UNIVERSITY
OF TURKU

MODELING AND PREDICTION OF ADVANCED PROSTATE CANCER

Teemu Daniel Laajala

TURUN YLIOPISTON JULKAISUJA – ANNALES UNIVERSITATIS TURKUENSIS
SARJA - SER. AI OSA - TOM. 594 | ASTRONOMICA - CHEMICA - PHYSICA - MATHEMATICA | TURKU 2018

UNIVERSITY
OF TURKU

MODELING AND PREDICTION OF ADVANCED PROSTATE CANCER

Teemu Daniel Laajala

TURUN YLIOPISTON JULKAISUJA – ANNALES UNIVERSITATIS TURKUENSIS
SARJA - SER. AI OSA - TOM. 594 | ASTRONOMICA - CHEMICA - PHYSICA - MATHEMATICA | TURKU 2018

University of Turku

Faculty of Science and Engineering
Department of Mathematics and Statistics (Mathematical Modeling)
Drug Research Doctoral Programme (DRDP; Faculty of Medicine)

Supervised by

Professor Tero Aittokallio
Department of Mathematics and Statistics
University of Turku
Turku, Finland

Professor Matti Poutanen
Institute of Biomedicine
University of Turku
Turku, Finland

Adjunct Professor Laura L. Elo
Turku Centre for Biotechnology
University of Turku & Åbo Akademi
Turku, Finland

Reviewed by

Professor Matti Nykter
Faculty of Medicine and Life Sciences
University of Tampere
Tampere, Finland

PhD Terry Meehan
European Bioinformatics Institute
Wellcome Genome Campus
Cambridge, United Kingdom

Opponent

Associate Professor Manuela Zucknick
Oslo Center for Biostatistics and Epidemiology
Department of Biostatistics, Institute of Basic Medical Sciences
University of Oslo
Oslo, Norway

The originality of this thesis has been checked in accordance with the University of Turku quality assurance system using the Turnitin Originality Check service.

ISBN 978-951-29-7443-6 (PRINT)
ISBN 978-951-29-7444-3 (PDF)
ISSN 0082-7002 (Print)
ISSN 2343-3175 (Online)
Grano Oy - Turku, Finland 2018

In celebration of the life that blooms in my family,

and in memory of Kati Aarniala

Table of Contents

ABSTRACT

Background: Prostate cancer (PCa) is the most commonly diagnosed cancer and second leading cause of cancer-related deaths for men in Western countries. The advanced form of the disease is life-threatening with few options for curative therapies. The development of novel therapeutic alternatives would greatly benefit from a more comprehensive and tailored mathematical and statistical methodology. In particular, statistical inference of treatment effects and the prediction of time-dependent effects in both preclinical and clinical studies remains a challenging yet interesting opportunity for applied mathematicians. Such methods are likely to improve the reproducibility and translatability of results and offer possibility for novel holistic insights into disease progression, diagnosis, and prognosis.

Methods: Several novel statistical and mathematical techniques were developed over the course of this thesis work for the *in vivo* modeling of PCa treatment responses. A matching-based, blinded randomized allocation procedure for preclinical experiments was developed that provides assistance for the statistical design of animal intervention studies, e.g., through power analysis and accounting for the stratification of individuals. For the post-intervention testing of treatment effects, two novel mixed-effects models were developed that aim to address the characteristic challenges of preclinical longitudinal experiments, including the heterogeneous response profiles observed in animal studies. Subsequently, a Finnish clinical PCa hospital registry cohort was inspected with a strong emphasis on prostate-specific antigen (PSA), the most commonly used PCa marker. After exploring the PSA trends using penalized splines, a generalized mixed-effects prediction model was implemented with a focus on the ultra-sensitive range of the PSA assay. Finally, for metastatic, aggressive PCa, an ensemble Cox regression methodology was developed for overall survival prediction in the DREAM 9.5 mCRPC Challenge based on open datasets from controlled clinical trials.

Results: The advantages of the improved experimental design and two proposed statistical models were demonstrated in terms of both increased statistical power and accuracy in simulated and real preclinical testing settings. Penalized regression models applied to the clinical patient datasets support the use of PSA in the ultra-sensitive range together with a model for relapse prediction. Furthermore, the novel ensemble-based Cox regression model that was developed for the overall survival prediction in advanced PCa outperformed the state-of-the-art benchmark and all other models submitted to the Challenge and provided novel predictors of disease progression and treatment responses.

Conclusions: The methods and results provide preclinical researchers and clinicians with novel tools for comprehensive modeling and prediction of PCa. All methodology is available as open source R statistical software packages and/or web-based graphical user interfaces.

Keywords: Prostate cancer, Castration resistance, PSA, Preclinical, Clinical, Experimental design, Reproducibility, Translatability, Open data, Open source, Regression modeling, Mixed-effects models, Feature selection, Statistical inference, Machine learning

TIIVISTELMÄ

Tausta: Eturauhassyöpä on yleisin diagnosoitu syöpätyyppi ja länsimaisten miesten toiseksi yleisin syövästä johtuva kuolinsyy. Taudin etenevä muoto on hengenvaarallinen, ja parantavia hoitoja on tarjolla rajatusti. Kehitettävien hoitokeinojen arviointiin tarvitaan tämän sovelluskohteen erityispiirteet huomioivia matemaattisia menetelmiä. Heterogeenisten hoitovasteiden mallintaminen ja aikariippuvaisten vaikutusten mallintaminen ovat haastavia mutta kiinnostavia soveltavan matematiikan kohteita sekä esikliinisessä että kliinisessä syöpätutkimuksessa. Hyvällä matemaattisella mallinnuksella pystytään parantamaan esikliinisten tulosten toistettavuutta ja tulkittavuutta sekä kattavampaan diagnosointiin ja taudin etenemisen ennustamiseen.

Menetelmät Tässä väitöskirjatyössä kehitettiin useita uusia tilastollisia ja soveltavan matematiikan menetelmiä eturauhassyövän *in vivo* -vasteiden mallintamiseksi. Koesuunnittelun tueksi kehitettiin uusi lähtötilanteen samankaltaisuuteen perustuva sovitusmenetelmä, jolla jaettiin eläimiä tasaisesti tutkittaviin hoitoryhmiin säilyttäen satunnaistetun sokkohoitokokeen edut. Kehitetty menetelmä tukee tilastollista testivoima-analyysia ja vähentää yksilöiden välisten ei-toivottujen aliryhmien vaikutusta päättelyssä. Hoitovasteiden mallintamista varten kehitettiin kaksi uutta sekamallia, joissa otettiin huomioon esikliinisessä tutkimuksessa esiintyviä ilmiöitä kuten tuumorien kasvuominaisuuksien spontaania vaihtelua xenografti-kokeissa. Kolmannessa osatyössä mallinnettiin turkulaisen potilasaineiston eturauhaselle ominaista antigeeniä (PSA), joka on laajalti käytössä ko. syövän diagnoosissa ja seurannassa. PSA:ta tutkittiin penalisoitujen käyrämallien avulla, minkä jälkeen rakennettiin yleistetty sekamalli syövän biokemiallisen relapsin ennustamiseksi painottaen ns. ultrasensitiivisen mittausalueen PSA:ta diagnostisena työkaluna. Viimeisessä osatyössä ennustettiin potilaiden selviytymistä aggressiivisesti etäpesäkkeitä lähettävässä eturauhassyövässä. Ennustamiseen käytettiin Cox:n penalisoituihin regressiomalleihin pohjautuvaa ensemble-kokoelmamallia. Menetelmä kehitettiin osana julkista DREAM 9.5 mCRPC analyysikilpailua, jossa jaettiin osanottajille avoimesti useita suuria kontrolloituja kliinisiä tutkimusaineistoja.

Tulokset: Aiempaa paremman koesuunnittelun ja kehitettyjen tarkempien sekamallien edut näkyivät nousseena tilastollisena voimana sekä parantuneena päättelyn tarkkuutena simuloiduissa ja todellisissa esikliinisissä kokeissa. Turkulaiseen kliiniseen aineistoon sovelletut penalisoidut regressiomallit tukivat ultrasensitiivisen mittausalueen PSA:n hyödyllisyyttä yhdessä relapsia ennustavan sekamallin kanssa. Viimeisessä osatyössä kehitetty ensemble-malli ennusti pitkälle edenneen eturauhassyövän potilaiden elinaikoja huomattavasti tarkemmin kuin alan tämän hetken huippumalli sekä tarkemmin kuin muut kilpailuun lähetetyt ennustemallit. Lisäksi löydettiin uusia tekijöitä, joita voidaan hyödyntää potilaiden selviytymisen ennustamisessa.

Johtopäätökset: Kehitetyt menetelmät ja niistä johdetut tutkimustulokset auttavat esikliinisiä tutkijoita sekä kliinistä työtä tekeviä lääkäreitä tarjoamalla uusia työkaluja eturauhassyövän moniulotteiseen ymmärtämiseen. Avoimen lähdekoodin periaatetta noudattaen kaikki kehitetyt mallit ovat käytettävissä R-paketteina tai verkossa toimivina graafisina käyttöliittyminä.

Avainsanat: Eturauhassyöpä, Kastraatioresistenssi, PSA, Esikliininen, Kliininen, Koesuunnittelu, Toistettavuus, Tulkittavuus, Avoin data, Avoin lähdekoodi, Regressiomallinnus, Sekamallit, Piirteiden valinta, Tilastollinen päättely, Koneoppiminen

SYMBOLS AND NOTATION

- d: Dimensionality of the data.
- N: Sample size.
- α: The L1/L2 norm control parameter in penalized/regularized regression.
- β: Estimated (population-wide) regression coefficients.
- δ: Covariance
- ε: Error term in regression, which is assumed to be normally and independent and identically distributed.
- λ: A sequence of penalization values to be tested for the objective function in penalized regression or the magnitude of second-order integral penalization in cubic splines.
- μ: Mean.
- σ: Standard deviation.
- σ^2: Variance.
- θ: Latent variable estimated using the EM algorithm for growing or poorly growing tumors in the method proposed in publication **II**.
- γ: Normally distributed random-effects term in mixed-effects modeling.
- Σ: Sum of values or the covariance-variance matrix in multivariate normal distribution.
- $||\beta||_1$: L1-norm (LASSO) for coefficients β
- $||\beta||_2^2$: L2-norm (Ridge Regression) for coefficients β
- $x \sim y$: x is distributed as y.
- $N(\mu, \sigma)$ / $MVN(\boldsymbol{\mu}, \Sigma)$: Univariate and multivariate normal distributions, respectively.

ABBREVIATIONS

- 4T1: A mouse-derived BCa cancer cell line in **II**.
- ADT: Androgen Deprivation Therapy.
- AIC: Akaike Information Criterion.
- ART: Adjuvant Radiation Therapy.
- AR: Androgen Receptor.
- ARN-509: An anti-androgen with a similar structure to enzalutamide
- B&B: Branch and bound, a discrete optimization solving framework utilized in publication **I**
- BCa: Breast Cancer.
- BCR: Biochemical relapse, the main end-point event in publication **III**.
- BIC: Bayesian Information Criterion.
- CRAN: Central R Archive Network, the main repository for user-contributed R packages.
- CRPC: Castration-Resistant Prostate Cancer.
- CV: Cross-validation.
- DMBA: 7,12-dimethylbenz(a)anthracene, a carcinogenic compound for inducing tumors.
- DPN: Diarylpropionitrile or 2,3-bis(4-hydroxyphenol)-propionitrile, intervention in **II**.
- DREAM: Dialogue for Reverse Engineering Assessments and Methods, a research initiative that organizes crowd-sourced data analysis challenges.
- EN: Elastic Net.
- ENL: Enterolactone, a dietary intervention in **II**.

- EM: Expectation-Maximization algorithm, the iterative two-step framework used for model fitting and estimating the latent growth variable in publication **II**.
- ER: Estrogen receptor (ERα and ERβ).
- ePCR: ensemble-based Penalized Cox Regression, the methodology developed for publication **IV** (available as an R-package in CRAN with the same name; Laajala et al. 2018).
- FDA: US Food and Drug Administration.
- GA: Genetic Algorithm, a generalized problem solving framework utilized in publication **I**.
- GEM: Genetically engineered mouse (models).
- hamlet: Hierarchical Optimal Matching and Machine Learning Toolbox, the R-package accompanying the methodology developed in publication **I** (available in CRAN).
- iAUC: integrated Area Under Curve, the main scoring metric in DREAM 9.5 mCRPC Challenge in publication **IV**.
- i.i.d.: Independent and identically distributed random variable.
- LAR: Lariciresinol, a dietary intervention in **II**.
- LASSO: Least Absolute Shrinkage and Selection Operator.
- LNCaP: Lymph Node Carcinoma of the Prostate, a human prostate cancer cell line.
- NCI: National Cancer Institute (US).
- MD: Doctor of Medicine (Lat. Medicinae Doctor).
- NIH: National Institute of Health (US).
- MCF-7: Michigan Cancer Foundation-7, a human breast cancer cell line.
- mCRPC: metastatic Castration-Resistant Prostate Cancer.
- MEM: Mixed-effects model.
- MDS: Multidimensional scaling.
- MDV3100: An anti-androgen of enzalutamide.
- MSE: Mean-Squared-Error (common in literature) or Median-Squared-Error (here).
- Murine: Rodent related experiment in this context, typically mouse or rat.
- OS: Overall Survival.
- ORX: Orchiectomy, the surgical removal of testes.
- REML: Restricted Maximum Likelihood, a method for fitting MEM.
- ROC: Receiver-Operator Curve, a method for evaluating sensitivity/specificity of a model.
- ROC-AUC: Area Under Curve of a ROC.
- RP: Radical Prostatectomy.
- RR: Ridge Regression.
- RT: Radiation Therapy.
- PCa: Prostate Cancer.
- PDS: Project Data Sphere, open data sharing platform with a focus on clinical data.
- PDX: Patient-derived xenograft.
- PSA: Prostate-Specific Antigen, a prominent biomarker for prostate cancer.
- R: An open source free statistical/mathematical programming language built on S/S-plus.
- Tx: An undisclosed intervention, which remains to be described by Huhtaniemi et al.
- t-PSA: Traditional-range PSA assay (0.1 ng/mL $\leq x$).
- u-PSA: Ultrasensitive-range PSA assay ($0.001 \leq x < 0.1$ ng/mL).
- VCaP: Vertebral-Cancer of the Prostate, a human prostate cancer cell line.

LIST OF ORIGINAL PUBLICATIONS

I. **Laajala TD**, Jumppanen M, Huhtaniemi R, Fey V, Kaur A, Knuuttila M, Aho E, Oksala R, Westermarck J, Mäkelä S, Poutanen M, Aittokallio T. *Optimized design and analysis of preclinical intervention studies in vivo*. Scientific Reports. 2016;6:30723.

II. **Laajala TD**, Corander J, Saarinen NM, Mäkelä K, Savolainen S, Suominen MI, Alhoniemi E, Mäkelä S, Poutanen M, Aittokallio T. *Improved statistical modeling of tumor growth and treatment effect in preclinical animal studies with highly heterogeneous responses in vivo.* Clinical Cancer Research. 2012;18(16):4385-96.

III. **Laajala TD***, Seikkula H*, Seyednasrollah F, Mirtti T, Boström PJ, Elo LL. *Longitudinal modeling of ultrasensitive and traditional prostate-specific antigen and prediction of biochemical recurrence after radical prostatectomy*. Scientific Reports. 2016;6:36161.

IV. Guinney J*, Wang T*, **Laajala TD***, Winner KK, Bare JC, Neto EC, Khan SA, Peddinti G, Airola A, Pahikkala T, Mirtti T, Yu T, Bot BM, Shen L, Abdallah K, Norman T, Friend S, Stolovitzky G, Soule H, Sweeney CJ, Ryan CJ, Scher HI, Sartor O, Xie Y, Aittokallio T, Zhou FL, Costello JC, Prostate Cancer Challenge DREAM Community. *Prediction of overall survival for patients with metastatic castration-resistant prostate cancer: development of a prognostic model through a crowdsourced challenge with open clinical trial data*. Lancet Oncology. 2017;18(1):132-142.

* Equal contribution as first authors.

1. INTRODUCTION

Cancer research and anticancer drug discovery is a multilayered process that typically involves *in vitro* preclinical testing (i.e., cell lines) (Breslin et al. 2013), *in vivo* preclinical testing (i.e., rodent models) (Valkenburg et al. 2015), and clinical modeling - either in randomized clinical trials or using real-world registry data, with an increasing emphasis on open data (Bender 2016). However, in recent years, there has been a considerable debate regarding the low reproducibility of preclinical results and the high attrition when translating preclinical findings into clinical applications (Hutchinson et al. 2011).

It has been proposed that the methodology used in preclinical research should mimic more closely that of clinical research to overcome challenges in translatability (Muhlhausler et al. 2013). It is also evident that both the preclinical and clinical stages would benefit from a more comprehensive mathematical and statistical modeling (Heitjan 2011; van der Worp 2010). Although novel methodology has been proposed for experimental design (Kasturi et al. 2011) and regression modeling of tumor growth profiles (Zhao et al. 2011), reports indicate that neither of these methods is enforced or implemented in current research (Baker et al. 2014). Thus, the preclinical motivation of this work was to confront yet unaddressed aspects in experimental design and post-intervention statistical inference in preclinical studies and promote the integrity of research work spanning the whole range of the drug discovery process. In this thesis, further clinical mathematical and statistical modeling aspects are considered in the context of PCa and particularly in its advanced forms. The presented thesis follows the chronological order of the drug discovery process; preclinical research with immunodeficient rodents represents the mainstay of early cancer research (Cunningham et al. 2015) (**I** – **II**). The focus then shifts into the use of PSA in clinical research with patient registry data and finally to a crowd-sourced clinical data analysis challenge of developing predictive models in open data (**III** – **IV**).

In the presented preclinical research, emphasis was first placed on developing a robust and improved experimental design in preclinical cancer intervention testing to lay the foundation to the subsequent statistical inference of treatment effects (**I**). After presenting this methodology, the focus is shifted to post-intervention statistical testing with an emphasis on capturing the underlying latent tumor heterogeneity typical of such experiments (**II**). In the presented clinical research, emphasis was first placed on the diagnosis of and the modeling of the disease by utilizing a commonly used cancer marker, PSA, which has challenges related to overdiagnosis and reliability (**III**). Furthermore, while PSA is a key factor in the diagnosis and monitoring of early-phase PCa, its utility in later-stage PCa is limited. Therefore, the focus was broadened to data-driven machine learning without a bias toward any particular marker in lethal, castration-resistant prostate cancer (CRPC). This methodology was applied in the metastatic form of CRPC (mCRPC), for which life expectancy and treatment options are severely limited. A novel ensemble-based machine learning methodology was developed in this context with the aim of providing clinicians with novel insight into survival-associated biomarkers with a practical and accurate prediction tool. This model development was conducted in a crowd-sourced data analysis challenge offering several open clinical trials with a high number of potential predictors (**IV**).

2. LITERATURE REVIEW

2.1. Cancer statistics and general overview of prostate cancer

Prostate cancer (PCa) is the most commonly diagnosed cancer type for men in Finland and the second leading cause of cancer-related deaths after lung cancer (Finnish Cancer Registry, 2014 Consensus; Torre et al. 2015). Similarly, in European countries, PCa has the highest incidence and the third highest mortality among all cancer types in men (Ferley et al. 2013). Concordantly, PCa is the most common cancer type and the second leading cause of cancer-related deaths (after lung cancer) in the USA (Siegel et al. 2016). In Asia, PCa has significantly lower incidence and mortality, but in many lower income countries, PCa has significantly higher mortality. Especially in Africa, this increased mortality may be partly explained by poor access to screening and underlying genetic susceptibility (Torre et al. 2015). The majority of initial PCa diagnoses involve relatively benign disease; however, at first screening, some patients present with advanced disease (Figure 1a-b). Thus, the clinical significance of PCa remains undisputed among men, especially given the aging population in Western countries. PCa is counted among the four most prominent cancer subtypes that include breast cancer (BCa) in women and lung and colorectal cancers in both genders (Siegel et al. 2016).

Obtaining consensus for the trends in the incidence of PCa remains challenging due to the dramatic changes in PCa screening and diagnosis over the past decades. This phenomenon is largely due to the emergence of the biomarker known as prostate-specific antigen (PSA), which was discovered independently by multiple scientists in 1970s (Rao et al. 2008). Applying PSA as a screening marker introduced a sharp increase in PCa diagnoses with a subsequent steady decline in new diagnoses since the early 1990s (Figure 1c). However, PCa-related survival rate has not strictly followed a similar trend. The measurement of serum PSA revolutionized the screening and detection of PCa and it has since served as the main tool both in PCa screening and monitoring by clinicians; it has also been applied in various forms of PCa research. PSA is secreted by the luminal epithelial cells of the prostate and correlates with PCa size and functions similarly in human and in murine models of PCa (Lilja 1985). However, it is not specific to PCa; elevated PSA levels can result from benign prostate hyperplasia, older age, and non-PCa related inflammation (Barry 2001; Oesterling 1991). In the event of PCa or related malignancies, PSA is prone to leak into normal blood circulation, thus, acting as a PCa biomarker. It became apparent that under standard curative interventions in early-stage PCa, elevated PSA concentrations should quickly decline to undetectable quantities and that subsequently re-elevated serum PSA concentrations would indicate disease relapse (Stamey et al. 1987). Advocates of extensive PSA screening point to the advantages of early PCa detection, which can prevent morbidity from local symptoms, such as bleeding or urinary tract obstruction, and progression to the metastatic form of the disease (Barry 2001). The current consensus is that absolute PSA levels should not be used as the sole marker for PCa therapy or diagnosis but should be complemented by patho-physiological factors such as biopsies, age, ethnicity, and the overall evaluation of the disease and patient health (Lilja et al. 2008; Hernández et al. 2004).

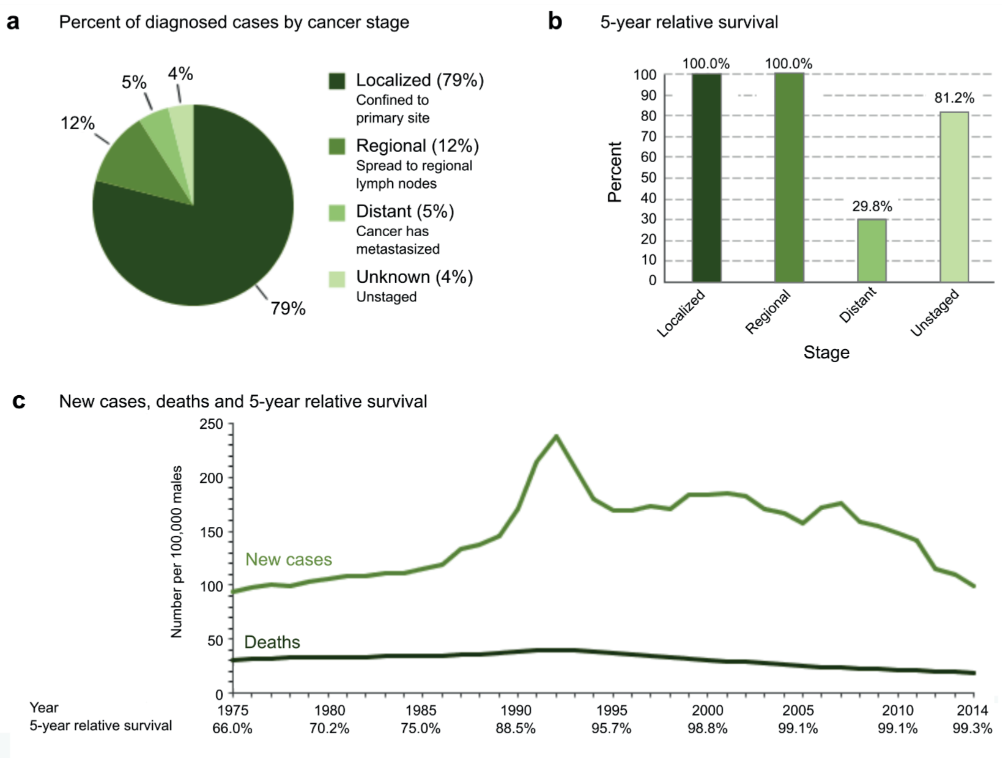


Figure 1: Prostate cancer statistics as provided by NCI / NIH. (a): The percentages for stages of detected PCa subtypes at initial diagnosis. (b): 5-year survival for the presented PCa subtypes, with the distant (metastasized) subtype as a clear outlier. (c): Overall number of PCa cases in the US since 1975, with new diagnosed PCa cases in olive green and PCa-related deaths in dark green. Relative 5-year survival rates are shown together as a function of the time-axis. (Modified and collected from National Institute of Health (NIH, US), Cancer Stat Facts: Prostate Cancer. URL: https://seer.cancer.gov/statfacts/html/prost.html ; Accessed 4th of January, 2018)

In the vast majority of cases (>95%), prostate adenocarcinoma presents with malignant transformations in the epithelial cells of the prostate gland, whereas other prostate-related malignancies, such as sarcomas or lymphomas, account for only a negligible portion of identified PCa (Epstein et al. 1994). PCa is a hormone-driven cancer and androgens play a key role in its development. As such, genetic alterations to the androgen receptor (AR) or perturbations altering androgen synthesis are known genetic factors contributing to PCa (Torre et al. 2016). Pharmacological interventions have been extensively developed to prevent the progression and recurrence of PCa or to provide curative treatment. Diet has been proposed as one of the environmental factors contributing to PCa, and thus, multiple dietary preventive measures have also been proposed, including for example, derivatives of natural food substances (Trottier et al. 2010).

It is worth highlighting the similarity of the hormone-driven nature of breast cancer (BCa) and PCa. BCa mirrors PCa in the sense that estrogen plays a similar key role in BCa development as androgen does in PCa. In this thesis, BCa is also briefly considered, especially in the retrospective analyses conducted in publication **II**, although the clear focus is on PCa. It is not surprising that treatments for both cancers revolve heavily around affecting the binding of sex steroids to their receptors, blocking the main signaling molecules or their downstream signaling pathways, or directly removing the corresponding organs producing the hormones. Furthermore, the quantity of estrogen and androgen precursors and their downstream metabolic compounds or chemical variants increase the difficulty of tackling these diseases. Significant overlap between the two diseases subtypes have been suggested, and these cancers have been researched in a mirrored manner to an increasing extent (Risbridger et al. 2010).

2.2. Clinical progression and treatment of PCa

A generalized overview of the clinical progression of PCa is presented through serum PSA concentration in Figure 2. The standard first-line therapy typically consists of radical prostatectomy (RP) and/or radiation therapy (RT), which are commonly referred to as local therapy (Figure 2a). Subsequently, a drastic decrease in PSA is observed. When PSA reaches undetectable levels, this is referred to as the PSA nadir. Depending on the initial response and patient-specific circumstances, the following therapies may involve a mixture of androgen deprivation therapy (ADT) as well as additional interventions aiming to lower the risk of recurrence such as adjuvant radiation therapy (ART). To further reduce the risk of biochemical recurrence (BCR), second-line hormonal therapy may be given even if PSA remains at low or undetectable levels (Figure 2b). Despite the initial response observed in PSA, many patients relapse to detectable levels of PSA. This disease is considered castration-resistant prostate cancer (CRPC), due to the failure of local therapy and following adjuvant therapies (Figure 2 middle panel). Notice, however, that the disease pattern presented in Figure 2 is a highly simplified version of the difficult process of a tailored treatment process, which depends on multiple clinical factors such as the pT-class, Gleason grade, spread to lymph nodes, and surgical margins (Gillessen et al. 2017). Up to 20% of patients are diagnosed with CRPC at initial diagnosis and treatment, of whom over 80% present with metastatic CRPC already at early progression; further, approx. 1/3 of the non-metastatic CRPC metastasize during follow-up (Figure 2a-c; Kirby et al. 2011). CRPC has become standard nomenclature in PCa literature describing the progressed state of the disease that does not respond to standard therapy. More accurate terminology, such as endocrine-resistant prostate cancer, has been proposed, as this phrase decreases emotional connotation for patients and distinguishes subtypes that may be treated with hormonal agents (e.g. abiraterone or MDV3100) from non-hormonal interventions such as immunotherapy or chemotherapies (Crawford et al. 2010).

A significant portion of the CRPC patients present with metastases (mCRPC), which are typically detected by imaging or symptoms such as bone pain (Figure 2b-d) (Sartor et al. 2013; Albala 2017). At this stage, the disease typically presents with increasing severity of symptoms that may be partially explained by the applied interventions. The mainstay treatment of mCRPC is docetaxel (Figure 2c-d). Although docetaxel increases the median expected survival time (Amato et al. 2009), it is rarely curative, and the CRPC or mCRPC continues to progress with increased lethality (Figure 2d). More than 90% of patients with CRPC subsequently develop bone metastases (Misra 2015), which are commonly considered one of the most lethal end-points for the disease. Late-stage treatment options include androgen synthesis inhibitors and anti-androgens (Agarwal et al. 2014), which have been also shown to increase median survival rates. However, the majority of patients in this late-stage disease die within a relatively short period of follow-up. PCa-related deaths at this point may occur not only due to disease symptoms but due to adverse events (AE) from aggressive treatment options such as docetaxel (Seyednasrollah et al. 2017a).

