and Changha Hwang. 2021. "Predicting Successes and Failures of Clinical Trials With

- Outer Product–Based Convolutional Neural Network.” *Frontiers in Pharmacology* 12. <https://www.frontiersin.org/articles/10.3389/fphar.2021.670670>.
- Shameer, Khader, Youyi Zhang, Dan Jackson, Kirsty Rhodes, Imran Khan A. Neelufar, Sreenath Nampally, Andrzej Prokop, et al. 2021. “Correlation Between Early Endpoints and Overall Survival in Non-Small-Cell Lung Cancer: A Trial-Level Meta-Analysis.” *Frontiers in Oncology* 11. <https://www.frontiersin.org/articles/10.3389/fonc.2021.672916>.
- Shameer, Khader, Youyi Zhang, Andrzej Prokop, Sreenath Nampally, Imran Khan A. N, Jim Weatherall, Renee Bailey Iacona, and Faisal M. Khan. 2022. “OSPred Tool: A Digital Health Aid for Rapid Predictive Analysis of Correlations Between Early End Points and Overall Survival in Non–Small-Cell Lung Cancer Clinical Trials.” *JCO Clinical Cancer Informatics*, no. 6 (December): e2100173. <https://doi.org/10.1200/CCI.21.00173>.
- Sheppard, N.N., S. Hemington-Gorse, O.P. Shelley, B. Philp, and P. Dziewulski. 2011. “Prognostic Scoring Systems in Burns: A Review.” *Burns* 37 (8): 1288–95. <https://doi.org/10.1016/j.burns.2011.07.017>.
- Siegfried, Sandra, Stephen Senn, and Torsten Hothorn. 2023. “On the Relevance of Prognostic Information for Clinical Trials: A Theoretical Quantification.” *Biometrical Journal* 65 (1): 2100349. <https://doi.org/10.1002/bimj.202100349>.
- Simon, Noah, Jerome Friedman, Trevor Hastie, and Rob Tibshirani. 2011. “Regularization Paths for Cox’s Proportional Hazards Model via Coordinate Descent.” *Journal of Statistical Software* 39 (5): 1–13. <https://doi.org/10.18637/jss.v039.i05>.
- Singh, Ritesh, and Keshab Mukhopadhyay. 2011. “Survival Analysis in Clinical Trials: Basics and Must Know Areas.” *Perspectives in Clinical Research* 2 (4): 145. <https://doi.org/10.4103/2229-3485.86872>.
- Sondhi, Arjun, Janick Weberpals, Prakirthi Yerram, Chengsheng Jiang, Michael Taylor, Meghna Samant, and Sarah Cherng. 2023. “A Systematic Approach towards Missing Lab Data in Electronic Health Records: A Case Study in Non-Small Cell Lung Cancer and Multiple Myeloma.” *CPT: Pharmacometrics & Systems Pharmacology* 12 (9): 1201–12. <https://doi.org/10.1002/psp4.12998>.
- Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. “Dropout: A Simple Way to Prevent Neural Networks from Overfitting.” *Journal of Machine Learning Research* 15 (56): 1929–58.
- Stuart, Elizabeth A., Brian K. Lee, and Finbarr P. Leacy. 2013. “Prognostic Score–Based Balance Measures Can Be a Useful Diagnostic for Propensity Score Methods in Comparative Effectiveness Research.” *Methods for Comparative Effectiveness Research/Patient-Centered*

- Outcomes Research: From Efficacy to Effectiveness* 66 (8, Supplement): S84-S90.e1. <https://doi.org/10.1016/j.jclinepi.2013.01.013>.
- Sudlow, Cathie, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, et al. 2015. "UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age." *PLOS Medicine* 12 (3): e1001779. <https://doi.org/10.1371/journal.pmed.1001779>.
- Suissa, Samy. 2007. "Immortal Time Bias in Pharmacoepidemiology." *American Journal of Epidemiology* 167 (4): 492–99. <https://doi.org/10.1093/aje/kwm324>.
- Sullivan, Daniel Carl, Lawrence H. Schwartz, and Binsheng Zhao. 2013. "The Imaging Viewpoint: How Imaging Affects Determination of Progression-Free Survival." *Clinical Cancer Research* 19 (10): 2621–28. <https://doi.org/10.1158/1078-0432.CCR-12-2936>.
- Sung, Hyuna, Jacques Ferlay, Rebecca L. Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. 2021. "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries." *CA: A Cancer Journal for Clinicians* 71 (3): 209–49. <https://doi.org/10.3322/caac.21660>.
- Taleb B, Amina. 2019. "Tumour Flare Reaction in Cancer Treatments: A Comprehensive Literature Review." *Anti-Cancer Drugs* 30 (9): 953–58. <https://doi.org/10.1097/CAD.0000000000000814>.
- Tan, Aidan, Raphael Porcher, Perrine Crequit, Philippe Ravaud, and Agnes Dechartres. 2017. "Differences in Treatment Effect Size Between Overall Survival and Progression-Free Survival in Immunotherapy Trials: A Meta-Epidemiologic Study of Trials With Results Posted at ClinicalTrials.Gov." *Journal of Clinical Oncology* 35 (15): 1686–94. <https://doi.org/10.1200/JCO.2016.71.2109>.
- Tan, Katherine, Jonathan Bryan, Brian Segal, Lawrence Bellomo, Nate Nussbaum, Melisa Tucker, Aracelis Z. Torres, et al. 2022. "Emulating Control Arms for Cancer Clinical Trials Using External Cohorts Created From Electronic Health Record-Derived Real-World Data." *Clinical Pharmacology & Therapeutics* 111 (1): 168–78. <https://doi.org/10.1002/cpt.2351>.
- Taylor, David. 2015. "The Pharmaceutical Industry and the Future of Drug Development." In *Pharmaceuticals in the Environment*, edited by R E Hester and R M Harrison. The Royal Society of Chemistry. <https://doi.org/10.1039/9781782622345-00001>.
- Telloni, Stacy M. 2017. "Tumor Staging and Grading: A Primer." In *Molecular Profiling: Methods and Protocols*, edited by Virginia Espina, 1–17. New York, NY: Springer New York. https://doi.org/10.1007/978-1-4939-6990-6_1.

- Temple, R. 2010. "Enrichment of Clinical Study Populations." *Clinical Pharmacology & Therapeutics* 88 (6): 774–78. <https://doi.org/10.1038/clpt.2010.233>.
- Tibshirani, Robert. 1997. "The Lasso Method for Variable Selection in the Cox Model." *Statistics in Medicine* 16 (4): 385–95. [https://doi.org/10.1002/\(SICI\)1097-0258\(19970228\)16:4<385::AID-SIM380>3.0.CO;2-3](https://doi.org/10.1002/(SICI)1097-0258(19970228)16:4<385::AID-SIM380>3.0.CO;2-3).
- Tomašev, Nenad, Natalie Harris, Sebastien Baur, Anne Mottram, Xavier Glorot, Jack W. Rae, Michal Zielinski, et al. 2021. "Use of Deep Learning to Develop Continuous-Risk Models for Adverse Event Prediction from Electronic Health Records." *Nature Protocols* 16 (6): 2765–87. <https://doi.org/10.1038/s41596-021-00513-5>.
- Ton, Thanh G.N., Navdeep Pal, Huong Trinh, Sami Mahrus, Michael T. Bretscher, Robson J.M. Machado, Natalia Sadetsky, Nayan Chaudhary, Michael W. Lu, and Gregory J. Riely. 2022. "Replication of Overall Survival, Progression-Free Survival, and Overall Response in Chemotherapy Arms of Non–Small Cell Lung Cancer Trials Using Real-World Data." *Clinical Cancer Research* 28 (13): 2844–53. <https://doi.org/10.1158/1078-0432.CCR-22-0471>.
- Tran, Manuel, Yashin Dicente Cid, Amal Lahiani, Fabian Theis, Tingying Peng, and Eldad Klaiman. 2024. "Training Transitive and Commutative Multimodal Transformers with LoReTTa." *Advances in Neural Information Processing Systems* 36.
- Turkson, Anthony Joe, Francis Ayiah-Mensah, and Vivian Nimoh. 2021. "Handling Censoring and Censored Data in Survival Analysis: A Standalone Systematic Literature Review." Edited by Niansheng Tang. *International Journal of Mathematics and Mathematical Sciences* 2021 (September): 9307475. <https://doi.org/10.1155/2021/9307475>.
- U. S. Food and Drug Administration. 2019a. "Step 2: Preclinical Research." FDA. FDA. April 18, 2019. <https://www.fda.gov/patients/drug-development-process/step-2-preclinical-research>.
- . 2019b. "Step 3: Clinical Research." FDA. FDA. April 18, 2019. <https://www.fda.gov/patients/drug-development-process/step-3-clinical-research>.
- Unger, Joseph M., Elise Cook, Eric Tai, and Archie Bleyer. 2016. "The Role of Clinical Trial Participation in Cancer Research: Barriers, Evidence, and Strategies." *American Society of Clinical Oncology Educational Book*, no. 36 (May): 185–98. https://doi.org/10.1200/EDBK_156686.
- Uno, Hajime, Tianxi Cai, Michael J. Pencina, Ralph B. D'Agostino, and L. J. Wei. 2011. "On the C-Statistics for Evaluating Overall Adequacy of Risk Prediction Procedures with Censored Survival Data." *Statistics in Medicine* 30 (10): 1105–17. <https://doi.org/10.1002/sim.4154>.

- U.S. Food and Drug Administration. 2018. “Clinical Trial Endpoints for the Approval of Cancer Drugs and Biologics.” <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/clinical-trial-endpoints-approval-cancer-drugs-and-biologics>.
- . 2019. “Enrichment Strategies for Clinical Trials to Support Approval of Human Drugs and Biological Products.” FDA. <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/enrichment-strategies-clinical-trials-support-approval-human-drugs-and-biological-products>.
- Viele, Kert, Scott Berry, Beat Neuenschwander, Billy Amzal, Fang Chen, Nathan Enas, Brian Hobbs, et al. 2014. “Use of Historical Control Data for Assessing Treatment Effects in Clinical Trials.” *Pharmaceutical Statistics* 13 (1): 41–54. <https://doi.org/10.1002/pst.1589>.
- Villaruz, Liza C., and Mark A. Socinski. 2013. “The Clinical Viewpoint: Definitions, Limitations of RECIST, Practical Considerations of Measurement.” *Clinical Cancer Research* 19 (10): 2629–36. <https://doi.org/10.1158/1078-0432.CCR-12-2935>.
- Wang, Xiaomeng, Flavio Dormont, Christelle Lorenzato, Aurélien Latouche, Ramon Hernandez, and Roman Rouzier. 2023. “Current Perspectives for External Control Arms in Oncology Clinical Trials: Analysis of EMA Approvals 2016–2021.” *Journal of Cancer Policy* 35 (March): 100403. <https://doi.org/10.1016/j.jcpo.2023.100403>.
- Wang, Xuan, Raquel Lahoz, Shantanu Jawla, Raymond Przybysz, Kristijan H. Kahler, Lisa Burdukova, Shiva Kumar Venkata, Maria Nassim, Anil Jalapu, and Nahila Justo. 2019. “Identification and Mapping of Worldwide Sources of Generic Real-World Data.” *Pharmacoepidemiology and Drug Safety* 28 (7): 899–905. <https://doi.org/10.1002/pds.4782>.
- Warr, Julia, Amanda E. Hird, Carlo DeAngelis, Angie Giotis, and Yoo-Joung Ko. 2013. “Baseline Blood Work Before Initiation of Chemotherapy: What Is Safe in the Real World?” *Journal of Oncology Practice* 9 (5): e182–85. <https://doi.org/10.1200/JOP.2012.000719>.
- Weberpals, Janick, Tim Becker, Jessica Davies, Fabian Schmich, Dominik Rüttinger, Fabian J. Theis, and Anna Bauer-Mehren. 2021. “Deep Learning-Based Propensity Scores for Confounding Control in Comparative Effectiveness Research: A Large-Scale, Real-World Data Study.” *Epidemiology* 32 (3): 378–88. <https://doi.org/10.1097/EDE.0000000000001338>.
- Weberpals, Janick, Lina Jansen, Myrthe P. P. van Herk-Sukel, Josephina G. Kuiper, Mieke J. Aarts, Pauline A. J. Vissers, and Hermann Brenner. 2017. “Immortal Time Bias in Pharmacoepidemiological Studies on Cancer Patient Survival: Empirical Illustration for Beta-Blocker Use in Four Cancers with Different Prognosis.” *European Journal of Epidemiology* 32 (11): 1019–31. <https://doi.org/10.1007/s10654-017-0304-5>.
- Webster-Clark, Michael, Til Stürmer, Tiansheng Wang, Kenneth Man, Danica Marinac-Dabic, Kenneth J. Rothman, Alan R. Ellis, et al. 2021. “Using Propensity Scores to Estimate Effects of

- Treatment Initiation Decisions: State of the Science.” *Statistics in Medicine* 40 (7): 1718–35. <https://doi.org/10.1002/sim.8866>.
- Weiss, Romano, Sanaz Karimijafarbigloo, Dirk Roggenbuck, and Stefan Rödiger. 2022. “Applications of Neural Networks in Biomedical Data Analysis.” *Biomedicines* 10 (7). <https://doi.org/10.3390/biomedicines10071469>.
- West, Howard, Michael McCleod, Maen Hussein, Alessandro Morabito, Achim Rittmeyer, Henry J Conter, Hans-Georg Kopp, et al. 2019. “Atezolizumab in Combination with Carboplatin plus Nab-Paclitaxel Chemotherapy Compared with Chemotherapy Alone as First-Line Treatment for Metastatic Non-Squamous Non-Small-Cell Lung Cancer (IMpower130): A Multicentre, Randomised, Open-Label, Phase 3 Trial.” *The Lancet Oncology* 20 (7): 924–37. [https://doi.org/10.1016/S1470-2045\(19\)30167-6](https://doi.org/10.1016/S1470-2045(19)30167-6).
- West, Stephen G., Heining Cham, Felix Thoemmes, Babette Renneberg, Julian Schulze, and Matthias Weiler. 2014. “Propensity Scores as a Basis for Equating Groups: Basic Principles and Application in Clinical Treatment Outcome Research.” *Journal of Consulting and Clinical Psychology* 82 (5): 906–19. <https://doi.org/10.1037/a0036387>.
- Westreich, Daniel, Stephen R. Cole, Michele Jonsson Funk, M. Alan Brookhart, and Til Stürmer. 2011. “The Role of the C-Statistic in Variable Selection for Propensity Score Models.” *Pharmacoepidemiology and Drug Safety* 20 (3): 317–20. <https://doi.org/10.1002/pds.2074>.
- Wharton, Gerold T., Claudia Becker, Dimitri Bennett, Mehmet Burcu, Greta Bushnell, Carmen Ferrajolo, Sigal Kaplan, et al. 2023. “Overview of Global Real-World Data Sources for Pediatric Pharmacoepidemiologic Research.” *Pharmacoepidemiology and Drug Safety* n/a (n/a). <https://doi.org/10.1002/pds.5695>.
- Williamson, Elizabeth J., and Andrew Forbes. 2014. “Introduction to Propensity Scores.” *Respirology* 19 (5): 625–35. <https://doi.org/10.1111/resp.12312>.
- Wilson, Brooke E., Michelle B. Nadler, Alexandra Desnoyers, and Eitan Amir. 2021. “Quantifying Withdrawal of Consent, Loss to Follow-Up, Early Drug Discontinuation, and Censoring in Oncology Trials.” *Journal of the National Comprehensive Cancer Network* 19 (12): 1433–40. <https://doi.org/10.6004/jnccn.2021.7015>.
- Wong, Chi Heem, Kien Wei Siah, and Andrew W Lo. 2019. “Estimation of Clinical Trial Success Rates and Related Parameters.” *Biostatistics* 20 (2): 273–86. <https://doi.org/10.1093/biostatistics/kxx069>.
- Wouters, Olivier J., Martin McKee, and Jeroen Luyten. 2020. “Estimated Research and Development Investment Needed to Bring a New Medicine to Market, 2009-2018.” *JAMA* 323 (9): 844–53. <https://doi.org/10.1001/jama.2020.1166>.

- Wu, Stephen, Kirk Roberts, Surabhi Datta, Jingcheng Du, Zongcheng Ji, Yuqi Si, Sarvesh Soni, et al. 2020. "Deep Learning in Clinical Natural Language Processing: A Methodical Review." *Journal of the American Medical Informatics Association* 27 (3): 457–70. <https://doi.org/10.1093/jamia/ocz200>.
- Wyss, Richard, Sebastian Schneeweiss, Mark van der Laan, Samuel D Lendle, Cheng Ju, and Jessica M Franklin. 2018. "Using Super Learner Prediction Modeling to Improve High-Dimensional Propensity Score Estimation." *Epidemiology* 29 (1): 96–106. <https://doi.org/10.1097/ede.0000000000000762>.
- Yap, Timothy A., Ira Jacobs, Elodie Baumfeld Andre, Lauren J. Lee, Darrin Beaupre, and Laurent Azoulay. 2022. "Application of Real-World Data to External Control Groups in Oncology Clinical Trial Drug Development." *Frontiers in Oncology* 11. <https://doi.org/10.3389/fonc.2021.695936>.
- Ye, Jiabu, Xiang Ji, Phillip A. Dennis, Hesham Abdullah, and Pralay Mukhopadhyay. 2020. "Relationship Between Progression-Free Survival, Objective Response Rate, and Overall Survival in Clinical Trials of PD-1/PD-L1 Immune Checkpoint Blockade: A Meta-Analysis." *Clinical Pharmacology & Therapeutics* 108 (6): 1274–88. <https://doi.org/10.1002/cpt.1956>.
- Yoshida, Kazuki, Daniel H. Solomon, and Seoyoung C. Kim. 2015. "Active-Comparator Design and New-User Design in Observational Studies." *Nature Reviews Rheumatology* 11 (7): 437–41. <https://doi.org/10.1038/nrrheum.2015.30>.
- Zhu, Jie, and Blanca Gallego. 2020. "Targeted Estimation of Heterogeneous Treatment Effect in Observational Survival Analysis." *Journal of Biomedical Informatics* 107 (July): 103474. <https://doi.org/10.1016/j.jbi.2020.103474>.
- Zhuang, Sen H., Liang Xiu, and Yusri A. Elsayed. 2009. "Overall Survival: A Gold Standard in Search of a Surrogate: The Value of Progression-Free Survival and Time to Progression as End Points of Drug Efficacy." *The Cancer Journal* 15 (5). <https://doi.org/10.1097/PPO.0b013e3181be231d>.
- Zou, James, Mikael Huss, Abubakar Abid, Pejman Mohammadi, Ali Torkamani, and Amalio Telenti. 2019. "A Primer on Deep Learning in Genomics." *Nature Genetics* 51 (1): 12–18. <https://doi.org/10.1038/s41588-018-0295-5>.

Appendix

Full length manuscript I - Artificial Intelligence for Prognostic Scores in Oncology: A Benchmarking Study

Hugo Loureiro, Tim Becker, Anna Bauer-Mehren, Narges Ahmidi* and Janick Weberpals* (* contributed equally)

The manuscript was peer-reviewed and is published under the Creative Commons Attribution License (CC BY 4.0) license in the *Frontiers in Artificial Intelligence* journal. Hence, the article was published open-access, and the published version is included.



Artificial Intelligence for Prognostic Scores in Oncology: a Benchmarking Study

Hugo Loureiro^{1,2,3}, Tim Becker¹, Anna Bauer-Mehren^{1*}, Narges Ahmidi^{2†} and Janick Weberpals^{1†}

¹Data Science, Pharmaceutical Research and Early Development Informatics (pREDI), Roche Innovation Center Munich (RICM), Penzberg, Germany, ²Institute of Computational Biology, Helmholtz Zentrum Munich, Munich, Germany, ³TUM School of Life Sciences Weihenstephan, Technical University of Munich, Freising, Germany

OPEN ACCESS

Edited by:

Enkelejda Miho,
University of Applied Sciences and
Arts Northwestern Switzerland,
Switzerland

Reviewed by:

Gregory R Hart,
Yale University, United States
Raghendra Mall,
Qatar Computing Research Institute,
Qatar

*Correspondence:

Anna Bauer-Mehren
anna.bauer-mehren@roche.com

[†]These authors have contributed
equally to this work

Specialty section:

This article was submitted to
Medicine and Public Health,
a section of the journal
Frontiers in Artificial Intelligence

Received: 03 November 2020

Accepted: 19 January 2021

Published: 16 April 2021

Citation:

Loureiro H, Becker T, Bauer-Mehren A,
Ahmidi N and Weberpals J (2021)
Artificial Intelligence for Prognostic
Scores in Oncology: a
Benchmarking Study.
Front. Artif. Intell. 4:625573.
doi: 10.3389/frai.2021.625573

Introduction: Prognostic scores are important tools in oncology to facilitate clinical decision-making based on patient characteristics. To date, classic survival analysis using Cox proportional hazards regression has been employed in the development of these prognostic scores. With the advance of analytical models, this study aimed to determine if more complex machine-learning algorithms could outperform classical survival analysis methods.

Methods: In this benchmarking study, two datasets were used to develop and compare different prognostic models for overall survival in pan-cancer populations: a nationwide EHR-derived de-identified database for training and in-sample testing and the OAK (phase III clinical trial) dataset for out-of-sample testing. A real-world database comprised 136K first-line treated cancer patients across multiple cancer types and was split into a 90% training and 10% testing dataset, respectively. The OAK dataset comprised 1,187 patients diagnosed with non-small cell lung cancer. To assess the effect of the covariate number on prognostic performance, we formed three feature sets with 27, 44 and 88 covariates. In terms of methods, we benchmarked ROPRO, a prognostic score based on the Cox model, against eight complex machine-learning models: regularized Cox, Random Survival Forests (RSF), Gradient Boosting (GB), DeepSurv (DS), Autoencoder (AE) and Super Learner (SL). The C-index was used as the performance metric to compare different models.

Results: For in-sample testing on the real-world database the resulting C-index [95% CI] values for RSF 0.720 [0.716, 0.725], GB 0.722 [0.718, 0.727], DS 0.721 [0.717, 0.726] and lastly, SL 0.723 [0.718, 0.728] showed significantly better performance as compared to ROPRO 0.701 [0.696, 0.706]. Similar results were derived across all feature sets. However, for the out-of-sample validation on OAK, the stronger performance of the more complex models was not apparent anymore. Consistently, the increase in the number of prognostic covariates did not lead to an increase in model performance.

Discussion: The stronger performance of the more complex models did not generalize when applied to an out-of-sample dataset. We hypothesize that future research may benefit by adding multimodal data to exploit advantages of more complex models.

Keywords: electronic health records, machine learning, prognostic scores, real world data, survival analysis

INTRODUCTION

With an estimated incidence of 18.1 million new cases and 9.6 million deaths worldwide in 2018, cancer is still one of the biggest healthcare challenges today (Ferlay et al., 2019). New paradigms such as cancer immunotherapy have led to an increase in survival for several hematological (Sant et al., 2014) and solid tumors (Pulte et al., 2019). Still, drug development in general, including in oncology, suffers from a high attrition rate. Most drugs (97%) fail during early development phases, a process that is both time-consuming (median duration of phase one clinical is 1.6 years) and costly (as much as \$42,000 per patient) (Fogel, 2018; Wong et al., 2019). One of the reasons for such failures may be rooted in a suboptimal enrollment of patients in clinical trials. As a consequence, patients may dropout early due to adverse events, lack of tolerability and/or lack of efficacy which might lead to an early failure of potentially effective drugs (Fogel, 2018). In this context, an accurate characterization of the patients' recovery (or response to medications) given their prognostic factors is key. Currently, the patients' prognostic factors are used to determine 1) clinical trial eligibility, 2) toxicity monitoring and 3) treatment decisions. Furthermore, prognostic factors allow us to gain a deeper understanding of disease biology and thus may contribute to the development of more effective treatments (Bhimani et al., 2019).