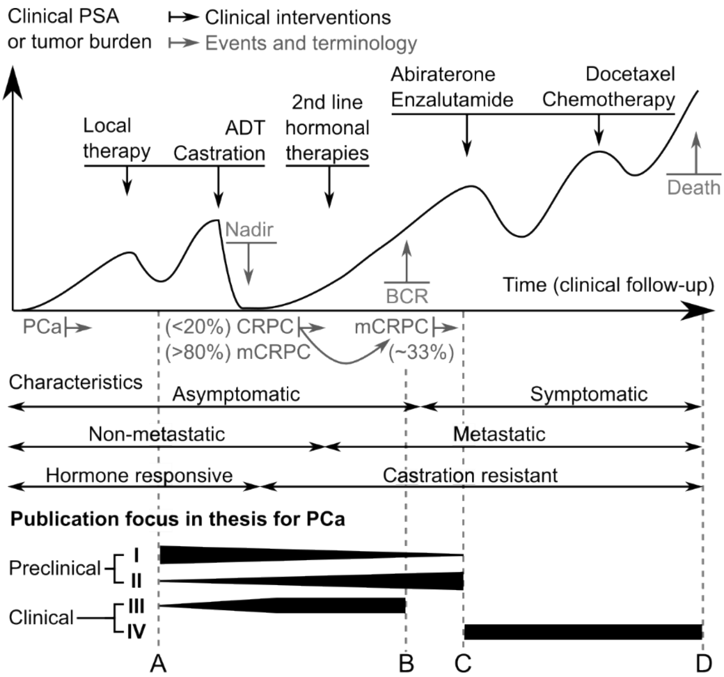


Figure 2. A highly generalized overview to PCa progression. (Top): a PSA overview to the aggressive progression of clinical PCa and potential treatment options as a simplified consensus based on available literature (Abrahamsson 2009; Kohli et al. 2010; Kirby et al. 2011; Misra 2015; Gillessen et al. 2017). (Middle): Disease characteristics. (Bottom): Focus of this thesis. (a) Local therapy typically includes radiation therapy and/or radical prostatectomy. After this, androgen deprivation therapy may take place. (b) Biochemical recurrence occurs for some patients, even if adjuvant chemotherapy and/or further hormonal interventions are utilized. (c) Docetaxel presents the current staple treatment to the aggressive mCRPC form of PCa. Additional treatment is evaluated on a case-basis. (d) Patients with PCa that progresses to CRPC and especially mCRPC have a low survival rate.

Should PSA levels increase after RP and other possible supplementary interventions, the disease relapse is castration resistant, and this recurrence event is called biochemical relapse (BCR). Two distinct PSA thresholds are used clinically to detect PCa relapse. After RP and reaching a PSA nadir, two consecutive measurements of PSA increasing at ≥ 0.2 ng/mL are sufficient to indicate that the PCa is developing BCR. For patients who have also undergone primary RT, any post-intervention PSA detectable at ≥ 2 ng/mL is considered an indicator for BCR (Cornford et al. 2017). Although the formal definition of BCR may be based on rather arbitrary thresholds, this rough division into non-relapsing and relapsing patients has proven useful, and it has been adopted as a worldwide standard for establishing an early and relatively reliable diagnosis of relapse (Lilja et al. 2008). Practicing clinicians have varying opinions of the use of PSA for mass screening in initial diagnoses and/or active surveillance, and majority of the current practices vary over countries and are often left at the discretion of the MD (Tikkinen et al. 2018). A recent meta-review of an aggregated test for four kallikrein-family PSA-related proteins, *4Kscore*, concluded that utilizing the said PSA-related test for screening for high Gleason grade aggressive PCa can have a highly consistent accuracy (Zappala et al. 2017). Thus, regardless of the presented criticism, PSA and its derivatives still retain potential for even large-scale screening and follow-up, as long as the tests and scoring metrics are formulated suitably.

2.3. Preclinical drug development, emerging practices, and validation

2.3.1. Preclinical experiments in drug discovery

The typical drug discovery pipeline often starts from preclinical *in vitro* experiments, such as drug-sensitivity testing in cell cultures, and is followed by *in vivo* experiments typically involving rodents (Cunningham et al. 2015). Murine species such as rats or mice are widely accepted as the first line of *in vivo* testing. One of the main advantages of this *in vivo* testing is the cost efficiency of xenograft experiments in immunodeficient nude mice, in which a well-characterized and suitable cell line is injected to produce subcutaneous or orthotopic tumors. The host presents a living, relatively realistic microenvironment for tumor growth, and the specific cell line is chosen to represent the main characteristics of a human cancer subtype of interest (Day et al. 2015). The animal strains in xenograft experiments are typically well established, genotyped, and bred under a controlled environment and are therefore considered a homogeneous platform for *in vivo* testing. Regardless of the broader aims of the animal study, the experimental design involves allocating animals into treatment groups (for different doses or treatment combinations) and observing the relative differences compared to a control group with no intervention or a standard treatment. The tumors are typically followed over a predefined time interval until the animals die, become moribund, or reach a preset date for sacrifice according to ethical criteria or to capture the onset of the representative human disease (Heitjan 2011; van der Worp 2010). PSA provides a convenient surrogate marker for measuring tumor growth in preclinical PCa because it only requires a relatively small blood sample.

The principles of maximizing humane treatment of laboratory animals have been laid out in the so-called 3R principles: *Replacement*, *Reduction*, and *Refinement* (Russell et al. 1959). Replacement refers to the desire to replace the use of sentient beings with a suitable surrogate, such as a cell or a tissue culture. Reduction aims to minimize the number of used animals, which is largely a factor of reliable estimation of, for example, the statistical power of a preclinical experiment and its corresponding power analysis. Finally, refinement refers to any practical methodology that aims to reduce unnecessary suffering to any sentient being, whether it is related to the experiment (e.g., surgical or sacrifice operations) or handling (e.g., placement in isolation to minimize the risk of animals scratching each other).

The disadvantages of animal studies include issues in translatability to the clinic, which arise from differences in the species' expected life span, their inherent physiology, the ability to capture the relevant window of disease onset, and the micro- and macro-environment of implanted tumors (van der Worp 2010). For example, cage effects have recently emerged as a concern, because cage-specific stratification with respect to the surrounding microbiota has been reported (Hasty et al. 2014; Hildebrand et al. 2013). Furthermore, male mice may express violent social behavior, and smaller male mice can suffer wounds caused by their larger brethren. Finally, moving animals from cage to cage resets their normal social surrounding or can result in severe stress if an animal is placed in isolation. Maintaining both the physical and social wellbeing of the animals reduces the risk of introducing confounding variation to the study design, but these factors are often non-intuitive and hard to identify (Reardon 2016).

2.3.2. Reproducibility, transparency, and reporting of animal studies

Given the mounting costs in the clinical phase of drug development coupled with the low translation rates of preclinical findings (Freedman et al. 2015; Landis 2011; Collins et al. 2014), increasing the reliability of preclinical findings in the cascading design of drug discovery is of utmost importance. It has been suggested that some of the irreproducibility issues in rodent experiments can be attributed to seemingly simple environmental conditions. As an example, diet can contain hormonal precursors or other disrupting chemicals that severely impact intervention studies, particularly in hormonally sensitive cancers. The strain of mice or even the choice of vendor providing the animals or the nutrition has been shown to exhibit drastic differences in their microbiota, and lab conditions such as exposure to light affect the behavior and biology of the animals (Reardon 2016).

The criticism of current preclinical experimentation practices involves the incomplete reporting of experimental design, power calculations, and effect size evaluation, or even the complete omission of these concepts (Day et al. 2015). A recent review identified the continued lack of power calculations, randomized allocation, and blinding of outcome, and to the surprise of the authors, these factors were not improved by the impact factor of the journal (Macleod et al. 2015). Slightly improving trends in these particular fields over time were noted, but the only reported factor found to correlate positively with the impact factor was the conflict of interest statement. Increasing concern regarding the attrition rate for preclinical result translation to the clinic, not only in oncology, has been reported prominently and has driven a change to more standardized reporting practices and better design (Landis et al. 2011; Hutchinson et al. 2011; Macleod et al. 2015); for example, in case examples of stroke treatment, only 36% of studies reported randomized allocation, and 29% reported blinding (Couzin-Frankel 2013). Furthermore, it was also noted that studies lacking proper reporting of experimental design claimed substantially higher estimates of intervention efficacy increased by 2-fold.

The translation of results from preclinical cancer experiments is especially challenging compared to other preclinical fields; 5% of promising oncological preclinical findings translate into phase III clinical trials, in contrast to 20% in cardiovascular research (Hutchinson et al. 2011). In published murine oncological experiments, 69% of publications lacked key information needed to replicate the experiment, and only 14% used appropriate statistical approaches (Sugar et al. 2012). To encourage standardization, the ARRIVE guidelines (Kilkenny et al. 2010) have been introduced to improve preclinical experiment reporting. These guidelines are in accord with the trend toward making preclinical experimentation and reporting similar to that of randomized clinical trials, for which the CONSORT guidelines are widely adopted (Schulz et al. 2010). Blinding, randomization, and masking reduce the risk of experimenter or journal-driven cognitive bias (Eisen et al. 2014), and are fundamental principles of experimental design together with sample size estimation (i.e., power analysis). Unfortunately, journals are not effectively enforcing the use of the ARRIVE guidelines (Baker et al. 2014). A meta-review by Henderson et al. of over one hundred research papers estimated that due to subpar experimental design and lack of transparency, the studies overestimated the efficacy of sunitinib intervention by 45%, therefore demonstrating publication bias in preclinical literature (Henderson et al. 2015).

2.3.3. Aims in preclinical testing

The aims during preclinical intervention testing can be roughly categorized into two classes of experiments: *exploratory* experimentation, in which a relatively large number of hypotheses are tested for potential drugs, combinations, or other interventions to narrow down a feasible set of downstream hypotheses; and *confirmatory* experimentation, in which a smaller set of drugs or even a single intervention is tested against a more refined approach, for example, to narrow down efficient dosage, the relevance of biological pathways, biological functionality, and toxicity (Kimmelman et al. 2014). While this rough division offers a rather good overview of the drug development process, there is a wide range of experiments that do not fall strictly into either of these categories. An example is drug repurposing, during which already approved drugs or drug-like compounds are explored computationally or experimentally for new disease applications (Wilkinson et al. 2015).

Preclinical researchers face a dilemma in creating reliable results that are translatable to the clinic; animal models that are based on a single strain of mouse, or arguably even a single species, coupled with a well-characterized but rather simplified representation of a general cancer type using a single cell line are inevitably under-representative of the variety of both benign and malignant PCa subtypes encountered in the clinic (Begley et al. 2012). Naturally, sources of variation that are known and can be controlled present an opportunity rather than a threat to the validity of the experiments. A particularly interesting novel application is patient-derived xenograft (PDX) experimentation, in which the aim is not necessarily to generalize to a broader population of patients but to study a single patient's tumor with case-specific tailoring (Valkenburg et al. 2015). Genetically engineered mouse (GEM) models have also emerged in the past decade as an interesting alternative to the traditional xenografts. The immunocompetent GEM mice offer a natural platform by presenting the tumors at the correct physiological location and aim to represent the clinical disease characteristics, e.g. by introducing known clinical genomic PCa drivers such as PTEN deletion or amplification and over-expression of MYC (Grabowska et al. 2014). However, while the PDX and GEM models have offered novel powerful tools for preclinical researchers, the choice of an animal model is still highly sensitive to the particular research question with unique strengths and weaknesses rising from the multitude of potential perturbations and the inherent differences in human and mouse longevity and physiological differences in the prostate (Irshad & Abate-Shen 2013). Further, it is possible to refine the design of such sophisticated PDX and GEM experiments prior to actual *in vivo* experimentation by exploring pathway-level hypotheses generated through purely hypothetical mathematical models. As an example, Boolean networks have been utilized in artificial settings to represent the corresponding biological pathways while aiming to generate clinically substantiated simulations for cancer initiation, progression, or mechanisms that drive treatment resistance (Ross et al. 2018).

2.4. Statistical testing in preclinical and clinical experiments

Previously, oversimplified statistical methods have been applied at a single time point or using tumor doubling time as readout with traditional statistical approaches such as the *t*-test, ANOVA, or their respective nonparametric versions (Ribonson et al. 1987; Shusterman et al. 2001; Saarinen et al. 2002; Galaup et al. 2003; Saarinen et al. 2006; Terada et al. 2010; Takahara et al. 2011). It is well established that *in vivo* treatment responses are often better modeled using either repeated measures or regression methods because most preclinical studies do not assess differences only at the sacrifice end point (Heitjan 2011). A significant number of preclinical studies may have used inappropriate statistical methods; the most common issues involve disregarding the time dependency in tumor growth experiments (Sugar et al. 2012). Furthermore, preclinical and clinical studies have suffered from issues that may be evident to statisticians, such as utilizing multiple testing correction, avoiding selection bias, or accounting for structured or random missing information (Smith et al. 2002).

2.4.1. Mixed-effects models

To model potential intervention effects in longitudinal follow-up measurements of either PSA as a tumor size surrogate or tumor dimensions measured under the skin, mixed-effects models (MEMs) were chosen due to their versatility. Mixed-effects models comprise of 3 main components (Pinheiro et al. 2000; Gelman et al. 2007): (i) Fixed effects (β) which capture population-level trends; (ii) Random effects (γ) which capture individual-level effects $\gamma \sim N(0, \sigma_\gamma)$; and (iii) Error term ($\varepsilon$), which is assumed to be homoscedastic and independently and identically distributed (i.i.d.) with $\varepsilon \sim N(0, \sigma_\varepsilon)$.

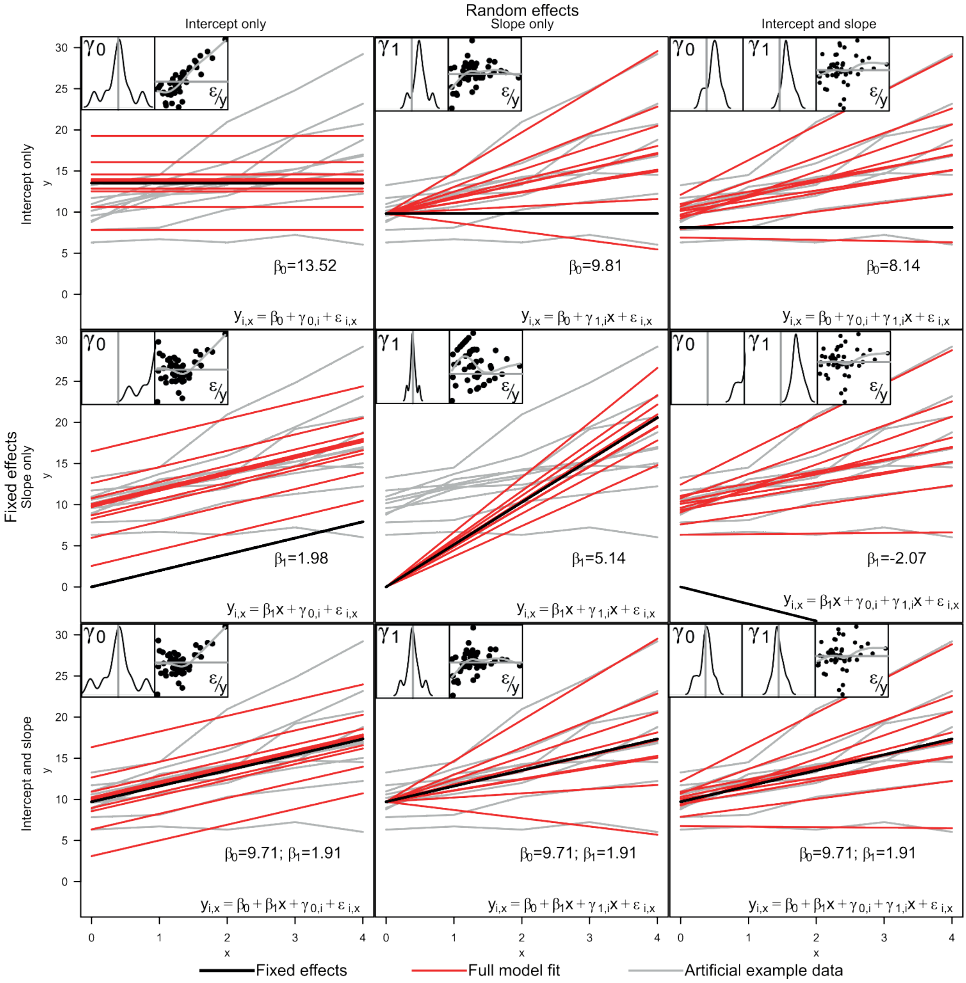


Figure 3: Combinations of fixed and random effects with mixed-effects models in an artificially generated dataset, with 10 example tumor growth profiles shown in grey. Horizontal panels vary population-level fixed effects (β, black line) by including only intercept, only slope, or both. Vertical panels vary the incorporation of individual-level random effects by including only intercept, slope, or both. The insets display model diagnostics for kernel density of the random effects or the residual plots as scatterplots with loess smoothed trend lines.

Figure 3 shows a representative combination of potential fixed and random effects formulations in the context of longitudinal MEMs in artificial tumor growth data. The main underlying assumption for the random effects is that they are normally distributed with zero mean, inspected by the distribution insets. The vertical and horizontal panels in Figure 3 portray both the flexibility related to the formulation of MEMs; the various random effects (shown in red) capture individual variation, whereas the fixed effects (shown in thick black) capture population-level trends in longitudinal tumor response profiles.

The presented work focuses on linear mixed-effects models. In this application, fixed effects were utilized to test population level differences between an intervention and a control group. Furthermore, the tumor response was followed over time, and random effects were incorporated to model the offset (intersect point at zero time point) and the slope (the coefficient of tumor growth as a function of time). As such, the random effects accounted for the necessary intra-individual correlation of observations while allowing variation over all individuals for these parameters.

2.4.2. Latent growth as a modeling variable

Heterogeneity-incorporating MEMs have been proposed in the past, in which the desired variability in the data - in principle - follows, for example, the assumptions of normality, but a hidden variable underlying the phenomenon divides individuals into varying latent substrata, resulting in multimodality or other severe issues in model parameters and inference (Verbeke et al. 1996). In particular, in preclinical research, there have been multiple accounts of issues with the inoculation of tumor cells or the spontaneous suppression of tumor growth, even when no intervention has been introduced; this can be observed as tumor heterogeneity depicting normally growing, rather benign, or even spontaneously shrinking tumors (Bhatia et al. 2012; Fisher et al. 2013). These indirectly observed underlying variables can be incorporated into statistical inference by assuming the existence and form of such latent variables and extending the standard MEMs for such latent effects.

For this purpose, the expectation-maximization (EM) algorithm (Dempster et al. 1977) provides a general framework in model fitting that can be coupled with the readily established well-suited characteristics of MEMs. The EM algorithm is a well-known and widely applied algorithm in various contexts that allows an iterative convergence of challenging modeling tasks consisting of predicting the expected outcome given the current state (expectation step), optimizing the likelihood of observing the expected outcome (maximization step), and then repeating these steps until the model parameters converge. Thus, coupled with MEM, the EM algorithm offers a powerful tool for tackling challenges in preclinical experiments that include substantial latent heterogeneity in their response profiles.

2.4.3. Power calculations

Sample size estimations, and thus power calculations, are necessary to ensure that the experiment has sufficient statistical power to detect true effects while utilizing minimal resources and ensuring the ethical aspects of experimenting on living beings (Couzin-Frankel 2013). Underpowered studies are unlikely to detect a true intervention effect (statistical significance) and may provide a skewed view of the effect size (clinical significance), thus leading to wasted animal lives. Furthermore, power analyses should be conducted prior to the final experiment to guide future experimental design; the misguided practice of interpreting statistically insignificant results using post-experiment power analyses is inherently flawed (Hoenig et al. 2001).

While conventional statistical tests, such as the *t*-test or one-way ANOVA, have straightforward methodology for evaluating sufficient sample size to achieve desired statistical power, this often involves a subjective evaluation of the inter-individual variance and the expected difference in means between the case and control. These estimates are abstract concepts that may be hard to grasp or estimate for a preclinical experimenter, especially if no pilot or representative studies are available or the estimates are not reported properly in representative literature.

Two fundamentally different approaches are explored here for conducting power calculations in MEMs. Both approaches require *a priori* data, but they do not require the user to evaluate parameter estimates for the power calculations. This helps reduce experimenter bias and allow estimation in a data-driven manner. Preliminary data can be extracted from a pilot study or even artificially generated observations can be utilized. After such data are generated, one of two approaches can then be utilized: (i) A mixed-effects model is fit to the data. After this, the parameter distributions from the fitted MEM are sampled, and a large quantity of simulated datasets are generated from the model. The same model structure is fitted to the simulated datasets and power is calculated as the fraction of statistically significant fixed effects for the hypothesis of interest. (ii) Alternatively, simulated datasets can be generated from the *a priori* data through stratified bootstrapping (sampling with replacement), in which the observations belonging to a single individual are always sampled together. These bootstrapped datasets are re-fitted using MEMs with the same formulation, and statistical power is defined as the fraction of MEMs with statistically significant fixed effects. The power threshold of 0.8, which was also used in this thesis, has been generally accepted as a feasible cutoff.

2.4.4. Bias-variance tradeoff and model complexity

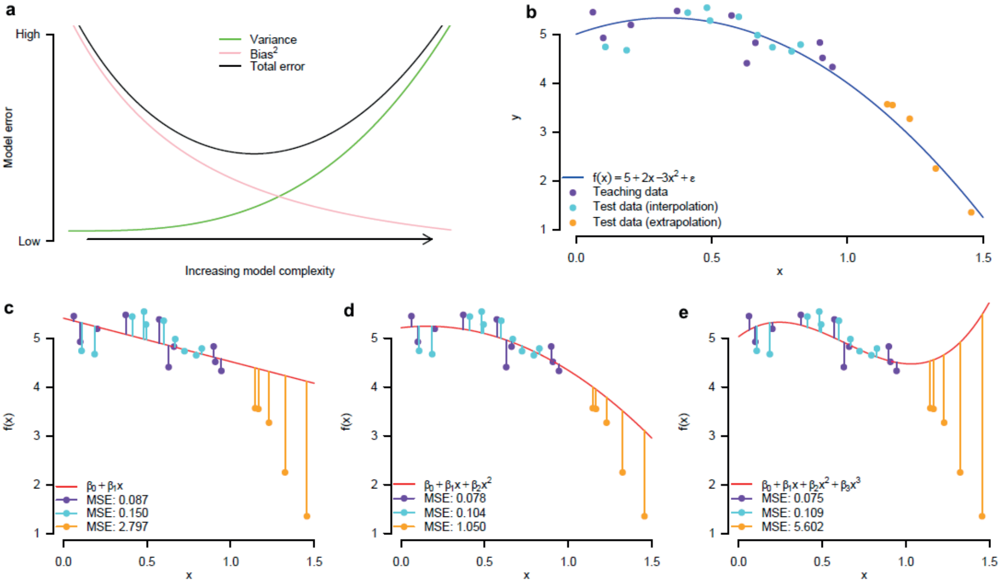


Figure 4: Visualization of the bias-variance tradeoff. (a): Bias-variance decomposition consists of two error components. The optimal model minimizes the total error. (b): A synthetic dataset was generated from a second order polynomial function, with a normally distributed error term. Three types of observations were extracted: training data that was to be used for model fitting, validation data interpolating between the observed regions

of training data, and validation data extrapolating outside the regions of training data. Mean-squared error was used as the error measure. (c): An underfitting model, with high bias but low variance. (d): Optimal model complexity, though estimated model parameters do not fully reproduce the original model due to observation error. (e): An overfitting model, with the lowest model fitting error. However, the validation data error reveals an elevated prediction error especially in the extrapolated region.

Careful considerations regarding model complexity (Figure 4) are a key ingredient in building statistical models that explain the observed variation in the training data while generalizing to future. This generalization capacity is typically analyzed using a decomposition of two error terms, bias and variance, henceforth referred to as the bias-variance tradeoff or dilemma (Figure 4a) (Hastie et al. 2001). Bias refers to an oversimplified formulation or a model formulated with systematic error that fails to incorporate a key component of the underlying true phenomenon of interest. Modeling of this type that suffers from lack of sufficient complexity is commonly known as *underfitting* (Figure 4b-c). Variance component occurs due to the phenomenon in which a model is overly complex and models the measurement error in the data. Therefore, models that also fit extensively to various sources of undesired noise in the data are very sensitive to the selection of measurements, and even small changes result in a high variance in the type of model predictions. This phenomenon is typically referred to as model *overfitting* (Figure 4e).

The gold standard in the field of machine learning for determining model complexity typically includes a suitable cross-validation (CV) schema, which aims to iteratively reveal and set aside parts of data for model training (Figure 5). The simplest version of this process involves splitting the training data into k separate bins of observations where one bin is left out at a time as a test set while the remaining bins are combined to train the model (k-fold CV). Notably, over all the folds, each observation serves as part of the test set only once, and the average scoring metric aims at optimal complexity over the tested models (i.e., Figure 4d). Some of the other notable alternatives to this computationally intensive approach are based on information theory, where Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC) are among the most popular measures in comparing models. The criteria provide measures that penalize the likelihood of a model by its parameterization complexity, therefore favoring simpler models at the expense of the goodness-of-fit.

2.4.5. Regression model family and feature selection

In many applications, the specific family of regression models is of great interest. The naïve example presented in Figure 5 shows a parametric regression method (polynomial), which is composed of parameters of interest (β) that are estimated typically by maximum likelihood or least squares fit. Splines were used in this thesis as an explorative methodology due to their ability to capture a wide range of trends from linear to nonlinear regression. However, because model parameters in preclinical and clinical research are often of great interest, e.g., in intervention testing or in identifying predictive markers, heavy emphasis was placed on easily interpretable parametric methods. For this purpose, MEMs and penalized linear regression models offer great potential.

A wide range of statistical regression methods exist that attempt to explain an observed (continuous) response as a function of predictors $\boldsymbol{x}$ through some underlying functional form $f(\boldsymbol{x})$. However, in a vast amount of modeling problems, $\boldsymbol{x}$ is multidimensional and it is not known which dimensions (predictors) are truly informative. Penalization and regularization have been introduced as sophisticated techniques for embedding feature selection within the regression model fitting, whereas traditional techniques, such as the backwards-elimination or forward-selection of variables (Sayes et al. 2007), utilize a stepwise approach in which the regression model is iteratively either extended or pruned until the model can no longer be improved with further variable selection steps. The latter portion of this thesis, especially after shifting from PSA to more open research questions, therefore carefully considered this extremely important modeling challenge.

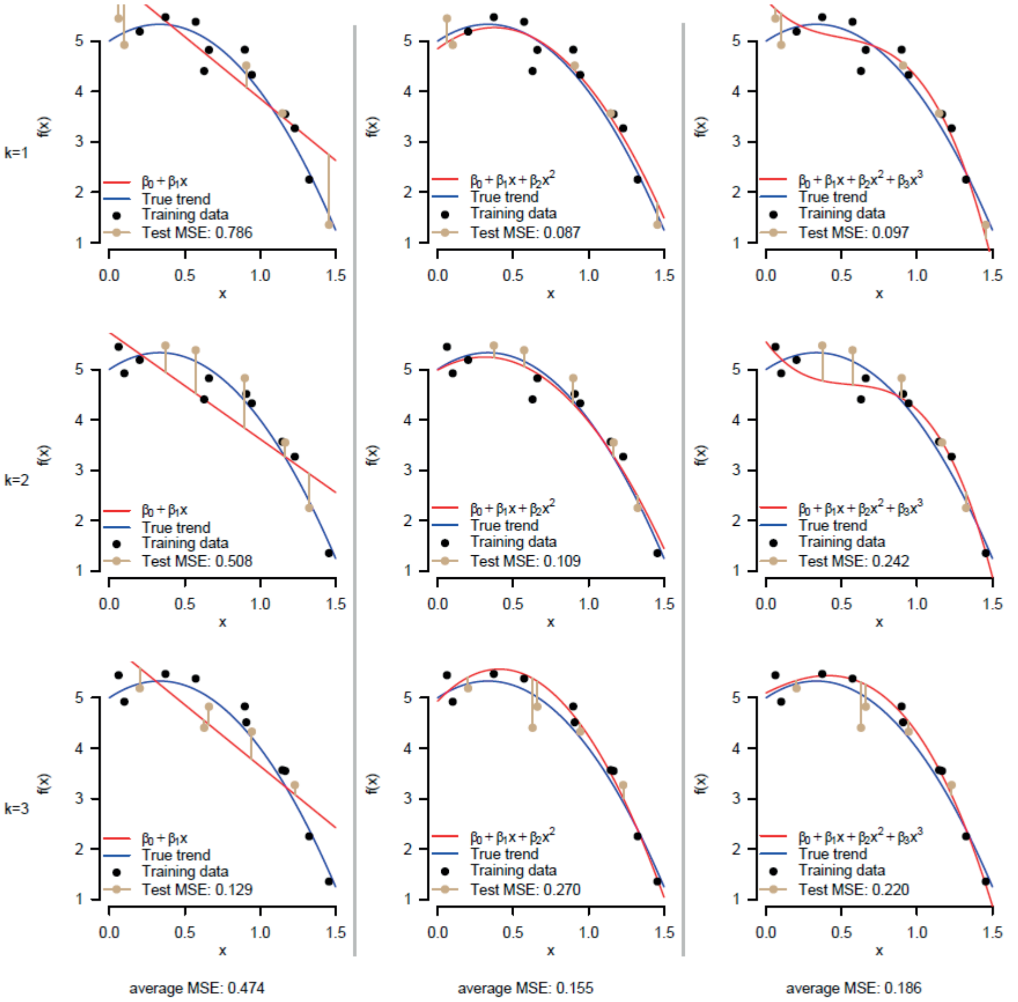


Figure 5: Example of 3-fold cross-validation. Observations were generated with noise (true trend in blue). Data were randomly assigned to 3 subsets (bins), and each of the bins is left out one at a time over the folds while the rest (black) are used for training the model (red curve). The left-out bin is used for testing the error (tan) within the fold. (Left panel): underfitting model. (Middle panel): correct formulation, suggested to be the optimal model complexity based on the lowest average CV error over the 3-folds. (Right panel): overfitting model.