To date, several prognostic scores in oncology have been published, such as the Royal Marsden Hospital Score (Arkenau et al., 2009), the international prognostic index (International Non-Hodgkin's Lymphoma Prognostic Factors Project, 1993), the IMDC risk model (Ko et al., 2015) or the Glasgow prognostic score (Kinoshita et al., 2013). Due to prior lack of access to large-scale patient data, the previous prognostic scores were significantly limited on the modeling approaches. Additionally, previous databases also usually contained a small number of covariates, which typically were cast into a simple counting scheme (number of covariates above a threshold).

As a major enhancement, the ROPRO was introduced recently (Becker et al., 2020). The ROPRO is a new pan-cancer prognostic score developed from more than 125k patients in the EHR-derived de-identified database which consists of 27 highly prognostic covariates for overall survival. This prognostic score is based on the Cox proportional hazards model (in the following referred to as Cox model) (Becker et al., 2020), a widely used survival analysis model. In (Becker et al., 2020), ROPRO showed an increased prognostic power when compared to the aforementioned scores and was validated in independent clinical data. In general, the Cox model cannot model nonlinearities or interaction effects, unless all of these effects are explicitly specified (Harrell et al., 1996). While the ROPRO is a multivariate model it does not include covariate interactions and possibly could have missed nonlinearities in the covariates.

To overcome the Cox model's limitations, recent models such as the regularized Cox model (Tibshirani 1997; Simon et al., 2011), random survival forests (Ishwaran et al., 2008), gradient

boosting (Ridgeway 1999) and DeepSurv (Katzman et al., 2018) a deep neural network-modified version of the Cox model have been introduced.

Several studies (Chen et al., 2019; Christodoulou et al., 2019; Desai et al., 2020; Kim et al., 2019; Steele et al., 2018) have been published that compare the prognostic/predictive performance of some of these new survival models. Still, there remains the need for a more systematic and direct comparison. Hence, the objective of this study is to compare the prediction performance of a set of models with respect to model complexity and automated covariate selection. We aimed to address model complexity by implementing more complex survival models (regularized Cox (Tibshirani 1997; Simon et al., 2011), Random Survival Forests (Ishwaran et al., 2008), Gradient Boosting (Ridgeway 1999), DeepSurv (Katzman et al., 2018), a new autoencoder based model (Goodfellow et al., 2016) and Super Learner (van der Laan et al., 2007)) and compared them against the classical model (ROPRO (Becker et al., 2020)). To address the automated covariate selection, we investigated whether an increase in the covariate number, even though not present for all patients, led to an increase in model performance.

MATERIALS AND METHODS

Datasets

In this study we used two databases: 1) the nationwide Flatiron Health (FH) electronic health record (EHR)-derived de-identified database and 2) OAK clinical trial database. During the study period, the FH database included de-identified patient-level structured and unstructured data, curated via technology-enabled abstraction (Birnbaum et al., 2020; Ma et al., 2020) and includes data from over 280 cancer clinics (~800 sites of care); Institutional Review Board approval of the FH study protocol was obtained prior to study conduct, and included a waiver of informed consent. The OAK dataset was derived from a

TABLE 1 | Number of patients per cohort in the FH dataset. Includes both train and test datasets.

Cohort	Patient number
Advanced endometrial	1,641
Advanced melanoma	4,332
Advanced non-small-cell lung cancer	38,201
Acute myeloid leukemia	2,232
Bladder cancer	5,363
Chronic lymphocytic leukemia	9,544
Diffuse large B-cell lymphoma	3,969
Breast cancer	655
Follicular cancer	1,958
Gastric cancer	6,212
Head and neck cancer	4,917
Metastatic breast cancer	14,429
Metastatic colorectal cancer	16,788
Metastatic renal cell carcinoma	5,116
Multiple myeloma	7,293
Ovarian cancer	4,407
Pancreatic cancer	6,212
Small-cell lung cancer	4,918

TABLE 2 | Summary statistics of the datasets.

	FH train	FH test	OAK
Number of patients	121,644	15,075	1,187
Time [months] (median (95% CI))	19.33 (19.10–19.57)	19.83 (19.33–20.57)	11.43 (10.40–12.67)
Event = death (%)	72,068 (59.2)	8,875 (58.8)	854 (71.9)
Age at baseline [years] (mean (SD))	66.47 (10.98)	66.47 (11.05)	62.79 (9.57)
History of smoking [yes/no] (mean (SD))	0.84 (0.37)	0.84 (0.37)	0.83 (0.37)
Group stage (mean (SD))	3.31 (0.85)	3.31 (0.85)	3.43 (0.89)
ECOG value (mean (SD))	0.81 (0.80)	0.81 (0.80)	0.64 (0.49)
Neutrophils-lymphocytes ratio (NLR) [%] (mean (SD))	4.90 (4.86)	4.82 (4.66)	6.59 (6.31)
Body Mass index (BMI) [kg/m ²] (mean (SD))	27.05 (5.96)	27.06 (5.92)	25.17 (4.80)
Number of metastasis sites (mean (SD))	0.37 (0.79)	0.36 (0.76)	1.46 (0.94)
Gender = male (%)	60,674 (49.9)	7,467 (49.5)	737 (62.1)
Alanine aminotransferase [enzymatic activity/volume] in serum or plasma [U/L] (mean (SD))	26.44 (29.47)	26.36 (29.24)	21.05 (13.80)
Calcium [mass/volume] in serum or plasma [mg/dL] (mean (SD))	9.33 (0.63)	9.33 (0.63)	9.40 (0.57)
Bilirubin total [mass/volume] in serum or plasma [mg/dL] (mean (SD))	0.57 (0.69)	0.56 (0.63)	0.47 (0.51)
Glucose [mass/volume] in serum or plasma [mg/dL] (mean (SD))	117.58 (34.19)	117.61 (34.35)	114.87 (33.02)
Protein [mass/volume] in serum or plasma [g/L] (mean (SD))	68.72 (7.16)	68.70 (7.18)	71.66 (6.61)
Urea nitrogen [mass/volume] in serum or plasma [mg/dL] (mean (SD))	17.87 (9.16)	17.81 (8.99)	26.37 (22.34)
Alkaline phosphatase [enzymatic activity/volume] in serum or plasma [U/L] (mean (SD))	114.71 (96.77)	114.57 (97.61)	118.84 (81.31)
Hemoglobin [mass/volume] in blood [g/dL] (mean (SD))	12.06 (1.97)	12.06 (1.96)	12.25 (1.67)
Chloride [moles/volume] in serum or plasma [mmol/L] (mean (SD))	101.17 (4.39)	101.14 (4.32)	101.18 (3.99)
Eosinophils/100 leukocytes in blood [%] (mean (SD))	2.54 (2.24)	2.55 (2.20)	2.59 (2.45)
Platelets [# /volume] in blood by automated count [10 ⁹ /L] (mean (SD))	264.80 (108.88)	265.96 (108.73)	281.13 (95.46)
Albumin [mass/volume] in serum or plasma [g/L] (mean (SD))	37.86 (5.39)	37.88 (5.38)	38.61 (5.70)
Lactate dehydrogenase [enzymatic activity/volume] in serum or plasma [U/L] (mean (SD))	278.18 (187.27)	276.51 (188.68)	295.28 (181.16)
Lymphocytes/100 leukocytes in blood by automated count [%] (mean (SD))	21.35 (13.11)	21.37 (13.14)	19.43 (9.43)
Monocytes [# /volume] in blood by automated count [10 ⁹ /L] (mean (SD))	0.68 (0.45)	0.68 (0.43)	0.65 (0.34)
Systolic blood pressure (mean (SD))	128.58 (19.36)	129.00 (19.19)	123.94 (16.92)
Heart rate (mean (SD))	83.18 (15.98)	83.25 (16.08)	84.38 (13.86)
Oxygen saturation in arterial blood by pulse oximetry [%] (mean (SD))	96.32 (2.39)	96.35 (2.35)	^a
AST/ALT ratio [%] (mean (SD))	1.25 (0.63)	1.25 (0.63)	1.31 (0.61)

^aThis covariate was not available in OAK.

phase III clinical trial (Rittmeyer et al., 2017) that evaluated the efficacy and safety of Atezolizumab monotherapy against a Docetaxel monotherapy in 1,187 patients with locally advanced or metastatic non-small cell lung cancer (NSCLC) after the failure of platinum based chemotherapy.

From FH we derived a cohort with 136,719 patients across 18 different primary cancers (Table 1). The majority of patients were diagnosed with advanced non-small cell lung cancer (38,201–26.7%), followed by metastatic colorectal cancer (16,788–12.1%) and metastatic breast cancer (14,429–10.4%). We randomly split the samples in the FH dataset into train (90% - 121,644) and in-sample test (10% - 15,075) sets. In case the model required a validation dataset (e.g., neural network based models), the training set was further divided into subsets of 90% for training and 10% for validation. The OAK study (1,187 patients) was used exclusively for out-of-sample testing.

In terms of covariates used per sample, we created three feature sets with differing numbers of covariates that could be used for modeling by the respective method. The first feature set contained 27 covariates of FH inspired from (Becker et al., 2020) (Table 2). The second feature set consisted of 44 covariates that were present in at least 30% of patients in FH, and the third feature set comprised almost all covariates (88 covariates present for at least 1% of the FH patients). The 88 and 44 feature sets included all the covariates of the 44 and 27 feature sets, respectively (a complete list of the covariates in each set is

available in **Supplementary Table S1**). The OAK dataset contained all the covariates used in the 27 covariates feature set except oxygen saturation in blood. In the 44 and 88 feature sets it was in addition lacking information on some covariates as compared to the FH dataset (for a complete list see **Supplementary Table S2**).

To prepare the data for methods that require a full data matrix, all datasets were imputed with random forests by using the R package missForest (Stekhoven and Bühlmann, 2011). To prevent leakage of information between train and test sets, the imputation (random forest) was trained only on the train sets, and then applied to the FH test set and OAK test set.

Models

One of our objectives in this paper was to determine if more complex survival models, that capture nonlinearities and feature dependence, are capable of predicting the patient's risk better than the state of the art prognostic scores that are based on the classical Cox model. We selected the ROPRO (a Cox based model) as our baseline model and compared it against the regularized Cox model (Tibshirani 1997), random survival forest (Ishwaran et al., 2008), gradient boosting (Ridgeway, 1999), DeepSurv (Katzman et al., 2018) and a (to our knowledge) new autoencoder-based survival model (Goodfellow et al., 2016). In addition, we extended the super learner (van der Laan et al., 2007) framework to survival analysis

problems and used it to aggregate the previous models into an ensemble that combined the predictions of all models, yielding a new weighted prediction.

Generally speaking, in survival analysis, the response variable is the time until an event occurs, such as death (Kalbfleisch and Prentice, 2002). If T represents a non-negative random variable that represents the time until the event, then the cumulative distribution function of T is called the survival function S . This function measures the probability of the event occurring after time t and is defined as

$$S(t) = P(T \geq t), t \geq 0.$$

The hazard function h is an alternative representation of the distribution of T . It represents the instantaneous rate of occurrence of the event at time t and is defined as

$$h(t) = \lim_{dt \rightarrow 0} \frac{P(t \leq T < t+dt | T \geq t)}{dt}.$$

The selected models in this paper all follow the underlying structure, but estimate the hazard function using different techniques.

Multivariate Cox Model on Main Effects (ROPRO)

The ROPRO, introduced in (Becker et al., 2020), is a prognostic score based on the Cox model. The Cox model (Cox, 1972) is a widely used model in survival analysis that estimates the hazard function based on a set of given covariates of the population. It assumes that the hazard function $h(t)$ is composed of two terms: a baseline hazard $h_0(t)$ that does not depend on the covariates and an exponential risk term $e^{r(X)} = e^{\beta X}$:

$$h(t|X) = h_0(t) \cdot e^{\beta X},$$

where X is the covariate vector and β are the model weights. The risk term integrates the interaction between the covariates and the hazard of each patient. In the case of the Cox model, the fitting focuses on the risk $r(X) = \beta X$, which is a linear function, using the following partial likelihood cost function:

$$\begin{aligned} \log PL(\beta) &= \sum_{i=1}^n \delta_i \left[r(X_i) - \log \left(\sum_{l \in R(T_i)} e^{r(X_l)} \right) \right] \\ &= \sum_{i=1}^n \delta_i \left[\beta X_i - \log \left(\sum_{l \in R(T_i)} e^{\beta X_l} \right) \right], \end{aligned}$$

where δ_i is the censoring indicator. It is 1 if the patient has faced the event by the end of data collection and 0 otherwise. Naturally, being a linear function, it cannot implicitly deal with nonlinearities or interaction effects between the covariates (Harrell et al., 1996). This is one of the pitfalls of the Cox model and one of the reasons that motivated the creation of other more complex models (Ridgeway 1999; Katzman et al., 2018).

The authors of ROPRO started with a Cox model with 44 covariates and applied backward selection, removing the least

significant covariates, until a total of 27 covariates remained in the model. The 27 selected covariates are represented in **Table 2**. In this work, we used the ROPRO formula as published in (Becker et al., 2020).

Regularized Cox Model

The regularized Cox is a modification of the Cox proportional hazards algorithm where a regularization term is added to the cost function (Tibshirani 1997; Simon et al., 2011). The new regularized cost function has the form

$$\begin{aligned} \log PL(\beta)_{RC} &= \sum_{i=1}^n \delta_i \left[\beta X_i - \log \left(\sum_{l \in R(T_i)} e^{\beta X_l} \right) \right] \\ &+ \lambda \left(\alpha \sum_{j=1}^p |\beta_j| + \frac{1}{2} (1 - \alpha) \sum_{j=1}^p \beta_j^2 \right). \end{aligned}$$

The regularization term forces a penalization to the model weights β . The penalization depends on the type of regularization. The L_1 regularization (Lasso) performs covariate selection by setting some of the β values to 0, effectively removing them from the model (Tibshirani, 1997). L_2 regularization (ridge regression) scales the β values toward 0 but does not perform covariate selection, i.e. does not set the β to exactly 0. The elastic net combines L_1 and L_2 . The parameter α determines which type of regularization is used, $\alpha = 0$ is the ridge regression, $\alpha = 1$ is Lasso, and values in between are the elastic net. Naturally, for values of α closer to 0 and 1, elastic net behaves more similar to ridge regression and Lasso, respectively.

Gradient Boosting

Gradient boosting (GB) is a machine learning algorithm used in classification and regression problems (Friedman 2001). It builds the predictive model in an iterative fashion, in each iteration adding a weak learner that reflects the current residuals. By doing so, in each iteration the model should fit better to the data and consequently, reduce the prediction error.

GB can be applied to survival analysis by using the Cox partial likelihood (Cox 1972) as the cost function to determine the residuals (Ridgeway 1999). The new GB partial likelihood has the form

$$\log PL(\theta)_{GB} = \sum_{i=1}^n \delta_i \left[\hat{r}_{GB}(X_i) - \log \left(\sum_{l \in R(T_i)} e^{\hat{r}_{GB}(X_l)} \right) \right].$$

Notice that the Cox model risk $r(X)$ was substituted by $\hat{r}_{GB}(X)$, the predicting function fitted by GB. This predicting function is composed of multiple regression trees. Each of them fit on the residuals of the model of the previous iteration:

$$\hat{r}_{GB}(X) = \sum_{k=1}^K \rho_k f_k(X),$$

where $f_k(x)$ corresponds to the model added in iteration k . As more models are added to the predictive model $\hat{r}_{GB}(x)$, the hazard function is estimated better (Ridgeway 1999).

Random Survival Forest

Random survival forests (RSF) is a machine learning method that fits an ensemble of regression trees, a “forest”, that estimates the cumulative hazard function (Ishwaran et al., 2008). At each tree node, a covariate is used to separate the patients into groups. The RSF selects the split condition that maximizes the difference between the survival curves of the groups. Each tree is grown until it is not possible to create a new split that has more than a pre-specified number of unique events in each node.

DeepSurv

DeepSurv uses a feed-forward neural network to predict the patient’s hazard $h(t|X)$ (Katzman et al., 2018). It is composed of multiple fully connected layers that combine the covariates in a nonlinear way. In the final layer the predicted nonlinear risk function $\hat{r}_{DS}(t|x)$ is yielded. The loss function used to fit the model is based on the Cox partial hazard:

$$\log PL(\theta)_{DS} = \sum_{i=1}^n \delta_i \left[\hat{r}_{DS}(t|X_i) - \log \left(\sum_{l \in R(T_i)} e^{\hat{r}_{DS}(t|X_l)} \right) \right] + \lambda \| \beta \|_2^2.$$

Autoencoder

Autoencoders (AE) are unsupervised neural networks composed of two components: 1) an encoder function that transforms the input X into an latent representation Z and 2) a decoder that transforms Z to $X_{reconstructed}$ (Goodfellow et al., 2016). The autoencoder is trained to minimize the difference between $X_{reconstructed}$ and X .

Here we exploit the autoencoder to perform dimensionality reduction. By setting Z to a lower dimensionality than X the autoencoder learns a representation that can best reconstruct X .

The autoencoder does not model the hazard function directly. Therefore, we use a Cox model to estimate the hazard from the intermediate representation. The new hazard is given by

$$h(t|Z) = h_0(t) \cdot e^{Z\beta}.$$

Super Learner

Above we introduced multiple models that are capable of predicting the hazard function. All these algorithms have distinct structures, leading to different strengths and weaknesses in their estimation capability. The Super Learner (SL) (van der Laan et al., 2007) offers a framework to combine these models into a single model with the aim of combining the strengths and mitigating the weaknesses.

The SL was originally proposed to handle classification and regression problems. In this work we extend it to address time-to-event data and to use all the models described above.

Consider the dataset $O_i = (X_i, T_i, \delta_i) \sim P_0, i = 1, \dots, n$ and the parameter of interest $\psi_0(X)$ which minimizes the cost function $L(O, \psi)$ such that

$$\psi_0 = \arg \min_{\psi \in \Psi} E_0 L(O, \psi)$$

In this particular problem, $\psi_0(X)$ is a function that estimates the risk of a patient given its covariates. The SL framework uses

V -fold cross-validation to split the dataset O into V distinct train-validation sets denoted by $P(v)$ and $V(v)$, respectively.

We learn a hazard function $\hat{h}_{k,v}$ from a given model k (e.g. DeepSurv) and a given training set $P(v)$, and further test that model on the validation set $V(v)$ to acquire predictions for each patient in $V(v)$. Repeating this process for all v -s, the predicted hazards of model k are concatenated to form \hat{h}_k . This process is repeated for all $k = \{1, \dots, 11\}$ models in our study. See **Table 3** for the list of 11 models used to inform the SL model.

The next step in SL is to combine the predicted hazards of all k models to learn a new hazard function. This is done by using a linear model of the form

$$\hat{h}_{SL}(t|x) = \sum_{k=1}^K \alpha_k \cdot \hat{h}_k(t|X),$$

where \hat{h}_{SL} is the predicted SL hazard and α_k are the weights of the linear model. In the original SL, the weights α_k can be modeled in a variety of ways, e.g. Least Squares (van der Laan et al., 2007) or area under the curve (AUC) (LeDell et al., 2016). Our approach is based on (LeDell et al., 2016) but instead of using the AUC, we use the C-index (Harrell et al., 1982) as the objective function and maximize it using the L-BFGS-B algorithm (Byrd et al., 1995).

Hyperparameter Tuning

As listed in **Table 3**, the more complex models require hyperparameters that adjust their complexity. Depending on the model, these hyperparameters were tuned by either cross-validation (for the regularized Cox models) or grid-search (for GB, RSF, DS and AE) on the FH training set.

Model Testing

All models were fit using the training set and tested using the two distinct testing sets: FH in-sample test and OAK out-of-sample test set. The ROPRO model was taken pre-trained from the formula published in (Becker et al., 2020) and was not trained again, however it was tested equally against our test sets.

To assess the discrimination performance of the models, we used Harrell’s C-index (C-index) (Harrell et al., 1982). The C-index is a generalization of the AUC. It is a goodness of fit measure for survival models and measures the concordance between the risk/hazard values given by the model and the time-to-event. More specifically, it measures if patients that died earlier in time have a higher risk score than patients that died later. The statistic is defined from 0 to 1. Where 1 means perfect concordance, 0.5 means that the model is equivalent to a random guess and 0 represents perfect discordance. The C-index 95% confidence intervals (CI) were determined by bootstrapping. We use the confidence intervals to determine if one model has a significantly higher C-index than the other. In essence, this process is a comparison of two means, where the null hypothesis is $H_0 : Cindex_1 - Cindex_2 = 0$.

To further evaluate the discrimination of the models we used Uno’s C-index (Uno et al., 2011). Uno’s C-index is an extension of Harrell’s C-index that incorporates the censoring

TABLE 3 | Models used in the SL and their hyperparameters.

Model	Hyperparameters	Observations
ROPRO	—	
RSF	N = 500	
Regularized cox	$\alpha = 0$	Lasso
	$\alpha = 0.25$	Elastic net
	$\alpha = 0.5$	
	$\alpha = 0.75$	
	$\alpha = 1$	Ridge regression
GB	N = 100; L = 1	
	N = 100; L = 2	
	N = 500; L = 1	
	N = 500; L = 2	
	N = 1,000; L = 1	
DS	Activation = <i>tanh</i>	All DS models had 1 hidden layer and 90 neurons in that hidden layer
	Activation = SELU	
AE	N = 1; $\rho = 8$	All AE models had <i>RELU</i> and sigmoid activation functions in the encoder and decoder parts
	N = 1; $\rho = 14$	
	N = 3; $\rho = 8$	
	N = 3; $\rho = 14$	

In the GB models, “N” and “L” correspond to the number of trees and their length, respectively. The “Activation” in the DS models corresponds to activation function used in the perceptrons, “N” corresponds to the number of hidden layers and “L” to the number of hidden neurons per layer. In the AE models, “N” corresponds to the number of layers of the encoder. “ ρ ” corresponds to the encoded variable size.

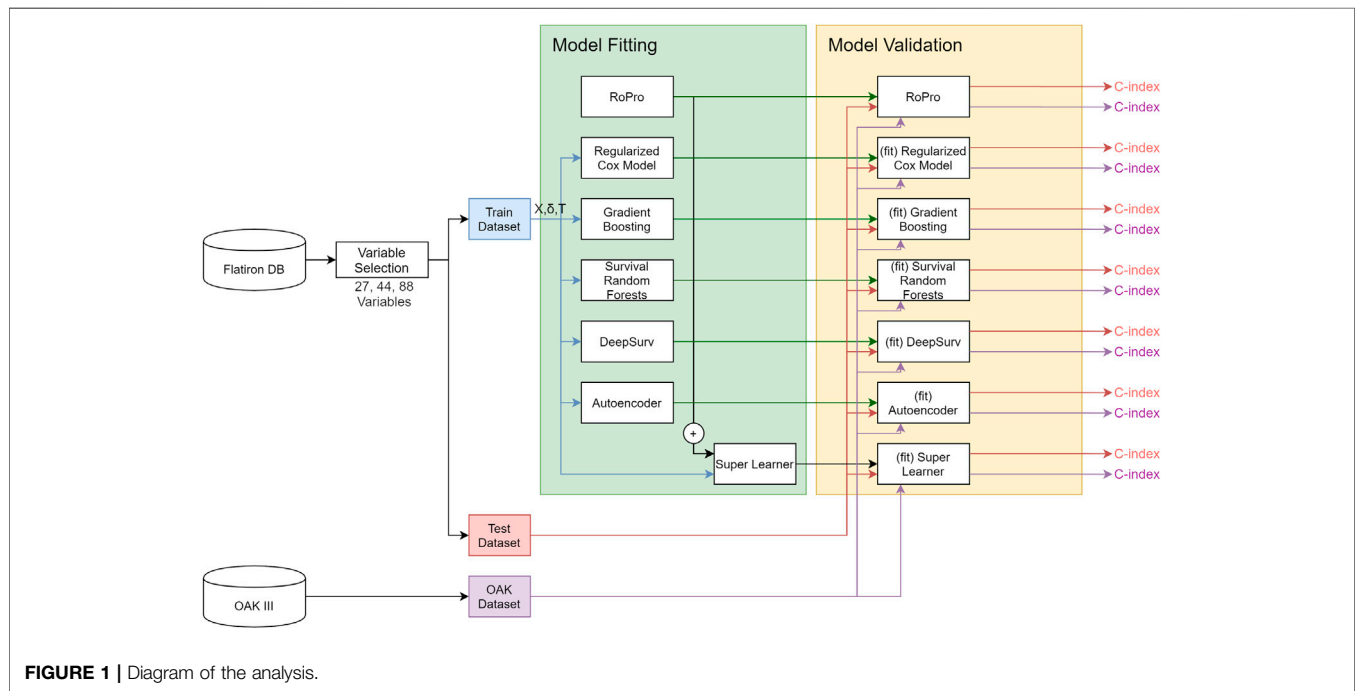


FIGURE 1 | Diagram of the analysis.

distribution into the score. This modification should make the C-index independent on the study’s censoring distribution (Uno et al., 2011).

Sensitivity Analyses

Given the differences between the FH test set and OAK, we performed additional analyses to validate our results. The

additional analyses include: 1) PCA analysis (Hastie et al., 2009) between FH test and OAK to verify differences between the datasets; 2) create an additional FH test set without the covariates not available in OAK and impute them to check the effect of these covariates. Further, 3) stratify the FH test set by cancer entity to check if C-index varies with cancer entity, and 4) permute FH train and test sets.

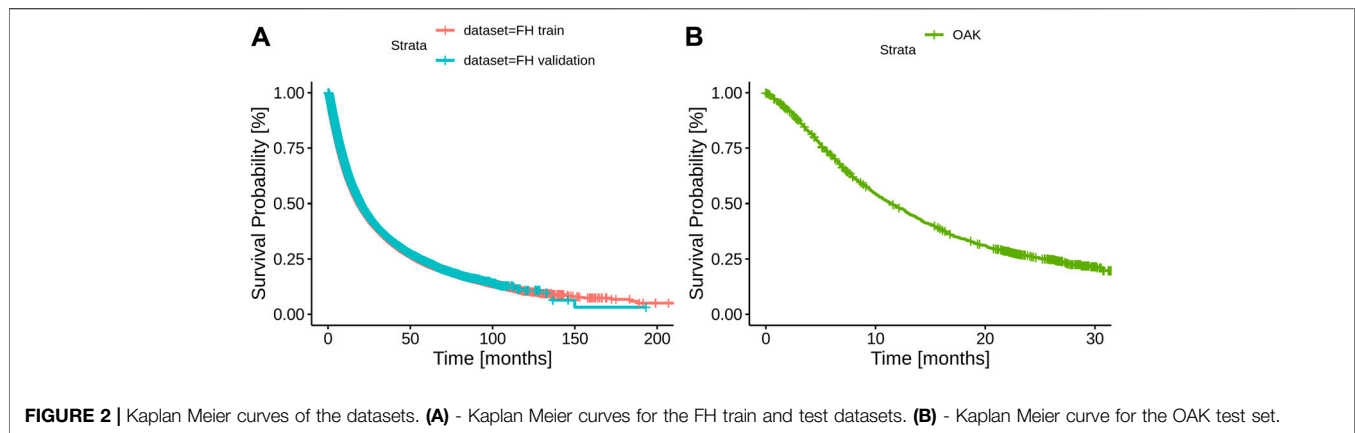


FIGURE 2 | Kaplan Meier curves of the datasets. **(A)** - Kaplan Meier curves for the FH train and test datasets. **(B)** - Kaplan Meier curve for the OAK test set.

Implementation

All the analyses were done using R 3.5.1 (R Core Team 2018) and Python 3.6. The Cox model and C-index were used as implemented in the survival library (Therneau, 2015). The GB, RSF, and SL were used as implemented in the R libraries, gbm (Greenwell et al., 2019), RandomForestSRC (Ishwaran et al., 2008) and SuperLearner (Polley et al., 2019), respectively. DeepSurv was implemented in the Python DeepSurv package (Katzman et al., 2018). The random forest imputation was implemented in the missForest R library (Stekhoven and Bühlmann, 2011). The full analysis diagram is illustrated in Figure 1.

We modified the SuperLearner, DeepSurv and missForest packages to add new features used in this work. In the SuperLearner package, we added the functionality to process survival analysis problems. More specifically, we added new models (ROPRO, regularized Cox, RSF, GB, RF and AE) and a new fitting algorithm based on the C-index. In DeepSurv we added some functions to assess the quality of fit of the models. Finally, in the missForest package we added the functionality to save the fitted model and use it to impute new data, e.g. test sets that have to remain independent to the training. The modified packages and analysis files are available in the **Supplementary materials**.

RESULTS

A total of three datasets were used in this analysis, FH train, FH test and OAK (see Methods) including cancer cohorts with a median follow-up time of 19.33 months (95% confidence interval (CI) 19.10–19.57), 19.83 (95% CI 19.33–20.57) and 11.43 (95% CI 10.40–12.67), respectively (Table 2; Figure 2).

Table 2 illustrates the summary statistics for the covariates in the 27 covariate feature set. The summary statistics for the 44 and 88 covariate feature sets are available **Supplementary Table S2**.

Individual Model Development

We benchmarked the ROPRO against a set of eight more complex models - regularized cox with lasso, ridge regression and elastic net, GB, RSF, AE, DS and SL - across a total of three different

feature sets, each with 27, 44 and 88 covariates yielding a total of 27 models.

After hyperparameter tuning (see **Supplementary Table S3** for a list of tested hyperparameters), the optimal shrinkage in the regularized cox resulted in the selection by lasso and elastic net of 23, 27 and 49 covariates in the 27, 44 and 88 covariate models, respectively. With grid search, we determined that the optimal α value for the elastic net model was close to 1. To avoid having two lasso models, we fixed $\alpha = 0.5$. The optimal number of weak learners in GB and trees in RSF was 1,000. In DS, the optimal number of hidden layers and number of neurons in the hidden layer was (1 and 120), (1 and 150) and (1 and 180) for the 27, 44 and 88 covariate feature sets, respectively. An increase in hidden layer size did not lead to an improvement in DS performance, resulting in shallow models. The optimal activation function was the *SELU* for all feature sets. Lastly, in AE the C-index values for all the hyperparameter combinations are depicted in **Figure 3**. Overall, a higher bottleneck size resulted in a higher C-index value. The optimal bottleneck sizes were 20, 36 and 84 for the 27, 44 and 88 covariate sets. In terms of total number of layers, the optimal values were five layers for the 27 and 44 covariate sets and three for the 88 covariate set). We used *RELU* and sigmoid activation functions in the encoder and decoder parts, respectively.

K-fold cross-validation was used in the SL to calculate the contribution of each model (listed in **Table 3**) to the final score. Results show that independent from the feature set (27, 44 or 88 covariates) the only models that contributed to the SL score were the ROPRO, RSF and two versions of DS, one with tanh and another with *SELU* activation functions. Each model contributed to the SL distinctively (see **Figure 4** for the models' risk distributions). The models' hazard value distributions varied for example, in center RSF (median 0.40–0.64) vs. ROPRO (median -0.161 to -0.0678), and in spread ROPRO (IQR 0.05–0.13) vs. DS with tanh (IQR 0.399–0.573). Additionally, the predicted risk values were stratified by the time-to-event of the patients (see **Figure 5**). In the 27 and 44 covariate models, RSF had the most sizable contribution for lower time-to-event (TTE). As the TTE increased, the contribution of RSF subsided while the contribution of both DS models increased. As a result, for later TTE, the model with the highest contribution changed from RSF

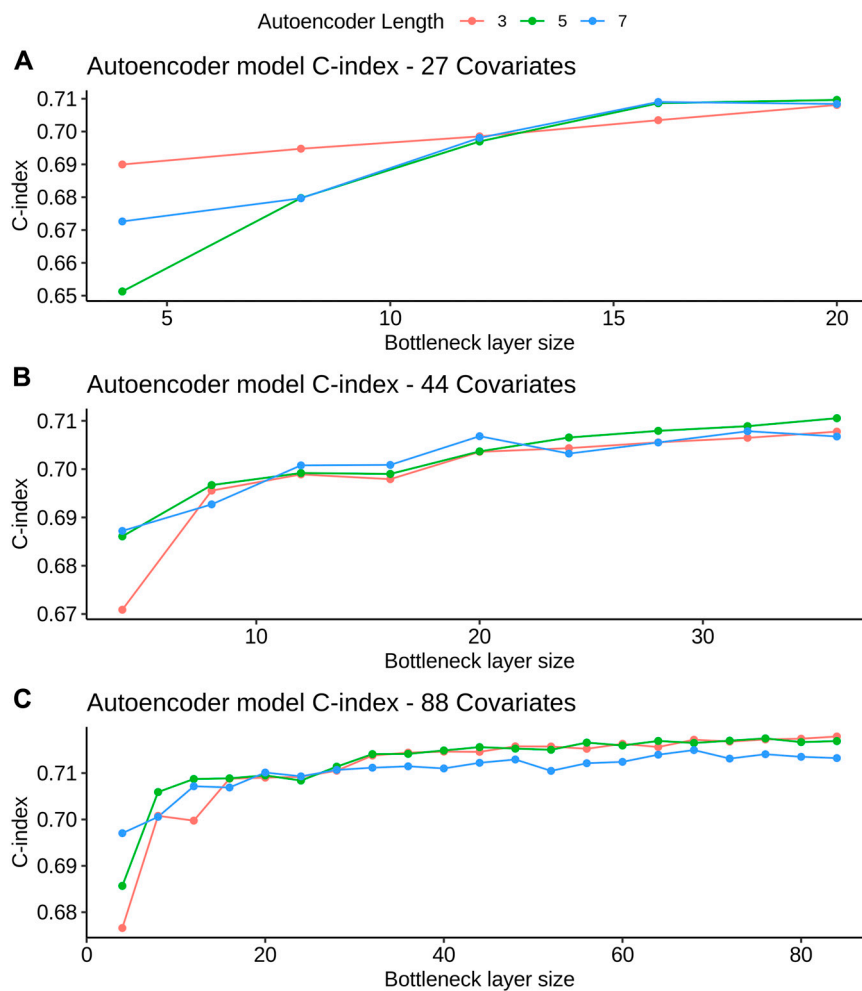


FIGURE 3 | AE model C-index values for different bottleneck layer sizes and total layer sizes. All C-index values are referent to the validation set derived from FH train (see *Datasets* section for more details). Figures A, B and C refer to the 27, 44 and 88 covariate models, respectively.

to DS with *tanh*. Conversely, in the 88 covariate feature set, there was not a clear separation of the most contributive models.

Model Performance

The C-index and corresponding 95% confidence intervals (CI) for the FH test dataset and the OAK test dataset are displayed in **Figure 6** and **Table 4**, and **Table 5**. **Figure 6** contains the C-index distributions for all models, datasets (FH test and OAK test) and feature sets (27, 44 and 88 covariates). **Table 4** offers a more granular view of the C-index distributions from **Figure 6**, with information on each models' C-index and 95% CI. Furthermore, **Table 5** includes the Uno C-index and corresponding 95% CI.

FH Test Set

As we observed similar patterns across all feature sets, we report here only results corresponding to the 44 covariate feature set. In the FH test dataset, the ROPRO achieved C-index values [95% CI] of 0.701 [0.696, 0.706]. In comparison, more complex models obtained slightly higher C-index values than ROPRO. Across all

ML-derived models, the AE consistently yielded the lowest C-index values (0.708 [0.703, 0.713]), followed by lasso and elastic net (C-index 0.708 [0.704, 0.714]) and ridge regression (C-index 0.709 [0.704, 0.714]). The model performances improved using RSF (c-index 0.720 [0.716, 0.725]), GB (C-index 0.722 [0.718, 0.727]), DS (C-index 0.721 [0.717, 0.726]) and lastly, SL (C-index 0.723 [0.718, 0.728]). However, given their 95% CI only GB, RSF, DS and SL obtained significant increases in C-index for all feature sets when compared to ROPRO. As an exception, in the 88 covariates feature set, the regularized Cox models (C-index [95%CI] lasso and elastic net 0.711 [0.706, 0.716]; ridge regression 0.711 [0.706, 0.715]) also had significant increases in C-index when compared with ROPRO (C-index [95% CI] 0.698 [0.693, 0.702]). The increases in C-index for the remaining models were not significantly different.

All Uno C-index values were lower than the respective (Harrell) C-index. Regardless, the models that obtained significant (Harrell) C-index increases also had significant Uno

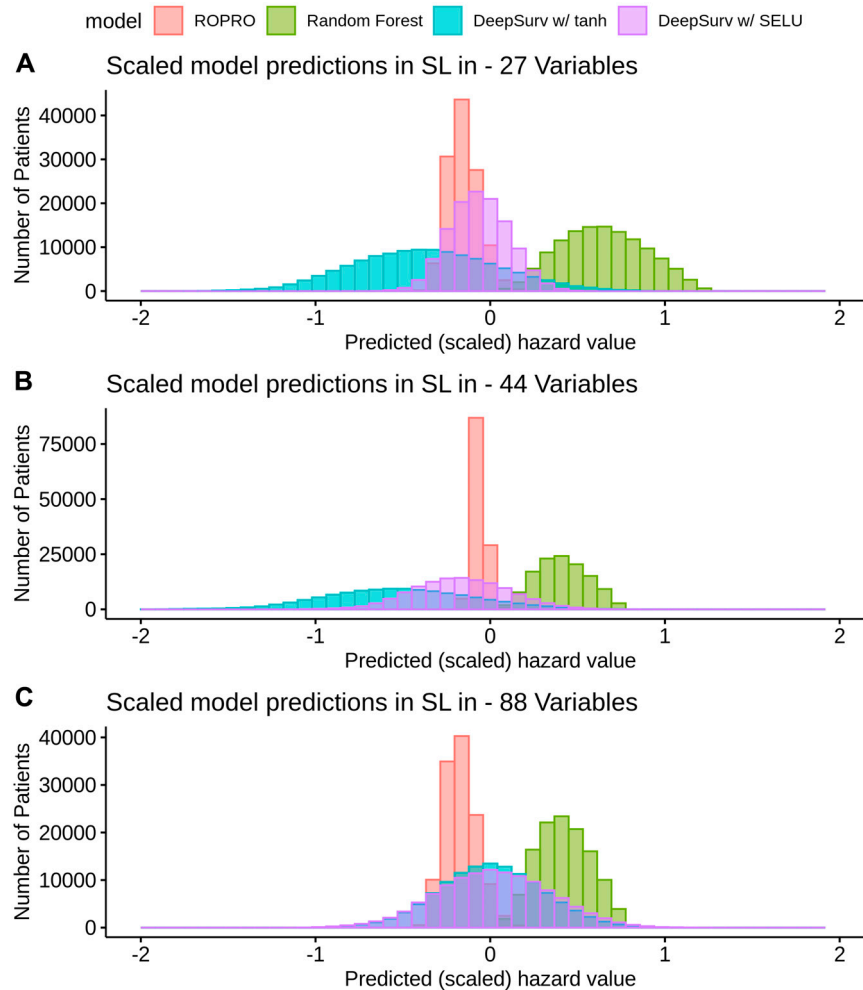


FIGURE 4 | Histogram of the risk predictions for each model in the SL in the FH training dataset. The risk values correspond to the risk yielded by the original model, i.e., by ROPRO or DS. The risk was multiplied by the α_k value of the model. The α_k value scales the risks of each of the models in the SL. In the risk of the SL, only four models are represented, i.e. are not scaled down to zero. Those four models are both DS models, ROPRO and RSF.

C-index increases. For the 44 covariate feature set, the GB (Uno C-index [95% CI] 0.697 [0.691, 0.701]), RSF (Uno C-index [95% CI] 0.693 [0.688, 0.698]), DS (Uno C-index [95% CI] 0.693 [0.688, 0.699]) and SL (Uno C-index [95% CI] 0.695 [0.690, 0.701]) obtained significantly higher Uno C-index values than ROPRO (Uno C-index [95% CI] 0.672 [0.667, 0.676]). Additionally, in the 88 covariates feature set, the regularized Cox models also had significantly higher Uno C-index than ROPRO.

OAK Test

In OAK we observed similar patterns for the different feature sets. For easier reporting the following results similarly correspond to the 44 covariate feature set. The ROPRO resulted in a C-index value [95% CI] of 0.670 [0.657, 0.685]. In comparison, the model that yielded the highest C-index was SL 0.677 [0.662, 0.695]. Nevertheless, we observed that, contrary to the results in the FH test set, the confidence intervals between ROPRO and SL (and all

the remaining models) overlapped, hence no statistically significant difference was found. Likewise, in the OAK dataset no model obtained a significantly higher Uno C-index than the ROPRO.

FH Test Set-Sensitivity Analyses

The PCA analysis with the first two principal components is shown in the **Supplementary Figure S1**. The C-index and 95% CI are displayed in **Supplementary Tables S4–7** for FH test set without the covariates unavailable in OAK, FH test stratified by cancer entity, and FH train and test permutations, respectively.

The PCA analysis illustrated that the FH test distribution exhibited a higher variance than OAK, with the FH test set having a variance of (3.704, 2.137) while the OAK population exhibited a variance of (2.690, 1.421).

There were only minor changes in the C-index values between the original FH test set and the FH test set without the covariates not present in OAK. Ultimately, the same models, GB, RSF, DS,

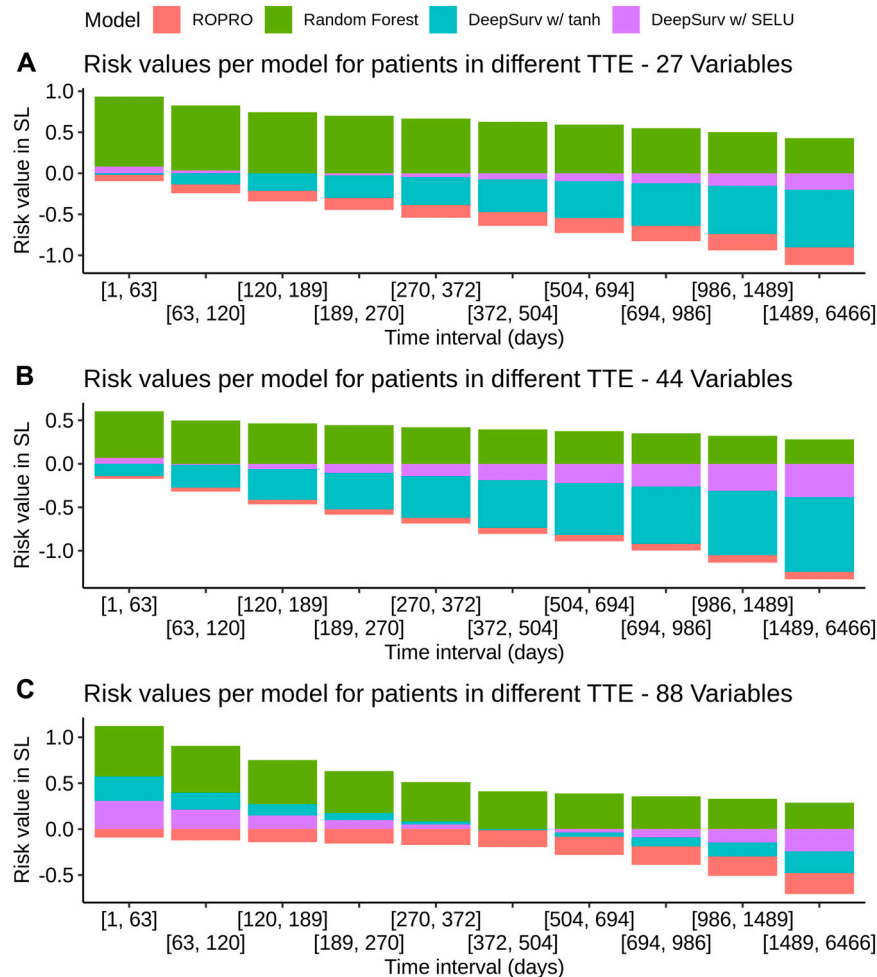


FIGURE 5 | Individual model contribution to the SL risk by time-to-event in the FH training set. To create this visualization, the patients were split into groups based on their time to event (TTE). Each of these groups is represented in the x-axis. Then, for each group the median risk value per model was calculated and is displayed on the y-axis. The contribution changes over time because the models correctly assign higher risk for lower times-to-event and lower risk for later times-to-event.

and SL, obtained significantly higher C-index values than ROPRO for all feature sets. Furthermore, in the 88 covariate feature set, the regularized Cox models obtained significant increases in C-index.

The C-index values showed a considerable variation between cancer entities. Most cancer entities had lower C-index values than the complete FH test (that had C-index values between 0.698 and 0.723). Only diffuse large B-cell lymphoma (C-index between 0.715 and 0.741), and follicular cancer (C-index between 0.771 and 0.788) had C-index values higher than the whole FH test dataset. Acute myeloid leukemia, breast cancer, gastric cancer, head and neck cancer, and metastatic breast cancer had the lowest discriminative power with C-index values close to 0.650. Advanced non-small cell lung cancer, the largest cohort in the FH test set, had C-index values between 0.673 and 0.687. In all cohort/feature set combinations, none of the more complex models obtained a significant increase in C-index against the ROPRO.

We reshuffled the original FH dataset twice, generating two extra sets of FH train and test. There were only minor changes in

C-index between the results of the primary analysis (in the section above) and the results from these two extra sets. More specifically, in the FH test, GB (C-index [95% CI] 0.724 [0.718, 0.729]; 0.723 [0.718, 0.728]), RSF (C-index [95% CI] 0.722 [0.717, 0.728]; 0.720 [0.715, 0.725]), DS (C-index [95% CI] 0.724 [0.718, 0.730]; 0.723 [0.718, 0.728]), and SL (C-index [95% CI] 0.725 [0.720, 0.731]; 0.724 [0.719, 0.729]) obtained significant C-index increases when compared with ROPRO (C-index [95% CI] 0.701 [0.695, 0.707]; 0.701 [0.695, 0.707]). As above, the C-index values refer to the 44 covariate feature set although we observed similar patterns for all feature sets. We presented for each model two C-index and 95% CI, each of which refers to one set of new FH train and test sets. In the 88 covariate feature set of the two new sets of FH train and test, the regularized Cox models also obtained significant increases in C-index. The only deviation in C-index significance observed compared to the primary analysis was that the AE model had a significant increase in C-index against ROPRO in the 88 covariate feature set of one of the sensitivity analyses.