2.5. Aims of the study

The main issues tackled over the course of this thesis were:

- Improve the overall experimental design, reproducibility, translatability, and reporting practices of preclinical cancer experiments (publication **I**), with a focus exclusively on PCa.
- Develop statistical modeling frameworks for accurate identification of post-intervention effects in preclinical experiments (**I** and **II**), with a primary focus on murine models of PCa.
- Investigate the specific role, reliability and trends of PSA in clinical PCa (**III**), or utilized as a marker together with other available biomarkers in the later stage of mCRPC (**IV**).
- Build improved predictive models for overall survival of mCRPC patients based on clinical data while exploring the available marker candidates, benchmarked in open data available from multiple controlled clinical trials (**IV**).
- Implement open-source tools available as R statistical software packages (**I**, **II**, **IV**) and provide web-based graphical user interfaces (**I**, **III**, **IV**), with the aim to facilitate the wide use of the novel methodology for experimenters with limited bioinformatics expertise.

3. MATERIALS AND METHODS

3.1. Datasets

3.1.1. Publication I

The first publication (**I**) made use of datasets of preclinical xenografts for prostate cancer. Of note, the statistical methodology was developed simultaneously as the back-to-back biological publications. Therefore, the methodology has the advantage of not being developed over the course of retrospective analysis of readily published data. Both biological publications focused on the mechanism of androgen dependency of castration-resistant prostate cancer.

The study by (Knuuttila et al. 2014) explored the properties of VCaP tumors as a novel animal model for castration-resistant prostate cancer. VCaP cells inoculated in intact mice progress with high likelihood to the castration-resistant form (Cunningham et al. 2015), making it an ideal model for CRPC research. Our published study (Knuuttila et al. 2014) showed that after the androgen production from testis was halted, an intratumoral androgen biosynthesis initiated, resulting in the castration-resistant growth. The efficacy of two anti-androgen compounds, ARN-509 and MDV3100, were tested to hinder tumor growth. RNA expression was analyzed for $N = 12$ tumors in post-sacrifice time point to study mechanisms for developing the castration resistance and this post-intervention RNA expression was further subjected here to testing potential correlations to baseline conditions. The processing and generation of the RNA expression data is described in detail in its corresponding biological publication (Knuuttila et al. 2014).

In the study by Huhtaniemi et al. (submitted), we analyzed the effects of orchiectomy on VCaP inoculated mice coupled with a novel, undisclosed intervention. Much of the experimental design aspects from the previous experiment, such as the sample sizes proposed by power analyses conducted for (Knuuttila et al. 2014), were fully utilized in this study.

3.1.2. Publication II

In the second publication (**II**), four preclinical studies representing a wide range of varying settings were analyzed retrospectively. To gain deeper understanding to the underlying mechanisms as well as to better model the available data, novel regression modeling was developed with a focus on accurately determining statistical significance, effect sizes, and connections to supporting markers. Notably, while the presented main LNCaP study represented PCa, the other datasets consisted of BCa studies. The LNCaP tumors were treated with diarylpropionitrile (DPN) and enteralactone (ENL) and was the primary study of interest (unpublished in-house study). These four studies were:

1. DMBA was introduced to female rats to produce spontaneous mammary gland carcinogenesis (Saarinen et al. 2002).
2. MCF-7 (a human BCa cell line) was introduced to immunodeficient female mice (Saarinen et al. 2008).
3. LNCaP (a human PCa cell line) was introduced to immunodeficient male mice (unpublished in-house pilot experiment)
4. 4T1 (a mouse BCa cell line) was introduced to immunocompetent female mice to demonstrate efficacy of known antitumoral compounds (Suominen et al. 2010)

All the analyzed datasets were provided by and analyzed in close collaboration with the experts from the Turku Center for Disease Modeling (TCDM), University of Turku, Finland. Additional details and study characteristics are available in Table 1 in publication **II**.

3.1.3. Publication III

The aims of study (**III**) included researching whether ultrasensitive PSA (u-PSA) measurements present with a meaningful signal that could predict later PSA behavior, especially in the traditional PSA (t-PSA) range. The u-PSA has been considered to contain a relatively large quantity of noise and unreliable signal, and thus its use in the clinic has been modest (Ferguson et al. 1996). While refining the ultimate aims based on preliminary results on the differences between u-PSA and t-PSA, the focus in studying sensitivity of low quantity PSA was shifted toward a clinically relevant question of predicting future biochemical recurrence (BCR). A real-word cohort of PCa patients operated with RP ($N = 503$) from the Turku University Hospital (TYKS) was collected for exploring the potential use of u-PSA assays that are capable of detecting much lower quantities than t-PSA assays. For modelling, two thirds of the full cohort was included into the model training set. $N = 52$ patients presented with BCR during a multi-year follow-up and $N = 279$ remained BCR-free. A total of 522 longitudinal t-PSA measurements and 2663 u-PSA measurements were present in the training set with a patient median follow-up time of 68.6 months. PSA nadir was chosen to be the lowest point in PSA within a three-month window from the primary operation. To evaluate the generalization ability of the prediction, the remaining one third of the data was left out as a validation data (Publication **III**: Table 1). This validation set was to be later tested by an independent researcher utilizing the readily fitted models.

3.1.4. Publication IV

The DREAM Challenges (Dialogue for Reverse Engineering Assessments and Methods, URL: http://dreamchallenges.org) is a research initiative that started in 2005, initially presenting unique but simulated research questions to interested participants. It has since grown to present large scale real-life biomedical problems to a growing community of participants. It has a strong focus on co-operation and promoting open science and currently collaborates with the Sage Bionetworks (URL: https://www.synapse.org/ProstateCancerChallenge) and the Project Data Sphere (PDS, URL: https://www.projectdatasphere.org/projectdatasphere/html/pcdc). DREAM Challenges aims to help researchers to improve in their specific expertise, share ideas, and to develop novel methodology to acute and relevant biomedical research questions.

The DREAM 9.5 mCRPC Prostate Cancer Challenge was launched with two main research subchallenges: 1) Prediction of overall survival (OS) for mCRPC patients; 2) Prediction of adverse effects (AE) occurring due to docetaxel intervention in mCRPC patients. A generic overview to this competitive phase for OS prediction is provided in Figure 6. The various clinical parameters were made available from multiple high profile pharmaceutical randomized controlled trials (RCTs) in mCRPC (Figure 6): ASCENT2 ($N = 476$) (Scher et al. 2011), MAINSAIL ($N = 526$) (Petrylak et al. 2015), VENICE ($N = 598$) (Tannock et al. 2013), and ENTHUSE 33 ($N = 470$) (Fizazi et al. 2013). ENTHUSE M1 ($N = 266$) (Nelson et al., 2011) was provided as an independent 5th dataset, which was used to test the top methodology against the gold standard in the field (Halabi et al. 2014). This benchmarking Halabi model was a LASSO-based regression model for OS-prediction in mCRPC, and had been trained on a prior study with over 700 participants.

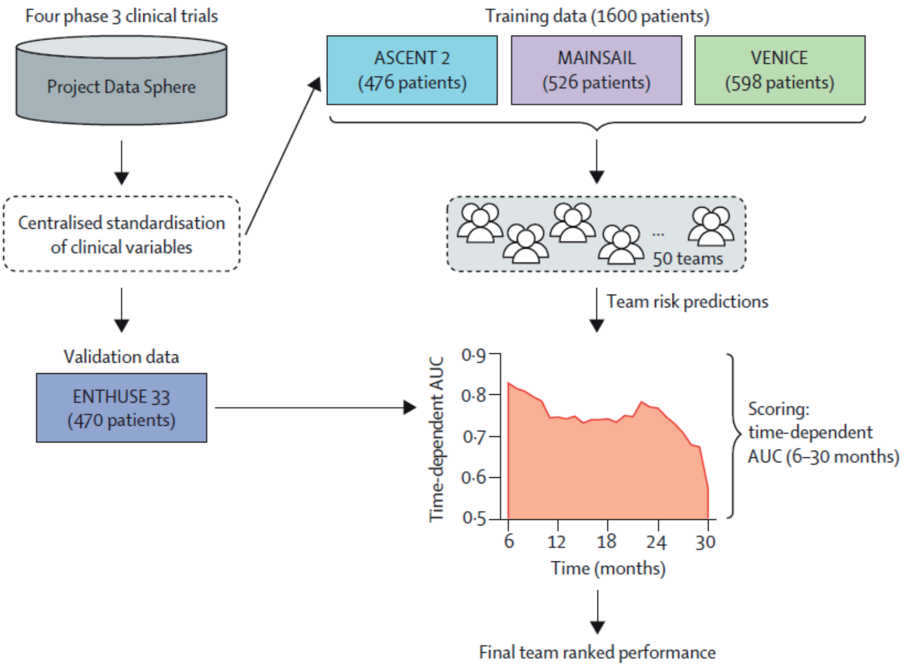

Four phase 3 clinical trials
Project Data Sphere
Centralised standardisation of clinical variables
Validation data
ENTHUSE 33 (470 patients)
Training data (1600 patients)
ASCENT 2 (476 patients)
MAINSAIL (526 patients)
VENICE (598 patients)
50 teams
Team risk predictions
Time-dependent AUC
0·9
0·8
0·7
0·6
0·5
6
12
18
24
30
Time (months)
Scoring: time-dependent AUC (6–30 months)
Final team ranked performance

Figure 6: Overview to the DREAM 9.5 mCRPC Challenge. Project Data Sphere functioned as the data depository. The organizers provided the registered participants with a standardized data table and raw data tables for further exploration. Three of the trials were offered as training data, while ENTHUSE 33 was held out as a leaderboard benchmark and test set. ENTHUSE M1 (not shown) was reserved for post-Challenge testing. (Adopted with permission from Publication **IV**: Figure 1)

The implemented methodology in publication **IV** for the OS-prediction was named ePCR (ensemble-based Penalized Cox Regression), and it was later subjected to more refined examination as well as contributed to the "*wisdom of the crowds*" meta-analysis of all the participating models. This principle is widely applied in the DREAM Challenges to examine if constructing a consensus prediction of the top-performing models can improve beyond performance of just the top-performing method (Costello et al. 2013).

3.2. Preclinical modeling methodology

3.2.1. Experimental design

To assure rigorous standards in the experimental design of preclinical studies, the following practices were emphasized:

- The researchers conducting and analyzing the experiment were blinded to the intervention groups
- Animals were allocated to the intervention arms in a balanced manner based on relevant baseline variables
- A stochastic, random component was included in the allocation method to avoid fully deterministic allocation

Balancing potentially predictive variables - such as the baseline value of the main response before the introduction of interventions and health criteria - should be conducted in animal allocation to ensure that there is no systematic bias in the starting conditions of the experiment. Due to criticism directed toward fully deterministic allocation procedures (Pond 2011), the presented baseline animal allocation balancing involves a stochastic random component (Figure 7).

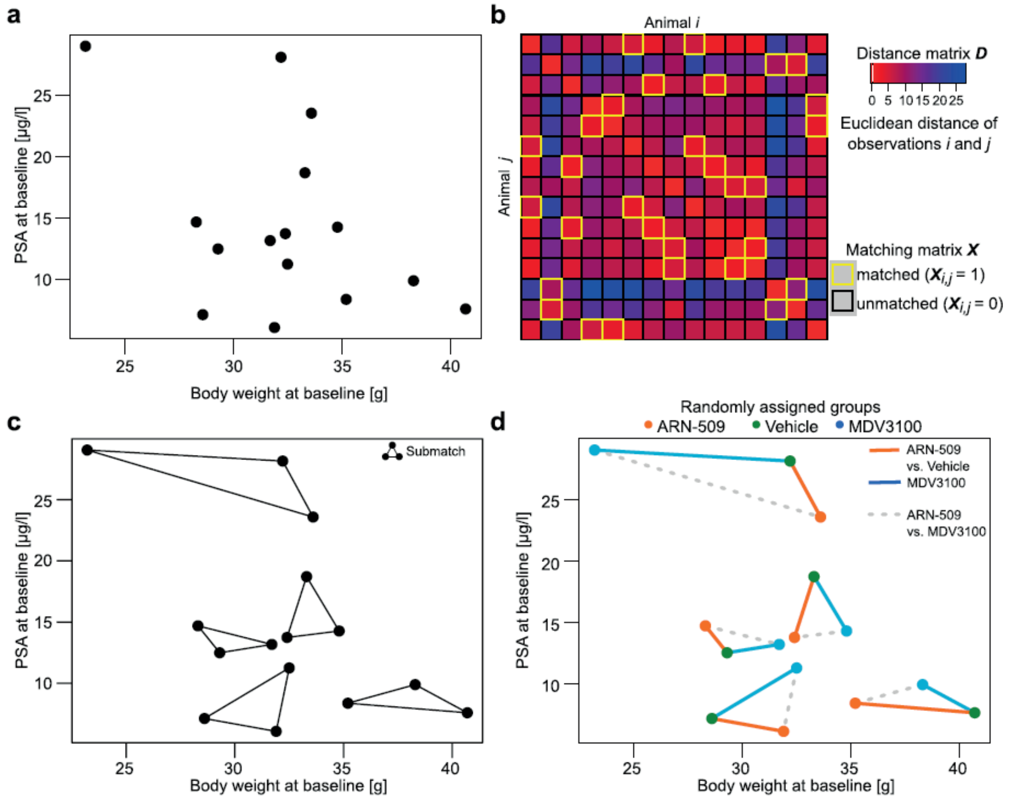


Figure 7: An example subset from the experiment by (Knuuttila et al. 2014) for the matching-based allocation. (a): Two variable baseline matching problem, where both the body weight (x-axis) and the PSA at baseline (y-axis) are considered equally important. (b): A naïve Euclidean distance matrix together with the corresponding matching matrix highlighting the submatches. (c): The submatches that minimize sum of the Euclidean distances. (d) Members of each submatch are randomly allocated to different intervention arms. (Adopted with permission from Publication **I**: Figure 2)

A randomly selected subset of 15 animals with PCa from (Knuuttila et al. 2014) at baseline is shown in Figure 7a. Two predictive baseline variables are provided in this hypothetical scenario: the body weight of the animal (x-axis) and the PSA level at baseline (y-axis). A distance or a dissimilarity matrix is constructed in Figure 7b, which depicts the amount of similarity between individuals. Notice that the standard Euclidean distance was utilized in this naïve two-dimensional example, whereas the choice of a particular similarity metric and/or weighting of specific variables is a conscious choice in an experiment. In Figure 7c, subgroups are identified through a deterministic optimization of a target function that minimizes the sum of all intra-group dissimilarities. Lastly (Figure 7d), within each of these subgroups, treatment labels are randomly assigned, and the intervention groups can be derived from the resulting labels. Notice that all intergroup differences are considered (Figure 7c); therefore, no group is fixed yet (e.g., control group) and can therefore be masked for the experimenter. By default, this procedure assumes balanced experimental design (equal number of animals in each group) because this is predominantly the desired experimental setup and is most likely to have optimal statistical modeling properties in downstream analyses. This grouping of similar individuals is conducted without any prior information regarding interventions, and thus the predictive similarity over such groups can be used to empower post-intervention statistical analyses (later referred to as matched analyses).

3.2.2. Distance/dissimilarity measures

Consider individuals *i* and *j*, which in this application denote animals with indices *i* and *j* that are to be matched. A distance/dissimilarity matrix $\boldsymbol{D}$ is a measure for the degree of dissimilarity between *i*:th and *j*:th individual presented as $D_{i,j}$. For this application, the following distance/dissimilarity criteria are required:

$$\begin{cases} D_{i,j} = D_{j,i} \\ D_{i,i} = 0 \end{cases} \forall\, i,j \qquad \text{Eq. } 1$$

These criteria indicate that there is no directionality in the distance/dissimilarity metric and that an individual is perfectly similar to its self. Supplementary Table 1 in Publication **I** lists commonly used distance/dissimilarity metrics, of which majority can be derived as special cases of the Minkowski or the Mahalanobis distance. By default, the proposed approach for depicting similarity was standardized Euclidean distance. The only distance metric directly applicable for mixed data reported here is the Gower's dissimilarity (Gower 1971).

3.2.3. Non-bipartite multigroup matching

The novel approach to experimental design through matching is based on the well-established principles of non-bipartite matching (Lu et al. 2011), which has been suggested as a method for randomized allocation (Greevy et al. 2004). In most clinical settings, the matching of individuals commonly refers to bipartite matching (Figure 8a). In this setting, two predefined groups (usually cases and controls) are matched to form intergroup pairs, and predictive covariates are used to identify pairs with similar expected risk of outcome. This less known variant of a similar matching problem, non-bipartite matching (Figure 8b), aims to identify optimal pairs of individuals within a single population. The non-bipartite matches were identified from a baseline population, and predictive markers for disease progression or treatment responsiveness were used to measure how similar individuals were within this single population. As a novel extension, instead of utilizing only matched non-bipartite pairs, the pairwise formulation was extended to *submatches*, which can be composed of three or more individuals. Within submatches, the sum of all intra-submatch dissimilarities are minimized (Figure 7c). To solve this particular problem, no applicable algorithm was found, and therefore two different approaches were developed - an exact optimization algorithm based on the branch and bound (B&B) approach (Clausen 1999) and a heuristic optimization algorithm based on the genetic algorithm (GA) (De Jong 1988). Although the former could be guaranteed to provide the global optimum, its search space may expand to computationally unfeasible size, and thus a computationally less extensive local solution was provided in the latter approach. Of note, both approaches are generalized problem solving frameworks rather than readily applicable methodologies *per se*; thus, their refinement and fine-tuning into this particular setting was extensively required.

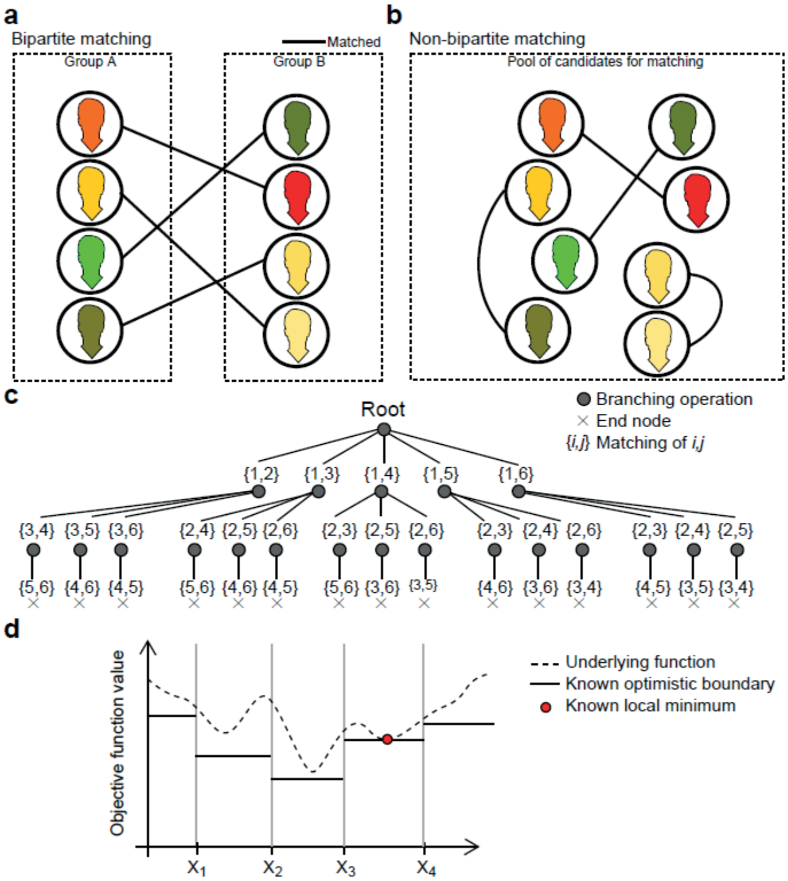


Figure 8: Overview of pairwise matching problems and base solution methods. (a): In bipartite matching, matched individuals come from two separate groups (typically case and control). (b): In non-bipartite matching, subgroups of similar individuals are identified from a single pool of candidates. (c): The branch and bound algorithm is capable of identifying a global solution through implicit coverage of the solution space. Nodes (grey circles) indicate branching and ×-symbols indicate complete solutions. (d): The bounding function guarantees an optimistic boundary in a certain area of the solution space (i.e. a certain branch in the solution tree). (Adopted with permission from Publication **I**: Supplementary Figure S7)

Extending the formulation by (Lu et al. 2011), the submatch-based non-bipartite matching is as follows. Consider a binary symmetric matching matrix $\boldsymbol{X}$:

$$X_{i,j} = \begin{cases} 1, & if\ V_i,\ V_j\ \in M_k \\ 0, & if\ V_i \in M_l,\ \ V_j \in M_k, l\ \neq k \end{cases} \qquad \text{Eq. 2}$$

where *i* and *j* are running indices for individuals in the population, V are vertices corresponding to the individuals, and M are the submatches. The vertices correspond to the observations in Figure 7c, in which the submatches are the identified connecting groups by minimizing the sum of all intra-group dissimilarities. The objective function to be minimized is:

$$\min_{X} \sum_{i=1}^{N} \sum_{j=1}^{N} X_{i,j} D_{i,j} \qquad \text{Eq. 3}$$

That is, the sum of all dissimilarities for the connected vertices should be minimized by identifying a suitable binary matching matrix $\boldsymbol{X}$ given the following constraints:

$$\sum_{i=1}^{N} X_{i,j} = G - 1 \ \forall\, j \in \{1,2,\dots,N\} \qquad \text{Eq. 4}$$

$$\sum_{j=1}^{N} X_{i,j} = G - 1 \,\forall\, i \in \{1,2,\dots,N\} \qquad \text{Eq. 5}$$

$$X_{i,j} = X_{j,i} \qquad \text{Eq. 6}$$

That is, each individual has $G - 1$ matching counterparts in the symmetric binary matching matrix, where G is the number of desired intervention groups ($G = 3$ in Figure 7b-d). After the submatches have been identified by minimizing the objective function in Eq. 3, the randomized allocation assigns random blinded intervention labels within each submatch (Figure 7d). The subgrouping guarantees that similar individuals are divided evenly among the G treatment arms.

Because the procedure assumes by default a balanced design, so-called *sinks* are utilized if N is not divisible by G. These sinks may be one of the following: i) averaged artificial individuals added to the data matrix before computing $\boldsymbol{D}$; ii) zero rows and columns added to $\boldsymbol{D}$. Both approaches add sinks to the symmetric distance matrix until the dimension is divisible by G, thus ultimately satisfying the balanced design condition.

3.2.4. Branch & Bound

The principles of the B&B algorithm can be depicted as a top-down tree (Figure 8c). There are two key steps in B&B: i) *branching* starts from the root at the top and at each node depicts possible choices for the discrete optimization task at hand. The branching should exhaustively cover all possible solutions to the optimization task. In this particular example, it can be observed that the pairwise matching of six individuals produces a solution space of 15 possible outcomes, as can be counted from the leaves at the bottom of the tree; ii) An optimistic *bounding* function is utilized to alleviate the vast solution space produced by such combinatory challenges. A bounding function is formulated to give an optimistic boundary on the best possible solution that can be found from a particular part of the solution space (Figure 8d), i.e., in this case, underneath a particular node in the top-down tree (Figure 8c).

Here, the notation $\{i, j\}$ was used to indicate that the *i*:th and *j*:th individuals i and j were part of the same submatch. Although the B&B method has a wide variety of applications in discrete optimization (Clausen 1999), it is important to note that the solution tree branching suffers from combinatory explosion as a function of the number of individuals and from complex nodes if the matching is extended beyond just pairs, i.e., allowing nodes with matching of triplets $\{i, j, k\}$ (as in Figure 7), quadruplets $\{i, j, k, l\}$, and beyond. Although the bounding function example in Figure 8d focuses on a continuous solution space, the principle is the same in discrete optimization; the bounding function dictates in this particular example that potential solutions found from $x \leq X_1$ or $x \geq X_4$ cannot improve the readily identified optimal solution. Therefore, these parts of the solution space do not need to be explored. This step is interpreted as the bounding or pruning of the tree-like solution space. In this example, the solution space in $X_1 \leq x \leq X_2$ or $X_2 \leq x \leq X_3$ may not be pruned based on the bounding function because a better solution may still be found in these ranges.

Effective pruning of the tree is largely dependent on how arbitrarily closely the bounding function can reach the true limits in the solution space. Furthermore, heuristic adjustments can be added to the search algorithm, such as employing depth-first, breadth-first, or combinations of the two search strategies that are most suitable for the task at hand. Finally, if a well-educated initial optimum can be proposed, e.g., by running a computationally light greedy search for the first optimum, the bounding function quickly prunes large portions of the search tree, and much of the solution space remains only implicitly covered. Nevertheless, in this particular application, it was noticed that as G (the number of intervention arms) was increased together with the number of animals N, the B&B algorithm became overwhelmingly computationally intensive. This computational requirement posed practical challenges because the matching-based allocation task often needs to be runnable within hours or within a day during a real experiment. Therefore, further effort was given to developing an alternative heuristic algorithm guaranteed to run in a linear time as a function of N.

3.2.5. Genetic Algorithm

The GA is a generic optimization framework that takes its inspiration from the way that genotypes are passed on through various mechanisms in living creatures, with the underlying assumption that a particular genotype with better fitness will more likely survive within the given environment and resist potential perturbations (i.e., mutations) (De Jong K 1988). GA-based solutions have been used previously in similar experimental design settings, such as in (Kasturi et al. 2011), and were therefore chosen as the primary alternative to B&B. The GA implemented for the purpose of optimizing Eq. 3 utilizes the following steps and iterations:

1. Initialize the algorithm with a user-defined number of initial, completely random solutions that fulfill the constraints set for optimizing Eq. 3; this is the population size in the GA.
2. Start looping through the generations; in each generation, the population size is kept constant. In each step, a positive event is more likely to occur to solutions that are better, while negative events are more likely to occur to worse solutions. The possible events at each generation are:
   - *Death*: Each individual solution has a possibility to die out, weighted by its fitness.
   - *Breeding*: To replace individuals in the population that have died out, two random parent solutions are picked with such weighting that more favorable solutions have higher chance of becoming a parent. The produced child is given the common rows/columns of the parents' matching matrices following the constraints Eq. 4, Eq. 5, and Eq. 6.
   - *Mutations*: Random permutations of the matching matrix $\boldsymbol{X}$ are introduced at a user-defined rate, with the likelihood of a mutation higher for non-favorable solutions in the population. A single mutation here is the swap of two columns and their corresponding rows in the matching matrix in order to retain symmetry.
3. After iterating through a pre-set number of generations, the algorithm is stopped. The best found solution over all generations is suggested as the solution to the optimization problem.

The progress of the population fitness (here, objective function values of the sums in Eq. 3) is visualized in Figure 9 as a function of the simulated generations. The initial spread of solutions is wide, because the starting population is completely random. In this example, over the first 50 generations the algorithm breeds better solutions to replace dying solutions, which can be observed as a consistent trend in the reduction of the various quantiles in the optimization function values (i.e, fitness) (left-half of Figure 9). Notice that in this example lower fitness is better, as the target function of the GA is a minimization problem. After the population has converged toward a stable state, mutations introduce the potential to explore the solution space further in an attempt to avoid local minima. This can be observed as the spiking of the solution quantiles in the right half of Figure 9. After the preset number of generations has been simulated, the algorithm stops and returns the best identified solution.

The advantages of the GA include that it is a very versatile optimization framework with great opportunity for fine-tuning various parameters in the simulation of generations, and the desired number of generations increases the run time of the algorithm linearly. The main drawback is that the algorithm is not guaranteed to identify the global optimum. At worst, the mechanisms for and balancing of death, breeding, and mutations may be unintuitive. Furthermore, if the population size (number of solutions participating in the algorithm) is too low or the population remains too stagnant at every new simulated generation, the algorithm may be prone to identify suboptimal local optima.

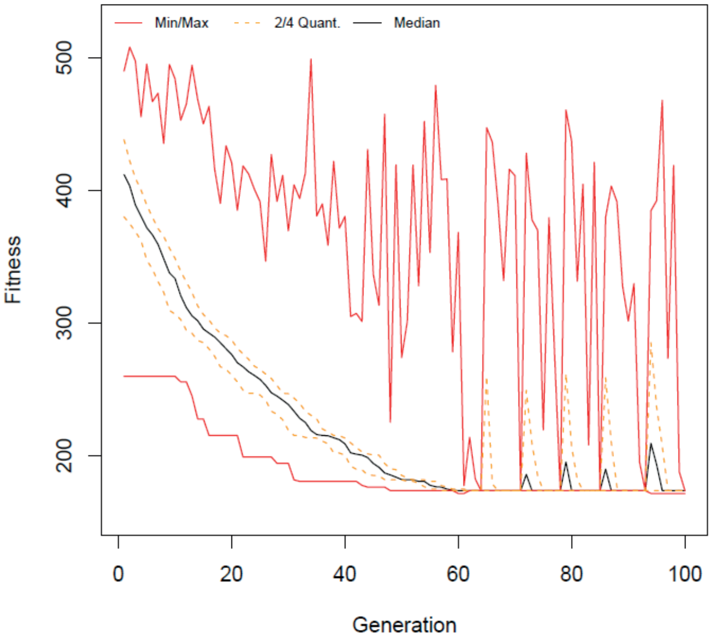


Figure 9: An example run of GA, where each new generation undergoes GA-related events with a weighted probability. Better solutions with a lower objective function value (i.e., fitness) are more likely to breed and survive to the next generation, as well as have a smaller chance of undergoing mutations. (Adopted with permission from Publication **I**: Supporting R-vignette, Figure 3)

3.2.6. Stratified matching and allocation

As preclinical experiments may present intrinsic groups or batches of individuals rising from e.g. animals arriving to the experiment at different times, interventions conducted at varying intervals or cage-effects due to differences in the microbiota environment (Hasty et al. 2014; Hildebrand et al. 2013, such stratification should be accounted for in the allocation phase. In order to account for stratification, two different approaches were proposed.