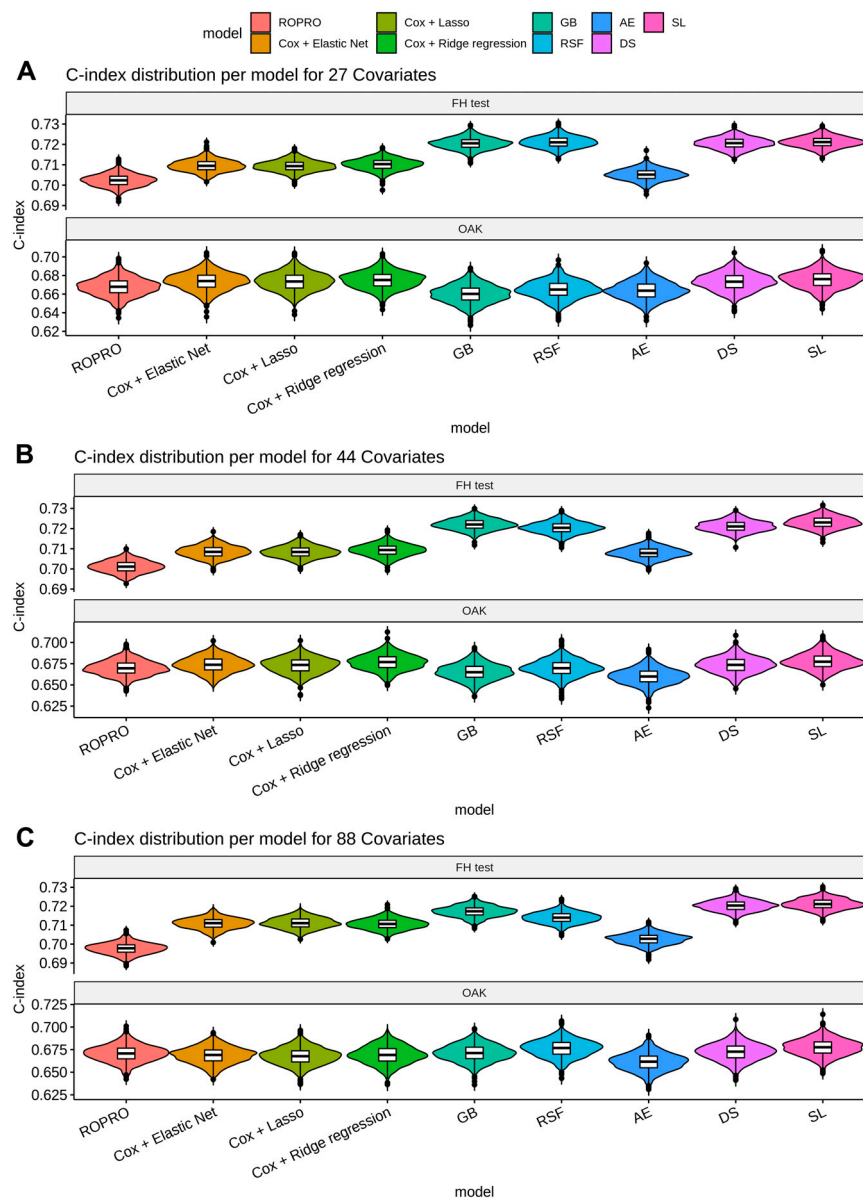


FIGURE 6 | Violin plot of the C-index in the FH test dataset (**top**) and OAK III (**bottom**). The C-index results for the 27, 44 and 88 covariate sets are illustrated in Figure (A)–(C), respectively. The plot displays the distribution and a box-plot of the C-index. Bootstrap was used to determine the distribution of the C-index.

DISCUSSION

We conducted an extensive benchmarking study to investigate: 1) whether the predictive power of prognostic scores in oncology could be improved by replacing the Cox model with more complex machine learning models and 2) whether increasing the number of covariates from 27 model-selected to 44 and 88 would increase the models' performance. To that end, we performed a comprehensive head-to-head comparison between a classic Cox model-based approach (ROPRO) and more complex ML-based survival models including two novel methods employing autoencoder and super learner algorithms. Overall, our

analysis suggests that neither increasing the number of covariates nor using complex machine learning models increases the performance of prognostic scores in oncology. In part, this might be explained by the absence in baseline clinical data (like blood work data and patient/disease characteristics) of complex covariate interactions that would have otherwise been learned by the more complex models. We hypothesize that the addition of rich patient/disease information in the form of imaging, genomics or longitudinal data could be the key to improving prognostic scores in cancer. These more complex data types, apart from adding prognostic factors to the models, should also contain information that are not easily extractable by classical methods

TABLE 4 | C-index and corresponding 95% confidence interval (CI) for all the models (ROPRO, regularized Cox models, Gradient Boosting (GB), Random Survival Forests (RSF), autoencoder (AE), DeepSurv (DS) and Super Learner (SL)) and covariate sets. Significant increases in C-index are in bold. Please refer to *Individual Model Development* section for the complete model hyperparameters.

#Covariates	Model	FH test		OAK	
		C-index	95% CI	C-index	95% CI
27 covariates	ROPRO ^a	0.702	[0.698, 0.707]	0.668	[0.652, 0.683]
	Cox + elastic net	0.709	[0.705, 0.714]	0.674	[0.657, 0.689]
	Cox + lasso	0.709	[0.705, 0.714]	0.674	[0.657, 0.690]
	Cox + ridge regression	0.710	[0.706, 0.715]	0.675	[0.659, 0.690]
	GB	0.721	[0.716, 0.725]	0.660	[0.644, 0.676]
	RSF	0.721	[0.716, 0.726]	0.665	[0.649, 0.680]
	AE	0.705	[0.700, 0.710]	0.664	[0.648, 0.680]
	DS	0.721	[0.716, 0.725]	0.673	[0.658, 0.689]
	SL	0.721	[0.717, 0.726]	0.676	[0.659, 0.691]
44 covariates	ROPRO ^a	0.701	[0.696, 0.706]	0.670	[0.657, 0.685]
	Cox + elastic net	0.708	[0.704, 0.714]	0.674	[0.658, 0.689]
	Cox + lasso	0.708	[0.704, 0.714]	0.674	[0.657, 0.687]
	Cox + ridge regression	0.709	[0.704, 0.714]	0.677	[0.661, 0.692]
	GB	0.722	[0.718, 0.727]	0.665	[0.650, 0.681]
	RSF	0.720	[0.716, 0.725]	0.670	[0.654, 0.686]
	AE	0.708	[0.703, 0.713]	0.660	[0.645, 0.676]
	DS	0.721	[0.717, 0.726]	0.674	[0.658, 0.689]
	SL	0.723	[0.718, 0.728]	0.677	[0.662, 0.695]
88 covariates	ROPRO ^a	0.698	[0.693, 0.702]	0.671	[0.656, 0.686]
	Cox + elastic net	0.711	[0.706, 0.716]	0.669	[0.653, 0.684]
	Cox + lasso	0.711	[0.706, 0.716]	0.668	[0.653, 0.683]
	Cox + ridge regression	0.711	[0.706, 0.715]	0.669	[0.652, 0.685]
	GB	0.717	[0.712, 0.722]	0.671	[0.656, 0.686]
	RSF	0.714	[0.709, 0.719]	0.677	[0.660, 0.692]
	AE	0.703	[0.698, 0.707]	0.662	[0.646, 0.678]
	DS	0.720	[0.716, 0.725]	0.673	[0.656, 0.688]
	SL	0.721	[0.717, 0.726]	0.678	[0.662, 0.692]

^aThe ROPRO was applied to all feature sets (27, 44 and 88 covariates). In all feature sets the ROPRO only uses 27 covariates (it was not refit) but since each dataset was separately imputed, the C-index value changes between feature sets.

(like the Cox model) which should lead to an increase in performance of the more complex models and therefore better prognostic performance.

To our knowledge, this is to date the largest benchmarking study of prognostic scores in oncology both in terms of number of models and patients. Previous analyses that compared the performance of simple and complex machine learning models have yielded rather inconsistent results. Some of these studies have demonstrated improvements in using complex models against the classic Cox model. For instance, a recent study challenged the Cox model against random survival forests (RSF) and DeepSurv (DS) to derive prognostic scores among patients with oral cancer (Kim et al., 2019). The DS was overall the model with the highest C-index. Yet, the study was limited by the low number of 255 patients and nine covariates. A separate study applied the Cox model, RSF and regularized cox to a larger dataset comprising a population of 80,000 patients with cardiovascular disease (Steele et al., 2018). The authors concluded that the elastic net model (C-index 0.801) using 600 covariates performed better than the Cox model (C-index 0.793) using 27 covariates, but the overall improvement was only moderate. In comparison, three other studies did not find any noticeable improvements by employing machine learning models (Chen et al., 2019; Christodoulou et al., 2019; Desai et al., 2020). However, these studies compared the use of logistic regression

with a binary endpoint against machine learning methods instead of the Cox model with a time-to-event endpoint. Our results showed that some of the more complex models, that could model covariate interactions and non-linear effects, did obtain significantly higher C-index values when compared to ROPRO. Although the C-index improvement size was still only moderate. Hence, the main results of the here presented study may contextualize the findings from (Chen et al., 2019; Christodoulou et al., 2019; Desai et al., 2020) to a survival analysis framework concluding that more complex machine learning models may not lead to a significant increase in performance over the Cox model.

Additionally, some of these studies also analyzed the effect of the covariate number in the prognostic score performance. The study by Kim et al. (Kim et al., 2019) analyzed models employing a range of five–nine covariates and found that the model performance increased with an increase in the covariate number. The same increase in performance was also observed when the performance of established prognostic scores (Arkenau et al., 2009; International Non-Hodgkin’s Lymphoma Prognostic Factors Project 1993; Ko et al., 2015; Kinoshita et al., 2013), which used a maximum of six covariates, was compared to the more recently developed ROPRO (Becker et al., 2020) that reported a number of 27 highly prognostic and independent covariates. Therefore, there is evidence that increasing the covariate size from small (less than 10 covariates) to a larger, but still moderate,

TABLE 5 | Uno C-index and corresponding 95% confidence intervals. Significant increases in C-index over the ROPRO model are in bold.

#Covariates	Model	FH test		OAK	
		Uno C-index	95% CI	Uno C-index	95% CI
27 covariates	ROPRO	0.674	[0.669, 0.679]	0.653	[0.637, 0.668]
	Cox + elastic net	0.683	[0.678, 0.688]	0.659	[0.643, 0.674]
	Cox + lasso	0.683	[0.678, 0.688]	0.659	[0.643, 0.674]
	Cox + ridge regression	0.683	[0.678, 0.688]	0.660	[0.644, 0.675]
	GB	0.695	[0.690, 0.700]	0.645	[0.629, 0.661]
	RSF	0.697	[0.692, 0.701]	0.650	[0.635, 0.666]
	AE	0.679	[0.674, 0.684]	0.650	[0.634, 0.665]
	DS	0.694	[0.689, 0.699]	0.658	[0.644, 0.674]
	SL	0.695	[0.690, 0.700]	0.660	[0.645, 0.676]
	44 covariates	ROPRO	0.672	[0.667, 0.676]	0.655
Cox + elastic net		0.680	[0.675, 0.686]	0.659	[0.644, 0.675]
Cox + lasso		0.680	[0.675, 0.686]	0.659	[0.644, 0.673]
Cox + ridge regression		0.681	[0.675, 0.686]	0.662	[0.647, 0.676]
GB		0.697	[0.691, 0.701]	0.651	[0.635, 0.666]
RSF		0.693	[0.688, 0.698]	0.655	[0.640, 0.670]
AE		0.681	[0.675, 0.686]	0.647	[0.631, 0.662]
DS		0.693	[0.688, 0.699]	0.659	[0.644, 0.674]
SL		0.695	[0.690, 0.701]	0.665	[0.648, 0.679]
88 covariates		ROPRO	0.669	[0.664, 0.674]	0.655
	Cox + elastic net	0.684	[0.679, 0.689]	0.654	[0.640, 0.670]
	Cox + lasso	0.684	[0.679, 0.689]	0.654	[0.639, 0.669]
	Cox + ridge regression	0.684	[0.679, 0.689]	0.655	[0.640, 0.670]
	GB	0.692	[0.687, 0.697]	0.658	[0.643, 0.674]
	RSF	0.687	[0.682, 0.692]	0.663	[0.648, 0.678]
	AE	0.677	[0.672, 0.682]	0.648	[0.633, 0.664]
	DS	0.695	[0.690, 0.700]	0.658	[0.643, 0.674]
	SL	0.695	[0.690, 0.700]	0.664	[0.648, 0.680]

number (30 covariates) leads to an increase in the prognostic score performance. This finding is not unexpected as the addition of more covariates increases the chances that some of them contain prognostic information that could be used by the models to increase their performance. In our analysis, we built upon the progress made in (Becker et al., 2020) by increasing the feature set from 27 to 44 and 88 covariates. However, the addition of these extra covariates did not lead to an increase in performance of the models as in previous studies. This could be caused by multiple reasons: First, the higher missingness in the 44 and 88 feature sets could have led to an erroneous imputation of the covariates with high missingness. Second, the 27 covariates included had been previously selected as the most relevant in (Becker et al., 2020), hence any additional covariates could have lower prognostic value. Both reasons should contribute to the lack of performance improvement. Yet, since the regularized Cox models did incorporate additional covariates from the 44 and 88 feature sets it gives evidence that there was some prognostic value in them although they did not lead to an increase in performance.

Conversely, in (Steele et al., 2018) two datasets with differing covariate numbers were studied: an expert-selected covariate dataset with 27 covariates vs. a much larger dataset with 600 covariates. The results demonstrated that the best 600 covariate model (elastic net) obtained a slightly higher C-index value than the best 27 covariate model (Cox model). The elastic net model added covariates such as prescription of cardiovascular medication (that should indicate severe cardiovascular

problems) and prescription of laxatives/home visits (that might indicate general frailty). All these covariates are possibly associated (proxies) with cardiovascular disease but were not identified by the experts as prognostic, which may explain the increase in performance of the elastic net model. This result illustrates the need to incorporate more diverse data into prognostic scores. As explained above, we followed a different approach in this analysis and instead focused on increasing the number of biomarkers (blood work/patient characteristics) from 27 model-selected to 44 and 88 feature sets. Our results showed that this addition did not result in an increase in performance. We hypothesized that, perhaps, we had exhausted the available information in the blood work/patient characteristics in the 27 covariate dataset and the covariates added in the 44 and 88 feature sets did not carry prognostic information. Therefore, these results might suggest that perhaps there is a hard limit on the predictive power of baseline blood work/patient characteristics. To further increase the performance it might be necessary to incorporate other types of covariates as suggested by Steele et al. (Steele et al., 2018) or data with increased richness, like images, genomics or longitudinal biomarkers. Although, we would suggest that further research in this area is still needed.

Furthermore, all models had a comparable internal performance (C-index 0.70–0.72 within the FH test and C-index 0.66–0.68 within OAK) while the performance between datasets, which may be an indicator for model generalizability, was less strong. Particularly when the same models were compared between datasets, the C-index

differences were more apparent. Some models had a considerable loss in performance with a decrease in C-index between FH test and OAK as high as 0.060 for GB or 0.056 for RSF. The ROPRO showed the most stable performance between datasets with a C-index difference as low as 0.027. These results suggest that the slight gains in performance achieved by the more complex models in the FH test dataset are not generalizable to other datasets. We hypothesize that this could happen due to multiple reasons: First, differences in the cohort number between FH test and OAK could cause differences in the C-index as the model performance could depend on the type of cancer. Second, the lack of some of the covariates in OAK, e.g., blood oxygen or granulocytes, could lead to a decrease in performance in the OAK dataset. Third, since OAK is a clinical trial, it is likely that the patient population is more homogeneous than in FH. Therefore, more extreme values in highly prognostic covariates (e.g. ECOG > 1) should be inexistent or rare, making it harder for prognostic prediction. Additionally, the study start date was defined differently between FH (first day of first line of treatment) and OAK (first day of second or third lines of treatment). We investigated some of these hypotheses in the sensitivity analyses above. For the first hypothesis, we tested whether the models had different performance for different cohorts. The FH test set C-index for advanced non-small cell lung cancer (the only cohort in OAK) ranged between 0.68–0.69, which is closer to the C-index in OAK. For the second hypothesis, we removed the covariates inexistent in OAK from the FH test and imputed them. This had little effect on the C-index of the FH test, therefore, we discard the effect of the absent OAK covariates. For the third hypothesis, we performed a PCA analysis where we compared both datasets, which supported the hypothesis that OAK has less extreme values. Given the sensitivity analyses, we argue that a combination of the first and third hypotheses is more likely. The C-index for the advanced non-small cell cancer in FH test was closer to the OAK value. Additionally, the other differences introduced in the third hypothesis might further decrease the C-index in the OAK dataset. Furthermore, there could have also been some overfitting to the FH test set that caused the decrease in OAK. Unfortunately, the compared prognostic scores in literature utilized test sets from the same data-source as the training set which makes a valid comparison not feasible. We suggest that further studies should be performed to investigate the true cause of this effect.

In general, we argue that in order to develop better prognostic scores in oncology, rather than focusing on more complex models on the same dataset, we should focus on getting access to larger and optimally multimodal data describing the patients in more detail. In particular, adding data about tumor biology via rich data types, e.g., via imaging, genomics or longitudinal data might be more beneficial and could lead to improved clinical decision-making when using prognostic scores. Consequently, these rich data types should contain complex information that the classical models cannot interpret, in that case, the more complex models tested in this work should demonstrate increased performance. Another area for improvement is related to the response of

patients to treatments. By combining the patients' treatments with longitudinal data, e.g. biomarkers, it might be possible to model the disease progression, leading to models that could offer real-time decision-making support. Overall, there remains an unmet clinical need for precise survival prediction to enable improved toxicity monitoring, treatment selection and assessment of clinical trial eligibility and hence further work is required to improve prognostic scores in oncology.

CONCLUSION

Prognostic scores are important clinical decision-making tools for treatment decisions, monitoring adverse events, and clinical trial eligibility. Our results show that complex machine learning-derived models did not improve prognostic scores in oncology compared to a classical Cox-based framework. We argue that further research should focus on the impact of adding other data types (e.g. imaging, genomics or longitudinal biomarkers) describing complementary features of disease biology. In these scenarios, complex machine learning architectures might still prove beneficial.

DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: The data presented is owned by either Flatiron Health Inc. (train and in-sample test datasets) or F. Hoffmann-la Roche LTD (test dataset). Access to Flatiron Health may be made available upon request, and are subject to a license agreement with Flatiron Health. Requests to access these datasets should be directed to DataAccess@flatiron.com.

ETHICS STATEMENT

Institutional Review Board approval of the Flatiron Health study protocol was obtained prior to study conduct, and included a waiver of informed consent. For the OAK dataset, an independent data monitoring committee reviewed safety. Protocol approval was obtained from independent ethics committees for each site.

AUTHOR CONTRIBUTIONS

HL wrote the codes, performed the analysis and wrote the draft of the publication. TB contributed to the conceptualization of this study, supported the data preparation and contributed significantly to the manuscript. AB-M contributed to the conceptualization of this study, supervised the study project with respect to clinical application and contributed to the manuscript. NA supervised ML aspects of the project, edited the manuscript significantly. JW contributed to the conceptualization of this study, supervised the study project and contributed to the manuscript.

FUNDING

This study was funded by F. Hoffmann-la Roche LTD. No grant number is applicable.

ACKNOWLEDGMENTS

The authors thank Fabian Schmich, Roche Innovation Center Munich; Sarah McGough, Svetlana Lyalina and Devin Incerti from PHC RWDS Analytics, Genentech Inc., San Francisco, United States and Stefka Tyanova from Data Science, Roche

Innovation Center Basel for their valuable input regarding methodology and results. HL is supported by the Helmholtz Association under the joint research school “Munich School for Data Science - MUDS”.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/frai.2021.625573/full#supplementary-material>.

REFERENCES

- Arkenau, H. T., Barriuso, J., Olmos, D., Ang, J. E., de Bono, J., Judson, I., et al. (2009). Prospective validation of a prognostic score to improve patient selection for oncology phase I trials. *J. Clin. Oncol.* 27 (16), 2692–2696. doi:10.1200/JCO.2008.19.5081
- Becker, T., Weberpals, J., Jegg, A. M., So, W. V., Fischer, A., Weisser, M., et al. (2020). An enhanced prognostic score for overall survival of patients with cancer derived from a large real-world cohort. *Ann. Oncol.* 31 (11), 1561–1568. doi:10.1016/j.annonc.2020.07.013
- Bhimani, J., Philipps, L., Simpson, L., Lythgoe, M., Soutati, A., Webb, A., et al. (2020). The impact of new cancer drug therapies on site specialized cancer treatment activity in a UK cancer network 2014–2018. *J. Oncol. Pharm. Pract.* 26 (1), 93–98. doi:10.1177/1078155219839445
- Birnbaum, B., Nathan, N., Seidl-Rathkopf, K., Agrawal, M., Estevez, M., Estola, E., et al. (2020). Model-assisted cohort selection with bias analysis for generating large-scale cohorts from the EHR for oncology research. Available at: <http://arxiv.org/abs/2001.09765>.
- Byrd, R. H., Lu, P., Nocedal, J., and Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM J. Sci. Comput.* 16 (5), 1190–1208. doi:10.1137/0916069
- Chen, D., Liu, S., Kingsbury, P., Sohn, S., Storie, C. B., Habermann, E. B., et al. (2019). Deep learning and alternative learning strategies for retrospective real-world clinical data. *NPJ Digit. Med.* 2 (1), 43. doi:10.1038/s41746-019-0122-0
- Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., and Van Calster, B. (2019). A systematic Review shows No performance benefit of machine learning over logistic regression for clinical prediction models. *J. Clin. Epidemiol.* 110, 12–22. doi:10.1016/j.jclinepi.2019.02.004
- Cox, D. R. (1972). Regression models and life-tables. *J. R. Stat. Soc. Ser. B* 34 (2), 187–202. doi:10.1111/j.2517-6161.1972.tb00899.x
- Desai, R. J., Wang, S. V., Vaduganathan, M., Evers, T., and Schneeweiss, S. (2020). Comparison of machine learning methods with traditional models for use of administrative claims with electronic medical records to predict heart failure outcomes. *JAMA Netw. Open* 3, e1918962. doi:10.1001/jamanetworkopen.2019.18962
- Ferlay, J., Colombet, M., Soerjomataram, I., Mathers, C., Parkin, D. M., Piñeros, M., et al. (2019). Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods. *Int. J. Cancer* 144 (8), 1941–1953. doi:10.1002/ijc.31937
- Fogel, D. B. (2018). Factors associated with clinical trials that fail and opportunities for improving the likelihood of success: a Review. *Contemp. Clin. Trials Commun.* 11, 156–164. doi:10.1016/j.conctc.2018.08.001
- Friedman, J. H. (2001). Machine. *Ann. Statist.* 29 (5), 1189–1232. doi:10.1214/aos/1013203451
- Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. Cambridge, MA: MIT Press.
- Grenewell, B., Bradley, B., Cunningham, J., and Developers, G. B. M. (2019). Gbm: generalized boosted regression models. Available at: <https://CRAN.R-project.org/package=gbm>.
- Harrell, F. E., Jr, Califf, R. M., Pryor, D. B., Lee, K. L., and Rosati, R. A. (1982). Evaluating the yield of medical tests. *JAMA* 247 (18), 2543–2546. doi:10.1001/jama.1982.03320430047030
- Harrell, F. E., Lee, K. L., and Mark, D. B. (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat. Med.* 15 (4), 361–387. doi:10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4
- Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. 2nd Edn. New York, NY: Springer Science & Business Media.
- International Non-Hodgkin's Lymphoma Prognostic Factors Project (1993). A predictive model for aggressive non-hodgkin's lymphoma. *New Engl. J. Med.* 329 (14), 987–994. doi:10.1056/NEJM199309303291402
- Ishwaran, H., Kogalur, U. B., Blackstone, E. H., and Lauer, M. S. (2008). Random survival forests. *Ann. Appl. Stat.* 2, 841–860. doi:10.1214/08-aos169
- Kalbfleisch, J. D., and Prentice, R. L. (2002). *The statistical analysis of failure time data*. New Jersey: John Wiley and Sons.
- Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., and Kluger, Y. (2018). DeepSurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC Med. Res. Methodol.* 18 (1), 24. doi:10.1186/s12874-018-0482-1
- Kim, D. W., Lee, S., Kwon, S., Nam, W., Cha, I. H., and Kim, H. J. (2019). Deep learning-based survival prediction of oral cancer patients. *Sci. Rep.* 9 (1), 6994. doi:10.1038/s41598-019-43372-7
- Kinoshita, A., Onoda, H., Imai, N., Iwaku, A., Oishi, M., Tanaka, K., et al. (2013). The Glasgow prognostic score, an inflammation based prognostic score, predicts survival in patients with hepatocellular carcinoma. *BMC Cancer* 13 (1), 52. doi:10.1186/1471-2407-13-52
- Ko, J. J., Xie, W., Kroeger, N., Lee, J. L., Rini, B. I., Knox, J. J., et al. (2015). The international metastatic renal cell carcinoma database consortium model as a prognostic tool in patients with metastatic renal cell carcinoma previously treated with first-line targeted therapy: a population-based study. *Lancet Oncol.* 16 (3), 293–300. doi:10.1016/S1470-2045(14)71222-7
- LeDell, E., van der Laan, M. J., and Petersen, M. (2016). AUC-maximizing ensembles through metalearning. *Int. J. Biostat* 12 (1), 203–218. doi:10.1515/ijb-2015-0035
- Ma, X., Long, L., Moon, S., Blythe, J., Adamson, S., and Baxi, S. S. (2020). Comparison of population characteristics in real-world clinical oncology databases in the US: flatiron health, SEER, and NPCR. Available at: <https://www.medrxiv.org/content/10.1101/2020.03.16.20037143v2>.
- Polley, E., LeDell, E., Kennedy, C., and van der Laan, M. (2019). SuperLearner: super learner prediction. Available at: <https://github.com/ecpolley/SuperLearner>.
- Pulte, D., Weberpals, J., Jansen, L., and Brenner, H. (2019). Changes in population-level survival for advanced solid malignancies with new treatment options in the second decade of the 21st century. *Cancer* 125 (15), 2656–2665. doi:10.1002/cncr.32160
- R Core Team (2018). *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
- Ridgegway, G. (1999). The state of boosting. *Comput. Sci. Stat.* 31, 172–181.
- Rittmeyer, A., Barlesi, F., Waterkamp, D., Park, K., Ciardiello, F., von Pawel, J., et al. (2017). Atezolizumab versus Docetaxel in patients with previously

- treated non-small-cell lung cancer (OAK): a phase 3, open-label, multicentre randomised controlled trial. *Lancet* 389 (10066), 255–265. doi:10.1016/S0140-6736(16)32517-X
- Sant, M., Minicozzi, P., Mounier, M., Anderson, L. A., Brenner, H., Hollecsek, B., et al. (2014). Survival for hematological malignancies in europe between 1997 and 2008 by region and age: results of EURO CARE-5, a population-based study. *Lancet Oncol.* 15 (9), 931–942. doi:10.1016/S1470-2045(14)70282-7
- Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2011). Regularization paths for cox's proportional hazards model via coordinate descent. *J. Stat. Softw.* 39 (5), 1–13. doi:10.18637/jss.v039.i05
- Steele, A. J., Denaxas, S. C., Shah, A. D., Hemingway, H., and Luscombe, N. M. (2018). Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease. *PLOS ONE* 13 (8), e0202344. doi:10.1371/journal.pone.0202344
- Stekhoven, D. J., and Bühlmann, P. (2011). MissForest--non-parametric missing value imputation for mixed-type data. *Bioinformatics* 28 (1), 112–118. doi:10.1093/bioinformatics/btr597
- Therneau, T. M. (2015). A package for survival analysis in S. Available at: <https://CRAN.R-project.org/package=survival>.
- Tibshirani, R. (1997). The lasso method for variable selection in the cox model. *Stat. Med.* 16 (4), 385–395. doi:10.1002/(SICI)1097-0258(19970228)16:4<385::AID-SIM380>3.0.CO;2-3
- Uno, H., Cai, T., Pencina, M. J., D'Agostino, R. B., and Wei, L. J. (2011). On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Stat. Med.* 30 (10), 1105–1117. doi:10.1002/sim.4154
- van der Lann, M. J., Polley, E. C., and Hubbard, A. E. (2007). Super learner. *Stat. Appl. Genet. Mol. Biol.* 6 (1), 25. doi:10.2202/1544-6115.1309
- Wong, C. H., Siah, K. W., and Lo, A. W. (2019). Estimation of clinical trial success rates and related parameters. *Biostatistics* 20 (2), 273–286. doi:10.1093/biostatistics/kxx069

Conflict of Interest: HL, AB-M and JW are employed by F. Hoffmann-La Roche. JW and AB-M hold shares of F. Hoffmann-La Roche. TB is a contractor paid by F. Hoffmann-La Roche.

The remaining author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2021 Loureiro, Becker, Bauer-Mehren, Ahmidi and Weberpals. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Full length manuscript II - Matching by OS prognostic score to construct external controls in lung cancer clinical trials

Hugo Loureiro, Andreas Roller, Meike Schneider, Carlos Talavera-López, Tim Becker, Anna Bauer-Mehren

The manuscript was peer-reviewed and was published in the *Clinical Pharmacology & Therapeutics* journal in 2023. The manuscript is published under the Creative Commons Attribution-NonCommercial-NoDerivs license (CC BY-NC-ND 4.0), the published version is included.

Matching by OS Prognostic Score to Construct External Controls in Lung Cancer Clinical Trials

Hugo Loureiro^{1,2,3} , Andreas Roller⁴, Meike Schneider⁴, Carlos Talavera-López², Tim Becker^{1,†} and Anna Bauer-Mehren^{1,*,†}

External controls (eControls) leverage historical data to create non-randomized control arms. The lack of randomization can result in confounding between the experimental and eControl cohorts. To balance potentially confounding variables between the cohorts, one of the proposed methods is to match on prognostic scores. Still, the performance of prognostic scores to construct eControls in oncology has not been analyzed yet. Using an electronic health record-derived de-identified database, we constructed eControls using one of three methods: ROPRO, a state-of-the-art prognostic score, or either a propensity score composed of five (5Vars) or 27 covariates (ROPROvars). We compared the performance of these methods in estimating the overall survival (OS) hazard ratio (HR) of 11 recent advanced non-small cell lung cancer. The ROPRO eControls had a lower OS HR error (median absolute deviation (MAD), 0.072, confidence interval (CI): 0.036–0.185), than the 5Vars (MAD 0.081, CI: 0.025–0.283) and ROPROvars eControls (MAD 0.087, CI: 0.054–0.383). Notably, the OS HR errors for all methods were even lower in the phase III studies. Moreover, the ROPRO eControl cohorts included, on average, more patients than the 5Vars (6.54%) and ROPROvars cohorts (11.7%). The eControls matched with the prognostic score reproduced the controls more reliably than propensity scores composed of the underlying variables. Additionally, prognostic scores could allow eControls to be built on many prognostic variables without a significant increase in the variability of the propensity score, which would decrease the number of matched patients.

Study Highlights

WHAT IS THE CURRENT KNOWLEDGE ON THE TOPIC?

✓ Prognostic scores have been suggested as an appropriate method to match patients. Still, they have not been applied to match external controls in oncology.

WHAT QUESTION DID THIS STUDY ADDRESS?

✓ Whether prognostic scores could adequately match patients into lung cancer external controls, and how they compare to propensity scores composed of the same covariates.

WHAT DOES THIS STUDY ADD TO OUR KNOWLEDGE?

✓ ROPRO (a prognostic score) showed a high performance in matching lung cancer external controls. The performance

was highest for phase III studies, which include more patients. Additionally, the ROPRO method matched more patients than propensity scores with an equivalent number of variables.

HOW MIGHT THIS CHANGE CLINICAL PHARMACOLOGY OR TRANSLATIONAL SCIENCE?

✓ Prognostic scores are a valid method to match lung cancer phase III external controls. Additionally, more research must be done in external controls for smaller studies, such as phases I and II, which are of high interest.

Randomized clinical trials (RCTs) are the standard method to determine the efficacy of new treatments. In each RCT, a new cohort of patients is recruited to receive either the new treatment or the standard of care. Recruiting patients for RCTs in oncology is difficult for multiple reasons, such as the lack of access to participating cancer clinics or the fear of random

treatment.¹ Hence, external control arms (in short, eControls) are considered an additional support tool to determine the efficacy of new medications.² The utility of eControls span across clinical trial phases, from small single-arm early-stage clinical trials to late-stage development.³ The eControls are composed of either historical data^{3,4} from previous clinical trials and/or

¹Data and Analytics, Pharma Research and Early Development, Roche Innovation Center Munich (RICM), Penzberg, Germany; ²Computational Health Center, Helmholtz Munich, Munich, Germany; ³TUM School of Life Sciences Weihenstephan, Technical University of Munich, Freising, Germany; ⁴Early Development Oncology, Pharma Research and Early Development, Roche Innovation Center Basel (RICB), Basel, Switzerland. *Correspondence: Anna Bauer-Mehren (anna.bauer-mehren@roche.com)

[†]These authors contributed equally to this work.

Received August 9, 2023; accepted November 8, 2023. doi:10.1002/cpt.3109

real-world data (RWD) sources, or of data collected concurrently to the clinical trial.^{3,5} However, given that eControls are not randomized, possible biases in patient selection and confounding must be accounted for.

One popular method to account for the possible confounding in eControls are propensity scores.^{6,7} Propensity scores model the probability of treatment assignment given a set of covariates. Hence, propensity scores can be used to reduce the possible confounding bias on observed covariates by balancing the eControl cohort to the treated population. The issue of which variables to use and how to include them in the propensity score model has been an active research topic. Brookhart *et al.*⁸ suggested that covariates related to the outcome alone or the outcome and exposure should be included in the model. Another approach, suggested by Stuart *et al.*,⁹ uses prognostic scores to balance the dataset. Prognostic scores, conversely to propensity scores, model the risk of an event (e.g., death) given a set of covariates. Although several subsequent analyses followed the Stuart *et al.* methodology,¹⁰⁻¹² prognostic scores have not been used to construct oncology eControls yet.

Hence, in this work, we applied prognostic scores to construct eControls in oncology following the initial analysis by Stuart *et al.*⁹ We selected the pan-cancer ROPRO prognostic score¹³⁻¹⁵ to construct the eControls. ROPRO comprises 27 covariates describing the host, the patient's lifestyle, and the tumor. All ROPRO covariates are associated with the mortality of oncology patients and hence with the outcome.

In this study, we performed a retrospective analysis of the performance of ROPRO to create eControl arms. Additionally, we compared the performance of ROPRO with two propensity score models composed of subsets of ROPRO's covariates. We analyzed 11 recent lung cancer clinical trials sponsored by Roche/Genentech, and constructed eControls using data from a large real-world oncology database.

METHODS

Lung cancer clinical trials

In this work, we retrospectively constructed eControls for advanced non-small cell lung cancer (advNSCLC) clinical trials. We extracted a list of 212 recent advNSCLC clinical trials sponsored by Roche/Genentech from [ClinicalTrials.gov](#). From this list, we excluded trials (Figure 1) whose control population would not be reproducible with the real-world database (described below) due the trial start date, location covered, or considered medications. After applying the exclusion criteria, there were 11 eligible clinical trials (the full list is available in [Table 1](#)). We split all trials comprising more than one experimental and one control arm into separate trials (e.g., NCT02366143 was split into 3 individual trials), yielding 16 different experimental-control comparison clinical trials.

Real-world database to generate the eControls

We used data from the Flatiron Health (FH) electronic health record-derived de-identified database to generate the eControl cohorts. The FH database is a longitudinal database, comprising de-identified patient-level structured and unstructured data, curated via technology-enabled abstraction.^{16,17} We extracted from FH the information on the first, second, and third lines of patients with advNSCLC. All lines of therapy were oncologist-defined and rule-based lines of therapy. Additionally, we selected patients that had started treatment before May 2020 so that all patients had a possible follow-up window of at least 2 years. The selected cohorts comprised 46,595 patients diagnosed with advNSCLC between

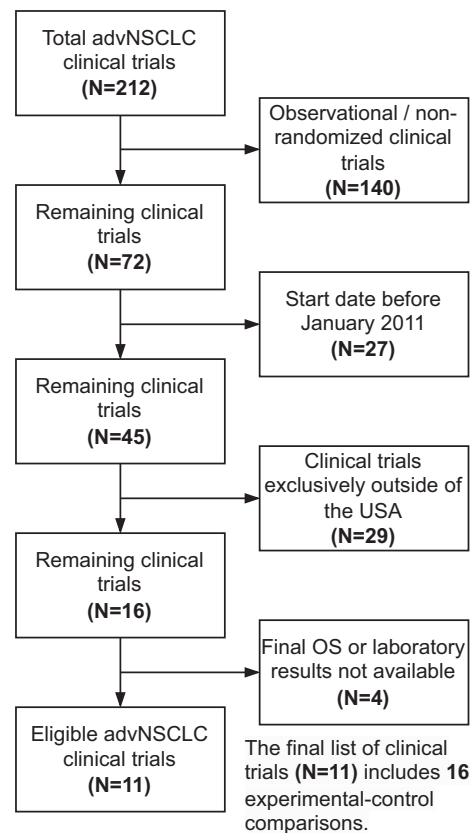


Figure 1 Trial selection process. AdvNSCLC, advanced non-small cell lung cancer; OS, overall survival.

January 2011 and May 2020. These patients were followed in a network of ~280 cancer clinics (~800 sites of care) in the United States. The majority of patients in the database originate from community oncology settings; relative community/academic proportions may vary depending on study cohort.

eControl construction

To create the eControls for each of the selected clinical trials, we first selected patients from the RWD cohort that had been prescribed the relevant medication considered in the trial's control cohort. Next, we applied the trial inclusion/exclusion criteria to the cohorts, so that non-eligible patients were not included in the eControls. Most inclusion/exclusion criteria based on prior treatment and blood tests could be easily applied to the RWD cohort. Still, other criteria, such as comorbidities, were usually not available in the RWD cohort and could not be applied.

ROPRO, the prognostic score used in this analysis, is based on a Cox model,¹⁸ and considers death as the event of interest. The ROPRO model was trained on a large RWD database, and is composed of 27 prognostic covariates that were selected using a backward selection approach. Most ROPRO covariates stem from routine blood work (e.g., white and red blood cell counts or blood chemistry markers). Additionally, the ROPRO variables also comprise the body mass index, smoking history, tumor, node, and metastasis staging, and number of metastatic sites. For more details on ROPRO, including the full list of covariates, refer to the original publication.¹³

To match the patients from the RWD cohort with the clinical trial patients, we used three different methods. First, we used the ROPRO as such. Additionally, we used a propensity score composed of the five most prognostic variables (which we refer to as 5Vars) from ROPRO (age, Eastern Cooperative Oncology Group (ECOG) performance status, albumin, chloride, and alkaline phosphatase (ALP), all lab-measures in serum

Table 1 List of selected clinical trials

Clinical Trial ID	Phase	Treatment	Control
NCT02008227 ³⁴	III	Atezolizumab	Docetaxel
NCT01903993 ³⁵	II	Atezolizumab	Docetaxel
NCT02366143 ABCP-BCP ³⁶	III	Atezolizumab + bevacizumab + carboplatin + paclitaxel	Bevacizumab + carboplatin + paclitaxel
NCT02366143 ACP-BCP ³⁶	III	Atezolizumab + carboplatin + paclitaxel	Bevacizumab + carboplatin + paclitaxel
NCT01519804	II	Onartuzumab + cisplatin/carboplatin + paclitaxel	Cisplatin/carboplatin + paclitaxel
NCT01496742	II	Onartuzumab + cisplatin/carboplatin + pemetrexed	Cisplatin/carboplatin + pemetrexed
NCT01496742 Bev	II	Onartuzumab + bevacizumab + cisplatin/ carboplatin + paclitaxel	Bevacizumab + cisplatin/carboplatin + pemetrexed
NCT01366131 ³⁷	II	Parsatuzumab + bevacizumab + carboplatin + paclitaxel	Bevacizumab + carboplatin + paclitaxel
NCT01493843 A-B	II	Pictilisib + carboplatin + paclitaxel	Carboplatin + paclitaxel
NCT01493843 C-D	II	Pictilisib + bevacizumab + carboplatin + paclitaxel	Bevacizumab + carboplatin + paclitaxel
NCT01493843 E-F	II	Pictilisib + bevacizumab + carboplatin + paclitaxel	Bevacizumab + carboplatin + paclitaxel
NCT02367781 ³⁸	III	Atezolizumab + carboplatin + nab-paclitaxel	Carboplatin + nab-paclitaxel
NCT02367794 ³⁹	III	Atezolizumab + carboplatin + paclitaxel	Carboplatin + nab-paclitaxel
NCT02367794 NAB ³⁹	III	Atezolizumab + carboplatin + nab-paclitaxel	Carboplatin + nab-paclitaxel
NCT02657434 ⁴⁰	III	Atezolizumab + carboplatin/cisplatin + pemetrexed	Carboplatin/cisplatin + pemetrexed
NCT02409342 ⁴¹	III	Atezolizumab	Carboplatin/cisplatin + pemetrexed/gemcitabine

or plasma), and the last approach was a propensity score composed of all the ROPRO variables (henceforth referred to as ROPROVars). For the ROPROVars approach, we fit the propensity score model with a regularized logistic regression (elastic net¹⁹), given the high number of covariates. Then, for each of these methods, we performed 1:1 matching with a 0.2 caliper (as suggested by Austin²⁰).

The ROPRO and ROPROVars methods compare the performance of a prognostic score against a propensity score, composed of its underlying variables, to match eControls. Still, the high number of variables considered in the ROPROVars propensity score can increase the variability of the ROPROVars propensity score, leading to a reduced performance.²¹ Hence, we have included the 5Vars prognostic score (or simply, 5Vars), which comprises the most prognostic variables in ROPRO, and can help characterize the effect of increasing the number of prognostic variables in the propensity score.

eControl performance metrics

Our main outcome of interest was overall survival (OS). We defined the survival time as the time from the randomization date (for clinical trial data), or the start of treatment date (for RWD) and the date of death. We calculated the OS hazard ratio (HR) using the Cox proportional hazards model.¹⁸ For each trial, we calculated the trial HR (obtained from the trial participants), and HR obtained with the eControls. We calculated the average OS HR error with the median absolute deviation (MAD) by pooling the OS HR errors from all trials. Additionally, we estimated the confidence interval (CI) of the OS HR error with bootstrap²² by resampling the trial-level error results.

Implementation

The analysis was performed using R 4.1.3.²³ We imputed missing variables at baseline using the missForest package.²⁴ To implement the regularized logistic regression, propensity score matching, and bootstrap we used the glmnet,²⁵ Matching,²⁶ and boot²² packages.

RESULTS

Characteristics of eligible trials

From the list of 16 experimental-control comparison clinical trials (see **Table 1**), half of the trials were phase III, and the remaining

half were phase II. All included trials except two (NCT02008227 and NCT01903993) studied the first-line of therapy. The remaining two trials included patients only after the failure of the first-line of therapy. Most considered clinical trials studied monoclonal antibody treatments (13 in total), although some (3 trials) considered small molecule treatments (e.g., Pictilisib). The experimental medications included approved medications (e.g., atezolizumab) and non-approved medications (e.g., parsatuzumab), as well as trials with heterogeneous sample sizes. Specifically, NCT02008227 had the highest patient number of all trials ($N = 1,187$), whereas NCT01493843 E-F included the lowest number of patients ($N = 91$). Additionally, all trials assigned the control population to a standard of care medication compatible with the patient's disease. For first-line of treatment, the control treatment was always platinum-doublet chemotherapy with or without bevacizumab. For the second-line of therapy trials, the patients were treated with docetaxel.

After applying the inclusion/exclusion criteria to the initial RWD population, the number of patients eligible to be included in the eControl was reduced by an average of 47%. The number of available patients for the eControls was below the number of treated patients in the clinical trial in two instances (NCT02008227 and NCT02367781; see **Table 2**). Therefore, independently of the matching procedure, the eControls for these two trials would always contain fewer patients than the experimental arm.

eControl cohorts

The two considered matching techniques yielded eControl cohorts with contrasting patient numbers (**Table 2**). First, the ROPRO eControls included a comparable number of patients with the trial controls. Specifically, for 14 out of 16 trials, the ROPRO eControls had a patient number higher than 90% of the number of patients of the trial control. The eControls of the remaining two trials (NCT02008227 and NCT02367781) included over

Table 2 Comparison of the patient number in the clinical trial and in the eControls

Trial	RW cohort after IE	Original trial		ROPRO			ROPROvars			5Vars		
	C	E	C	E	C	% C	E	C	% C	E	C	% C
NCT02008227	527	609	578	438	438	75.8	317	317	54.8	423	423	73.2
NCT01903993	416	142	135	132	132	97.8	110	110	81.5	133	133	98.5
NCT02366143 ABCP-BCP	741	394	394	374	374	94.9	306	306	77.7	336	336	85.3
NCT02366143 ACP-BCP	741	400	394	368	368	93.4	313	313	79.4	335	335	85
NCT01519804	2,136	55	54	54	54	100	55	55	102	55	55	102
NCT01496742	3,588	59	61	59	59	96.7	59	59	96.7	59	59	96.7
NCT01496742 Bev	818	69	70	69	69	98.6	67	67	95.7	66	66	94.3
NCT01366131	781	52	52	52	52	100	49	49	94.2	49	49	94.2
NCT01493843 A-B	1,259	126	124	126	126	102	115	115	92.7	119	119	96
NCT01493843 C-D	742	79	79	79	79	100	68	68	86.1	70	70	88.6
NCT01493843 E-F	742	61	30	61	61	203	59	59	197	59	59	197
NCT02367781	195	481	239	195	195	81.6	156	156	65.3	179	179	74.9
NCT02367794	620	332	334	315	315	94.3	236	236	70.7	249	249	74.6
NCT02367794 NAB	677	331	334	312	312	93.4	256	256	76.6	265	265	79.3
NCT02657434	3,147	292	282	291	291	103	268	268	95	265	265	94
NCT02409342	3,840	284	286	284	284	99.3	269	269	94.1	273	273	95.5

C, number of patients assigned to control group; "% C", percentage of patients in the eControl compared with the trial control; E, number of patients assigned to experimental group; IE, inclusion/exclusion criteria; RW, real-world.

75% of the number of patients of the control cohort. The eControls matched with the 5Vars propensity scores included, in general, less patients than the ROPRO eControls. Nevertheless, for 9 out of 16 trials, it included more than 90% of the number of patients in the original trial. Last, the ROPROvars eControls had an overall lower number of matched patients than the ROPRO and 5Vars matching methods. Still, for the majority of trials, the ROPROvars eControls had at least 75% of the number of patients of the control cohort, except for 3 trials NCT02367794, NCT02367781, and NCT02008227, which had 70.7%, 65.3%, and 54.8% of the trial control, respectively.

For some clinical trials, the eControl cohorts included more patients than the trial controls. Specifically, the number of patients of the NCT01493843 A-B, NCT01493843 E-F, and NCT02657434 ROPRO eControls had over 102%, 203%, and 103% of the number of trial control patients, respectively. Additionally, for the NCT01493843 E-F and NCT01519804, both the ROPROvars and 5Vars eControl cohorts had 197% and 102% of the number of patients of the trial control, respectively.

Hazard ratio results

The eControl cohorts constructed with the ROPRO, 5Vars, and ROPROvars methods were able to reproduce the trial control reasonably well in most cases (Table 3). There was an overall lower OS HR error for the ROPRO eControls (MAD: 0.072, bootstrap CI: 0.036–0.185) than for the 5Vars (MAD: 0.081, bootstrap CI: 0.025–0.283), or ROPROvars (MAD: 0.087, bootstrap CI: 0.054–0.383). Moreover, the ROPRO eControls obtained OS HR error lower than 0.05 for a total of 7 clinical trials (specifically, the error was 0 for NCT01496742 Bev, and NCT02366143

ABCP-BCP, 0.01 for NCT02008227, 0.03 for NCT02657434, 0.04 for NCT02409342, and 0.05 for NCT02366143 ACP-BCP and NCT02367794). Strikingly, all but one of the aforementioned clinical trials are phase III. The 5Vars and ROPROvars eControls had an error lower than 0.05 for 6, and 4 trials, respectively. Specifically, for the 5Vars, the OS HR error was 0, 0.01, 0.01, 0.02, 0.03, and 0.04 for NCT02367794 NAB, NCT01493843 C-D, NCT02409342, NCT02367794, NCT01903993, and NCT02008227. Last, for ROPROvars, the OS HR error of 0, 0.01, 0.04, 0.04, and 0.05 for the NCT02657434, NCT02366143 ACP-BCP, NCT02366143 ABCP-BCP, NCT01493843 E-F, and NCT02409342 trials, respectively.