In the *strict* approach, individuals rising from separate substrata may never be matched to each other. Effectively, this separates the optimization task into two or more separate problems, requiring solutions to multiple objective functions independently for each substrata. Coupled with the random-allocation step following this matching step (Figure 7c-d), the major advantage of this strict approach is that it guarantees that the number of individuals from each substrata are divided evenly among the desired interventions groups. However, a major disadvantage is that separating the allocation into smaller sub-problems considers where the inter-individual variation may not be represented well enough. This problem may be severe especially if the *N* count in each substrata is relatively small in comparison to the desired number of intervention groups *G*.

In the *relaxed* approach, the stratification is incorporated into the dissimilarity metric itself, often as a categorical variable with expert-curated weighting, and it is subsequently treated similarly to any other variable depicting baseline differences between individuals. Such is easily possible in dissimilarity metrics such as the Gower's dissimilarity (Gower 1971) or with expert-tailoring of conventional distance metrics. Effectively, this can be seen as a penalization procedure with a cost related to allowing individuals from different substrata to be part of the same submatch.

3.3. Regression modeling in preclinical and clinical applications

3.3.1. Right-censored responses, missing observations, and suitable model families

The nonrandom nature of the missing observations in the preclinical setting arises from the fact that larger tumors are likely to be more lethal. In the case of death, all subsequent measurements for the individual will be missing. This phenomenon is referred to as *right-censoring*. In preclinical trials, the loss of animals due to death or a pre-set sacrifice threshold affects the inference of subsequent tumor growth curves. This right-censoring effect can have severe confounding consequences, especially if no individual-level observation data are provided and only averaged curves over the remaining individuals are provided together with the standard deviations or standard errors. This situation may give a false impression of a tumor growth plateau effect, whereas in reality, this occurs mainly due to the structured nature of missing observations. Naïve imputation methods that are sometimes utilized in the field, such as projecting the last observed tumor burden to the later time points, introduce further spurious effects into downstream statistical inferences. To minimize the effect of right-censored preclinical experiments herein and to avoid imputation, the main emphasis on statistical inference of intervention effects was based on examining differences in the growth slopes of tumor burden curves. These longitudinal models take into account only the truly observed data, and each individual obtains a random effects estimate for their tumor burden growth slope. A population consensus in the intervention versus the control arm is then inferred as a growth coefficient after this individual variation is accounted for.

Regarding right-censoring in clinical trials or hospital registry data, a typical end-point is a binary outcome (event or no event, such as death or biochemical recurrence). A characteristic trait for this field of survival analysis is that the response consists of two components: the first component depicts the time t until censoring or observed event. The second component is a binary indicator for the mathematical modeling, which indicates whether the patient was censored and that we only know that the event did not occur within the given time-period (typically denoted as $event = 0$) or that event was observed at the exact time ($event = 1$). Examples of popular model families that aim to address this inherent 2-column nature of the response are Cox regression and random survival forests (RSFs), both of which are known to perform well in survival modeling (Omurlu et al. 2009). The former was applied over the course of these publications.

3.3.2. Linear mixed-effects models

Linear mixed-effects models (MEMs) can be presented in the following form (Pinheiro et al. 2000):

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e} \qquad \text{Eq. 7}$$

where the left-hand side $\boldsymbol{y}$ is the vector of response values, which is typically a continuous variable. $\boldsymbol{X}$ is the model matrix of size $N \times p$, which depicts the model formulation in connection to the observations. $\boldsymbol{b} = \{\beta_0, \beta_1, \dots, \beta_{p-1}\}$ is a vector of the fixed effects, providing estimates that are common to all individuals and therefore effectively model population averages. $\boldsymbol{Z}$ is the random effects' model matrix, which groups individual-level observations. $\boldsymbol{u} = \{\boldsymbol{\gamma_0}, \boldsymbol{\gamma_1}, \dots\}$ is a set of random effects that offer individual-level estimates with the underlying assumption that these are normally distributed with zero mean. This setup is especially useful in longitudinal mixed-effects models because the multiple associated observations are grouped in $\boldsymbol{Z}$. The error term $\boldsymbol{e}$ is assumed to be normally distributed, zero mean, and i.i.d. The random effects may hold very complex normally distributed structures that attempt to capture partially overlapping variation in the data; however, the complexity rising from the random effects requires mixed-effects models to be estimated using restricted maximum likelihood (REML) rather than ordinary maximum likelihood based fitting (Pinheiro et al. 2000).

The following aspects were emphasized in the presented longitudinal modeling:

- There is typically a single main regression response variable, which is usually the PSA concentration or the physically measured tumor volume or area. Thus, univariate regression is suitable.
- Time-dependency exists between the observations, and the measurements are conducted at fixed time points for all individuals. This structure is incorporated in the random effects.
- Measurements are not censored randomly because high tumor response values are more likely to result in death, and thus all subsequent observations would be missing.
- Two or more groups are to be compared using fixed effects when testing if there is a statistically significant effect on tumor growth due to an intervention. Furthermore, clinical significance, i.e., effect size estimated on such a coefficient, is of great interest.

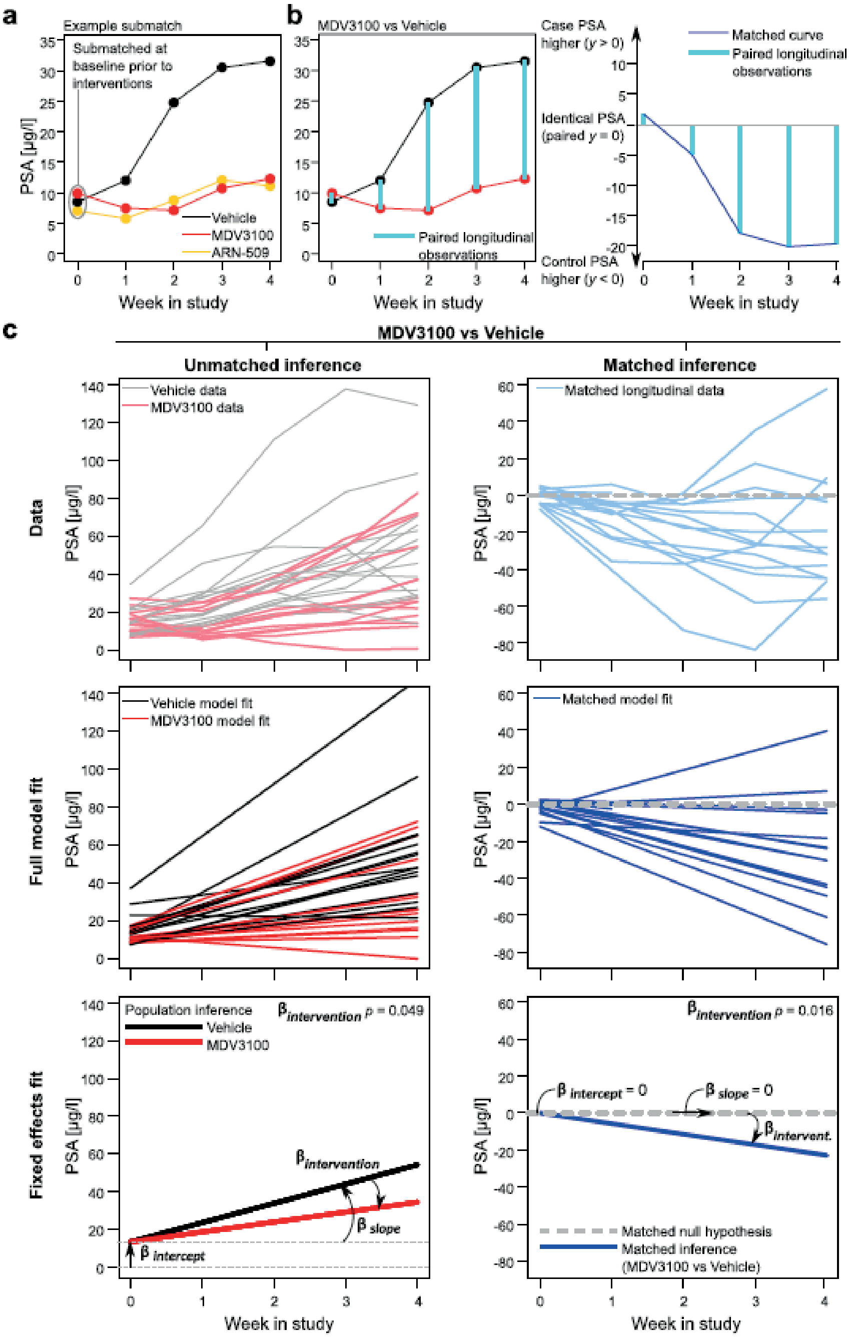


Figure 10: Example of the conventional and matched models for the MDV3100 intervention. (a): Baseline submatches were used to group individuals with a similar prognosis. (b): Matched observations were constructed using pairwise differences at the equal time points. (c): The pairwise differences were used as the growth response for the matched inference. (Left panel): Conventional inference. (Right panel): Matched inference. (Adopted with permission from Publication **I**: Figure 3)

3.3.3. Matched longitudinal analysis

Because the experimental design presented in **I** inherently incorporates matching that identified individuals with a similar prognosis, a mixed-effects model was designed that could utilize this aspect in the analysis of the longitudinal growth patterns. Due to the controlled trial setting of such preclinical experiments, each individual is measured at equidistant time points t. A model was formulated for paired observations (left-hand side) with a suitable model matrix (right-hand side) that utilizes submatches in capturing interesting longitudinal intervention effects:

$$y_{i,t,g=control} - y_{j,t,g=case} = y_{\{i,j\},t} \qquad \text{Eq. 8}$$

$$y_{\{i,j\},t} = \beta_{intercept} + \beta_{slope} x_t + \beta_{intervention} x_t + \gamma_{0,\{i,j\}} + \gamma_{1,\{i,j\}} x_t + \varepsilon_{\{i,j\},t} \qquad \text{Eq. 9}$$

where $y_{\{i,j\},t}$ refers to the paired difference in observed tumor response at time point t, x_t is the longitudinal time point, and individuals of indices $\{i,j\}$ were part of the same submatch at baseline (Figure 10a), $\gamma_{0,\{i,j\}}$ is the random effects for the pairwise intercept difference, $\gamma_{1,\{i,j\}}$ is the random effects for pairwise slope difference, and $\varepsilon_{\{i,j\},t}$ is the normally distributed error. However, we proposed fixed effects in this respect to focus solely on $\beta_{intervention}$ because the other fixed effects are redundant due to the pairing of observations. The interpretation of these observations changes dramatically when paired observations are modeled rather than individual growth curves. The final model, as shown in right panel of Figure 10c, is:

$$y_{\{i,j\},t} = \beta_{intervention} x_t + \gamma_{0,\{i,j\}} + \gamma_{1,\{i,j\}} x_t + \varepsilon_{\{i,j\},t} \qquad \text{Eq. 10}$$

To provide a benchmarking model that does not utilize the baseline matching information, the following analogous formulation was utilized:

$$y_{i,g,t} = \beta_0 + \beta_1 x_t + \beta_2 x_t g_i + \gamma_{0,i} + \gamma_{1,i} x_t + \varepsilon_{i,g,t} \qquad \text{Eq. 11}$$

where i is the index for the i:th individual, $y_{i,g,t}$ is the tumor response (i.e., PSA) for the i:th individual belonging to the group g at the specific time point t, x_t is the t:th time point, g_i is a binary indicator for the intervention group of individual i (value 1 for case and 0 for control), β_0 is the population-wide intercept, β_1 is the control-specific growth slope (due to the presence of a binary group indicator), β_2 is the intervention testing growth slope that differentiates according to the binary indicator for groups, $\gamma_{0,i}$ is the random effect allowing individualized variation at the intercept, $\gamma_{1,i}$ is the random effects slope allowing individual growth variation regardless of intervention group, and $\varepsilon_{i,g,t}$ is the normally distributed error. Both the conventional (Eq. 11) and the matched models (Eq. 9) are shown in a full linear mixed-effects model fit as a sum of the β and γ terms in Figure 10c middle panel, respectively.

3.3.4. Heterogeneity-incorporating MEM coupled with the EM algorithm

Multiple studies have indicated that some of the difficulties in analyzing preclinical response profiles can be attributed to inherent subgroups of tumors and intratumoral heterogeneity. This heterogeneity has gathered increasing interest due to its relationship to the development of drug resistance or its ability to partially explain the heterogeneous response profiles (Bhatia et al. 2012; Fisher et al. 2013). To this end, a mathematical framework was developed in publication **II** with the underlying assumption of growing (latent variable $\theta_i = 1$) or poorly growing (latent variable $\theta_i = 0$) spontaneous tumor growth distributed similarly over intervention arms. To estimate this latent variable, an adaptation of the expectation-maximization (EM) algorithm was developed in which θ is estimated parallel to the rest of the MEM coefficients with two varying versions having either binary or continuous uniform θ, as shown in Figure 11. The EM algorithm framework consists of two looped steps, which are run until convergence for the parameter of interest (in this case θ) is reached. These key steps for the EM algorithm are as follows (Dempster et al. 1977):

- Initialize conditions for the EM algorithm, i.e., all tumors are assigned θ_0 or θ_1.
  (E), Expectation step: For each tumor's latent θ_i, evaluate the relative likelihood of generating its observed profile from either of the latent growth classes $\theta_i \in \{0,1\}$.
  (M), Maximization step: Assign the more likely candidate θ_i for each tumor, and then re-estimate the mixed-effects model parameters β and γ using REML. Return to (E) if θ_i has changed.
- Iterate E and M steps until convergence of the parameters, complemented by multi-start.

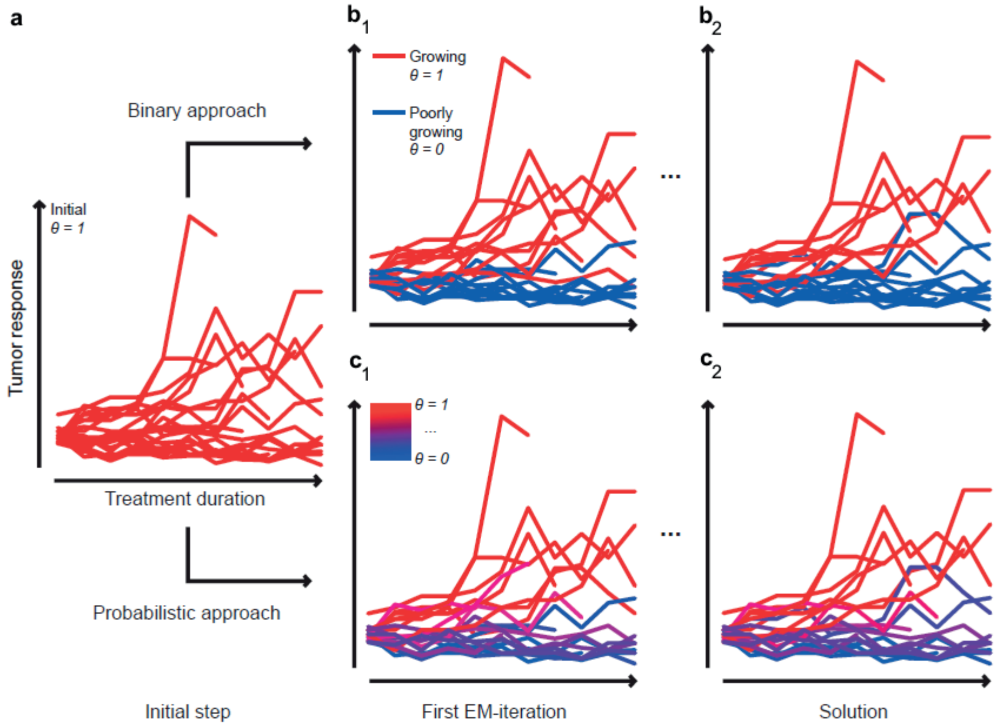


Figure 11: Progression of the EM algorithm with two variations for the latent growth variable. (a): Initially all tumors are set to the growing latent category. (b): By default, the latent variable $\theta \in \{0,1\}$ was modeled with binary classes: growing or poorly growing tumors. (c): As a probabilistic alternative $\theta \in [0,1]$, the algorithm was allowed more flexibility in estimating the continuous growth characteristics. (Adopted with permission from Publication **II**: Supplementary Figure S2)

These steps E and M are conducted iteratively, thus explaining the EM algorithm nomenclature; the ultimate aim is providing feasible estimates for the latent variables (here θ_i for all tumors i). Here, two comparable MEMs were utilized - the latent-variable categorizing model and a conventional model without the latent variable. The latent-variable formulation was as follows:

$$y_{i,t} = \beta_1 + \beta_2 g + \beta_3 x_t \theta_i + \beta_4 x_t \theta_i g + \gamma_{1,i} + \gamma_{2,i} x_t + \varepsilon_{i,t} \qquad \text{Eq. 12}$$

where $y_{i,t}$ is the tumor response (i.e., PSA) for individual i at time point t, β_1 is the control-specific intercept, $\beta_2 g$ is the intervention-specific vertical shift (which may counter-intuitively also model intervention effects due to the presence of θ_i), β_3 is the control growth slope given the estimated growth characteristics, and β_4 is the potential intervention effect of interest given the growth characteristics. $\gamma_{1,i}$ and $\gamma_{2,i}$ are the individual-level random effects allowing variation for the intercept and growth slopes, respectively, and $\varepsilon_{i,t}$ is the normally distributed error. We refer to this particular MEM in Eq. 12 as the *categorizing* model. The *conventional* benchmarking model is obtained by assuming $\theta_i = 1$ for every tumor, in which case the mixed-effects model formulation simplifies to:

$$y_{i,t} = \beta_1 + \beta_2 g + \beta_3 x_t + \beta_4 x_t g + \gamma_{1,i} + \gamma_{2,i} x_t + \varepsilon_{i,t} \qquad \text{Eq. 13}$$

This model formulation is analogous to Eq. 12 with the exception that it lacks the individualized θ and the inclusion of EM algorithm, but the interpretation for the model coefficients is the same. The models presented in Eq. 12 and Eq. 13 are visualized in Figure 12 with respect to the fixed effects inference. The conventional model is obtained if one omits the growing ($\theta_i = 1$) or poorly growing ($\theta_i = 1$) latent variable, in which case all tumors are treated equally except with regards to the intervention group g.

To expand the EM algorithm utilized in publication **II**, a probabilistic alternative was offered (Figure 11c). Although the binary categorizing model offers a more easily interpreted and biologically motivated premise, the probabilistic latent variable allows more flexibility for the modeling task. The continuous approach may alleviate some of the challenges regarding the strong assumption of the binarization of the growth categories or practical issues in the algorithm convergence if the interventions present very homogeneous profiles. In the expectation step of the EM algorithm, the latent variable θ_i is inferred based on the whole range of observations belonging to the i:th tumor, similarly to how one might interpret a random effect for the i:th individual:

$$p(\boldsymbol{\theta}|data) = \frac{p(\boldsymbol{\theta} = 1) \cdot p(data|\boldsymbol{\theta} = 1)}{p(\boldsymbol{\theta} = 0) \cdot p(data|\boldsymbol{\theta} = 0) + p(\boldsymbol{\theta} = 1) \cdot p(data|\boldsymbol{\theta} = 1)} \qquad \text{Eq. 14}$$

with equal priors for both binary $\boldsymbol{\theta}$ values resulting in:

$$p(\boldsymbol{\theta}|data) = \frac{p(data|\boldsymbol{\theta} = 1)}{p(data|\boldsymbol{\theta} = 0) + p(data|\boldsymbol{\theta} = 1)} \qquad \text{Eq. 15}$$

where the likelihood for either case are computed as a proportion of the chance of observing a growing tumor in comparison to the sum of both likelihoods. The likelihoods for a particular tumor of index i are combined over all the available longitudinal time points t:

$$p(\theta_i = 1|\boldsymbol{y}_i) = \frac{p(y_{i,t=1}|\theta_i = 1) \cdot \ldots \cdot p(y_{i,t=T}|\theta_i = 1)}{p(y_{i,t=1}|\theta_i = 0) \cdot \ldots \cdot p(y_{i,t=T}|\theta_i = 0) + p(y_{i,t=1}|\theta_i = 1) \cdot \ldots \cdot p(y_{i,t=T}|\theta_i = 1)} \qquad \text{Eq. 16}$$

The fixed effects fit of the MEM is used for predicting expected observations, thus providing the expectation step in the EM algorithm:

$$p\left(y_{i,t}\middle|\theta_i\right) \sim N(\mu = E(\boldsymbol{Xb}|i,t,\theta_i), \sigma^2 = var(\boldsymbol{y})) \quad \text{Eq. 17}$$

After this, the θ_i are assigned to the more likely binary class, and the algorithm refits the whole MEM in the maximization step using REML. This iterative two-step procedure continues until θ_i converge and therefore none of the model parameters change thereafter. Compared to the existing heterogeneity-incorporating MEMs (Verbeke et al. 1996), the previous models have focused on incorporating a latent variable heterogeneity component into the random effects portion of the model. In the context of preclinical experimentation, however, here the latent variable is introduced to the fixed effects growth component to accurately infer intervention effects that are population-wide effects.

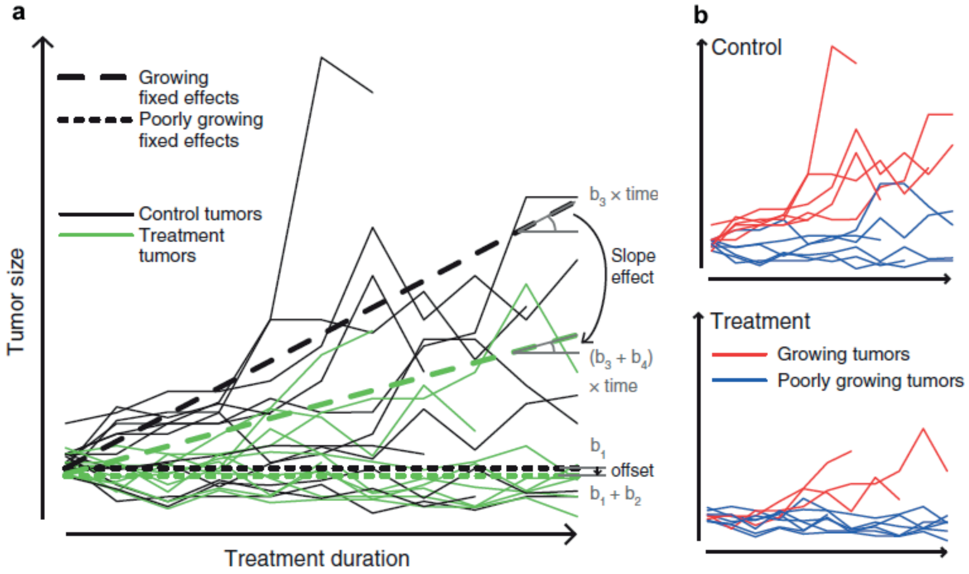


Figure 12: Overview into the categorizing model. (a): Four fixed-effects coefficients were included in the model: β_1 and β_2 model population-wide vertical effects without and with intervention, respectively. Similarly, β_3 and β_4 are the estimated growth coefficients without and with intervention, respectively. (b): The underlying assumption for the latent growth variable was that it was present in both intervention arms and was estimated using the EM algorithm in parallel with MEM fitting. (Adopted with permission from Publication **II**: Figure 1)

Figure 12 presents an overview to the categorizing MEM with the fixed effects coefficients $\beta_1 - \beta_4$ together with the underlying latent variable ($\theta_i \in \{0,1\}$ for the poorly growing or growing tumors, respectively). Because the latent variable θ_i in Eq. 12 affects the inference for the growth slope in β_3 (control growth slope) and β_4 (slope effect), the inference for β_1 and β_2 is not trivial. Although β_1 (intercept) would appear to be the intercept at $x_t = 0$, the inference for β_2 (offset) is not merely a baseline difference in the intercept between the intervention arms. Because θ_i is incorporated into the growth coefficients (Figure 12b), should $\theta_i = 0$ become the dominant latent variable for most tumors, the burden of explaining potential intervention effects rests on the offset term β_2 (horizontal lines in Figure 12a). Therefore, if there is an intervention effect that prevents vast majority of tumor growth, β_2 captures this drastic decrease in the tumor response over time. The intervention effect in β_4 tests a slope effect in comparison to the control growth captured in β_3 (slopes in Figure 12a).

3.3.5. Power analyses

To perform power calculations through sampling from an estimated MEM, the following approach was utilized in publication **II**. Because the random effects are considered to be normally distributed with zero mean and the fixed effect β is considered to be representative means of the whole population, the following sampling scheme was used:

$$\begin{bmatrix}\widehat{\gamma_{1,J}}\\ \widehat{\gamma_{2,J}}\end{bmatrix} \sim MVN(\boldsymbol{\mu} = \begin{bmatrix}0\\0\end{bmatrix}, \Sigma = \begin{bmatrix}\sigma_{\gamma_1}^2 & \delta_{1,2}\cdot\sigma_{\gamma_1}\cdot\sigma_{\gamma_1}\\ \delta_{1,2}\cdot\sigma_{\gamma_1}\cdot\sigma_{\gamma_1} & \sigma_{\gamma_2}^2\end{bmatrix}) \qquad \text{Eq. 18}$$

To calculate the power for the categorizing MEM, the latent growth variable was assumed to be binomially distributed equal proportion of $\theta_i = 1$:

$$\widehat{\theta_j} \sim Binom(n = 1, p = \frac{\sum_{i=1}^{N}\theta_i}{N}) \qquad \text{Eq. 19}$$

where $\theta_i \in \{0,1\}$ were obtained using the EM algorithm coupled with the categorizing mixed-effects model (i.e. Figure 11b). One can produce simulated responses for power analysis with the estimates of fixed effects held constant at their expected mean values:

$$\widehat{y_{J,t}} = \beta_1 + \beta_2 g + \ \beta_3 x_t \widehat{\theta_J} + \beta_4 x_t \widehat{\theta_J} g + \widehat{\gamma_{1,J}} + \widehat{\gamma_{2,J}} x_t + \varepsilon_{j,t} \qquad \text{Eq. 20}$$

where the time t should be representative of the desired experimental design. $\varepsilon_{j,t}$ is assumed to follow the normal and i.i.d. residual assumptions, and the variance is chosen according to the residual variance in the original MEM fit. This model-driven sampling was chosen as the primary power analysis method in publication **II**.

Monte Carlo methods, which include simulation-based solutions (as described above) or bootstrapping (resampling with replacement), have been presented with increased utility in settings that are hard to capture in closed-form analytical solutions. For example, Monte Carlo methods can incorporate complex structures that are difficult to describe using conventional parameters in statistical testing (such as simple difference in means and group-wise variance as in *t*-testing). Furthermore, conventional power analyses typically do not incorporate application-specific traits, such as the right-censoring in preclinical studies presented here.

In addition to propagating clinical methodology to preclinical studies (Muhlhausler et al. 2013), stratified bootstrap-based power analyses are convenient in addressing multiple challenges inherent to preclinical studies: right-censoring and its effect on subsequent statistical modeling as well as lower-censoring are incorporated. As a resampling technique it is less sensitive to underlying assumptions than model-driven sample generation and it gives more freedom to the researcher to process the resampled datasets as they wish. As such, stratified bootstrapping is agnostic to any subsequent modeling choices. Such applications have been recently introduced into e.g. cluster-based randomized clinical trials (Kleinman et al. 2017). In publication **I**, stratified bootstrapping was therefore chosen as the primary power analysis method. MEMs were re-fitted to stratified resampled datasets, and the proportion of statistically significant findings in a fixed effect of interest was studied as a function of resampled N.

3.4. Modeling methodology in clinical patient data

3.4.1. Modeling clinical PSA using splines

The transformation of the response vector values, e.g., through a logarithmic transformation or through the square root, are popular choices when, for example, the normality assumption of residuals does not hold. In the particular context of PSA kinetics, we found the $\log_2$-transformation to be convenient. Figure 13a displays the example PSA profile of 30 randomly selected patients from the hospital registry cohort in publication **III**; each curve is a single patient with PCa after RP and a subsequent PSA nadir. It is quite evident based on the visual inspection of the response profiles therein that the whole range of response profiles cannot be captured easily by a single model family, especially not any of the commonly used parametric ones such as linear, sigmoidal logistic, or exponential regression (Figure 13a). However, when the PSA response was $\log_2$-transformed (Figure 13b), the response vectors behaved in a way that was more easily captured by conventional modeling techniques. This modeling choice was initially motivated by the rather recently emerged interest in PSA kinetics, which in particular focuses on PSA doubling times (Vickers et al. 2009). The $\log_2$ transformation therefore introduced an interesting connection to the PSA kinetics because a unit increase in the $\log_2$-transformed PSA response corresponds to a doubling of the original PSA within the given time frame.

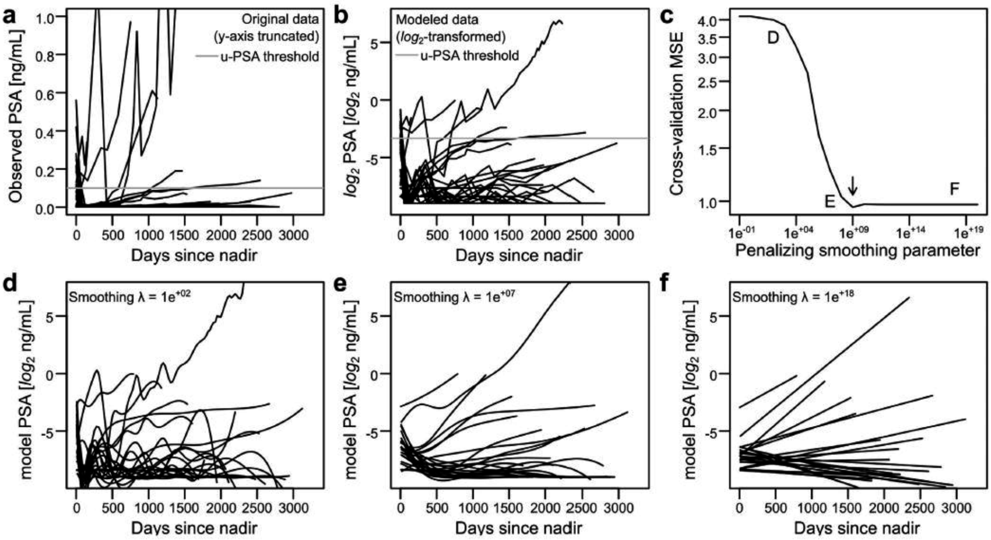


Figure 13: Penalized splines in modeling PSA response curves after nadir. Each curve represents PSA for a single patient. (a): Example data of 30 patients. (b): After $\log_2$ transformation of the PSA response, the response patterns followed linear trends based on visual inspection. (c): A wide range of nonlinear and approximately linear models were tested. CV median-squared error was used as a criterion for optimizing the spline penalization λ (log y-axis). Nonlinear models (inset D) presented with significantly higher error. The optimal model (arrow; inset E) was approximately linear. The linear models (inset F) were close to the optimal model. (d): A highly non-linear spline fit with large CV error. (e): The optimal spline penalization as suggested by CV. (f): Linear spline models performed almost as well as the model with minimized CV error. (Adopted with permission from Publication **III**: Figure 1)

Despite the more uniform form of log$_2$-transformed PSA curves, the question of which model family to utilize remained. Therefore, for the explorative modeling of the log$_2$-PSA, semiparametric penalized splines were chosen as the initial approach to investigate preliminary trends. The utilized natural cubic splines consist of the following third order polynomial form (Ripley 2013):

$$f(x) = \begin{cases} A_0x^3 + B_0x^2 + C_0x + D_0, & t_0 \leq x \leq t_1 \\ \qquad\qquad \dots & \\ A_{n-1}x^3 + B_{n-1}x^2 + C_{n-1}x + D_{n-1}, & t_{n-1} \leq x \leq t_n \end{cases} \qquad \text{Eq. 21}$$

Given the observed interval in time $t \in [a_i, b_i]$ for the predictor x, the smoothing splines are defined using intervals t with $a_i = t_0 < t_1 < \cdots < t_{n-1} < t_n = b_i$. The cubic smoothing spline is then fit by minimizing the following target function:

$$\sum (y - f(x))^2 + \lambda \int_a^b f''(x) \qquad \text{Eq. 22}$$

The Eq. 22 consists of two main components: 1) the sum of squared errors, a common measure of discrepancy between the true observed values y and the model prediction $f(x)$; 2) the penalization term with the smoothing coefficient λ indicating the magnitude of smoothing. The smoothed component is the second degree integral over the observed range of x. The smoothing parameter λ allows cubic penalized splines flexibility (Figure 13d-f) from highly nonlinear fits to linear fits, where only the first-order coefficients of x remain nonzero. As $\lambda \rightarrow \infty$, the model converges toward simple regression because the overwhelming penalization of the second-order integral drives both coefficients A and B to zero in Eq. 21 (Ramsay et al. 1997). Between these extremes lies a sequence of suitable λ for modeling, under which the parameter λ is usually determined using CV. For this purpose, median squared error (MSE) was utilized in this application (Figure 13c) instead of the more popular mean squared error (identically abbreviated as MSE) due to the challenging nature of extrapolating higher-order polynomials outside the original range of training x. Overall, higher values of the smoothing parameter λ generally resulted in better generalization ability in this application, even with the median-based error (Figure 13c). While the log$_2$-transformation did transform the responses to linear curves based on visual inspection and model evaluation, lower censoring presented a challenge regarding the error term assumptions for parametric models. As such, the lower threshold was expected to present a jagged-like effect in residual plots, but was deemed to present only a minor violation to assumptions, as splines highly supported the use of linear modeling trends.

3.4.2. Modeling clinical BCR using mixed-effects models

The spline smoothing parameter λ suggested by CV effectively relied on the linear components (Figure 13c), coefficients Cx^1 (slope) and $Dx^0 = D$ (intercept), for use in Eq. 21. Due to the model complexity favoring these linear trends, we chose to use linear MEMs similar to their use in modeling preclinical response patterns. The following model formulation was utilized:

$$y_{i,t} = \beta_0 + \beta_1 x_{i,t} + \gamma_{0,i} + \gamma_{1,i} x_{i,t} + \varepsilon_{i,t} \qquad \text{Eq. 23}$$

where $y_{i,t}$ is log$_2$-transformed PSA concentration at the time point t for the individual i, β_0 denotes a fixed effect for the population-wide average for the intercept at the PSA nadir in the log$_2$ scale, β_1 denotes a population-wide PSA doubling time (PSADT), and $x_{i,t}$ is the follow-up time in days for the patient i. $\gamma_{0,i}$ and $\gamma_{1,i}$ were the corresponding normally distributed zero mean random effects, which allowed lenience for the whole model fit to allow varying nadirs and PSADTs, respectively. Therefore, to obtain the individualized estimate for the patient with index i, one must sum the population average with the corresponding individual variation:

$$\text{Nadir}: \beta_0 + \gamma_{0,i} \qquad \text{Eq. 24}$$

$$\text{PSADT}: \beta_1 + \gamma_{1,i} \qquad \text{Eq. 25}$$

Notice that the $\log_2$ transformation has a convenient interpretation in regard to PSADT. Because the linear model was fitted to the $\log_2$ scale PSA, the estimate for a linear unit increase in the new scale corresponded to doubling of the original PSA concentration. Therefore, the obtained individualized estimates in Eq. 24 and Eq. 25 provided clinically relevant insight into the PSA kinetics in the original scale. To utilize this convenient connection, a generalized linear mixed-effects model was then constructed for predicting future patients' risk of BCR using the logistic link function:

$$log \frac{p(x)}{1-p(x)} = \beta_{base} + \beta_{nadir}(\beta_0 + \gamma_{0,i}) + \beta_{PSADT}(\beta_1 + \gamma_{1,i}) + \varepsilon_i \qquad \text{Eq. 26}$$

Here, β_{base} denotes the base chance of a BCR occurrence, i.e., an imbalance in the positive or negative cases in the binary prediction task. The β_{nadir} and β_{PSADT} take plug-in estimates from the original linear MEM that described a patient's $\log_2$ scale nadir and PSADT, respectively, and $p(x)$ denotes the probability of observing an event conditional to our input variables x, i.e., $p(x) = p(y = 1|x)$. Therefore, the estimated model in Eq. 26 can be utilized to also predict future patients' risk of BCR given that the patient's nadir and PSADT can be estimated. Inversely, the probability of observing a positive class in Eq. 26 can be derived as:

$$p(x) = \frac{e^{\beta_{base}+\beta_{nadir}(x_1)+\beta_{PSADT}(x_2)}}{1 + e^{\beta_{base}+\beta_{nadir}(x_1)+\beta_{PSADT}(x_2)}} \qquad \text{Eq. 27}$$

where x_1 and x_2 correspond to the individuals' estimated $\log_2$-nadir and PSADT, respectively. As $p(x) \in [0,1]$, the formula in Eq. 27 can be used to predict BCR for a new patient. By varying a threshold for this classifier between $[0,1]$, one may construct receiver-operator curves (ROC) and choose a suitable trade-off between sensitivity and specificity of the binary prediction. As an example, if one desires equal emphasis for both BCR and non-BCR patients, the classifier could be:

$$\begin{cases} p(x) \geq 0.5 \rightarrow BCR \\ p(x) < 0.5 \rightarrow non\ BCR \end{cases} \qquad \text{Eq. 28}$$

For this purpose, noticeable connections to simple linear regression were utilized. Since Eq. 23 required access to multiple individuals to reevaluate the variance in the random effects γ, we proposed the use of simple linear regression to estimate plug-in estimates required by Eq. 26:

$$y_t = \hat{\beta}_0 + \hat{\beta}_1 x_t + \varepsilon_t \qquad \text{Eq. 29}$$

where $\hat{\beta}_0$ and $\hat{\beta}_1$ serve as substitutes for the plug-in predictors that were originally obtained while also modeling population-wide variance in the random effects. These $\hat{\beta}$ estimates are quickly computed using a closed-form solution assuming normally distributed error. The connection to piecewise simple linear regression also appears in penalized splines, and an analogous formula to Eq. 29 is obtained when the spline penalization converges $\lambda \to \infty$ in optimizing Eq. 22.

3.4.3. Regularized regression for response modeling

Regularized or penalized regression typically refers to least absolute shrinkage and selection operation (LASSO), elastic net (EN) or ridge regression (RR). However, the terminology is rather ambiguous as penalization may refer to any methodology where the estimated model is a compromise between goodness of fit and model complexity. The most common form of regularized regression aims to minimize the following target function (Friedman et al. 2010):

$$\min_{\beta_0,\beta} \frac{1}{N} \sum_{i=1}^{N} w_i l(y_i, \beta_0 + \beta^T x_i) + \lambda \left[\frac{1}{2}(1-\alpha)\|\beta\|_2^2 + \alpha\|\beta\|_1 \right] \qquad \text{Eq. 30}$$

where w_i is a user-defined importance weighting for the i:th observation, $l(y_i, f(x))$ is the negative log-likelihood contribution for the i:th observation, λ is the magnitude of penalization for nonzero model coefficients in comparison to the goodness of fit, $1 \geq \alpha \geq 0$ is the regularization family tuning parameter, $\|\beta\|_2^2$ is the L_2-norm (also known as the Euclidean norm) squared, $\|\beta\|_1$ is the absolute L_1-norm (also known as the Manhattan norm), and β are the model coefficients. For the purposes of this thesis, no weighting was utilized; thus, $w_i = 1 \, \forall \, i$ and is omitted from further inspection. The goodness of fit term $l(y_i, f(x))$ was utilized in multiple different forms and was specified with a suitable formulation dependent on the application. Finally, the λ penalization term balances different characteristics of the norm-penalization, although in extreme cases ($\alpha = 1$, LASSO; $\alpha = 0$, RR), parts of it become redundant while the full L_1/L_2-norm combination is retained in the EN-case ($1 > \alpha > 0$). The estimated coefficients β initially correspond to a full set of variables that are iteratively converging as a function of λ toward zero starting from a full, nonpenalized model with sparse coefficients. This methodology belongs to the embedded family of feature selection techniques because the model is estimated simultaneously while selecting the contributing features. It is common knownledge that RR ($\alpha = 0$) retains multiple highly correlated β variables that converge arbitrarily close to zero as λ increases, whereas LASSO ($\alpha = 1$) picks a single one of the highly correlated variables and drives the coefficients of the others to exactly zero. EN presents a compromise between the two, and a typical approach in selecting α involves testing both extreme ends of the α spectrum along with some suitable EN-variants (e.g., conventional values such as $\alpha \in \{0.25, 0.50, 0.75\}$).

The choice of the goodness of fit measure $l(y, f(x))$ is dependent on the application. The $y \in \{0,1\}$ binary classifier error was used for predicting biochemical response in publication **III**, whereas the DREAM Challenge included a two-column survival y-response in publication **IV** that is commonly modeled as proportional hazards (also known as Cox model) due to its time-censoring dependent nature. The continuous, normally distributed y is perhaps the most dominant application, and for this purpose, the traditional measure for the goodness of fit is the sum of squared errors:

$$\min_{\beta_0,\beta} \frac{1}{N} \sum_{i=1}^{N} \frac{1}{2}(y_i - \beta_0 - \beta^T x_i)^2 + \lambda \left[\frac{1}{2}(1-\alpha)\|\beta\|_2^2 + \alpha\|\beta\|_1 \right] \qquad \text{Eq. 31}$$

Additional parameters such as the Gleason score, T class, or histological characteristics were subjected to feature selection together with the PSA nadir and kinetics in publication **III**; for this purpose, a penalized LASSO regression model was built for testing if the additional clinical parameters in **III** could complement the prediction of BCR in addition to the PSA (as derived from Eq. 30):

$$\min_{\beta_0,\beta} - \left[\frac{1}{N}\sum_{i=1}^{N} y_i(\beta_0 + \beta x_i^T) - \log(1 + e^{(\beta_0+\beta x_i^T)})\right] + \lambda[\|\beta\|_1] \qquad \text{Eq. 32}$$

where the goodness of fit is measured using the negative binomial log-likelihood. For survival responses in **IV**, an ensemble structure was built in which each ensemble member optimized a single regularized Cox regression model with the objective function (Simon et al. 2011):

$$\min_{\beta}\left[\frac{2}{N}\sum_{i=1}^{n}(X_{j(i)}^T\ \beta - \ln(\sum_{j\in R_i} e^{X_j^T\beta})) - \lambda(\alpha\sum_{i=1}^{d}|\beta_i| + \frac{1}{2}(1-\alpha)\sum_{i=1}^{d}\beta_i^2)\right] \qquad \text{Eq. 33}$$

where X is the model matrix of predictors, β are the model coefficients in Cox regression, d is the number of predictors (dimension), and N is the number of observations. Furthermore, inner loops exist inside the Cox regression goodness of fit; $j(i)$ is the index of observation event at time T_i, R_i is the set of patient indices j for which $y_j \geq T_i$ (set of patients at risk at time T_i), and y_j is the observed death or right-censoring time. The set R_i is iteratively redefined as a function of $j(i)$ given the outer loop's patient i by incorporating patients at risk at time T_i into the inner loop. This goodness-of-fit measure in the Eq. 33 is known as the Cox or the proportional hazards model. The above formulation of the Cox model is popular in survival response modeling because it does not require assumptions for the underlying base hazard function. Although the interpretation of the resulting regression model can be difficult, the set of assumptions results in an easily estimable model, in which the regression coefficient β estimates are interpreted as hazard ratios (Tibshirani 1997).

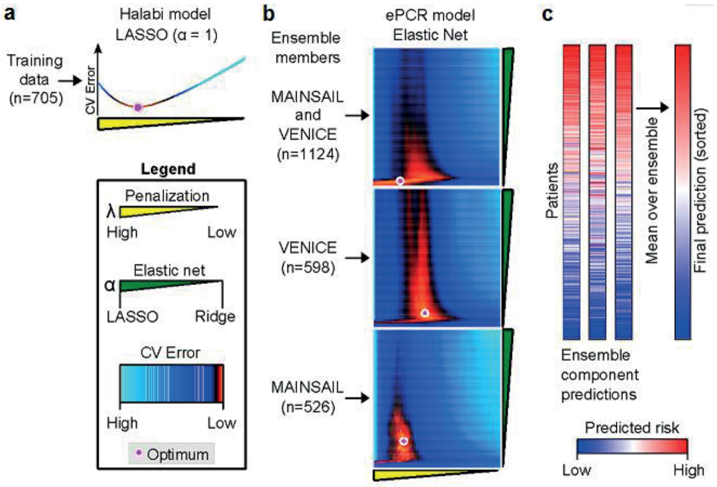


Figure 14: The novel ePCR-modeling approach in comparison to the benchmarking Halabi model. (a): The benchmarking Halabi model is a LASSO model with $\alpha = 1$. (b): The ePCR-methodology stratifies the most relevant trials as separate ensemble members. For each member, a two-dimensional grid $\{\lambda, \alpha\}$ is explored using multiple averaged runs of CV. (c): The ensemble prediction is obtained by averaging over the predicted ranks from ensemble members. (Adopted with permission from Publication **IV**: Supplementary Figure S1)

The overall novel methodology is presented in Figure 14. The implemented ePCR methodology can be seen as an extension of the previous state-of-the-art model (Halabi et al. 2014.); instead of utilizing LASSO regression as presented in the Halabi model (Figure 14a), the methodology allows a two-dimensional parameter grid to be explored (Figure 14b). First, $\alpha \in [0,1]$ is explored with EN models including the extreme ends of LASSO and RR. After this, a sequence of λ values is explored conditional for each given α. By default the CV procedure is run multiple times, and an averaged heatmap of the CV surface was plotted to examine the model performance in regards to integrated time-dependent AUC (iAUC) or other suitable survival performance metric. After the optimum for each ensemble member is identified, they are combined together into the ePCR-structure, and a consensus risk prediction is given as an average over all the predicted ranks from the ensemble members (Figure 14c). Furthermore, it should be emphasized that the input data to the ePCR model in DREAM 9.5 mCRPC Challenge - despite being linear by nature - included pairwise multiplications of the input variables, thus allowing for nonlinear trends to be incorporated. These pairwise interactions attempted to capture interesting clinical phenomena that do not occur only as a function of a single predictor.

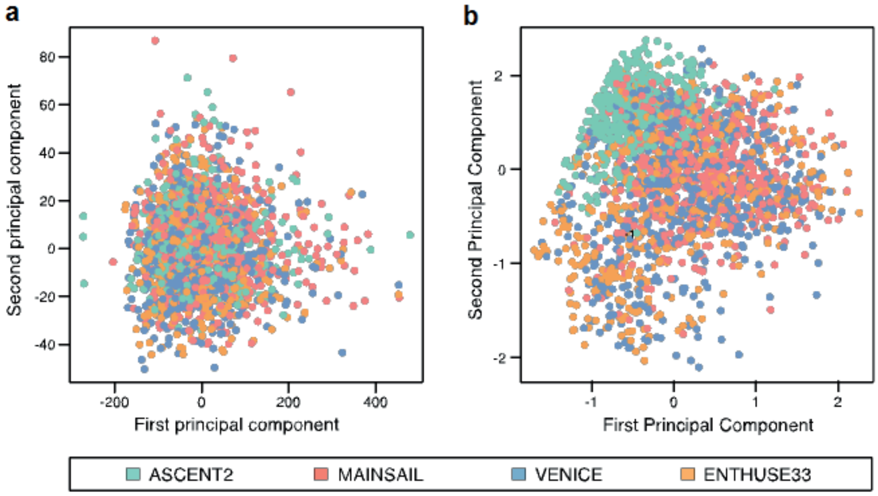


Figure 15: PCA plots with annotations for the clinical trials in the DREAM 9.5 mCRPC Challenge. (a): The mixture of continuous, ordinal and binary clinical variables resulted in a PCA plot with no visible stratification in respect to the trials. (b): For binary variables such as medication history, previous diseases, and metastatic lesion sites, the PCA plot displayed alarming systematic trends for the ASCENT2 trial. (Adopted with permission from Publication **IV**: Supplementary Figure S2)

Out of the three training sets (ASCENT2, MAINSAIL, and VENICE), ASCENT2 was dropped from the final ensemble model. ASCENT2 was clearly different in its characteristics; mainly, the binary indicator labels differentiated it from the other primary studies based on principal component analysis and other standard diagnostics (Figure 15). Furthermore, the survival response vector in ASCENT2 had a significantly shorter mean follow-up time and markedly lower count of observed events.

One open research question in penalized regression is how to interpret statistical significance of model coefficients as conventional *p*-values are typically reported (and required) by life science journals.

Because the final ePCR ensemble model members were close to being RR (Figure 14b), the number of nonzero coefficients in the model remained high, even if they contributed minimally to the final prediction. Bootstrapped *p*-values for regression coefficients give highly optimistic values for the variance, whereas the bias (λ penalization) drives the coefficients toward zero (Goeman et al. 2016). Therefore, the null hypothesis for the penalized coefficients is ill-defined. The interpretation for the statistical significance of these coefficients remains an open field, and prominent figures in the penalized regression field have contributed research aiming toward feasible interpretations of *p*-values in these penalized regression methods (Lockhart et at. 2014).

3.4.4. Network projection and meta-analysis in the DREAM 9.5 mCRPC Challenge

Because the input data matrix for the ePCR model also incorporated pairwise interactions (multiplication of columns in the data matrix), the visualization of the resulting ensemble structure pose challenges. For this purpose, the freely available *graphopt*-package was used (Gábor et al. 2006). The estimated model's single variables and pairwise interactions were first filtered using a bootstrap-based evaluation of the significance of the coefficients, and their importance was weighted by the absolute integrated area under/over the regularization curve of the coefficients to gain insight into their overall effect sizes. The *graphopt* algorithm takes the importance of the vertices and the importance of edges connecting the nodes as input. It then simulates an algorithm that is inspired by the behavior of electrons in molecules based on the attractiveness of atoms (vertex, single predictors) with the particular bond structure (edges, pairwise predictors). As such, the algorithm is data-driven, and the two-dimensional is not subjective to user bias; however, the algorithm does incorporate multiple tuning parameters that can be used to refine the produced illustration if desired.

As is typical for the DREAM Challenges, the "wisdom of the crowd" principle is applied in the meta-review phase of all final submitted models (Costello et al. 2013). Furthermore, in an application such as this, the meta-review aspect that emphasizes novel findings in single clinical predictors or groups of predictors offers practical utility to the clinical audience. As such, comprehensive surveys and meta-analysis of the competing models was conducted. Naturally, extensive inspection into the particular findings of the top-performing ePCR model was conducted in addition to the Challenge-wide surveys.

Due to the competitive nature of the Challenge, the statistical significance of the model performance in the validation dataset was evaluated relative to the benchmarking Halabi model (Halabi et al. 2014) using the Bayes factor (BF) (Lavine et al. 1999). Furthermore, in addition to ranking the submitted models in descending scoring order, the top-performing model was evaluated whether it outperformed the rest of the models just by chance based on the BF. The conventional threshold of BF ≥ 3 was used as an indicator of significant difference in the models' prediction accuracy.

4. RESULTS

4.1. Experimental design

Up to $N \leq 100$ animals were allocated using the proposed methodology in both the experiments by Knuuttila et al. (2014) and Huhtaniemi et al. (submitted). Multidimensional Scaling (MDS) was used for diagnostic inspection of the submatching in both of the experiments (Figure 16b for Knuuttila et al.; Supplementary Figure S4 in Publication **I** for Huhtaniemi et al.). After randomized allocation, neither of the experiments presented with statistically significant univariate differences in the baseline variables over the randomized groups (tested using one-way ANOVA). Furthermore, castration of the animals in two batches over subsequent weeks (Knuuttila et al. 2014) was successfully blocked out in the experiment by solving two separate matching subtasks (Figure 16). Mahalanobis distance and weighted Euclidean distance were chosen for the two experiments, respectively. Mahalanobis distance adjusted for the following correlated baseline variables: pre-castration PSA, latest post-castration PSA, relative change in PSA, and body weight. Weighted Euclidean distance was used because expert-curated weighting was requested by the experimenters.

Interestingly, the dissimilarity matrix computed from global RNA expression in post-sacrifice tumors after anti-androgen therapy showed statistically significant correlation with the baseline dissimilarity matrix in the VCaP experiment (Knuuttila et al. 2014) (*p*=0.0389, Mantel's test for matrix correlation; Publication **I:** Supplementary Figure S8c). In addition to the real experiments, the feasibility of matching-based allocation was subjected to simulated conditions. A varying number of predictively informative or non-informative variables were included in the computation of the distance matrix, and the simulated results supported the *a priori* expectation that the matching approach performs as good or better than conventional randomization and showed increased performance as additional informative baseline variables were included. (Publication **I:** Supplementary Figures S9 and S10; Supplementary material “Simulation study for predictive baseline covariates” and “Simulation study for baseline-adjusted or matched regression models”).

The bootstrap-based power calculations from Publication **I** are presented in Figure 17 for both the model that utilized matching and for the conventional model. The results were largely concordant over both approaches in the ARN-509 vs. vehicle (Figure 17a left panel) and ORX vs. intact tumors comparisons (Figure 17b left panel). However, in detecting an MDV3100 vs. vehicle intervention effect, the matched approach was considerably more powerful (Figure 17a right panel). The difference was notably greater in the ORX+Tx vs. ORX comparison (Figure 17b right panel), in which the original growth curves did not originally follow a linear trend (Publication **I**: Supplementary Figure S6b). When pairwise matching-based differences were utilized, these differences approximately followed a linear pattern, thus explaining why the inference was so different between the two approaches (Publication **I**: Supplementary Figures S5 and S6). This improved statistical power was due to modeling pairwise differences of equidistant observations between baseline-matched individuals (Eq. 8 and Eq. 9). Detection of statistically significant differences required slightly fewer than 10 tumors with both approaches in treatment arms with a notable effect size (left panels in Figure 17a-b). Identifying more subtle differences benefitted from utilizing the matching information (right panels in Figure 17a-b).

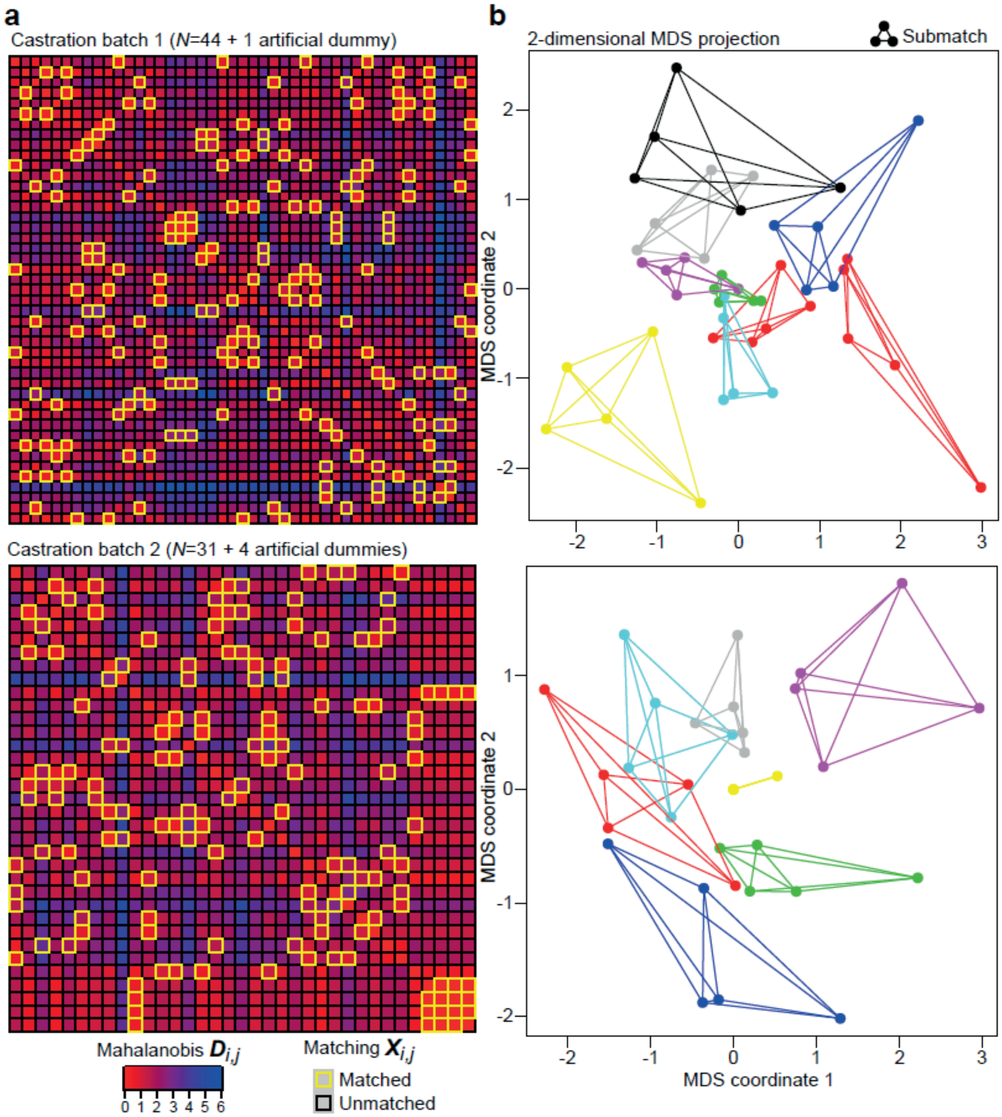


Figure 16: Baseline matching of five intervention arms for the VCaP experiment using Mahalanobis distance for baseline characteristics. (a): Distance matrix together with the matching matrix indicated with highlighting. (b): Two-dimensional projection of the identified submatches using multidimensional scaling (MDS). Each submatch is indicated with a different color. Within submatches each individual was randomized into a different treatment arm. (Adopted with permission from Publication **I**: Supplementary Figure S3)

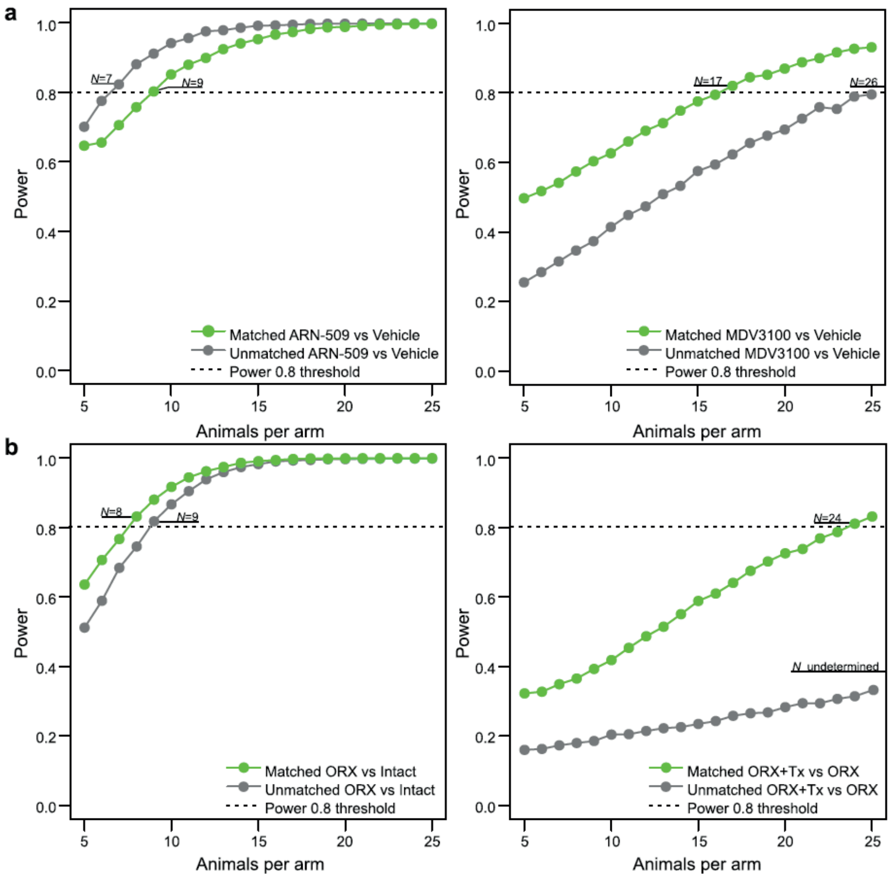


Figure 17: Bootstrap-based power simulations. (a): ARN-509 / MDV3100 experiment. (b): ORX / ORX + Tx experiment. (Adopted with permission from Publication **I**: Figure 4)

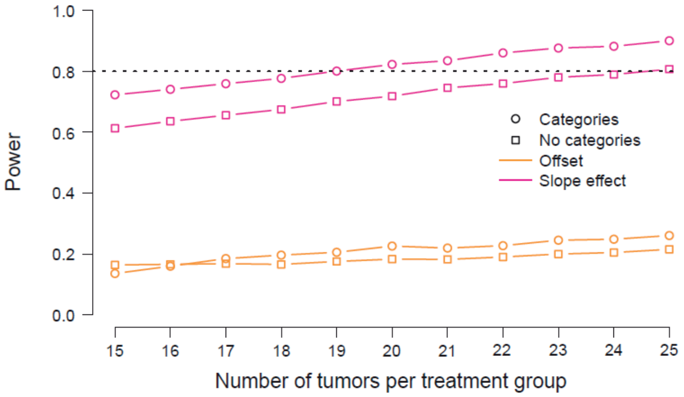


Figure 18: Model-based power simulations for the LNCaP study. The proposed sample sizes were $N = 19$ for the categorizing model (Eq. 12) and $N = 25$ for the conventional model (Eq. 13). (Adopted with permission from Publication **II**: Supplementary Figure S5)

The power analysis in publication **II** (Figure 18) concluded that the categorizing MEM maintained higher statistical power in the dataset in which LNCaP cells were treated with the compound DPN (unpublished in-house study). The slope coefficient effect using the conventional power threshold of 0.8 suggested an optimal number of animals of $N = 19$ for the categorizing model and $N = 25$ for the conventional, non-categorizing model. Although both suggestions involved a relatively high number of animals, the reduction of $N_{difference} = 6$ indicates that identifying readily known treatment interventions in novel animal model settings could be achieved using a significantly smaller number of individuals. The importance of an offset term in such settings remains open because it does not test the baseline difference between treatment arms as one might intuitively interpret; the entire growth profile of tumors using $\theta_i = 0$ affects the estimation of this coefficient and it did not show signs of identifying a treatment effect in the LNCaP study (Figure 18).