We selected 3 trials to perform an in-depth analysis: NCT02366143 ABCP-BCP (phase III) NCT01519804 (phase II), and NCT01493843 C-D (phase II; Figure 2). For the first trial, the ROPRO, 5Vars, and ROPROvars eControls accurately represented the trial control (OS HR error of 0, 0.07, and 0.04 for ROPRO, 5Vars, and for ROPROvars, respectively). The ROPRO eControl cohorts had a low standardized mean difference (SMD) for the ROPRO variable (below 0.01). The ROPROvars and 5Vars had low SMD for other prognostic variables, such as age, ECOG, or albumin levels. Additionally, the ROPROvars and 5Vars eControl cohorts had lower SMD (below 0.1) for more variables, as expected by design. Conversely, for the NCT01519804 clinical trial, the eControls of all methods diverged from the trial control (OS HR error above 0.3 for all eControls). For this clinical trial, although the ROPRO eControl had a low SMD in the ROPRO variable, the 5Vars had a low SMD in the 5 controlled variables, and the ROPROvars eControl adequately balanced many of the variables (11 variables had SMD < 0.1), the eControls did not

Table 3 OS HRs from the clinical trials and from the eControl analysis with ROPRO and the 27 ROPRO variables

Trial	Trial OS HR	ROPRO OS HR	ROPROvars OS HR	5Vars OS HR
NCT02008227	0.81 (0.70–0.92)	0.82 (0.70–0.95)	0.89 (0.75–1.06)	0.77 (0.66–0.90)
NCT01903993	0.73 (0.57–0.95)	0.82 (0.62–1.07)	0.79 (0.59–1.05)	0.70 (0.54–0.92)
NCT02366143 ABCP-BCP	0.80 (0.68–0.94)	0.80 (0.67–0.94)	0.76 (0.63–0.91)	0.73 (0.62–0.87)
NCT02366143 ACP-BCP	0.85 (0.72–0.99)	0.90 (0.76–1.06)	0.84 (0.70–1.01)	0.78 (0.65–0.93)
NCT01519804	0.97 (0.60–1.57)	1.39 (0.86–2.24)	1.80 (1.09–2.99)	1.27 (0.79–2.04)
NCT01496742	1.14 (0.72–1.81)	1.29 (0.81–2.05)	1.52 (0.95–2.44)	1.42 (0.89–2.27)
NCT01496742 Bev	1.29 (0.77–2.14)	1.29 (0.79–2.10)	0.82 (0.52–1.30)	0.82 (0.51–1.32)
NCT01366131	1.07 (0.52–2.19)	0.78 (0.41–1.47)	0.65 (0.33–1.28)	0.62 (0.32–1.19)
NCT01493843 A-B	1.01 (0.73–1.39)	1.45 (1.06–2.00)	1.40 (1.00–1.95)	1.29 (0.93–1.78)
NCT01493843 C-D	1.04 (0.72–1.50)	1.48 (1.00–2.17)	1.17 (0.78–1.76)	1.03 (0.70–1.52)
NCT01493843 E-F	1.25 (0.74–2.11)	1.32 (0.85–2.06)	1.29 (0.84–2.00)	1.71 (1.09–2.69)
NCT02367781	0.81 (0.66–1.00)	0.68 (0.53–0.86)	0.67 (0.51–0.88)	0.67 (0.52–0.87)
NCT02367794	0.84 (0.70–1.00)	0.89 (0.74–1.07)	0.76 (0.62–0.94)	0.86 (0.70–1.05)
NCT02367794 NAB	0.95 (0.80–1.13)	1.02 (0.85–1.22)	0.86 (0.71–1.05)	0.95 (0.78–1.15)
NCT02657434	0.83 (0.68–1.01)	0.80 (0.66–0.96)	0.83 (0.68–1.02)	0.74 (0.60–0.90)
NCT02409342	0.82 (0.67–1.01)	0.86 (0.70–1.05)	0.77 (0.62–0.95)	0.83 (0.67–1.02)

Note: The bold values highlight the matching method with lowest OS HR error for each trial. Abbreviations: HR, hazard ratio; OS, overall survival.

match the trial control. The low number of patients included in the trial (55 in the experimental arm) alongside other trial-specific aspects, such as the enrichment for mesenchymal-to-epithelial transition, should have contributed to the lower performance.²⁷ Finally, for NCT01493843 C-D, the eControl of the 5Vars matching method accurately reproduced the trial control (OS HR error of 0.01). Conversely, the ROPROvars and ROPRO eControls had higher OS HR errors of 0.13 and 0.44, respectively. For the ROPRO eControls, specifically, there was an appropriate balancing of the ROPRO variable. Still, some of the underlying variables were not fully matched. For example, the ROPRO eControl patients had lower ALP and alanine aminotransferase, which could contribute to the higher times-to-event observed in the eControl population. Although the ROPROvars and ROPRO eControls obtained a lower performance for the C-D comparison, all 3 methods performed worse than on average for the NCT01493843 treatment-control comparisons (A-B, C-D, and E-F). The higher errors could be explained by lower cohorts (phase II), as well as for enrichment for phosphoinositide 3-kinases in the clinical trial.

Performance for phase III trials

We observed a significantly lower OS HR error for the phase III trials regardless of the matching method. The OS HR error for the ROPRO eControls was much lower (MAD: 0.044, bootstrap CI: 0.020–0.067). Additionally, the ROPRO eControls had an absolute OS HR error below 0.05 for 6 out of 8 phase III trials (Table 3, Figure 3). The 5Vars and ROPROvars eControl OS HR errors were also lower for the subset of phase III trials: 0.05 (bootstrap CI: 0.009–0.095), and 0.063 (bootstrap CI: 0.006–0.089), respectively.

DISCUSSION

In this analysis, we investigated the use of prognostic scores to create external control arms (eControls) for advNSCLC clinical

trials. Specifically, we contrasted the error in predicting the clinical trial's OS HR of eControls matched with a prognostic score (ROPRO), or propensity scores composed of either 5 (5Vars method) or 27 prognostic variables (ROPROvars). We performed this analysis in many advNSCLC trials (11, with 16 individual experimental-control comparison cohorts). The ROPRO approach obtained an overall lower error than 5Vars and ROPROvars methods. In particular, when only phase III clinical trials were considered, there was an even lower overall error for the ROPRO eControls vs. the 5Vars or ROPROvars eControls.

There was a noticeable difference in the number of patients included in the eControls of the three analyzed methods. The ROPRO eControls included, on average, 6.54% and 11.7% more patients than the 5Vars and ROPROvars eControls, respectively. This difference was due to a higher overlap in the propensity score distributions between the experimental and control populations for the ROPRO matching method vs. the 5Vars and ROPROvars. The overlap between the distributions decreased as more variables were added to the propensity score (i.e., the overlap was highest for the ROPRO eControls, and decreased sequentially with the 5Vars eControls, and finally with the ROPROvars). Therefore, using prognostic scores could allow to include more factors in the eControl construction, whereas also matching a larger number of patients. The difference in the number of patients matched into the eControl can strongly impact the statistical power of the analysis, which might be an important aspect to consider in a prospective study. Hence, in studies with a limited size of the eControl population, using prognostic models to match patients might lead to more matches, ultimately increasing the statistical power of the analysis.

One of the most striking results was the significantly lower OS HR prediction error for the phase III trials of the ROPRO eControls (decrease in OS HR MAD from 0.072 to 0.044). We argue that the decrease in error is likely due to the significant increase in

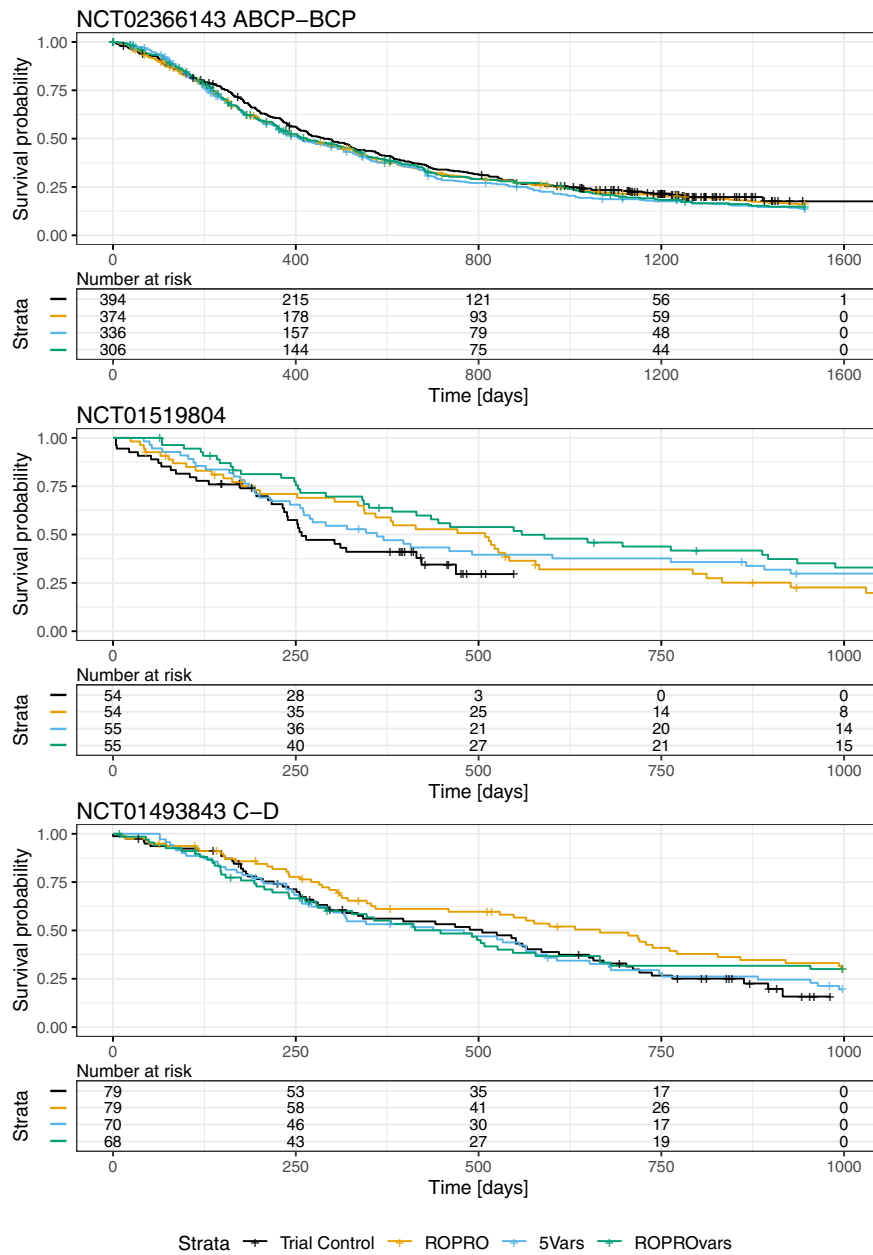


Figure 2 Kaplan-Meier curves of the trial control and eControls for the NCT02366143 (comparison of ABCP and BCP), NCT01519804, and NCT01493843 (comparison of the C and D arms) clinical trials.

patients between the phase II (average of 156 patients) and phase III trials (average of 745.5 patients). Phase II trials are usually not powered to observe a significant OS difference between the arms. The lower number of patients are expected to increase the error of the OS HR estimate. Another possible cause for the lower error could be that all the considered phase III trials studied cancer immunotherapy drugs. Still, some phase II trials also included immunotherapy drugs (e.g., onartuzumab). Therefore, we argue that this is a less likely cause. To further understand this effect, a future analysis, comprising many more studies, could analyze whether these methods have a higher performance for certain medication types.

The lower performance of the prognostic score model (ROPRO) for smaller sample sizes (phase II trials) is in line

with the results reported by Andrillon *et al.*²⁸ Andrillon *et al.* noted that matching on a smaller sample size can result in lower performance. Many of the methodological studies on matching^{8,9,20,29} rely on simulation studies that include large numbers of simulated patients (> 1,000 patients).²⁸ Although some phase III studies might include cohorts of this size, phase II and earlier will likely not include patient populations as large as this. Therefore, the current methodological literature might not be directly translatable to early-stage clinical trials, where the cohort size is closer to 100–200 patients. Thus, given the increased interest in eControls, more research should be devoted to the use of matching in settings with limited cohort sizes.

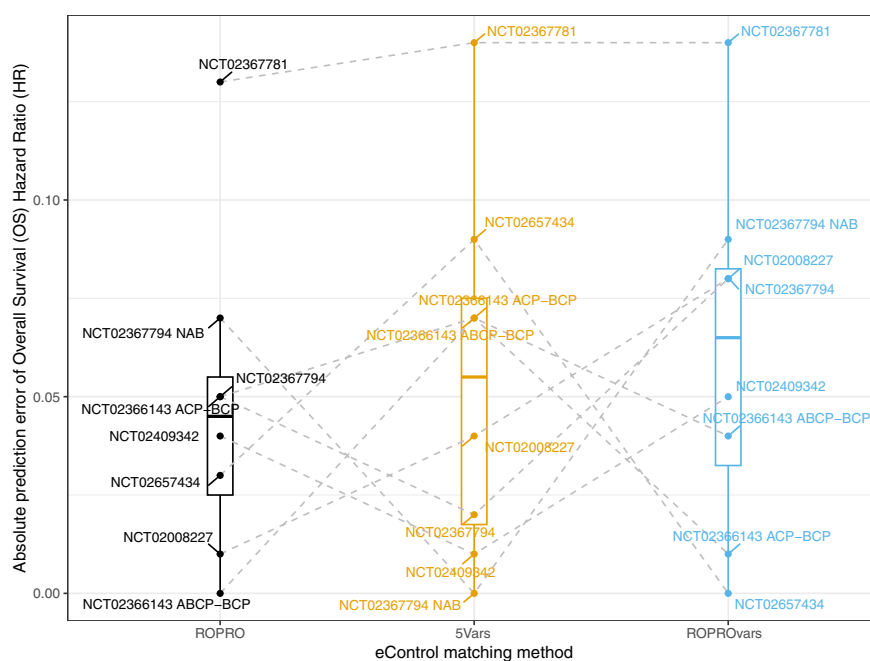


Figure 3 Prediction error of overall survival (OS) hazard ratio (HR) for each eControl matching method for the considered phase III clinical trials. The dashed lines connect the OS HR results for each trial over each matching method.

Apart from OS, other studies have successfully used eControls to derive important results for early phase trials.³⁰ For instance, Yin *et al.*³¹ matched 15 patients of a phase Ia clinical trial with historical clinical trial data, and were able to compare progression, as well as adverse events between the cohorts. Additionally, Ghione *et al.*³² created an eControl for the one-arm ZUMA-5 trial using RWD, and one historical clinical trial. With the eControl, the study team showed a substantially longer progression-free survival (PFS) for the patients in ZUMA-5. Both these examples provided study teams with additional evidence, outside of OS, to inform decisions during drug development. Still, especially in earlier phases, eControls must be correctly specified to avoid biased results.³

Limitations

The current analysis focused on advNSCLC. Analyses in other cancer indications are required to understand the applicability of eControls with and without prognostic scores. Additionally, most of the trials we considered tested cancer immunotherapy medications vs. standard chemotherapy. The analysis of trials of other medication comparisons should be performed to understand the suitability of these methods.

The use of prognostic scores instead of propensity scores carries additional limitations. First, the prognostic score is tied to the considered end point (in our analysis, it was mortality). In an analysis that considered multiple end points (e.g., OS as well as PFS, and adverse events), prognostic scores might not sufficiently balance the cohorts. Next, the use of prognostic scores does not lead to the same variable-level balance of propensity scores. Hence, in a prospective analysis, it might not be straightforward to assess the quality of balancing of the two cohorts. Therefore, more analyses are required to assess the quality of eControls created with

prognostic scores in these two situations. Additionally, more tools and guidelines on the use of prognostic scores to create eControls are necessary.

Here, we used only one data source to generate the eControls for these trials, but generalization to other RWD sources has to be analyzed in the future. The generalizability of this approach with other types of data, such as historical data from previous studies, should also be considered. The used data source was a limiting factor for two trials because after applying the inclusion/exclusion criteria there were already fewer patients than in the trial control. Additionally, some of the inclusion/exclusion criteria applied in the clinical trials could not be accounted for in this study due to the limitations of the dataset. Information such as diagnoses prior to the start of cancer treatment or other comorbidities was not available. The unavailability of this information led to a flawed application of the trial inclusion and exclusion criteria. The RWD-based eControl cohorts can include patients with comorbidities, which will negatively impact their time-to-event, which can ultimately lead to biased OS results.³³ Additionally, although the majority of blood test-based inclusion/exclusion criteria were available in the data source, there was a high missingness of measurements at baseline, limiting the correct application of the inclusion/exclusion criteria. The high missingness is a typical challenge when working with longitudinal data from real-world databases compared with clinical trial settings. Moreover, some blood tests were frequently unavailable, such as coagulation assays (e.g., prothrombin time), and information, such as the number of metastatic sites was inconsistent with the data from clinical trials, which led us to not use this information. Therefore, applying the inclusion/exclusion criteria could not be as strict as in the original studies. Last, some clinical trials (e.g., NCT01519804 or NCT01493843) consider

populations enriched for specific research biomarkers. RWD cohorts or historical clinical trials might not have tested the population for the studied biomarker, limiting the applicability of eControls in these studies.

CONCLUSION

The choice of which variables to account for in external control arms (eControls) is an important aspect to minimize the bias in matching. All methods tested, one based on a prognostic score (ROPRO) and 2 others based on propensity scores composed of 5 (5Vars) and 27 covariates (ROPROvars), replicated most of the trials' OS HR reasonably well. Overall, the eControls matched with ROPRO obtained a lower OS HR error than the two propensity score models.

Our results suggest that matching on prognostic scores could yield good results in phase III trials of advNSCLC. Prognostic scores could represent a suitable approach to control for a high number of variables without greatly increasing the variability of the propensity score.

ACKNOWLEDGMENTS

The authors thank Dr. Michael Bretscher (F. Hoffmann-La Roche AG) and Dr. Thibaut Sanglier (F. Hoffmann-La Roche AG) for their valuable input. H.L. is supported by the Helmholtz Association under the joint research school "Munich School for Data Science – MUDS."

FUNDING

This study was funded by F. Hoffmann-la Roche LTD. No grant number is applicable.

CONFLICTS OF INTEREST

A.B.-M., A.R., H.L., and M.S. are employed by F. Hoffmann-La Roche. T.B. is a contractor paid by F. Hoffmann-La Roche. All other authors declared no competing interest for this work.

AUTHOR CONTRIBUTIONS

A.B.-M., A.R., C.T.-L., H.L., M.S., and T.B. wrote the manuscript and designed the research. H.L. performed the research and analyzed the data.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study have been originated by Flatiron Health, Inc. Requests for data sharing by license or by permission for the specific purpose of replicating results in this manuscript can be submitted to dataaccess@flatiron.com.

© 2023 Roche Diagnostics GmbH. *Clinical Pharmacology & Therapeutics* published by Wiley Periodicals LLC on behalf of American Society for Clinical Pharmacology and Therapeutics.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

1. Unger, J.M., Cook, E., Tai, E. & Bleyer, A. The role of clinical trial participation in cancer research: barriers, evidence, and strategies. *Am. Soc. Clin. Oncol. Educ. Book* **35**, 185–198 (2016).
2. Viele, K. *et al.* Use of historical control data for assessing treatment effects in clinical trials. *Pharm. Stat.* **13**, 41–54 (2014).
3. Mishra-Kalyani, P.S. *et al.* External control arms in oncology: current use and future directions. *Ann. Oncol.* **33**, 376–383 (2022).

4. Wang, X., Dormont, F., Lorenzato, C., Latouche, A., Hernandez, R. & Rouzier, R. Current perspectives for external control arms in oncology clinical trials: analysis of EMA approvals 2016–2021. *J. Cancer Policy* **35**, 100403 (2023).
5. Jahanshahi, M. *et al.* The use of external controls in FDA regulatory decision making. *Ther. Innov. Regul. Sci.* **55**, 1019–1035 (2021).
6. Schmidli, H., Häring, D.A., Thomas, M., Cassidy, A., Weber, S. & Bretz, F. Beyond randomized clinical trials: use of external controls. *Clin. Pharmacol. Ther.* **107**, 806–816 (2020).
7. Weberpals, J. *et al.* Deep learning-based propensity scores for confounding control in comparative effectiveness research: a large-scale, real-world data study. *Epidemiology* **32**, 378–388 (2021).
8. Brookhart, M.A., Schneeweiss, S., Rothman, K.J., Glynn, R.J., Avorn, J. & Stürmer, T. Variable selection for propensity score models. *Am. J. Epidemiol.* **163**, 1149–1156 (2006).
9. Stuart, E.A., Lee, B.K. & Leacy, F.P. Prognostic score–based balance measures can be a useful diagnostic for propensity score methods in comparative effectiveness research. *J Clin Epidemiol* **66**, S84.e1–S90.e1 (2013).
10. Leacy, F.P. & Stuart, E.A. On the joint use of propensity and prognostic scores in estimation of the average treatment effect on the treated: a simulation study. *Stat. Med.* **33**, 3488–3508 (2014).
11. Smith, A.R. & Schaubel, D.E. Time-dependent prognostic score matching for recurrent event analysis to evaluate a treatment assigned during follow-up. *Biometrics* **71**, 950–959 (2015).
12. Bohn, J. *et al.* Controlling confounding in a study of Oral anticoagulants: comparing disease risk scores developed using different follow-up approaches. *EGEMS (Wash DC)* **7**, 27 (2019).
13. Becker, T. *et al.* An enhanced prognostic score for overall survival of patients with cancer derived from a large real-world cohort. *Ann. Oncol.* **31**, 1561–1568 (2020).
14. Becker, T., Mailman, M., Tan, S., Lo, E. & Bauer-Mehren, A. Comparison of overall survival prognostic power of contemporary prognostic scores in prevailing tumor indications. *Med. Res. Arch.* **11** (2023). <https://doi.org/10.18103/mra.v11i4.3638>.
15. Loureiro, H., Becker, T., Bauer-Mehren, A., Ahmidi, N. & Weberpals, J. Artificial intelligence for prognostic scores in oncology: a benchmarking study. *Front. Artif. Intell.* **4**, 625573 (2021).
16. Ma, X., Long, L., Moon, S., Adamson, B.J.S. & Baxi, S.S. Comparison of population characteristics in real-world clinical oncology databases in the US: Flatiron Health, SEER, and NPCR. *medRxiv* (2020). <https://doi.org/10.1101/2020.03.16.20037143>.
17. Birnbaum, B. *et al.* Model-assisted cohort selection with bias analysis for generating large-scale cohorts from the EHR for oncology research. <<http://arxiv.org/abs/2001.09765>> (2020).
18. Cox, D.R. Regression models and life-tables. *J. R. Stat. Soc. Ser. B Methodol.* **34**, 187–220 (1972).
19. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67**, 301–320 (2005).
20. Austin, P.C. Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharm. Stat.* **10**, 150–161 (2011).
21. Callendo, M. & Kopeinig, S. Some practical guidance for the implementation of propensity score matching. *J. Econ. Surv.* **22**, 31–72 (2008).
22. Davison, A.C. & Hinkley, D.V. *Bootstrap Methods and their Application* (Cambridge University Press, New York, NY, 1997).
23. R Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, 2018). <<https://www.R-project.org/>>.
24. Stekhoven, D.J. & Bühlmann, P. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics* **28**, 112–118 (2011).
25. Friedman, J.H., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1–22 (2010).