4.2. Preclinical findings – Publication I

Table 1. Mixed-effects model fits for the fixed effects (population inference) and random effects (individual effects and the random error term)**.** Model estimates and their significance levels using the conventional unmatched and matching-based pairwise models are presented for each intervention comparison separately. The model term that explicitly tests for an intervention effect is highlighted in bold. N.S., not significant; * *p* < 0.05; ** *p* < 0.01; *** *p* < 0.001. (Adopted with permission from Publication **I**: Table 1)

Model		Fixed effects estimate (*p*-value)			Random effects SD		
		$\beta_{\text{intercept}}$	β_{slope}	$\beta_{\text{intervention}}$	$\gamma_{\text{intercept}}$	γ_{slope}	$\varepsilon_{\text{error}}$
ARN-509 vs Control	Unmatched	14.311 (<0.001)* * *	10.062 (<0.001) ***	**– 7.627 (<0.001) *****	8.234	5.163	5.749
	Matched	0 (-)	0 (-)	**– 7.962 (0.0047) ****	7.053	8.894	8.399
MDV3100 vs Control	Unmatched	13.536 (<0.001)* * *	10.188 (<0.001) ***	**– 4.940 (0.0494) ***	7.635	6.259	6.395
	Matched	0 (-)	0 (-)	**– 5.729 (0.0160) ***	7.013	7.401	11.247
ORX vs Intact	Unmatched	14.548 (<0.001)* * *	1.336 (<0.001) ***	**– 1.265 (0.0034) ****	14.578	0.997	8.518
	Matched	0 (-)	0 (-)	**– 1.931 (0.0063) ****	4.251	2.157	9.522
ORX+Tx vs Intact	Unmatched	9.998 (<0.001)* * *	0.122 (0.0675) N.S.	**– 0.101 (0.2704) N.S.**	10.476	0.167	9.977
	Matched	0 (-)	0 (-)	**– 0.112 (0.0457) ***	2.381	0.155	4.618

The results presented in Table 1 were generated to provide biological insight into the intratumoral synthesis of androgens in (Knuuttila et al. 2014) and insight into the effect of orchiectomy (ORX) coupled with a novel undisclosed treatment (Tx) by Huhtaniemi et al. (submitted). The results for ARN-509 and MDV3100 (upper half of Table 1) were concordant with the clinical observations of the therapeutic effect of these antiandrogens, and both are FDA-approved agents for PCa. The presented work in the preclinical stage supported the use of the VCaP cancer cell line as a sensitive platform for testing clinically translatable results in the preclinical phase. Furthermore, sacrifice-level PSA and supporting markers validated the inference beyond just longitudinal intervention testing presented above (Knuuttila et al. 2014).

The results from the experiment by Huhtaniemi et al. (lower half of Table 1) validated the expected drastic decrease in PSA due to ORX intervention. An undisclosed supplementing intervention, denoted here as Tx, further facilitated the inhibition of tumor growth. The biological motivation behind this novel intervention in a preclinical experiment remains to be published (Huhtaniemi et al., submitted). It should be noted that the ORX+Tx was particularly difficult for conventional regression modeling and presented highly nonlinear patterns both in the ORX and ORX+TX groups (Publication **I**: Supplementary Figure S6 right side). Utilizing the matched pairwise-difference analysis of the coupled observations, the problem could be transformed into the scope of linear MEMs. This effect had also been observed in the power analyses presented before (right panel in Figure 17b). This suggests the useful trait that the matched MEM approach may in special cases simplify nonlinear inference to linear inference by blocking out nonlinear trends.

4.3. Preclinical findings – Publication II

Table 2. Categorizing mixed-effects model versus the conventional modeling approach. In studies where the tumors had no target size before introducing interventions, the intercept (β_1) and offset (β_2) terms were set to zero. (Adopted with permission from Publication **I**: Table 2)

Dataset	Model	Fixed effect estimates (*p*-value)			
		β_1 intercept	β_2 offset	β_3 overall growth	β_4 slope effect
DMBA: ENL low dose	Categorizing	0 (-)	0 (-)	3.12 ***	-1.25 ***
	Conventional	0 (-)	0 (-)	0.989 ***	-0.174
DMBA: ENL high dose	Categorizing	0 (-)	0 (-)	3.26 ***	-1.60 ***
	Conventional	0 (-)	0 (-)	1.02 ***	-0.681
MCF-7: LAR low dose	Categorizing	21.4 ***	-2.66	8.71 ***	-0.599
	Conventional	21.0 ***	-4.18	8.37 ***	-1.04
MCF-7 : LAR high dose	Categorizing	21.4 ***	1.05	8.71 ***	-1.72
	Conventional	21.0 ***	-0.945	8.37 ***	-3.07 *
LNCaP: DPN	Categorizing	234 ***	-22.7	101 ***	-48.9 *
	Conventional	233 ***	-19.5	52.8 **	-41.0
LNCaP: ENL	Categorizing	234 ***	-8.31	101 ***	-81.1 **
	Conventional	233 ***	-5.19	52.7 **	-45.1
4T1: Doxorubicin	Categorizing	0 (-)	0 (-)	68.4 ***	-16.8 *
	Conventional	0 (-)	0 (-)	68.4 ***	-16.8 *
4T1: Cyclophosphamide	Categorizing	0 (-)	0 (-)	68.4 ***	-66.5 ***
	Conventional	0 (-)	0 (-)	68.4 ***	-66.8 ***

* $p<0.05$; ** $p<0.01$; ***$p<0.001$; Effect deemed not statistically significant otherwise.

The retrospective analyses for the three readily published studies (DMBA, MCF-7, and 4T1) successfully replicated key parts of the readily published inference (Table 2; Observations are shown in full in Publication **II**: Supplementary Figure S1). The DMBA study reproduced the original conclusions of the antitumoral effects of high dietary concentrations of enterolactone (ENL) (Saarinen et al. 2002), although the conventional longitudinal model failed to detect a statistically significant difference between the control and a high concentration of ENL. This suggests an increased sensitivity for the categorizing model. Retrospective analysis of the MCF-7 BCa study (Saarinen et al. 2008) was replicated by identifying the antitumoral properties of dietary lariciresinol (LAR). The effect in the categorizing model was detected in the *post hoc* Fisher's exact test for the identified latent growth categories in connection with the intervention arms (Table 3 MCF-7 LAR high dose left-side) together with novel insight into ERβ expression in BCa tumors supported by the literature (Hartman et al. 2006) (Table 3 MCF-7 LAR high dose right-side). The LNCaP xenograft PCa study was novel and presented no retrospective insight, but presented with statistically significant intervention effects for to the well-known anti-tumoral compounds ENL and DPN. The conventional modeling approach did not identify an intervention effect for these known compounds.

The numbers of tumors in the categories of the intervention arms (Table 3, LNCaP) did not show statistically significant differences and provided an estimate of the distribution of the inherent latent groups present in the PCa-related experiment. Artificial data simulations were conducted based on the obtained data, and we concluded that the categorizing model did not result in an elevation of type I error (falsely rejecting null hypothesis; Publication **II**: Supplementary Figure S8). In the syngeneic 4T1 mouse BCa cell line study, the treatment profiles for doxorubicin and cyclophosphamide were clearly distinct from the control group. In this case, the EM-algorithm classified all $\theta_i = 1$, which results in the categorizing model presented in Eq. 12 converging to the special case of the non-categorizing conventional model presented in Eq. 13. Therefore, both models resulted in identical conclusions.

Table 3: *Post hoc* testing of the latent subgroups and relevant markers. The 4T1 study was redundant and was omitted. Study specific tumor characteristics: DMBA: Histologic subtyping; MCF-7: ERβ-receptor expression; LNCaP: PSA concentration at the end of study. (Adopted with permission from Publication **II**: Table 3)

Tumor categories vs. intervention arms (% shown within latent category)				*Tumor characteristics (study specific)*			
DMBA	Control	Treatment	*p*	Poorly differentiated	Well-differentiated	Atrophic	*p*
ENL low dose							
$\theta = 1$	4 (44%)	5 (56%)		4 (67%)	2 (33%)	0 (0%)	
$\theta = 0$	9 (53%)	8 (47%)	1.000	2 (17%)	7 (58%)	3 (25%)	0.156
ENL high dose							
$\theta = 1$	4 (67%)	2 (33%)		3 (60%)	1 (20%)	1 (20%)	
$\theta = 0$	9 (45%)	11 (55%)	0.645	2 (13%)	10 (67%)	3 (20%)	0.069
MCF-7							
LAR low dose				ERβ expression per 1,000 cells[1]			
$\theta = 1$	14 (48%)	15 (52%)		248.1 ± 238.7			
$\theta = 0$	1 (17%)	5 (83%)	0.207	82.0 ± 56.6			0.115
LAR high dose							
$\theta = 1$	14 (56%)	11 (44%)		213.0 ± 127.0			
$\theta = 0$	1 (10%)	9 (90%)	0.022	329.7 ± 32.7			0.008
LNCaP							
DPN				PSA concentration at sacrifice (μg/L)[1]			
$\theta = 1$	6 (67%)	3 (33%)		97.3 ± 48.3			
$\theta = 0$	6 (46%)	7 (54%)	0.415	29.3 ± 17.7			0.005
ENL							
$\theta = 1$	6 (60%)	4 (40%)		99.1 ± 45.5			
$\theta = 0$	6 (60%)	4 (40%)	1.000	29.1 ± 15.4			0.001

[1]: Values shown as mean ± SD; underlining indicates statistically significant difference *p*<0.05; $\theta = 1$ depict tumors categorized as growing, $\theta = 0$ as poorly growing.

Although not a statistically significant finding (Fisher's exact test, *p* = 0.069), we observed a trend known from biological literature that well-differentiated DMBA-induced tumors were overexpressed in the poorly growing ($\theta = 0$) latent subgroup. Inversely, slight overrepresentation of poorly differentiated tumors in the growing latent subgroup ($\theta = 1$) was expected (Table 3; ENL high dose, right side). It has been proposed that estrogen receptor β (ERβ) plays a major role in the growth of BCa xenografts especially in the development of blood vessels (Hartman et al. 2006). Accordingly, we validated this effect of ERβ elevation in growing tumors in the MCF-7 study over the latent growth groups regardless of the intervention arms (Table 3; LAR high dose, right side). Lastly, sacrifice PSA was highly correlated in PCa for the identified latent growth groups both in the DPN and ENL interventions for the LNCap study regardless of the intervention arms (Table 3; LNCaP right side). Therefore, the identified latent tumor growth characteristics were extensively supported by these external factors in the *post hoc* testing.

4.4. Clinical findings – Publication III

Using the raw data and its $\log_2$-transformed formulation (Figure 13), the optimal model formulation with the penalized splines heavily favored linear and/or slightly nonlinear trends based on the CV (large values of λ penalization). The utility of these linear models is simplified compared with nonlinear models. The initial research aim in Publication **III** was to study the reliability of measurements in the t-PSA range ($x \geq 0.1$ ng/mL) versus the u-PSA range (0.1 ng/mL $> x \geq 0.001$ ng/mL), where u-PSA had been suspected to be unreliable. However, in our study, the u-PSA modeled using splines displayed consistent trends over the threshold, suggesting utility for u-PSA (Figure 19).

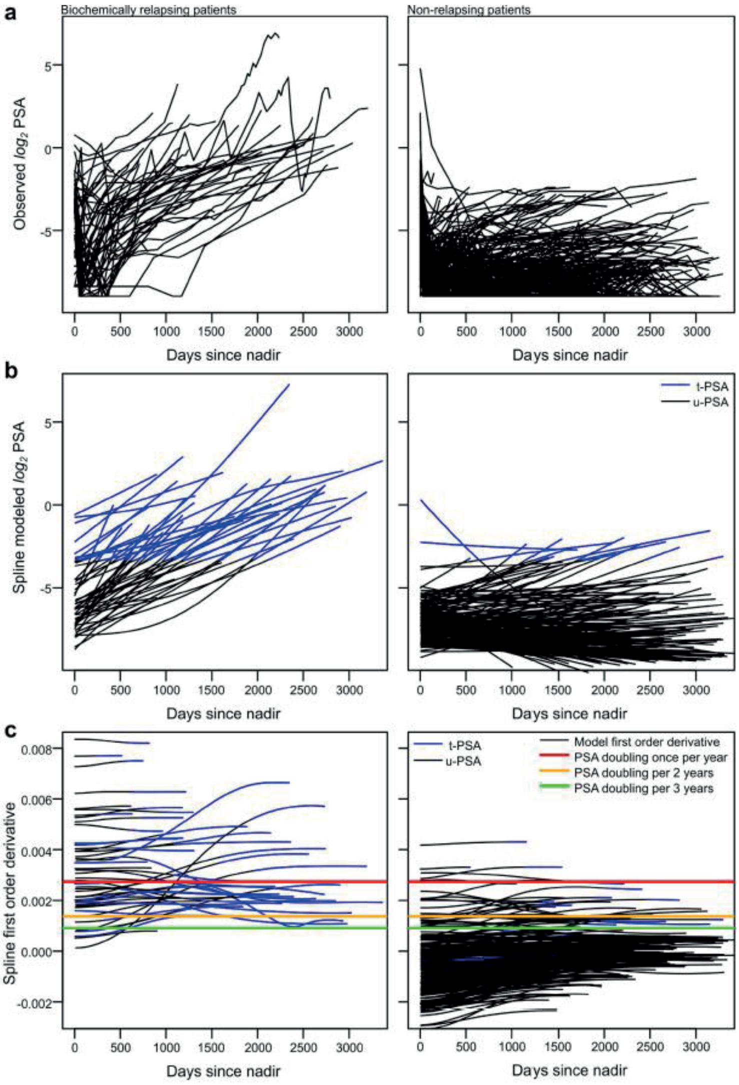


Figure 19: Data and spline modeling in the PSA study. (Left panel): Patients with BCR. (Right panel): Patients without BCR. (a): Log_2 scale longitudinal trends with each curve representing a patient. (b): Optimal spline fits for the $\log_2$ transformed data. (c): First order derivatives for the splines. (Adopted with permission from Publication **III**: Figure 2)

To determine whether the threshold between traditional and ultrasensitive PSA was functioning as a trend-setting threshold, both the spline model fits and their first order derivatives were visually inspected (Figure 19b-c). Annotating these ranges did not present systematic differences, except that patients who eventually experienced BCR had a larger portion of t-PSA measurements. Because BCR is detected via certain PSA thresholds, the models identified clear differences in their first-order derivatives (Figure 19c). For example, PSA doubling occurring more often than once per 2 years appeared as an indicator for heightened risk of BCR. Because the majority of u-PSA measurements were present mostly in patients who never had BCR and primarily presented with horizontal derivatives (Figure 19c; i.e., linear growth patterns), it appeared that these consistent PSA patterns could be connected further to BCR prediction.

Based on the results indicating that linear models could provide sufficient modeling capacity to capture major trends in the $\log_2$-transformed PSA measurements, generalized linear mixed-effects models were utilized with a logistic link function. To compensate for potential time-dependent effects in detecting BCR while retaining clinical relevance in the prediction window and interpretable coefficients, the generalized MEMs were trained limiting the observations to either to a 1-year or 3-year post-surgery window. These fitted models for these respective time windows are presented in Figure 20a-b. For single individuals arriving at the clinic, similar estimates could be generated with simple regression in Eq. 29 (e.g., Publication **III**: Supplementary Table S2). This approach made use of the underlying assumption that the expected value of random effects (γ) is zero, which is true if the zero mean normally distributed random effects assumption holds. Based on a logistic fit for the 1-year and 3-year windows (Figure 20c-d), a risk prediction surface between non-BCR and BCR was built as defined in Eq. 27.

To examine potential residual term heteroscedasticity that may appear due to differences in the traditional and ultrasensitive assays, a representative residual plot was drawn from the 1-year constrained MEM (Figure 20e). Two particular outliers stand out - one from the u-PSA range and the other from t-PSA range. Although residuals maintain the zero mean over the whole range of fitted values, there is a non-alarming decrease in residual variance toward the higher fitted values, suggesting a slight heteroscedasticity. This challenge in modeling is further elaborated by the jagged-like effect in the lower-left end of the residual plot (Figure 20e), where the low-censored ultrasensitive PSA measurements produced a slight artifact.

The held-out external validation set (1/3 of the original data; Publication **III**: Table 1) was tested using ROC-AUC by a researcher blinded to the original conclusions. Surprisingly, the validation dataset presented a very high ROC-AUC for the 3-year post-surgery window (Figure 20f, green) window and excellent ROC-AUC even when utilizing only 1-year post-surgery measurements (Figure 20f, orange). For interpreting these ROC-AUC results in a broader scope, one must remember that the BCR event itself is defined using PSA thresholds. Therefore, it is expected that early PSA trends have potential in predicting the BCR event. However, the very high ROC-AUC in a held-out validation set already at the 1-year window suggests that the data consisting predominantly of only u-PSA measurements presents an early prognostic signal. Given that 1-year of follow-up is a feasible time frame for practical clinical use in follow-up, the modeling results suggest that informative and useful signals from u-PSA can be extracted.

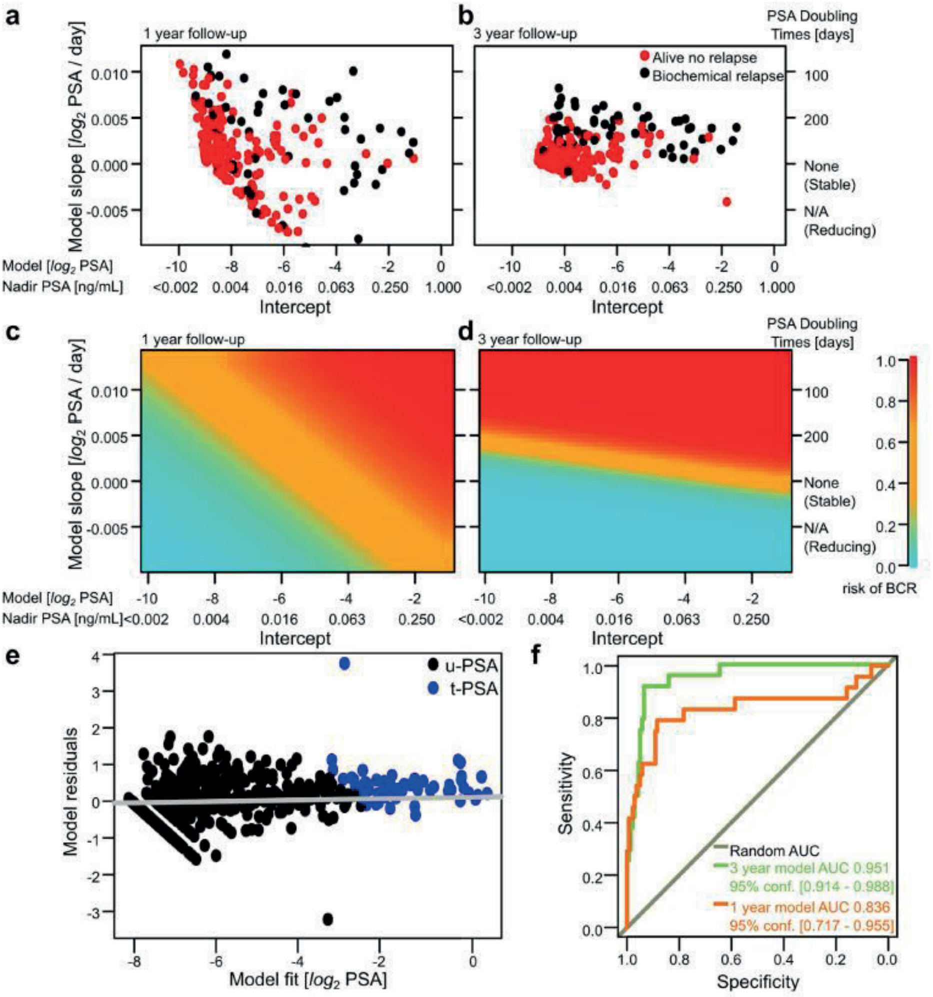


Figure 20: Generalized logistic regression for the 1-year and 3-year post-operative window, model diagnostics, and model validation. (a,b): Estimated patient-wise MEM parameters in post-operative 1-year and 3-year windows, respectively. (c,d): Logistic regression prediction surfaces corresponding 1-year and 3-year post-operative PSA measurements, respectively. (e): Representative residual plot of the 1-year generalized MEM fit. (f) ROC-AUC prediction accuracy based on the data set aside to serve as independent validation. (Adopted with permission from Publication **III**: Figure 3)

To test PSA trend specificity in the BCR prediction, we further subjected the individualized PSA nadir and PSADT characteristics to LASSO modeling (Eq. 32), complemented by all the available clinical parameters. In the regularization curve, PSADT was clearly the best predictor, closely followed by the PSA nadir. CV suggested the use of only these two parameters in BCR prediction (Publication **III**: Supplementary Figure S2).

4.5. Clinical findings – Publication IV

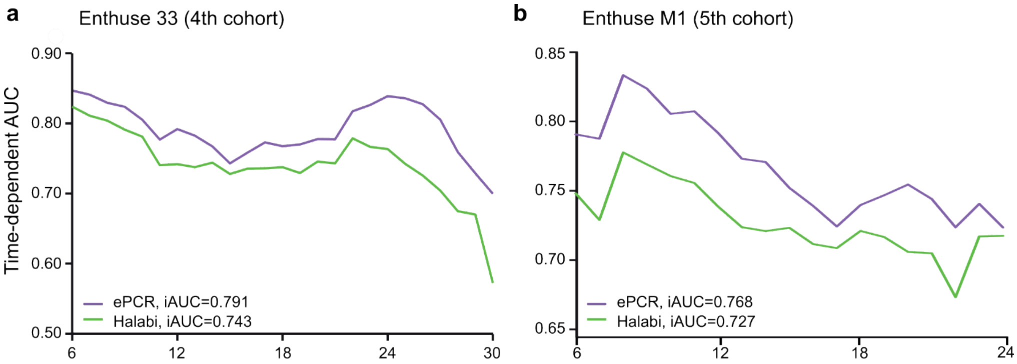


Figure 21: ePCR method performancein comparison to the Halabi benchmark. iAUC is the proportional area under the curves. (a): ENTHUSE 33 cohort utilized for leaderboards and validation. (b): ENTHUSE M1 cohort utilized only once for independent validation. (Adopted with permission from Publication **IV**: Figures 2 and 5)

Figure 21 displays the difference of the developed ePCR methodology with iAUC in comparison with the Halabi model benchmark. The novel ePCR methodology outperformed the state-of-the-art method for mCRPC OS prediction by a significant margin (BF > 20) and was consistently better both in the independent 4th (Figure 21a) and 5th validation cohorts (Figure 21b) at each time point. The submitted 2nd, 3rd, and 4th best performing models had the following iAUCs in the ENTHUSE 33 dataset: 0.7789, 0.7778, and 0.7758. The final submitted iAUC of the ePCR during the Challenge was 0.7915. In comparison with the 2nd best performing team, ePCR had an advantage of BF > 5.

Interestingly, when the ePCR ensemble model's most important predictors were projected to a two-dimensional graph using the data-driven, agnostic *graphopt* algorithm (Figure 22), the resulting graph presented clinically relevant subgroupings that were expert-curated for the purposes of publication **IV**. The most prominent singular markers were largely enzymes related to antitumor activity, such as aspartate aminotransferase (AST), lactate dehydrogenase (LDH), and alkaline phosphatase (ALP). Interestingly, PSA was not a top contributor, although it remained a significant single contributor, hinting that survival or tumor burden in the mCRPC form of the disease is perhaps no longer well represented by this surrogate marker. Furthermore, blood-related markers and general clinically relevant factors were among the top hits, although they may have been partially present as general predictors for survival and not necessarily specific for mCRPC. The former markers included hematocrit (HCT), hemoglobin (HB), red blood cell count (RBC), and albumin (ALB). The latter included the previous use of opioid analgesics, which was one of the central nodes and is possibly linked to the presence of various comorbidities. Furthermore, other general survival-related markers were present, such as the ECOG performance status (Oken et al. 1982), which is an ordinal-scale subjective evaluation of the patient's functionality at the workplace or with self-care.