26. Sekhon, J.S. Multivariate and propensity score matching software with automated balance optimization: the matching package for R. *J. Stat. Softw.* **42**, 1–52 (2011).
27. Carrigan, G. et al. Using electronic health records to derive control arms for early phase single-arm lung cancer trials: proof-of-concept in randomized controlled trials. *Clin. Pharmacol. Ther.* **107**, 369–377 (2020).
28. Andriillon, A., Pirracchio, R. & Chevret, S. Performance of propensity score matching to estimate causal effects in small samples. *Stat. Methods Med. Res.* **29**, 644–658 (2020).
29. Austin, P.C., Grootendorst, P. & Anderson, G.M. A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. *Stat. Med.* **26**, 734–753 (2007).
30. Yap, T.A., Jacobs, I., Baumfeld Andre, E., Lee, L.J., Beaupre, D. & Azoulay, L. Application of real-world data to external control groups in oncology clinical trial drug development. *Front. Oncol.* **11**, 695936 (2022).
31. Yin, X. et al. Historic clinical trial external control arm provides actionable GEN-1 efficacy estimate before a randomized trial. *JCO Clin. Cancer Inform.* **7**, e2200103 (2023).
32. Ghione, P. et al. Comparative effectiveness of ZUMA-5 (axi-cel) vs SCHOLAR-5 external control in relapsed/refractory follicular lymphoma. *Blood* **140**, 851–860 (2022).
33. Incerti, D., Bretscher, M.T., Lin, R. & Harbron, C. A meta-analytic framework to adjust for bias in external control studies. *Pharm. Stat.* **22**, 162–180 (2023).
34. Rittmeyer, A. et al. Atezolizumab versus docetaxel in patients with previously treated non-small-cell lung cancer (OAK): a phase 3, open-label, multicentre randomised controlled trial. *Lancet* **389**, 255–265 (2017).
35. Fehrenbacher, L. et al. Atezolizumab versus docetaxel for patients with previously treated non-small-cell lung cancer (POPLAR): a multicentre, open-label, phase 2 randomised controlled trial. *The Lancet* **387**, 1837–1846 (2016).
36. Socinski, M.A. et al. IMpower150 final overall survival analyses for Atezolizumab plus bevacizumab and chemotherapy in first-line metastatic nonsquamous NSCLC. *J. Thorac. Oncol.* **16**, 1909–1924 (2021).
37. von Pawel, J. et al. Randomized phase II trial of Parsatuzumab (anti-EGFL7) or placebo in combination with carboplatin, paclitaxel, and bevacizumab for first-line nonsquamous non-small cell lung cancer. *Oncologist* **23**, 654-e58 (2018).
38. West, H. et al. Atezolizumab in combination with carboplatin plus nab-paclitaxel chemotherapy compared with chemotherapy alone as first-line treatment for metastatic non-squamous non-small-cell lung cancer (IMpower130): a multicentre, randomised, open-label, phase 3 trial. *Lancet Oncol.* **20**, 924–937 (2019).
39. Jotte, R. et al. Atezolizumab in combination with carboplatin and nab-paclitaxel in advanced squamous NSCLC (IMpower131): results from a randomized phase III trial. *J. Thorac. Oncol.* **15**, 1351–1360 (2020).
40. Nishio, M. et al. Atezolizumab plus chemotherapy for first-line treatment of nonsquamous NSCLC: results from the randomized phase 3 IMpower132 trial. *J. Thorac. Oncol.* **16**, 653–664 (2021).
41. Herbst, R.S. et al. Atezolizumab for first-line treatment of PD-L1–selected patients with NSCLC. *N. Engl. J. Med.* **383**, 1328–1339 (2020).

Full length manuscript III - Correlation between early trends of a prognostic biomarker and overall survival in non-small-cell lung cancer clinical trials

Hugo Loureiro, Theresa M. Kolben, Astrid Kiermaier, Dominik Rüttinger, Narges Ahmidi, Tim Becker*, Anna Bauer-Mehren* (* contributed equally)

The manuscript was peer-reviewed and is published under the Creative Commons Attribution License (CC BY-NC-ND 4.0) license in the JCO Clinical Cancer Informatics journal. Hence, the article was published open-access, and the published version is included.

Correlation Between Early Trends of a Prognostic Biomarker and Overall Survival in Non–Small-Cell Lung Cancer Clinical Trials

Hugo Loureiro, MSc^{1,2,3} ; Theresa M. Kolben, MD, PhD⁴; Astrid Kiermaier, PhD⁵; Dominik Rüttinger, MD, PhD⁶; Narges Ahmadi, PhD² ; Tim Becker, PhD¹; and Anna Bauer-Mehren, PhD¹ 

DOI <https://doi.org/10.1200/CC1.23.00062>

ABSTRACT

PURPOSE Overall survival (OS) is the primary end point in phase III oncology trials. Given low success rates, surrogate end points, such as progression-free survival or objective response rate, are used in early go/no-go decision making. Here, we investigate whether early trends of OS prognostic biomarkers, such as the ROPRO and DeepROPRO, can also be used for this purpose.

METHODS Using real-world data, we emulated a series of 12 advanced non–small-cell lung cancer (aNSCLC) clinical trials, originally conducted by six different sponsors and evaluated four different mechanisms, in a total of 19,920 individuals. We evaluated early trends (until 6 months) of the OS biomarker alongside early OS within the joint model (JM) framework. Study-level estimates of early OS and ROPRO trends were correlated against the actual final OS hazard ratios (HRs).

RESULTS We observed a strong correlation between the JM estimates and final OS HR at 3 months (adjusted $R^2 = 0.88$) and at 6 months (adjusted $R^2 = 0.85$). In the leave-one-out analysis, there was a low overall prediction error of the OS HR at both 3 months (root-mean-square error [RMSE] = 0.11) and 6 months (RMSE = 0.12). In addition, at 3 months, the absolute prediction error of the OS HR was lower than 0.05 for three trials.

CONCLUSION We describe a pipeline to predict trial OS HRs using emulated aNSCLC studies and their early OS and OS biomarker trends. The method has the potential to accelerate and improve decision making in drug development.

ACCOMPANYING CONTENT

 [Data Supplement](#)

Accepted September 7, 2023

Published November 3, 2023

JCO Clin Cancer Inform

7:e2300062

© 2023 by American Society of

Clinical Oncology

Creative Commons Attribution
Non-Commercial No Derivatives
4.0 License

INTRODUCTION

In oncology clinical trials, the gold standard measure of efficacy is overall survival (OS). To get a reliable estimate of OS, a high number of patients and a long follow-up are required.^{1,2} These requirements constrain the estimation of OS across clinical trial phases: (1) in early phases (I/II) where both the number of patients and follow-up time are limited and (2) in interim analyses of late phases, where the follow-up time might still be a limiting factor.

To assist with early go/no-go decisions in the aforementioned settings, several statistical and machine learning tools have been proposed. Specifically, Shameer et al³ created a software tool that predicts the OS hazard ratio (HR) on the basis of progression-free survival (PFS) values from early interim analyses. Seo et al⁴ approached this problem by analyzing the relationship between the molecular structure, drug target, and drug success. In addition, Beinse et al⁵ and Hegge et al⁶ aimed to predict success of new molecules given

drug/trial characteristics and results from phases I and II, respectively. Finally, Schperberg et al⁷ created an algorithm to predict both OS and PFS results given drug, target, and trial characteristics. In this work, we explore whether early OS in combination with the early trend of oncology prognostic scores is predictive of final OS results and can be used to inform go/no-go decisions.

Oncology prognostic scores, such as ROPRO⁸ or DeepROPRO,⁹ that were recently introduced by our group are correlated with OS. Both models combine a set of 27 parameters describing the host (demographics, vitals, and blood test parameters), the lifestyle (BMI and smoking history), and the tumor characteristics, all of which are associated with cancer survival. We use the (prognostic score) risk trend from baseline (treatment start) in a joint modeling framework, which measures the patient-level deviation of these scores from the start of treatment. We hypothesize that the risk trend can represent the actual improvement/deterioration of the patient's condition/fitness over time.

CONTEXT

Key Objective

Early efficacy predictions in interim analyses of clinical trials are usually dependent on surrogate end points such as progression-free survival. In this introductory analysis, we explored the correlation between the longitudinal trend of prognostic scores (risk trend) at early time points (equivalent to interim analysis) and efficacy.

Knowledge Generated

We considered 12 clinical trials emulated with a large real-world database. The risk trend in interim analyses strongly correlated with the efficacy.

Relevance

The observed correlation suggests that the early risk trend could be an interesting additional tool for internal decision making. Still, further validation analyses in different types of data are necessary to develop the methodology.

Therefore, in a clinical trial setting, we expect that the treatment arm whose patients have the highest improvement in risk trend should be the arm with the highest OS benefit.

In this study, we performed retrospective analyses to investigate the applicability of the risk trends, combined with observed early OS results, to inform go/no-go decisions in interim analyses of late-stage clinical trials. For the benchmarking of our approach, we emulated 12 recent phase III clinical trials (covering a wide range of medication types) using data from a large real-world database and assessed the performance of our approach in these data.

METHODS

Ethics Statement

Institutional Review Board approval of the Flatiron Health (FH) study protocol for data collection from the real-world cohort was obtained before study conduct, including a waiver of informed consent. Additional details on Flatiron's institutional review board approval are outlined below:

- IRB name: WCG IRB
- Protocol No. and title: RWE-001: The Flatiron Health Real World Evidence Parent Protocol
- Registration No.: IRB00000533
- Protocol approval ID/tracking No.: 420180044

Joint Modeling of Risk Trend and OS

We modeled OS alongside the risk trends using Joint Models for Longitudinal and Survival Data (in short JM).¹⁰ JM couples the risk trend from baseline (the longitudinal variable) with the survival information, measuring the impact of the risk trend on survival and hence eliminating the possible bias.¹⁰ Specifically, we defined the JM as

$$h_i(t) = h_0(t) \exp\{\gamma \cdot \text{treatment}_i + \alpha \cdot \text{risk}_{\text{trend},i}(t)\},$$

$$\text{risk}_{\text{trend},i}(t) = \beta_0 + (b_{1,i} + \beta_1) \cdot t + \beta_2 t \cdot \text{treatment}_i. \quad (1)$$

where $h(t)$ is the hazard function, $h_0(t)$ is the baseline hazard, γ is the direct effect of the treatment on the hazard (ie, analogous to the classical OS HR of the treatment), β_0 and β_1 are the intercept and slope coefficients, β_2 is an additional slope coefficient that depends on the arm of treatment, and b_1 is a slope random effect. The $\text{risk}_{\text{trend}}$ value at time t for patient i is given by the difference between the risk at time t and the baseline risk: $\text{risk}_{\text{trend},i}(t) = \text{risk}_i(t) - \text{risk}_i(t=0)$. The risk values at all time points ($\text{risk}_i(t)$) were calculated using the selected prognostic scores.

As an auxiliary, we also visualize the average progression of the risk trends using the locally estimated scatterplot smoothing (LOESS) smoother¹¹ and provide an illustration of $\text{risk}_{\text{trend}}$. These visualizations are solely for illustrative purposes and are not used in the actual framework described below.

Figure 1 summarizes the risk trend framework.

Correlation Analysis

Within each study, we estimated the JM coefficient using either all information available until 3 months or, alternatively, until 6 months (all future patient information was censored) and extracted the JM coefficients (β_2 , γ). These two time points represent clinical trial interim analyses. Next, we investigated whether the early JM coefficients correlated with the final study OS HRs. Specifically, we performed a weighted linear regression

$$\log(\text{OS HR}_j) = \theta_0 + \theta_1 \beta_{2,j} + \theta_2 \gamma_j + \epsilon, j \in \{1, \dots, 12\} \quad (2)$$

between the JM coefficients (β_2 , γ) from early time points and the final OS HR of the 12 emulated studies (j is the trial

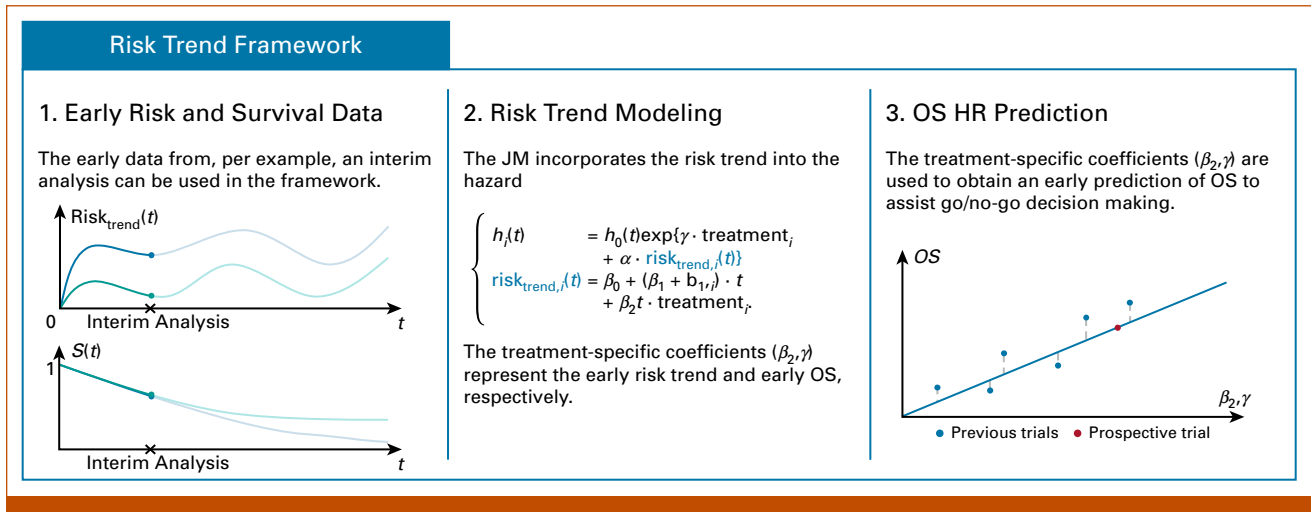


FIG 1. The risk trend framework workflow. HR, hazard ratio; JM, joint model; OS, overall survival.

iterator and θ are regression coefficients). The regression was weighted by the inverse variance of $OSHR_j$. The adjusted multiple correlation coefficient (R^2) of the linear regression and Kendall τ served as quality measures of the regression. In addition, we performed a leave-one-study-out analysis, where we predicted the final OS HR of the study from its specific JM coefficients using the regression formula (Eq 2) derived from the other 11 studies. In the leave-one-out analysis, we used the root-mean-square error (RMSE) to characterize the prediction performance. We estimated the CI of the R^2 and RMSE with bootstrap¹² by resampling the trial results and refitting Equation 2.

Finally, to describe the contribution of the β_2 and γ coefficients to the OS HR prediction, we performed an additional analysis where we fit the JM (Eq 1) without the γ coefficient. Next, we performed the same linear regression and leave-one-out analyses (Eq 2), including only the β_2 coefficient.

Clinical Trial Emulation with RWD

We emulated previously conducted clinical trials with RWD to obtain a data basis on which to evaluate our prediction framework since the actual study data were not available to us in the majority of cases. First, we gathered a comprehensive (Fig 2) list of phase III lung cancer clinical trials from ClinicalTrials.gov covering multiple medication types. Next, we emulated these clinical trials using the deidentified electronic health record (EHR)-derived FH database. The FH database is a longitudinal database, comprising deidentified patient-level structured and unstructured data, curated via technology-enabled abstraction.^{13,14} From FH, we extracted deidentified information collected between January 2011 and December 2020 from approximately 280 US cancer clinics (approximately 800 sites of care) about the first-line treatment of 34,061 patients diagnosed with advanced non-small-cell lung cancer (aNSCLC).

We focused on phase III aNSCLC studies since (1) aNSCLC is one of the most common types of cancers and (2) phase III trials usually report OS results. We extracted a list of 184 clinical trials from ClinicalTrials.gov (accessed on October 5, 2021). From the initial list, we excluded 171 clinical trials on the basis of the trial design and patient availability in FH (Fig 2 shows detailed criteria). The final list of 12 potentially reproducible clinical trials is included in Table 1 and the Data Supplement (Table S1; following the study by Yang et al,¹⁵ we incorporated LUX-Lung 3 and LUX-Lung 6 together, lowering the number of trials by one).

To emulate the clinical trials, we selected patients from FH who were prescribed the trial's medications. We applied the clinical trial-specific inclusion/exclusion criteria (such as tumor histology or specific tumor mutations). We relaxed bloodwork inclusion/exclusion criteria following the results from the study by Liu et al.¹⁶ Next, to reduce confounding, we applied propensity score matching^{17,18} on the baseline (before treatment) ROPRO value. Propensity score matching attempts to create experimental and control arms with reduced confounding bias. The approach of controlling for confounding with prognostic scores was introduced in the landmark study by Stuart et al,¹⁸ where they showed that prognostic score values controlled for the bias. Finally, to verify that our emulated clinical trials are concordant with the original clinical trials, we contrasted their OS HR confidence intervals.

Implementation

The analyses were performed using R 4.0.4¹⁹ and Python 3.6. The ROPRO⁸ and DeepROPRO⁹ were calculated as specified in the original publications. The propensity score matching, DeepSurv, and JM were implemented using the MatchIt,²⁰

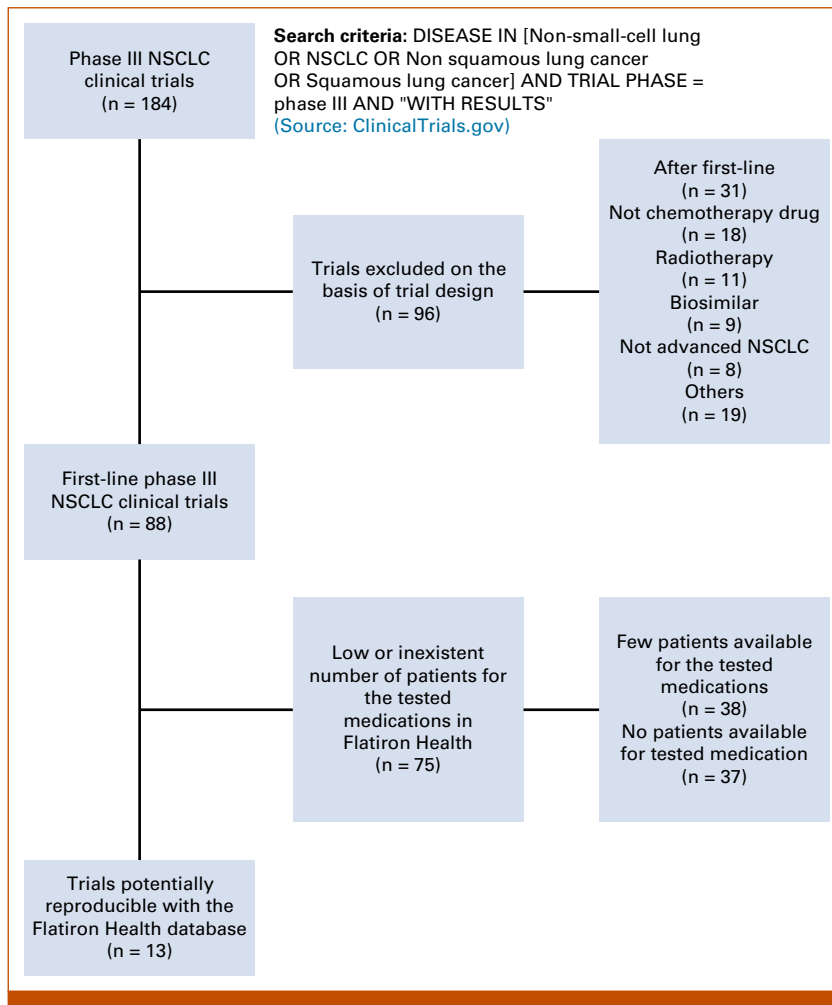


FIG 2. Flowchart of the clinical trial selection. NSCLC, non-small-cell lung cancer.

DeepSurv,²¹ and JM²² packages, respectively. The analysis code is available in GitHub.²³

RESULTS

Evaluation and Sensitivity of the Emulated Trials

The data sets of the 12 RWD-emulated clinical trials cover a wide range of patient numbers (from 190 to 2,156 patients) per treatment arm (baseline characteristics are available in the Data Supplement [Tables S3–S14]). The emulated clinical trials were generally more inclusive than the original trials, including more female patients (who are still under-represented in clinical trials,²⁴ and also slightly older patients [eg, median age of the KEYNOTE-189 control arm: 63.5 years; median age of the KEYNOTE-189 emulated control: 68 years]). Regardless of the slight differences in baseline characteristics, the OS results of the emulated and original clinical trials were consistent in the majority of trials. There were a high correlation between the emulated and actual HR (R^2 of 0.86) and moderate error (RMSE of 0.17). Still, there were some observed differences in the

emulated OS. For instance, in both LUX-Lung 3+6 and PROFILE 1014, there was a higher OS benefit in the emulated (OS HR LUX-Lung 3+6: 0.40, PROFILE 1014: 0.50) versus original (OS HR LUX-Lung 3+6: 0.81, PROFILE 1014: 0.67) clinical trials (the Data Supplement [Fig S1] contains a comparison of the actual and emulated HRs).

Correlation of JM Coefficients With Final OS HR

Next, we explored the correlation between the early treatment-specific JM coefficients (β_2 , γ) in our makeshift interim analyses and the final OS HR in the 12 considered trials. The JM coefficients at 3 months highly correlated with the final OS HR (ROPRO JM adjusted R^2 values [bootstrap CI]: 0.88 [0.62 to 0.98], Kendall τ : 0.82, and Fig 3). The 6-month JM coefficients similarly correlated with the final OS HR (ROPRO JM adjusted R^2 values [bootstrap CI]: 0.85 [0.52 to 0.98], Kendall τ : 0.82). In addition, the DeepROPRO JM coefficients had similarly high correlation with the final OS HR (3-month adjusted R^2 values [bootstrap CI]: 0.86 [0.52 to 0.98], and 6-month adjusted R^2 values [bootstrap CI]: 0.82 [0.40 to 0.98], Kendall τ : 0.70).

TABLE 1. Absolute Prediction Error of the Trial’s OS HR Using the ROPRO Risk Trend and Early OS

Trial Name	OS Treatment Benefit	Risk Trend Treatment Benefit	ROPRO Absolute Error of OS HR Prediction (3 months)	ROPRO Absolute Error of OS HR Prediction (6 months)
KEYNOTE-189	Platinum + pembrolizumab + pemetrexed	Platinum + pembrolizumab + pemetrexed	0.16	0.25
KEYNOTE-024	Pembrolizumab	Pembrolizumab	0.02	0.06
KEYNOTE-042	Pembrolizumab	Pembrolizumab	0.04	0.05
KEYNOTE-407	Carboplatin + pembrolizumab + paclitaxel/nab-paclitaxel	Carboplatin + pembrolizumab + paclitaxel/nab-paclitaxel	0.13	0.11
PRONOUNCE	–	Bevacizumab + carboplatin + paclitaxel	0.12	0.18
PointBreak	–	–	0.19	0.10
PROFILE 1014	Crizotinib	Crizotinib	0.10	0.10
FLAURA	Osimertinib	Osimertinib	0.05	0.02
LUX-Lung 3+6	Afatinib	Afatinib	0.15	0.19
NCT00540514	–	–	0.08	0.01
AURA3	Osimertinib	Osimertinib	0.03	0.02
NCT00520676	–	–	0.12	0.00

Abbreviations: HR, hazard ratio; OS, overall survival.

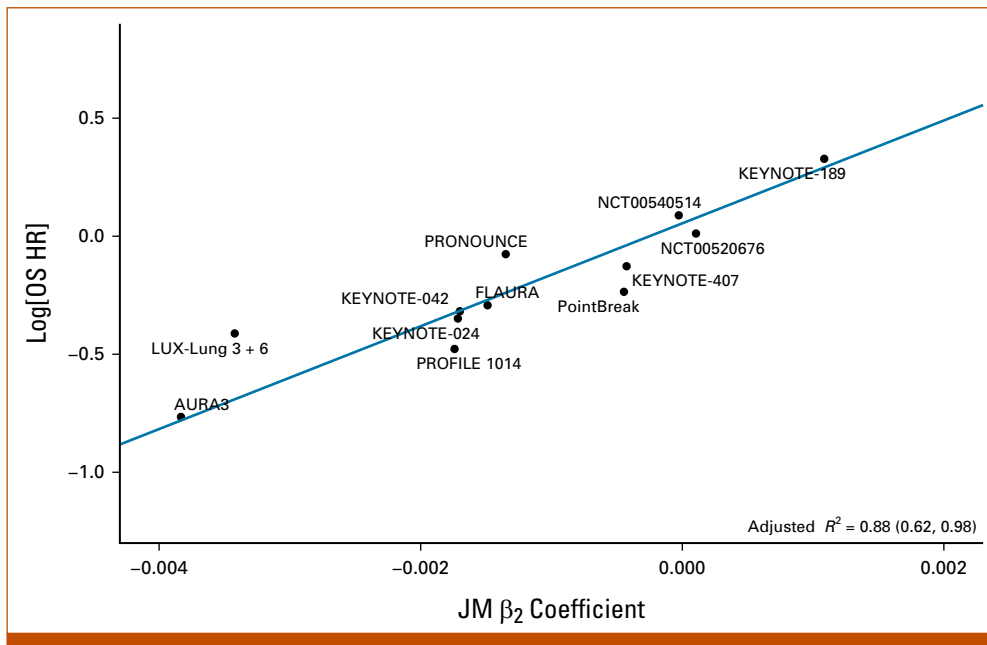


FIG 3. Scatter plot of the final OS HRs versus the ROPRO JM coefficients at 3 months. The plot includes the adjusted R^2 value and its bootstrapped CI of the regression. Our linear prediction model of OS HR depends on both β_2 and γ coefficients; here, we consider the plane $\gamma = 0$ to make a simple 2D representation of the prediction model and correlation. 2D, two-dimensional; HR, hazard ratio; JM, joint model; OS, overall survival.

As a sensitivity analysis, we performed the previous analysis without the two emulated studies that had a higher OS benefit than the original trials. The correlation results were similar to the those in previous analysis (3-month adjusted R^2 [bootstrap CI] 0.90 [0.46 to 0.99], Kendall τ 0.87).

Prediction of Final OS HR Using JM Coefficients

After the positive correlation between the JM coefficients and the OS HR, we investigated whether the JM coefficients (β_2 , γ) could predict the final OS HR values in a leave-one-out analysis. The JM coefficients obtained from only 3 months of data (Fig 4) predicted the final OS HR with low error (RMSE [bootstrap CI] for ROPRO JM: 0.11 [0.08 to 0.14] and DeepROPRO JM: 0.11 [0.08 to 0.14]). Remarkably, for the ROPRO JM models (considering both coefficients [β_2 , γ]), five studies had an absolute OS HR error of <0.1 and three trials had an absolute error lower than 0.05 (AURA3: 0.03, KEYNOTE-042: 0.04, and KEYNOTE-024: 0.02). When additional data were added to the models (up to 6 months), there was a similar overall prediction error (RMSE [bootstrap CI] for ROPRO: 0.12 [0.07 to 0.16] and for DeepROPRO: 0.13 [0.08 to 0.17]). Although, for the ROPRO risk trend, there were more studies (eight in total) that had an absolute OS HR error value lower than 0.1, five had an absolute OS HR error lower than 0.05 (specifically, FLAURA: 0.03, ClinicalTrials.gov identifier: NCT00540514: 0.01, AURA3: 0.02, ClinicalTrials.gov identifier: NCT00520676: 0.00). The full prediction errors for the ROPRO and DeepROPRO analyses are available in Table 1 and the Data Supplement (Table S2), respectively.

Characterization the Effect of the β_2 and γ Parameters

To determine which parameter had larger predictive performance, we performed an additional analysis without the γ parameter. The performance using only the β_2 parameter decreased only slightly in the 3-month (ROPRO adjusted R^2 [bootstrap CI]: 0.85 [0.53 to 0.97], Kendall τ : 0.88) and 6-month (ROPRO adjusted R^2 [bootstrap CI]: 0.86 [0.46 to 0.98], Kendall τ : 0.88) analyses when compared with the previous results.

Additional Analysis: Risk Trend Concordance With the OS Benefit

In addition, we verified that the risk trends (the signs of the β_2 coefficient) were concordant with the original clinical trial's OS benefit in 11 of 12 clinical trials (all but PRONOUNCE, Table 1, Fig 5). That is, the medications that had the lowest risk trend (ie, highest risk improvement over time) also had the highest OS. Only in PRONOUNCE, the carboplatin + pemetrexed treatment arm had significantly higher risk trend values although there was no difference in OS between the arms in the original study. A representation of the risk curves is available in the Data Supplement (Figs S2–S13).

DISCUSSION

We introduced a new research and development (R&D) decision support methodology leveraging the time course of an OS prognostic marker. It models the difference in the risk

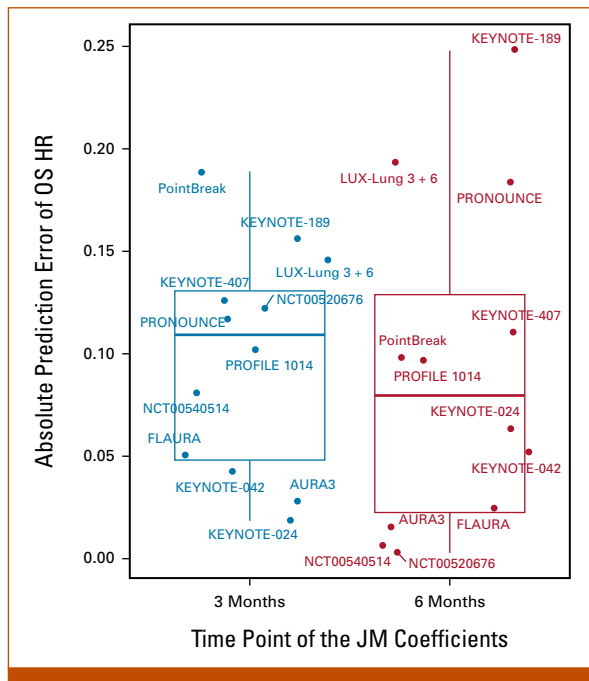


FIG 4. Absolute prediction error of the OS HRs obtained with the ROPRO JM coefficients. HR, hazard ratio; JM, joint model; OS, overall survival.

trend and early OS between treatment arms with JMs. Our results show that the early JM coefficients (at 3 and 6 months) correlated with the final OS HR and could predict the final OS HR of unseen trials with a small error (in most cases as low as 0.1).

The JM framework had an adequate performance when lower amounts of information were included. Specifically, the correlation performance with 3-month JM coefficients was similar to the performance later at 6 months (Kendall τ of 0.82 and 0.82, adjusted R^2 of 0.88 and 0.85, respectively). The prediction performance was also similar (RMSE of 0.11 and 0.12, respectively). In addition, the performance of the risk trend framework did not decrease substantially when the sample size was lower (PROFILE 1014 and ClinicalTrials.gov identifier: [NCT00520676](#), both studies with about 190 patients per arm). For both these studies, there was a low absolute prediction error of the final OS HR (always below 0.11 and as low as 0 for ClinicalTrials.gov identifier: [NCT00520676](#)).

The JM coefficients obtained higher correlation (adjusted R^2) results than other analyses that considered the correlation between early PFS and final OS HR. Specifically, the recent analysis by Shameer et al,²⁵ which also considered multiple mechanisms of action, reported overall R^2 values of 0.23 (and for PD-1/PD-L1 inhibitors of 0.86), whereas our adjusted R^2 was higher at 0.88. Still, we performed our analysis on a smaller number of studies and drugs, and therefore, our adjusted R^2 value is subject to change. We have to admit that

our analysis is based on a comparatively small number of studies, and we cannot rule out the possibility that the power of our approach is to some extent overestimated by mere chance. Still, we think that the results presented here are substantial and support the role of early OS/risk trend modeling as a further decision making tool. The method is not intended as a replacement, but as an add-on to prediction via progression results. We note that one could construct a combined framework in the future; progression, early OS, and risk trend could be simultaneously modeled in the JM framework to further improve OS HR prediction.

Following the moderate error obtained in the prediction analysis, we argue that one possible use for the risk trend framework could be to inform futility analyses in phase III interim analyses. In addition, another possible use of the risk trend framework could be as an initial indicator of OS benefit in early clinical trial phases (this setting was not considered in this work and needs to be studied in a future analysis). At the aforementioned stages, the risk trend framework could be used alongside other methods such as those in the studies by Beinse et al⁵ and Shameer et al³ to inform go/no-go decisions. All these methods consider different types of data, and hence, their joint use could more comprehensively describe the effects of the drug. In summary, our analysis suggests that the risk trend framework has the potential to serve as a valuable additional R&D support tool. Still, further analyses would be necessary to validate the risk trend framework in these settings.

At this stage of the risk trend framework, we did not attempt to formally prove its surrogacy to OS according to the guidelines introduced by Prentice.²⁶⁻²⁸ Although the framework might have the potential to generate a surrogacy end point, we focused here solely on its possible utility in early decision making. Proof of surrogacy would be an effort reaching beyond the scope of this manuscript, requiring the involvement of further types of data (eg, clinical studies and other RWD sources), other pharmaceutical companies, and academic institutions.

In addition, we focused on aNSCLC. The good performance at early time points (3 months) is likely partially due to the generally low median survival time observed in aNSCLC. Further analyses are required to validate the framework in other cancer indications, also with respect to the identification of optimal time points for interim analyses. Nevertheless, from our experience with the baseline pan-cancer ROPRO,^{8,29} we believe that the risk trend of ROPRO has the potential to perform well in other indications.

Since our analyses were conducted using RWD, it has to be shown that the results translate to clinical trials. More validation is needed, especially in early clinical trial phases. Finally, our analysis is biased toward trials testing efficacious medications as only these medications are available in RWD. We plan to investigate the risk trend framework further in drugs that failed to show efficacy.

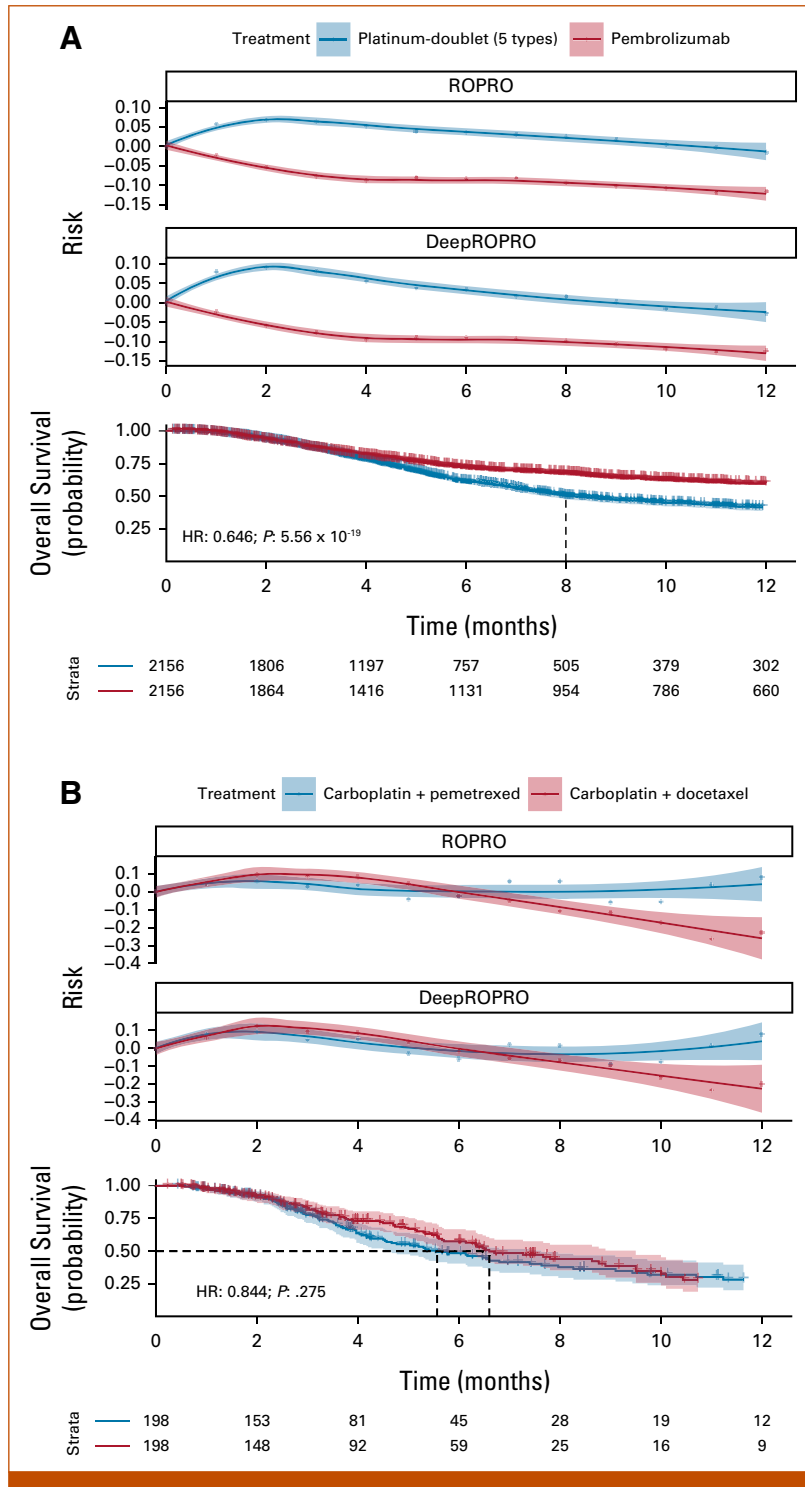


FIG 5. Superimposed risk trend and survival curves of the (A) KEYNOTE-024 and (B) NCT00520676 clinical trials. In each panel, the first two plots are the ROPRO and DeepROPRO trends, followed by the OS Kaplan-Meier curves and the risk table. The HR is referent to the effect of the pembrolizumab and carboplatin + docetaxel, respectively. HR, hazard ratio; OS, overall survival.

Finally, we used the ROPRO and DeepROPRO prognostic scores to calculate the risk. These prognostic scores are composed mainly of vital and blood test parameters. There are other, independent, prognostic biomarkers that were not included in these models (cancer-specific biomarkers, circulating tumor DNA, and C-reactive protein, among others). These can be investigated in their own right, using the risk trend framework we exemplified here, or be combined into a joined score to further increase the performance.

In conclusion, trustworthy estimates of OS are essential for precise decision making in clinical trials. Our results show

that the early OS/risk trend framework (using ROPRO/DeepROPRO) predicted treatment benefit for the majority of emulated clinical trials studied. In our analysis, prediction of the OS HR with a low error was possible at 3 and 6 months after the start of treatment for most considered trials.

The results of this initial analysis introduce the risk trend framework as a potential new R&D decision support tool for aNSCLC clinical trials. Further analyses in clinical studies and other RWD sources are necessary to further validate the risk trend framework in aNSCLC.

AFFILIATIONS

¹Data & Analytics, Pharmaceutical Research and Early Development, Roche Innovation Center Munich (RICM), Penzberg, Germany

²Computational Health Center, Helmholtz Munich, Munich, Germany

³TUM School of Life Sciences Weihenstephan, Technical University of Munich, Freising, Germany

⁴Early Clinical Development Oncology, Pharmaceutical Research and Early Development, Roche Innovation Center Munich (RICM), Penzberg, Germany

⁵Research and Early Development, Roche Innovation Center Basel (RICB), Basel, Switzerland

⁶Research and Early Development Oncology, Pharmaceuticals, Bayer AG, Berlin, Germany

CORRESPONDING AUTHOR

Anna Bauer-Mehren, Pharmaceutical Research and Early Development (pRED) Data & Analytics, Roche Diagnostics GmbH, Nonnenwald 2, 82377 Penzberg, Germany; e-mail: anna.bauer-mehren@roche.com.

EQUAL CONTRIBUTION

T.B. and A.B.-M contributed equally to this work.

SUPPORT

Supported by the Helmholtz Association under the joint research school Munich School for Data Science (MUDS), in which Hugo Loureiro is a doctoral researcher.

AUTHOR CONTRIBUTIONS

Conception and design: Theresa M. Kolben, Astrid Kiermaier, Narges Ahmidi, Tim Becker, Anna Bauer-Mehren

Collection and assembly of data: Hugo Loureiro, Tim Becker

Data analysis and interpretation: Hugo Loureiro, Theresa M. Kolben, Narges Ahmidi, Tim Becker

Manuscript writing: All authors

Final approval of manuscript: All authors

Accountable for all aspects of the work: All authors

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

The following represents disclosure information provided by authors of this manuscript. All relationships are considered compensated unless otherwise noted. Relationships are self-held unless noted.

I = Immediate Family Member, Inst = My Institution. Relationships may not relate to the subject matter of this manuscript. For more information about ASCO's conflict of interest policy, please refer to www.asco.org/rwc or ascopubs.org/cci/author-center.

Open Payments is a public database containing information reported by companies about payments made to US-licensed physicians ([Open Payments](http://OpenPayments)).

Hugo Loureiro

Employment: Roche

Theresa M. Kolben

Employment: Bayer HealthCare Pharmaceuticals, Roche

Stock and Other Ownership Interests: Bayer HealthCare Pharmaceuticals, Roche

Patents, Royalties, Other Intellectual Property: I am a patent holder based on my work at Roche

Astrid Kiermaier

Employment: Roche

Stock and Other Ownership Interests: Roche

Patents, Royalties, Other Intellectual Property: Patent applications in context of HER2 disease

Dominik Rüttinger

Employment: Bayer

Leadership: Bayer

Stock and Other Ownership Interests: Bayer

Patents, Royalties, Other Intellectual Property: I am a patent holder for patents derived out of my R&D work at Roche Diagnostics GmbH

Narges Ahmidi

Honoraria: Sanofi

Research Funding: Roche Diagnostics Penzberg (Inst)

Tim Becker

Consulting or Advisory Role: xValue GmbH

Anna Bauer-Mehren

Employment: Roche

Stock and Other Ownership Interests: Roche

No other potential conflicts of interest were reported.

ACKNOWLEDGMENT

The authors thank Carlos Talavera-López (Institute for Computational Health, Helmholtz Munich), Fabian Schmich (Roche Innovation Center Munich), and Bruno Gomes (Roche Innovation Center Basel) for their valuable input.

REFERENCES

1. Mushti SL, Mulkey F, Sridhara R: Evaluation of overall response rate and progression-free survival as potential surrogate endpoints for overall survival in immunotherapy trials. *Clin Cancer Res* 24:2268-2275, 2018
2. Zhuang SH, Xiu L, Elsayed YA: Overall survival: A gold standard in search of a surrogate: The value of progression-free survival and time to progression as end points of drug efficacy. *Cancer J* 15:395-400, 2009
3. Shameer K, Zhang Y, Prokop A, et al: OSPred tool: A digital health aid for rapid predictive analysis of correlations between early end points and overall survival in non-small-cell lung cancer clinical trials. *JCO Clin Cancer Inform* 10.1200/CCI.21.00173
4. Seo S, Kim Y, Han H-J, et al: Predicting successes and failures of clinical trials with outer product-based convolutional neural network. *Front Pharmacol* 12:670670, 2021
5. Beine G, Tellier V, Charvet V, et al: Prediction of drug approval after phase I clinical trials in oncology: RESOLVED2. *JCO Clin Cancer Inform* 10.1200/CCI.19.00023
6. Hegge S, Thunecke M, Krings M, et al: Predicting success of phase III trials in oncology. *medRxiv* 2020.12.15.20248240, 2020
7. Schperberg AV, Boichard A, Tsigelny IF, et al: Machine learning model to predict oncologic outcomes for drugs in randomized clinical trials. *Int J Cancer* 147:2537-2549, 2020
8. Becker T, Weberpals J, Jegg AM, et al: An enhanced prognostic score for overall survival of patients with cancer derived from a large real-world cohort. *Ann Oncol* 31:1561-1568, 2020
9. Loureiro H, Becker T, Bauer-Mehren A, et al: Artificial intelligence for prognostic scores in oncology: A benchmarking study. *Front Artif Intell* 4:9, 2021
10. Rizopoulos D: *Joint Models for Longitudinal and Time-to-Event Data: With Applications in R*. Boca Raton, FL, CRC Press, 2012
11. Cleveland WS, Devlin SJ: Locally weighted regression: An approach to regression analysis by local fitting. *J Am Stat Assoc* 83:596-610, 1988
12. Hastie T, Tibshirani R, Friedman J: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (ed 2). New York, NY, Springer Science & Business Media, 2009
13. Ma X, Long L, Moon S, et al: Comparison of population characteristics in real-world clinical oncology databases in the US: Flatiron Health, SEER, and NPCR. *medRxiv* 2020.03.16.20037143, 2020
14. Birnbaum B, Nussbaum N, Seidl-Rathkopf K, et al: Model-assisted cohort selection with bias analysis for generating large-scale cohorts from the EHR for oncology research. *ArXiv* 200109765 Cs. 2020. <http://arxiv.org/abs/2001.09765>
15. Yang J-C, Wu Y-L, Schuler M, et al: Afatinib versus cisplatin-based chemotherapy for EGFR mutation-positive lung adenocarcinoma (LUX-Lung 3 and LUX-Lung 6): Analysis of overall survival data from two randomised, phase 3 trials. *Lancet Oncol* 16:141-151, 2015
16. Liu R, Rizzo S, Whipple S, et al: Evaluating eligibility criteria of oncology trials using real-world data and AI. *Nature* 592:629-633, 2021
17. Ho DE, Imai K, King G, et al: Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Polit Anal* 15:199-236, 2007
18. Stuart EA, Lee BK, Leacy FP: Prognostic score-based balance measures can be a useful diagnostic for propensity score methods in comparative effectiveness research. *J Clin Epidemiol* 66:S84-S90.e1, 2013
19. R Core Team: *R: A Language and Environment for Statistical Computing*. Vienna, Austria, R Foundation for Statistical Computing, 2018. <https://www.R-project.org/>
20. Ho D, Imai K, King G, et al: MatchIt: Nonparametric preprocessing for parametric causal inference. *J Stat Softw* 42:1-28, 2011
21. Katzman JL, Shaham U, Cloninger A, et al: DeepSurv: Personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Med Res Methodol* 18:24, 2018
22. Rizopoulos D: JM: An R package for the joint modelling of longitudinal and time-to-event data. *J Stat Softw* 35:1-33, 2010
23. Loureiro H: Risk trend analysis Github repository. <https://github.com/loureirh/risktrend>
24. Bierer BE, Meloney LG, Ahmed HR, et al: Advancing the inclusion of underrepresented women in clinical research. *Cell Rep Med* 3:100553, 2022
25. Shameer K, Zhang Y, Jackson D, et al: Correlation between early endpoints and overall survival in non-small-cell lung cancer: A trial-level meta-analysis. *Front Oncol* 11:672916, 2021
26. Prentice RL: Surrogate endpoints in clinical trials: Definition and operational criteria. *Stat Med* 8:431-440, 1989
27. Alonso A, Bigirimurame T, Burzykowski T, et al: *Applied Surrogate Endpoint Evaluation Methods with SAS and R*. Chapman and Hall/CRC, 2016. <https://www.taylorfrancis.com/books/9781482249378>
28. Burzykowski T, Molenberghs G, Buyse M (eds): *The Evaluation of Surrogate Endpoints*. New York, NY, Springer New York, 2005
29. Becker T, Mailman M, Tan S, et al: Comparison of overall survival prognostic power of contemporary prognostic scores in prevailing tumor indications. *Med Res Arch* 10.18103/mra.v11i4.3638