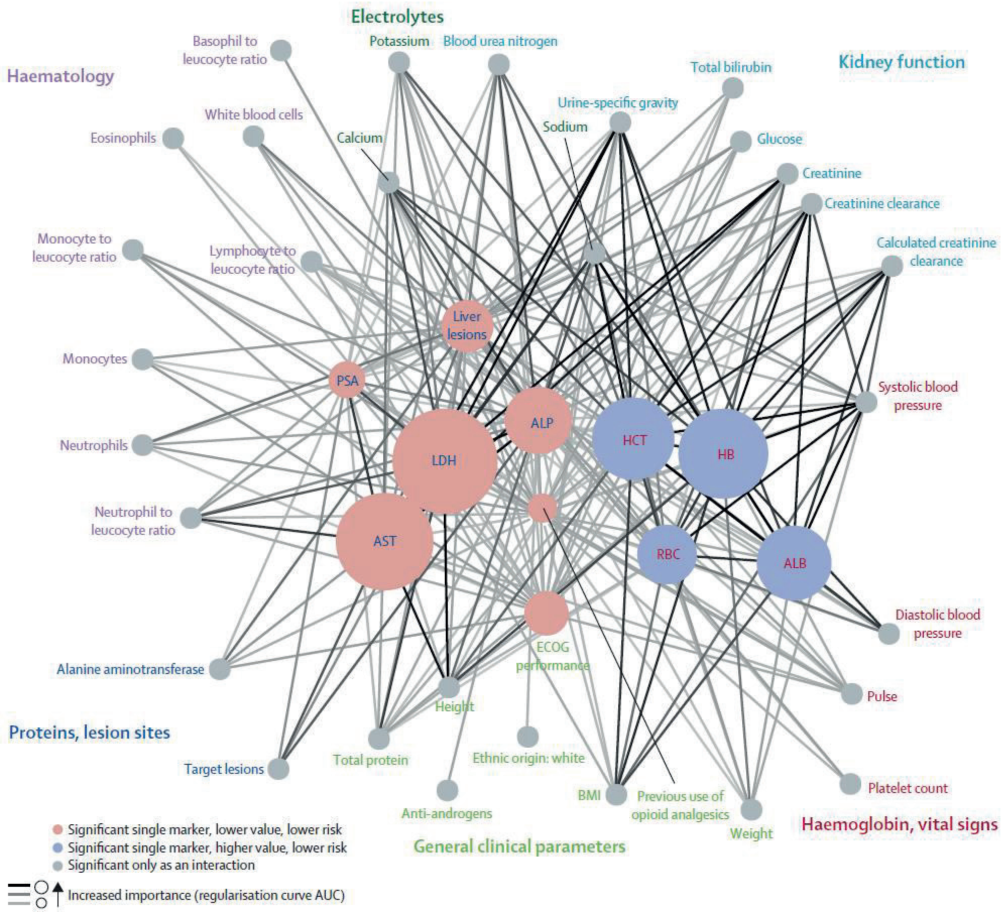


Figure 22: A data-driven graph of clinical predictors contributing most to the final top-performing ePCR model. When single predictors and their interactions were included, the two-dimensional agnostic projection algorithm grouped the predictors into physiologically and clinically relevant subsets that were expert-curated as follows: (red): Blood-related markers; (light blue): Kidney function related markers; (dark green): Electrolytes; (purple): Immunosystem function; (dark blue): Proteins/enzymes and lesion sites; (light green): General clinical parameters. (Adopted with permission from Publication **IV**: Figure 3)

It should be noted that the ensemble members in the proposed ePCR closely resembled RR, although they were still technically EN (Figure 14b). Because α was relatively close to zero, the RR-like behavior allowed multiple correlated coefficients to be included in the final model while dispersing the effect of the whole phenomenon among those coefficients. This result partly explains why multiple variables depicting the same biological phenomenon (e.g., kidney creatinine function in its various forms in upper-right section of Figure 22) remained in the model after embedded feature selection. While the CV procedure also tested LASSO, the objective criterion suggested the combined use of these correlated variables in EN instead of the LASSO-like behavior of picking the best among multiple correlated variables, which was the approach used in the Halabi model (Figure 14a). A tendency toward favoring RR could occur due to less technical variation contributing to the model after combining multiple independent technical measurements of the same biological phenomenon.

Alternatively, it is possible that the introduction of pairwise variable interactions captured some small but significant nuances in the mCRPC OS prediction. The network-level presentation, which is essential for visualizing such a complex Cox regression ensemble, may prove to be challenging to interpret for practicing clinicians (Davis 2017) even if improves the prediction accuracy. However, the presentation and model fitting procedures were conducted in a data-driven manner, and for this purpose, RR-like ensemble modeling with pairwise interactions appeared optimal.

Table 4 reports the top 5 single predictors in the ePCR model (panel a), the top 5 pairwise predictors in the ePCR model (panel b), and the top 5 predictors from the meta-analysis based on a survey of the 50 participating teams (panel c). The clinical novelty of AST was reported in Publication **IV**; AST was one of the top contributors to ePCR and was important for over half of the other models submitted to the challenge (Publication **IV**: Supplementary Figure S7). Furthermore, because the Halabi model did not consider marker interactions, the prominent presence of such interactions in the ePCR method suggested that successful models for OS prediction are not only based on single molecules or clinical attributes, but may instead involve their complex combinations.

Table 4: Top 5 reported single ePCR, pairwise ePCR, and meta-review predictors over all models submitted to the DREAM 9.5 mCRPC Challenge. (Adopted with permission from Publication **IV**: Supplementary Table S7 and Supplementary Figure S3)

a	Top 5 single predictors in ePCR				
Single predictor	LDH	AST	HB	HCT	ALB
Novelty vs. Halabi		✓		✓	

b	Top 5 pairwise predictors in ePCR				
Pairwise predictors	AST & LDH	ALP & LDH	ALP & AST	HB & SBP[1]	LDH & USG[2]
Novelty vs. Halabi	✓		✓	✓	✓

c	Top 5 predictors reported as important over all submitted models				
Meta-review predictors	ALP[3]	HB	ECOG	AST	LDH
(% reported as important)	(70%)	(60%)	(60%)	(50%)	(45%)
Novelty vs. Halabi				✓	

[1]: Systolic blood pressure. [2]: Urine-specific gravity. [3]: ALP was #6 as a single predictor for ePCR.

4.6. Method implementation and user-interfaces

Over the course of work leading to publications **I** - **IV**, a substantial amount of open source R code was created for the R Statistical Software (R Development Core Team, 2015). However, given the highly multidisciplinary nature of the work, it was important that the methods were also made easily accessible as well as the key results and the corresponding code reproducible and transparent. A considerable effort was therefore put into facilitating the use of the methods also by non-specialized experimenters, who may face limited or no access to suitable bioinformatics services.

An R package called *hamlet* (Hierarchical Optimal Matching and Machine Learning Toolbox; URL: https://CRAN.R-project.org/package=hamlet ; Accessed: 4th October 2017) was created for publication **I** and the pre- and post-intervention VCaP data is embedded into the package to exemplify both the pre-intervention design approach as well as the post-intervention analysis of preclinical experiments (Knuuttila et al. 2014). The results from the data, as well as the pre-intervention allocation, are reproducible in a supporting R vignette document. *hamlet* is located on CRAN (Comprehensive R Archive Network), which offers automated sanity checks on R-packages, manually curated support, download services, and readily hosts thousands of specialized R packages. A graphical user-interface was implemented for hamlet, where all key functionality of the R-package was made available for preclinical researchers in a web browser using the R Shiny platform (https://Rvivo.tcdm.fi ; Accessed 20th February 2018).

The methodology in Publication **II** was released as an R-package called *XenoCat*. However, after the Google Code repository services were taken down on 24th of August 2015, the package has been re-directed and its content supplemented into future extensions of the *hamlet* package (original Google Code URL: https://code.google.com/archive/p/r-xenocat/). Similar to *hamlet* in publication **I**, exemplifying data from Publication **II** was embedded into its R package *XenoCat* to illustrate its use. The package gathered interest at its release and subsequently users forked the original code and continued using and further developing the open-source R-package independently (e.g., fork by A. Borgman, URL: https://github.com/borgmaan/xeno-fix/ ; Accessed: 11th February 2016).

Both of the clinically focused publications **III** and **IV** aimed at providing practical R implementations as well as easy-to-use web-based user interfaces. The PSA modeling procedure was released as a web-application using the R Shiny platform (https://compbiomed.shinyapps.io/u-pa/), with a focus on the novelty aspect of predicting BCR using u-PSA. The anonymized PSA measurements are fully disclosed as an example dataset, but the user may also easily upload their own Excel, tab-delimited or similar format data to produce a template-based PDF report of predicted PSA-based trends.

Similarly, the R-code produced during the DREAM challenge resulting in publication **IV** has been made available on the Sage Bionetworks' website (https://www.synapse.org/ProstateCancerChallenge). This code requires the user to adhere to the data sharing policies of the PDS and Sage Bionetworks / Synapse before it can be downloaded. As per the DREAM rules, this code resulting in a top-performing model was manually inspected and re-run successfully by the DREAM Challenge organizers. Since its initial release, the ePCR methodology has been further developed and applied to new datasets. The R package of the method is available on CRAN (https://CRAN.R-project.org/package=ePCR ; Accessed 21st December 2017) and its manuscript has recently been accepted for publication (Laajala et al. 2018).

5. DISCUSSION

5.1. Preclinical considerations

The improved preclinical experimental design procedure developed in this thesis addresses the desire to implement clinical design practices in preclinical studies (Muhlhausler et al. 2013). This approach resulted in the stochastic allocation procedure that allows masked balancing and the use of multiple baseline covariates or strata. The developed matched methodology increased statistical power (Figure 17), and the simulation-based testing of the benefits of matching-based design showed that incorporating baseline information is highly useful (Publication **I**: Supplementary Figures S9 and S10). The advantages of patient matching are largely known and acknowledged in clinical trial experimental design (Greevy et al. 2004), but obtaining empirical preclinical evidence supporting the use of the developed methodology in this specific application would require repeated experiments comparing conventional methods with a matching-based approach, which we consider unethical because the matching-based approach is likely to lead to better treatment effect estimation with fewer animals. Overall, the proposed modeling approaches and the resulting experimental designs were tested by simulating and investigating potential pitfalls in these choices, and the methodologies were consistently as good as better than their conventional counterparts.

One shortcoming in the preclinical studies and methodology presented in publications **I** and **II** is that informative censoring was not systematically explored. Although it was acknowledged and incorporated into the underlying decision making, extended simulation studies coupled with true *in vivo* studies (e.g., by introducing artificial missingness) could provide further information regarding the potential pitfalls and relative strengths and weaknesses of each approach. Regression-based modeling that is based on growth coefficients should be robust in this aspect, but the upper-censoring due to moribund animals varied significantly over the experiments, and some of the experiments presented with a high quantity of lower-censoring of tumor burden or surrogate values lower than the detection limit. Research has been conducted through the simulation of the effects of lower-censoring due to challenges in measuring small tumors (Pierrillas et al. 2016a) and upper-censoring due to sacrifice or large tumors that may be necrotic (Pierrillas et al. 2016b). However, this challenging aspect remains an interesting venue for further research.

We propose a higher N in Publication **II** than what is conventionally used in many preclinical experiments. Although this suggestion may seem to directly contradict the 3Rs principle, we considered sufficient statistical power to be more important than minimal N because the latter may lead to a lack of generalization based on underpowered studies and thus the waste of animal lives. There are multiple reasons why one may consider it unethical to conduct underpowered intervention studies. The importance of sufficient statistical power was emphasized both in **I** and **II** and has also been discussed widely by preclinical experimenters who worry about underpowered studies that attempt to lower expenses while failing to deliver robust results. After performing the explorative power analysis in **II**, the experiments presented in **I** were conducted at a later date with a relatively high N. Subsequent to our original N counts suggested in **II**, multiple preclinical experts have highlighted the importance of power analyses and experiments conducted with sufficient statistical power (Couzin-Frankel 2013; Day et al. 2015; Macleod et al. 2015).

The presence of latent growth subgroups has been largely reported in an *ad hoc* manner in preclinical cancer studies (Enmon et al. 2003; Bedogni et al. 2006; Gutman et al. 1999; Saarinen et al. 2006; Saarinen et al. 2002; Galaup et al. 2003; Ribonson et al. 1987), whereas the proposed EM algorithm modeling identifies the subgroups without subjective decision making. The underlying assumption of these latent subgroups may appear strong, but the binary growth categories could be connected to either verifying findings such as PSA, tumor differentiation characteristics, or novel insight into estrogen receptor expression (Table 3). Furthermore, these verifying findings were identified regardless of the intervention arms, suggesting that the latent growth groups were inherently present in the experiments and were not caused by interventions. In summary, the underlying MEM assumptions in publications **I** and **II** were inspired by the literature in the field, and the models performed well both under real and simulated settings. Given the emerging trends of increasingly diverse features at baseline (i.e. varying genetic traits) in PDX and GEM preclinical models, the developed framework has potential to generalize also to future applications. As a practical example, distance metrics are widely used in 'omics analysis for clustering. In a similar manner, various other established clinical applications can therefore be back-propagated to refine the currently presented preclinical methodology, if such complex approaches become conventions also in preclinical research.

5.2. Clinical considerations

In the PSA study (**III**), we used an external validation dataset that was set aside prior to any modeling, and an independent researcher conducted the validation. Despite this approach, the ROC-AUCs were surprisingly high (Figure 20f; >0.9 ROC-AUC in 3-year time window, >0.8 ROC-AUC in 1-year time window). One might argue that these estimates are overly optimistic, but it should also be noted that a high ROC-AUC is to be expected from u-PSA predictions of BCR. This high ROC-AUC occurs because BCR is defined according to a clinically adopted threshold or subsequent rising PSA values (Cornford 2017), and thus the predictors are connected to the primary outcome by definition. However, because the majority of PSA measurements were in the u-PSA range (Table 1 in **III**) and the definition of BCR is relatively high in the t-PSA range, the study suggests that u-PSA has potential for accurate predictions for PCa-related events after RP. Future BCR modeling would benefit from extending studies to multiple hospitals and from a systematic comparison of u-PSA related signals with the experiment setting specifically designed to address potential differences between u-PSA and t-PSA. The results here suggest that u-PSA already holds prognostic value in early clinical time points, but these results are not conclusive. Further studies, possibly incorporating other end-points beyond BCR, are warranted. In meta-reviews of PSA kinetics, the most commonly used end-points in PCa relapse have been determined using biopsies rather than PSA alone (Vickers et al. 2009). Thus, providing an independent relapse indicator that is not directly tied to PSA would increase the validity of the inference presented in publication **III**.

The accurate detection of the PSA nadir is important because the reliable estimation of nadir timing helps follow PSA kinetics for predicting PCa recurrence (Zhao et al. 2011). There is no general well-defined consensus regarding the definition of PSA nadir, and for the purposes of publication **III**, PSA nadir was defined as the point of lowest measured PSA within a 3-month period after RP. Some statistically sophisticated methods, such as the Bayesian turn-point regression, place emphasis on the accurate determination of the PSA nadir and the subsequent doubling time post-nadir (Zhao et al. 2011). Furthermore, it has been proposed that PSA kinetics such as time-to-nadir from ADT to the nadir may provide an interesting predictor for disease relapse (Skove et al. 2017). Therefore, further research into the role and modeling of the PSA nadir still presents an important opportunity.

The aspects of missing data and censored information were only briefly presented here for the clinical studies in **III** and **IV**. This aspect has such a drastic impact on downstream conclusions that it would warrant its own extensive research effort. In each step, however, the considerations of structured versus non-structured and random versus non-random missing observations or information were embedded into the modeling choices, and effort was made to utilize the best possible practices. Mixed-effects models present robust properties that are resilient against non-balanced settings in the random effects (i.e., some individuals missing more observations than others), and they do not force the user to perform a subjective imputation methodology. Especially in the applications of modeling time-dependent in publications **I – III**, the statistical inference is less sensitive to accumulating missing information in the latter time points because growth slopes are utilized instead of single time points. The low levels of PSA that exist, especially at the nadir, present censored information such that the true measurable quantity is below the detection limit even of ultrasensitive the assay. This was the case for PSA for some observations in publication **III**, although this issue was alleviated by the use of an ultrasensitive assay in this application. Linear models were proposed as a suitable model family candidate according to the explored parameter space with CV (Figure 13c). Although a slight jagged effect in the final MEM residuals could be observed due to this lower threshold (Figure 20e), this informative censoring was deemed to be a negligible trend due to the proposed models' ability to generalize well to the set-aside validation data (Figure 20f).

Due to the competitive nature of the DREAM Challenges, it is difficult to evaluate the effect of modeling choices occurring due to the competitive context and not due to the more interesting biomedical research question alone. The time span of a challenge typically consists of roughly half a year, which is divided into milestones. Thus, each challenge poses severe limitations to the possibility of entirely novel method development, but also effectively discourages exhaustive testing or overfitting the models. In practice, the leading participants will typically be scored and ranked according to a single ranking metric, and this obviously may lead to techniques that fine-tune the scoring metric to gain even a slight advantage over competitors. Although this situation may not have been the case for the ePCR methodology, which outperformed its competitors by a statistically highly significant margin, the context under which the methodology was developed may have affected the choices for the method itself. DREAM Challenges or similar crowd-sourced challenges may be prone to produce models optimized for the specific setup of the challenge. This may lead to overfitting to e.g., the special traits of the studies, the presentation of the data, or scoring metric, rather than the primary objectives of the research question *per se*. For example, the final resulting ePCR ensemble consisted of penalized/regularized Cox models in which the L_1/L_2-norm parameter α was heavily favoring RR-like EN (α close to 0). To be able to precisely produce the same exact prediction as the original model, all the original available variables for the coefficients should be available, even if the contribution of a majority of predictors is arbitrarily close to zero. Therefore, optimal prediction from a computational perspective may be suboptimal for clinical use because the number of measured assays (i.e., predictors) factors heavily into the applicability of the method.

As an analogy, one could consider how principal component analysis is widely used in bioinformatics; to accurately display all of the variation, one should present all of the principal axes, but in practice, the useful functionality of PCA results from being able to project the majority of the variation into a lower dimension representation of the data. In order to advance clinical applicability, follow-up work refining the ePCR methodology could be steered to favor more LASSO-like (α = 1) ensembles or even be complemented by additional penalization of certain hard-to-access features (i.e., weighting by the real-life cost of assays). This alternative approach was made possible because the method development has evolved independently from the initial setting of the DREAM Challenge and its scoring metric (iAUC), which was agnostic toward the number and nature of the proposed features for prediction.

As has been demonstrated by my colleagues at the University of Turku, the practical applications of a model developed for a cohort of well-standardized, controlled clinical trial data may perform very differently in real-world cohort data (Seyednasrollah et al. 2017b). Although the model developed in **IV** outperforms the Halabi benchmark in the real-world data (Halabi et al. 2014) in their study, the prediction accuracy of ePCR steeply declines after one year of diagnosis. This result raises the question of how to translate these models to real-world cohorts, where the standardization of data, the timing of end points, and patient practices vary drastically. To reflect clinical reality, it is highly unlikely that a single prediction will be utilized after a year from the corresponding lab measurements. Instead, it is likely that models that can be updated based on new data, would be much more applicable to real-world data. Due to the severity of the disease, patients are undergoing intense follow-up checks, providing regular new information regarding acute disease progression. In this respect, an updating real-time mathematical modeling approach would be more suitable.

To better treat mCRPC patients based on their expected survival, it is important to assess the specificity of the resulting ensemble-based model and its predictors of mCRPC. Although the DREAM organizers censored factors known to affect general survivability, such as reported cigarette smoking, the current model may include covariates that predict survivability that is nonspecific to mCRPC, because the training and validation datasets consisted solely of clinical trials of mCRPC in **IV**. To better identify these specific factors, the current approach could be extended by studies in the nonmetastatic version of CRPC, other cancer types, and healthy control populations. For this purpose, many open-data platforms offer a wide variety of well-standardized datasets, and many of the currently utilized clinical markers are available. Because the current modeling framework offers a state-of-the-art baseline that improves upon the benchmark of the Halabi model (Halabi et al. 2014) , a natural next step is to narrow down the covariates that have specific clinical impact on survivability prediction in mCRPC or to extend the ensemble model toward more general and clinically-orientated use.

The focus of mathematical modeling here was heavily based on clinical features, while an important future expansion would be genomics. Various levels of 'omics in cancer research, such as mRNA expression, copy number alterations, mutations, and epigenetics, are all highly relevant in characterizing cancer (Cancer Genome Atlas Research Network 2015). There are multiple challenges in effectively utilizing these platforms, especially when harmonizing over different 'omics; however, due to the ensemble nature of the ePCR methodology, one can create tailored model formulations that are suitable for each platform. Further challenges are posed by the high dimensionality of 'omics data. To this end, regularized regression has robust properties for modeling high dimensional sparse data in such applications (Dasgupta et al. 2011). An on-going collaborative effort has already been established to continue toward this venue, i.e. poster presentation in (Laajala et al., 2018).

6. CONCLUSIONS

This multidisciplinary thesis traverses the colliding worlds of applied mathematics, statistics, machine learning, and systems medicine-oriented oncological research and experimentation. It addresses many current issues in drug discovery and experimentation, the reproducibility of preclinical findings, and the reporting of corresponding experimental design. The presented work seeks to aid in this the acute search for better treatments for challenging malignancies with a strong focus on prostate cancer and its aggressive forms. The work is roughly divided into two main subcategories with chronologically intuitive dissemination. First, publications **I** and **II** focus on preclinical experimentation, starting from fundamental issues in the design of preclinical experiments (publication **I**) and then shifting the focus into post-intervention statistical testing of the treatment effects (portions of publication **I**, and the whole of publication **II**). Second, the clinical window spans from generic PSA screening and recurrence prediction in PCa in a real-world cohort (publication **III**) to comprehensive, ensemble-driven modeling of the late-stage aggressive mCRPC form of PCa in a large cooperative initiative to connect open science and numerous sources of clinical trials (publication **IV**). It should also be noted that the presented publications follow the pattern of drug discovery that progresses from rodent *in vivo* experimentation (**I** and **II**) to clinical applications with patient data (**III** and **IV**). Although the *in vitro* phase of drug discovery was not explicitly considered here, it is still implicitly present in the experimental design choices for experiments in publications **I** and **II** because the cell lines were chosen to optimally represent the human disease, disease progression, and mechanisms under investigation.

The recently discussed issues regarding the reporting, reproducibility, and design of preclinical experiments were addressed in publication **I**. An improved comprehensive framework for designing preclinical experiments was proposed, spanning from controlling inter-individual variation to power calculations that justify the ethical and practical undertaking of future experiments. For preclinical statistical inference of intervention testing, three main categories of model formulations are offered for linear MEMs in time-dependent tumor growth: i) a conventional longitudinal model, which is trivial to fine-tune (publications **I** and **II**); ii) a matching-based paired response longitudinal model that makes use of baseline characteristics (publication **I**); iii) and a latent variable longitudinal mixed-effects model that utilizes the EM algorithm (publication **II**). The conclusions drawn from using the novel statistical inference in these preclinical experiments were concordant with the known literature regarding orchiectomy in the development of castration resistance in VCaP cells and the currently FDA-approved clinical use of anti-androgens. The retrospective analysis of several publications verified their original conclusions and provided novel insight into receptor-related activity regarding the proposed latent growth parameter. Further biology-related insights were derived in the side-by-side published articles focusing on, for example, the intratumoral neo-biosynthesis of androgen and its significance in developing castration resistance (Knuuttila et al. 2014; Huhtaniemi et al. submitted).

The current practices of PSA screening in Finland and the prediction of BCR were examined with an emphasis on the practicality of ultrasensitive assays and the possibility of earlier PCa recurrence detection (publication **III**). The ultrasensitive range of PSA measurements were found to contain interesting informative trends, especially in PSA kinetics, but this particular subfield of PSA research would benefit from further multicohort verification. After examining PCA kinetics in $\log_2$-transformed responses using splines, it was established that linear models could reliably capture the main trends, and generalized linear MEMs were constructed for predicting BCR risk. After studying BCR and the development of hormone-therapy resistance, patient survival in advanced forms of PCa was modeled in the prominent DREAM 9.5 mCRPC Challenge (publication **IV**). Novel biomarkers for OS prediction in mCRPC were identified in the meta-analysis of the challenge, which produced interesting insights from the machine learning methodology conducted by the groups from over 50 universities around the globe. The top-performing ensemble-based Cox regression methodology was published together with an emphasis on the network analysis of possible interactions for clinical predictors in Publication **IV**.

True to the spirit of open science, this compilation of scientific publications adheres to the pillars of the scientific method, including reproducibility and transparency. This aspect spans from reproducible code and open data embedded in R packages to the provision of extensive, accessible, and easily applicable novel methods intended to improve the translatability and interpretability of *in vivo* research. The published studies aim to set an example for confronting issues that have plagued the transition from the preclinical phase to the clinical phase. Although the targeted audience focuses on oncology, we have embarked on a journey to promote open science, data, and corresponding methods that extends beyond cancer, as discussed in the supporting editorials (Laajala et al. 2017; Sartor et al. 2017). While basking in the dawning light of a promising era of vast amounts of generated data (so-called big data challenge), international multidisciplinary collaboration, and accumulating scientific research (crowd-sourced research approach), one must always return to the fundamentals of the scientific method to verify that mankind's accumulating knowledge converges toward truth. To this idealistic end, which has been a driving motivator behind the conducted research, this thesis has contributed in the focused field of hormonally regulated cancer with a focus on PCa. Although the work is far from finished, the presented stepping stones now await the footprints of eager new seekers of knowledge.

ACKNOWLEDGEMENTS

I am thankful to the abundance of people who have affected me personally and my work during my PhD studies. Naturally, without the guidance of the three extremely gifted and dedicated supervisors, I would not be the young sprout of a scientist that I am today. I will elaborate my personal gratitude in chronological order in which I had the privilege to get acquainted with you:

Laura, thank you for first taking me in to conduct my BSc a decade ago in ChIP-seq data analysis when I was incredibly wet behind the ears. You've later integrated me into your group and offered me numerous interesting projects and opportunities that have widened my scope as a data scientist. You led me to my first door, and introduced to me just how deep the rabbit hole of science is – and it has truly been a wondrous journey thanks to your initiative and support.

Tero, thank you for subsequently adopting me into doing my MSc in preclinical research and leading me on the path to becoming an independent scientist. I know I haven't always been the easiest PhD student, but you've patiently suffered through my quirks and helped turning my weak points into strengths. You were wise and far-sighted in building me up, when I initially focused on a narrow field, and it is much due to your talent in observing beyond the obvious that has brought me here.

Matti, thank you for always believing in me and praising me up to the point of blushing, even for trivial things. You've always made me feel as if I was among or if not the most important person in a room full of professors, MDs, PhDs and other people upon whose giants' shoulders I've learned to build science. You've offered me an incredible network of collaboration over numerous projects, and I feel you justify bringing the word "applied" to my overall title as an applied mathematician.

I would like to express gratitude to the parties who funded the thesis work as well as supporting projects during my PhD: DRDP, the Finnish Cultural Foundation, Sigrid Juselius Foundation, Academy of Finland, as well as project funding from the Aittokallio-, Elo-, Poutanen-, and Westermarck-groups.

From the Department of Mathematics and Statistics at the University of Turku (UTU), I am especially grateful to my PhD-director Prof. Marko Mäkelä, whose door has always been open to me and who never fails to impress me with his optimism coupled with being top of his field. Further, without the assistance of Prof. Iiro Honkala as the Head of the Department, this dissertation would not have been possible. Numerous other people at the department have supported and encouraged me over the years coupled with extracurricular activities. To name just a few of the many, I would like to enumerate those who I shared room with or spent board game evenings with (in alphabetic order): Ilmari Ahonen, Anni Halkola, Bogdan Iancu, Maiju Pesonen, Sean Robinson, Markus Matilainen, and Joni Virta. There are numerous other groups in UTU to whom I am deeply grateful for their professional contribution to my scientific research as well as personal support especially during the early phases of my career. Naturally, Elo-group at the Turku Centre for Biotechnology (BTK) has offered me a crucial anchor, and I have been present to see the growth of the group all the way from when it was first established. Since the members are today too numerous to enumerate, I would like emphasize those who I have had the privilege to get to know also outside work and would thus like to extend my deepest gratitude to Fatemeh Seyednasrollah and Mehrad Mahmoudian, for whom I thankful that I may consider them also friends rather than just colleagues. You are both extremely talented scientists and admirable personalities, and I am sure that your talent and intuition will take you far. Furthermore, I am grateful

to have shared a journey in academia with Deepankar Chakroborty, Thomas Faux, Ye Hong, Maria Jaakkola, Riku Klén, Sohrab Saraei, Arfa Mehmood, and Aidan McGlinchey at BTK. I have also had the chance to work in close collaboration with Prof. Jukka Westermarck and his group at BTK. I am especially thankful to Otto Kauko and Amanpreet Kaur from the Westermarck-group, for all the support and interesting sessions as peer-PhD students. Simultaneously at BTK, thanks to Anni Vehmas and Prof. Garry Corthals I am no longer a stranger to the intriguing world of proteomics. Furthermore, thanks to Prof. Riitta Lahesmaa and her group, I was given the chance to the take my first steps in science starting with BSc. Multiple senior scientists have been kind to share their wisdom and experiences over the years. Anna Lipsanen and Prof. Jaakko Nevalainen from my PhD committee have been kind enough to give me insightful perspective to conducting my PhD. Prof. Harri Lähdesmäki and Prof. Jukka Corander from Aalto University and University of Helsinki, respectively, were especially helpful in the beginning of my PhD studies. Furthermore, I have always been very interested in exploring machine learning, and for this reason I am very grateful to Prof. Tapio Pahikkala, Antti Airola, and Mika Murtojärvi from the IT-Department at UTU who have been kind enough to extensively share their ideas and knowledge. Furthermore, I have had the privilege to work with the Turku Center for Disease Modeling and the Department of Physiology at the Faculty of Medicine; especially Vidal Fey, Michael Gabriel, Kaisa Huhtinen, Matias Knuuttila, Taija Saloniemi-Heinonen, and Emrah Yatkin.

I have been offered the chance to work with the private sector as well as be on the front in exploring new frontiers in many fields such as biobanks; from these, I'd like to extend my gratitude to Esa Alhoniemi, H-P Schukov, Samu Kurki, Perttu Terho, and Arho Virkki. After the DREAM Challenge, I was given the opportunity of working closely with senior scientists from abroad, namely Prof. James C. Costello (University of Colorado). Furthermore, during the Challenge I was privileged to share thoughts with the talented senior scientists at FIMM: Suleiman Khan, Gopal Peddinti, and Tuomas Mirtti. I haven't forgotten our gold-lined BSc/MSc days at the Aalto-university in Espoo near Meilahti and FIMM, for which I am in especially thankful to Olli Kotiranta and Marja Pitkänen, and Maria Osmala.

I have had the positive experience of being in the strange world of drug discovery albeit having a background in bioinformatics. Much of this is owed to the fact that the Drug Research Doctoral Programme (DRDP) took me in, and I have since been in debt to coordinator Eeva Valve, as well as the whole of DRDP steering committee members, especially Prof. Markku Koulu, Prof. Mika Scheinin, and Prof. Sari Mäkelä. This list would additionally include personnel already mentioned elsewhere herein. It is much thanks to DRDP's broad perspective in understanding the multidisciplinary nature of today's oncological research that I have found such an educative and fruitful environment. I always felt that the most rewarding parts of my work have been when I have been able to work directly with those who affect the lives of patients undergoing various diseases through clinical procedures or decision-making, and for this reason I'd like to highlight a few MDs/PhDs/Profs., to whom I'm grateful for offering this particular opportunity (in alphabetic order): Peter Boström, Otto Ettala, Jutta Huvila, Panu Jaakkola, Kalle Mattila, Antti Perheentupa, Heikki Seikkula, and Pia Suvitie.

I have been granted the privilege on multiple occasions to infiltrate the social bubbles of many people where one transcends the borderline of friendship, time spent together with mundane things and professional collaboration. Out of these people, I would especially like to thank Marja Heiskanen (extended to Jani), for numerous deep conversations as well as for being kind enough to play-test my board games, and Anna Pursiheimo for much needed extended moments by coffee when life has not exactly been like in Strömsö. Many old contacts from Ristin kilta at Aalto University campus have been there for me when my inner Mr. Hyde of a computer gamer wakes up or when I've needed the acute peer-review to various matters in life; I cannot over-emphasize how important this has been for me to stay on my feet and keep a healthy perspective on life. Amongst many others, I am thankful in particular to (in alphabetic nickname-order): Seppo "Ange" Tiilikainen, Antti "ankka" Mäkilä, Mikko "cpsof95" Luttinen, Ville "Drendil" Toivonen, Janne "Hoppeli" Wallenius, Johannes "mousefly" Haataja, Pekka "Nightpanda" Tiilikainen, Hannu "sebdul" Hartikainen, Otto-Ville "Tikru8" Sormunen, Tommi "Tspoon" Larjomaa, and Mikael "Zartek" Jumppanen. For spiritual nurturing, I am grateful to Calvary Chapel Turku and Varikko. Many friends have been kind to me riding along in life's highs and lows; old faithful friends from Salo, as well as the Eirola, Förbom, Helin, Oja and Perttula families.

I have always been lucky that I could count on the support of my family. Dad (extended to Leena Kluusteri), thank you for showing me through deeds what it means to be a humble, honest and hard-working man. I wish the world had more people like you, who persistently fight off tough odds and overcome, despite not even taking credit where it would be due. Such acts but may go silently unnoticed by the majority, but alter the course of our world. Mom (extended to Pekka Ahonen), thank you for being the embodiment of empathy. You are the light in the darkness for many, and have shown me the importance of being a human to my fellow neighbors. Had it not been for your motherly patience, love and wisdom, cynicism might be today my overwhelming trait. Sini (extended to Erkki), thank you for being the best sister in world. You are truly my kin, and I admire you and what you do. I want to see you grow happy and be in the life of my family ever-more. This also includes my in-laws, their spouses and children – you have welcomed me with a genuine warmth.

Last but not least, I am thankful to my immediate family. Essi, you've been my best friend for a decade now ever since we met on the first day at the Aalto University, and you have since become the wife of my dreams as well as the mother-lioness to my two daughters – the sort of a woman I feel I am not deserving of. Thank you for walking hand-in-hand with me through the thick and thin. Words are beyond capability of expressing the magnitude of feelings I have for you three. Elea and Vilja, you two little adorable rascals - you are eternally part of me, as I am part of you. No matter what happens, you'll always be my much desired and loved offspring, and if I can affect this stream of life, there will always be room for two (or more) long and warm father's embraces and aid for you. Maybe you will follow in my footsteps one day, or even if you don't I wish to be there to aid you in side-stepping some of the traps and sinkholes that life inevitably besets before you.

This thesis' dedication extends to the memory of Kati Aarniala, a dear sister-in-battle against the darkest pits of the human mind, who lost her battle during the course of my PhD. I firmly believe that mankind needs science in order to advance to a brighter future, but that science also needs humanity in order to function properly and to truly achieve its idealistic goals. I hope that at least for some of you I have gone beyond the implicit social constructions under which a colleague or a friend might be assumed to support you. Kati, should my musings somehow reach some conscious part of you somewhere, I wish that you have obtained the much deserved peace of mind that this realm ultimately could not offer you. I will wander yonder in search of mine.

In Turku, 10.10.2018

Teemu Daniel Laajala

REFERENCES

Albala DM. *Imaging and treatment recommendations in patients with castrate-resistant prostate cancer.* Rev Urol. 2017;19(3):200-202.

Amato RJ, Teh BS, Henary H, Khan M, Saxena S. *A retrospective review of combination chemohormonal therapy as initial treatment for locally advanced or metastatic adenocarcinoma of the prostate*. Urol Oncol. 2009 Mar-Apr;27(2):165-9. doi: 10.1016/j.urolonc.2007.12.004

Baker D, Lidster K, Sottomayor A, Amor S. *Two years later: journals are not yet enforcing the ARRIVE guidelines on reporting standards for pre-clinical animal studies*. PLoS Biol. 2014 Jan;12(1):e1001756. doi: 10.1371/journal.pbio.1001756

Barry MJ. *Clinical practice. Prostate-specific-antigen testing for early diagnosis of prostate cancer*. N Engl J Med. 2001 May 3;344(18):1373-7. doi: 10.1056/NEJM200105033441806

Bates D. et al. *lme4: Linear mixed-effects models using Eigen and S4*. R-package version 1.1–7. URL: http://cran.r-project.org/package=lme4 (Accessed: 21st June 2016).

Bedogni B, Welford SM, Kwan AC, Ranger-Moore J, Saboda K, et al. *Inhibition of phosphatidylinositol-3-kinase and mitogen-activated protein kinase 1/2 prevents melanoma development and promotes melanoma regression in the transgenic TPRas mouse model*. Mol Cancer Ther. 2006 Dec;5(12):3071-7. doi: 10.1158/1535-7163.MCT-06-0269

Begley CG, Ellis LM. *Drug development: Raise standards for preclinical cancer research*. Nature. 2012 Mar 28;483(7391):531-3. doi: 10.1038/483531a

Bender E. *Challenges: Crowdsourced solutions*. Nature. 2016 May 12;533(7602):S62-4. doi: 10.1038/533S62a

Bhatia S, Frangioni JV, Hoffman RM, Iafrate AJ, Polyak K. *The challenges posed by cancer heterogeneity*. Nat Biotechnol. 2012 Jul 10;30(7):604-10. doi: 10.1038/nbt.2294

Breslin S, O'Driscoll L. *Three-dimensional cell culture: the missing link in drug discovery*. Drug Discov Today. 2013 Mar;18(5-6):240-9. doi: 10.1016/j.drudis.2012.10.003

Cancer Genome Atlas Research Network. *The Molecular Taxonomy of Primary Prostate Cancer*. Cell. 2015 Nov 5;163(4):1011-25. doi: 10.1016/j.cell.2015.10.025

Collins FS, Tabak LA. *Policy: NIH plans to enhance reproducibility*. Nature. 2014 Jan 30;505(7485):612-3.

Clausen J. *Branch and bound algorithms-principles and examples*. Department of Computer Science, University of Copenhagen. 1999: 1-30. URL: https://imada.sdu.dk/~jbj/DM85/TSPtext.pdf (Accessed: 21st June 2016).

Couzin-Frankel J. *When mice mislead*. Science. 2013 Nov 22;342(6161):922-3, 925. doi: 10.1126/science.342.6161.922

Costello JC, Stolovitzky G: *Seeking the wisdom of crowds through challenge-based competitions in biomedical research*. Clin Pharmacol Ther. 2013 May;93(5):396-8. doi: 10.1038/clpt.2013.36

Cornford P, Bellmunt J, Bolla M, Briers E, De Santis M, et al. *EAU-ESTRO-SIOG Guidelines on Prostate Cancer. Part II: Treatment of Relapsing, Metastatic, and Castration-Resistant Prostate Cancer*. Eur Urol. 2017 Apr;71(4):630-642. doi: 10.1016/j.eururo.2016.08.002

Crawford ED, Petrylak D. *Castration-resistant prostate cancer: descriptive yet pejorative?* J Clin Oncol. 2010 Aug 10;28(23):e408. doi: 10.1200/JCO.2010.28.7664

Cunningham D, Zongbing Y. *In vitro and in vivo model systems used in prostate cancer research*. J Biol Methods. 2015; 2(1): doi:10.14440/jbm.2015.63

Day CP, Merlino G, Van Dyke T. *Preclinical mouse cancer models: a maze of opportunities and challenges*. Cell. 2015 Sep 24;163(1):39-53. doi: 10.1016/j.cell.2015.08.068

Dasgupta A, Sun YV, König IR, Bailey-Wilson JE, Malley JD. *Brief review of regression-based and machine learning methods in genetic epidemiology: the Genetic Analysis Workshop 17 experience*. Genet Epidemiol. 2011;35 Suppl 1:S5-11. doi: 10.1002/gepi.20642

Davis ID. *Challenges of data sharing: valuable but costly?* Lancet Oncol. 2017 Jan;18(1):15-16. doi: 10.1016/S1470-2045(16)30564-2

De Jong K. *Learning with Genetic Algorithms: An Overview*. Machine Learning. 1988;3:121-138.

Dempster AP, Laird NM, Rubin DB. (1977*). Maximum likelihood from incomplete data via the EM algorithm*. Journal of the royal statistical society. Series B (methodological):1-38.

Eisen JA, Ganley E, MacCallum CJ. *Open science and reporting animal studies: who's accountable?* PLoS Biol. 2014 Jan;12(1):e1001757. doi: 10.1371/journal.pbio.1001757

Enmon R, Yang WH, Ballangrud AM, Solit DB, Heller G, et al. *Combination treatment with 17-N-allylamino-17-demethoxy geldanamycin and acute irradiation produces supra-additive growth suppression in human prostate carcinoma spheroids*. Cancer Res. 2003 Dec 1;63(23):8393-9.

Epstein JI, Walsh PC, Carmichael M, Brendler CB. *Pathologic and clinical findings to predict tumor extent of nonpalpable (stage T1c) prostate cancer*. JAMA. 1994 Feb 2;271(5):368-74. doi: 10.1001/jama.1994.03510290050036

Ferlay J, Steliarova-Foucher E, Lortet-Tieulent J, Rosso S, Coebergh JW, et al. *Cancer incidence and mortality patterns in Europe: estimates for 40 countries in 2012*. Eur J Cancer. 2013 Apr;49(6):1374-403. doi: 10.1016/j.ejca.2012.12.027

Ferguson RA, Yu H, Kalyvas M, Zammit S, Diamandis EP. *Ultrasensitive detection of prostate-specific antigen by a timeresolved immunofluorometric assay and the Immulite immunochemiluminescent third-generation assay: potential applications in prostate and breast cancers*. Clin Chem. 1996 May;42(5):675-84.

Finnish Cancer Registry. URL: https://cancerregistry.fi/statistics/ (Accessed: 28th March 2018)

Fisher R, Pusztai L, Swanton C. *Cancer heterogeneity: implications for targeted therapeutics*. Br J Cancer. 2013 Feb 19;108(3):479-85. doi: 10.1038/bjc.2012.581

Fizazi K, Higano CS, Nelson JB, Gleave M, Miller K, et al. *Phase III, randomized, placebo-controlled study of docetaxel in combination with zibotentan in patients with metastatic castration-resistant prostate cancer*. J Clin Oncol. 2013 May 10;31(14):1740-7. doi: 10.1200/JCO.2012.46.4149

Freedman LP, Cockburn IM, Simcoe, TS. *The economics of reproducibility in preclinical research*. PLoS Biol. 2015 Jun 9;13(6):e1002165. doi: 10.1371/journal.pbio.1002165

Friedman J, Hastie T, Simon N, Tibshirani R. *glmnet: Lasso and elastic-net regularized generalized linear models*. R-package version 2.0–5. http://cran.r-project.org/package=glmnet (Accessed: 8th February 2016).

Friedman J, Hastie T, Tibshirani R. *Regularization Paths for Generalized Linear Models via Coordinate Descent*. J Stat Softw. 2010;33(1):1-22.

Gábor C, Tamás N. *The igraph software package for complex network research*. InterJournal Complex Systems. 2006;1695.

Galaup A, Opolon P, Bouquet C, Li H, Opolon D, Bissery MC, et al. *Combined effects of docetaxel and angiostatin gene therapy in prostate tumor model*. Mol Ther. 2003 Jun;7(6):731-40. doi: 10.1016/S1525-0016(03)00121-7

Gelman A, Hill J. In: Alvarez RM, Beck NL, Wu LL editors. *Data analysis using regression and multilevel/hierarchical models*. New York: Cambridge University Press; 2007.

Goeman J, Meijer R, Chaturvedi N. *penalized: L1 and L2 Penalized Regression Models*. R-package version 0.9-47 Vignette, Chapter 6: "A note on standard errors and confidence intervals". URL: https://cran.r-project.org/web/packages/penalized/vignettes/penalized.pdf (Accessed: 1st June 2016).

Gower JC. *A general coefficient of similarity and some of its properties*. Biometrics. 1971;27:857–871.

Grabowska MM, DeGraff DJ, Yu X, Jin RJ, Chen Z, et al. *Mouse models of prostate cancer: picking the best model for the question*. Cancer Metastasis Rev. 2014 Sep;33(2-3):377-97. doi: 10.1007/s10555-013-9487-8

Greevy R, Lu B, Silver JH, Rosenbaum P. *Optimal multivariate matching before randomization*. Biostatistics. 2004 Apr;5(2):263-75. doi: 10.1093/biostatistics/5.2.263

Gillessen S, Attard G, Beer TM, Beltran H, Bossi A, et al. *Management of Patients with Advanced Prostate Cancer: The Report of the Advanced Prostate Cancer Consensus Conference APCCC 2017*. Eur Urol. 2018 Feb;73(2):178-211. doi: 10.1016/j.eururo.2017.06.002

Gutman M, Couillard S, Labrie F, Candas B, Labrie C. *Effects of the antiestrogen EM-800 (SCH 57050) and cyclophosphamide alone and in combination on growth of human ZR-75-1 breast cancer xenografts in nude mice*. Cancer Res. 1999 Oct 15;59(20):5176-80.

Halabi S, Lin C-Y, Kelly WK, Fizazi KS, Moul JW, et al. *Updated prognostic model for predicting overall survival in first-line chemotherapy for patients with metastatic castration-resistant prostate cancer*. J Clin Oncol. 2014 Mar 1;32(7):671-7. doi: 10.1200/JCO.2013.52.3696

Hartman J, Lindberg K, Morani A, Inzunza J, Ström A, et al. *Estrogen receptor beta inhibits angiogenesis and growth of T47D breast cancer xenografts*. Cancer Res. 2006 Dec 1;66(23):11207-13. doi: 10.1158/0008-5472.CAN-06-0017

Hastie T, Tibshirani R, Friedman J. Springer Series in Statistics Springer New York Inc., New York, NY, USA: *The Elements of Statistical Learning* (2001). Chapter 2.9: *Model Selection and the Bias–Variance Tradeoff*

Hasty AH, Gutierrez DA. *What have we really learned about macrophage recruitment to adipose tissue?* Endocrinology. 2014 Jan;155(1):12-4. doi: 10.1210/en.2013-2027

Heitjan DF. *Biology, models, and the analysis of tumor xenograft experiments*. Clin Cancer Res 2011;17:949–52. doi: 10.1158/1078-0432.CCR-10-3279

Henderson VC, Demko N, Hakala A, MacKinnon N, Federico CA, et al. *A meta-analysis of threats to valid clinical inference in preclinical research of sunitinib*. Elife. 2015 Oct 13;4:e08351. doi: 10.7554/eLife.08351

Hernández J, Thompson IM. *Prostate-specific antigen: a review of the validation of the most commonly used cancer biomarker*. Cancer. 2004 Sep 1;101(5):894-904. doi: 10.1002/cncr.20480

Hoenig JM, Heisey DM. *The abuse of power: the pervasive fallacy of power calculations for data analysis*. Am Stat. 2001;55:19–24. doi: 10.1198/000313001300339897

Hildebrand F, Nguyen TL, Brinkman B, Yunta RG, Cauwe B, et al. *Inflammation-associated enterotypes, host genotype, cage and inter-individual effects drive gut microbiota variation in common laboratory mice.* Genome Biol. 2013 Jan 24;14(1):R4. doi: 10.1186/gb-2013-14-1-r4

Hutchinson L, Kirk R. *High drug attrition rates—where are we going wrong?* Nat Rev Clin Oncol. 2011 Mar 30;8(4):189-90. doi: 10.1038/nrclinonc.2011.34

Irshad S, Abate-Shen C. *Modeling prostate cancer in mice: something old, something new, something premalignant, something metastatic*. Cancer Metastasis Rev. 2013 Jun;32(1-2):109-22. doi: 10.1007/s10555-012-9409-1

Kasturi J, Geisler JG, Liu J, Kirchner T, Amaratunga D, et al. *IRINI: random group allocation of multiple prognostic factors.* Contemp Clin Trials. 2011 May;32(3):372-81. doi: 10.1016/j.cct.2010.12.009

Kilkenny C, Browne WJ, Cuthill IC, Emerson M, Altman DG. *Improving Bioscience Research Reporting: The ARRIVE Guidelines for Reporting Animal Research*. PLoS Biol. 2010;8(6): e1000412. doi:10.1371/journal.pbio.1000412

Kimmelman J, Mogil JS, Dirnagl U. *Distinguishing between exploratory and confirmatory preclinical research will improve translation*. PLoS Biol. 2014 May 20;12(5):e1001863. doi: 10.1371/journal.pbio.1001863

Kirby M, Hirst C, Crawford ED. Characterising the castration-resistant prostate cancer population: a systematic review. Int J Clin Pract. 2011 Nov;65(11):1180-92. doi: 10.1111/j.1742-1241.2011.02799.x

Kleinman K, Huang SS. *Calculating Power by Bootstrap, with an Application to Cluster-Randomized Trials*. EGEMS (Wash DC). 2017 Feb 9;4(1):1202. doi: 10.13063/2327-9214.1202

Knuuttila M, Yatkin E, Kallio J, Savolainen S, Laajala TD, et al. *Castration induces up-regulation of intratumoral androgen biosynthesis and androgen receptor expression in an orthotopic VCaP human prostate cancer xenograft model.* Am J Pathol. 2014 Aug;184(8):2163-73. doi: 10.1016/j.ajpath.2014.04.010

Laajala TD, Guinney J, Costello JC. *Community mining of open clinical trial data*. Oncotarget. 2017 Sep 13;8(47):81721-81722. doi: 10.18632/oncotarget.20853

Laajala TD, Murtojärvi M, Virkki A, Aittokallio T. *ePCR: an R-package for survival and time-to-event prediction in advanced prostate cancer, applied to real-world patient cohorts*. Bioinformatics. 2018 Jun 15. doi: 10.1093/bioinformatics/bty477

Laajala TD, Aden-Buie G, Gerke T, Creed J, Berglund A, et al. *Identifying genetic interactions that drive aggressive prostate cancer using an ensemble of penalized cox regression models*. Accepted abstract to be presented as a poster at the "*30th Anniversary AACR Special Conference Convergence: Artificial Intelligence, Big Data, and Prediction in Cancer*" at Newport (RI, US) on October 15th, 2018. (Poster PDF available upon request)

Landis SC, Amara SG, Asadullah K, Austin CP, Blumenstein R, et al. *A call for transparent reporting to optimize the predictive value of preclinical research*. Nature. 2012 Oct 11;490(7419):187-91. doi: 10.1038/nature11556

Lavine M, Schervish MJ. *Bayes Factors: What they are and what they are not*. Am Stat. 1999;53(2):119-122. doi: 10.2307/2685729

Lilja H. *A kallikrein-like serine protease in prostatic fluid cleaves the predominant seminal vesicle protein*. J Clin Invest. 1985 Nov;76(5):1899-903.

Lilja H, Ulmert D, Vickers AJ. *Prostate-specific antigen and prostate cancer: prediction, detection and monitoring*. Nat Rev Cancer. 2008 Apr;8(4):268-78. doi: 10.1038/nrc2351

Lockhart R, Taylor J, Tibshirani RJ, Tibshirani R. *A significance test for the lasso*. Ann Stat. 2014 Apr;42(2):413-468. doi: 10.1214/13-AOS1175

Lu B, Greevy R, Xu X, Beck C. *Optimal Nonbipartite Matching and Its Statistical Applications*. Am Stat. 2011;65(1):21-30. doi: 10.1198/tast.2011.08294

Nelson JB, Fizazi K, Miller K, Higano CS, Moul JW, et al. *Phase III study of the efficacy and safety of zibotentan (ZD4054) in patients with bone metastatic castration-resistant prostate cancer (CRPC)*. Proc Am Soc Clin Oncol 2011;29: abstr 117.

Macleod MR, Lawson McLean A, Kyriakopoulou A, Serghiou S, de Wilde A, et al. *Risk of Bias in Reports of In Vivo Research: A Focus for Improvement*. PLoS Biol. 2015 Oct 13;13(10):e1002273. doi: 10.1371/journal.pbio.1002273

Misra, K. *Guidelines Evolving for the Treatment of Bone Metastases in Castration-Resistant Prostate Cancer*. Targeted Oncology, 2015. URL: http://www.targetedonc.com/publications/special-reports/2015/prostate-issue1/guidelines-evolving-for-the-treatment-of-bone-metastases-in-castration-resistant-prostate-cancer (Accessed: 3rd January 2018)

Muhlhausler BS, Bloomfield FH, Gillman MW. *Whole animal experiments should be more like human randomized controlled trials*. PLoS Biol. 2013;11(2):e1001481. doi: 10.1371/journal.pbio.1001481

Oesterling JE. *Prostate specific antigen: a critical assessment of the most useful tumor marker for adenocarcinoma of the prostate*. J Urol. 1991 May;145(5):907-23. doi: 10.1016/S0022-5347(17)38491-4

Oken M, Creech R, Tormey D, Horton J, Davis TE, et al. *Toxicity and response criteria of the Eastern Cooperative Oncology Group*. Am J Clin Oncol. 1982;5:649-655.

Omurlu IK, Ture M, Tokatli F*. The comparisons of random survival forests and Cox regression analysis with simulation and an application related to breast cancer*. Expert Syst Appl. 2009;36(4):8582-8. doi: 10.1016/j.eswa.2008.10.023

Perrin S. *Preclinical research: make mouse studies work.* Nature. 2014 Mar 27;507(7493):423-5. doi: 10.1038/507423a

Petrylak DP, Vogelzang NJ, Budnik N, Wiechno PJ, Sternberg CN, et al. *Docetaxel and prednisone with or without lenalidomide in chemotherapy-naive patients with metastatic castration-resistant prostate cancer (MAINSAIL): a randomised, double-blind, placebo-controlled phase 3 trial*. Lancet Oncol. 2015 Apr;16(4):417-25. doi: 10.1016/S1470-2045(15)70025-2

Pierrillas PB, Tod M, Amiel M, Chenel M, Henin E. *Improvement of Parameter Estimations in Tumor Growth Inhibition Models on Xenografted Animals: a Novel Method to Handle the Interval Censoring Caused by Measurement of Smaller Tumors*. AAPS J. 2016 Mar;18(2):404-15. doi: 10.1208/s12248-015-9862-1

Pierrillas PB, Tod M, Amiel M, Chenel M, Henin E. *Improvement of Parameter Estimations in Tumor Growth Inhibition Models on Xenografted Animals: Handling Sacrifice Censoring and Error Caused by Experimental Measurement on Larger Tumor Sizes*. AAPS J. 2016 Sep;18(5):1262-72. doi: 10.1208/s12248-016-9936-8

Pinheiro JC, Bates DM. *Mixed effects models in S and S-PLUS* (eds. Chambers, J. et al.) Springer-Verlag, 2000.

Pond GR. *Statistical issues in the use of dynamic allocation methods for balancing baseline covariates*. Br J Cancer. 2011 May 24;104(11):1711-5. doi: 10.1038/bjc.2011.157

R Development Core Team (2015). *R: a language and environment for statistical computing.* Version ≥3.2.2. R Foundation for Statistical Computing, Vienna, Austria. URL: http://www.R-project.org (Accessed: 15th August 2015)

Ramsay JO, Heckman N, Silverman BW. *Spline smoothing with model-based penalties*. Behav Res Methods Instrum Comput. 1997;29: 99-106. doi: 10.3758/BF03200573

Rao AR, Motiwala HG, Karim OM. *The discovery of prostate-specific antigen*. BJU Int. 2008 Jan;101(1):5-10. doi: 10.1111/j.1464-410X.2007.07138.x

Reardon S. *A mouse's house may ruin experiments*. Nature. 2016 Feb 18;530(7590):264. doi: 10.1038/nature.2016.19335

Ribonson SP, Jordan VC. *Reversal of the antitumor effects of tamoxifen by progesterone in the 7,12-dimethylbenzanthracene-induced rat mammary carcinoma model*. Cancer Res 1987;47:5386–90.

Risbridger GP, Davis ID, Birrell SN, Tilley WD. *Breast and prostate cancer: more similar than different*. Nat Rev Cancer. 2010 Mar;10(3):205-12. doi: 10.1038/nrc2795

Ripley B. *psplines: Smoothing splines with penalties on order m derivatives*. R-package version 1.0–16. http://cran.r-project.org/package=pspline (Accessed: 21st June 2016).

Ross BC, Boguslav M, Weeks H, Costello JC. *Modeling heterogeneous populations using Boolean networks*. BMC Syst Biol. 2018 Jun 7;12(1):64. doi: 10.1186/s12918-018-0591-9

Russel WMS, Burch RL. *The principles of humane experimental technique*. (1959); Special edition (1992). URL: http://altweb.jhsph.edu/pubs/books/humane_exp/het-toc (Accessed: 4th January 2018).

Saarinen NM, Huovinen R, Wärri A, Mäkelä SI, Valentín-Blasini L, et al. *Enterolactone inhibits the growth of 7,12-dimethylbenz(a)anthracene-induced mammary carcinomas in the rat*. Mol Cancer Ther 2002;1:869–76.

Saarinen NM, Power K, Chen J, Thompson LU. *Flaxseed attenuates the tumor growth stimulating effect of soy protein in ovariectomized athymic mice with MCF-7 human breast cancer xenografts*. Int J Cancer 2006;119:925–31. doi: 10.1002/ijc.21898

Saarinen NM, Wärri A, Dings RP, Airio M, Smeds AI, et al. *Dietary lariciresinol attenuates mammary tumor growth and reduces blood vessel density in human MCF-7 breast cancer xenografts and carcinogen-induced mammary tumors in rats*. Int J Cancer. 2008 Sep 1;123(5):1196-204. doi: 10.1002/ijc.23614

Sartor O, Eisenberger M, Kattan MW, Tombal B, Lecouvet F. *Unmet Needs in the Prediction and Detection of Metastases in Prostate Cancer*. The Oncologist. 2013;18(5):549-557. doi:10.1634/theoncologist

Sartor O, Laajala TD, Guinney J, Wang T, Murphy MJ, et al. *Scientists don't have to travel alone; solutions can come from the crowd*. Atlas of Science. Sep, 2017. URL: https://atlasofscience.org/scientists-dont-have-to-travel-alone-solutions-can-come-from-the-crowd/ (Accessed: 3rd January 2018).

Saeys Y, Inza I, Larrañaga P. *A review of feature selection techniques in bioinformatics*. Bioinformatics. 2007 Oct 1;23(19):2507-17. doi: 10.1093/bioinformatics/btm344

Scher HI, Jia X, Chi K, de Wit R, Berry WR, et al. *Randomized, open-label phase III trial of docetaxel plus high-dose calcitriol versus docetaxel plus prednisone for patients with castration-resistant prostate cancer*. J Clin Oncol. 2011 Jun 1;29(16):2191-8. doi: 10.1200/JCO.2010.32.8815

Schulz KF, Altman DG, Moher D, the CONSORT Group. *CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials*. BMJ. 2010;340:c332. doi: 10.1136/bmj.c332

Seruga B, Ocana A, Tannock IF. *Drug resistance in metastatic castration-resistant prostate cancer*. Nat Rev Clin Oncol. 2011;8(1):12-23. doi: 10.1038/nrclinonc.2010.136

Seyednasrollah F, Koestler DC, Wang T, Piccolo SR, Vega R, et al. *A DREAM Challenge to Build Prediction Models for Short-Term Discontinuation of Docetaxel in Metastatic Castration-Resistant Prostate Cancer.* JCO Clin Cancer Inform. 2017;1:1–15. doi: 10.1200/CCI.17.00018

Seyednasrollah F, Mahmoudian M, Rautakorpi L, Hirvonen O, Laitinen T, et al. *How Reliable are Trial-based Prognostic Models in Real-world Patients with Metastatic Castration-resistant Prostate Cancer?* Eur Urol. 2017 May;71(5):838-840. doi: 10.1016/j.eururo.2017.01.043

Shusterman S, Grupp SA, Barr R, Carpentieri D, Zhao H, et al. *Angiogenesis inhibitor TNP-470 effectively inhibits human neuroblastoma xenograft growth, especially in the setting of subclinical disease*. Clin Cancer Res. 2001;7:977–84.

Siegel RL, Miller KD, Jemal A. *Cancer statistics, 2016*. CA Cancer J Clin. 2016 Jan-Feb;66(1):7-30. doi: 10.3322/caac.21332

Simon N, Friedman J, Hastie T, Tibshirani R. *Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent*. J Stat Softw. 2011 Mar;39(5):1–13. doi: 10.18637/jss.v039.i05

Skove SL, Howard LE, Aronson WJ, Terris MK, Kane CJ, et al. *Timing of Prostate-specific Antigen Nadir After Radical Prostatectomy and Risk of Biochemical Recurrence*. Urology. 2017 Oct;108:129-134. doi: 10.1016/j.urology.2017.07.009

Smith GD, Ebrahim S. *Data dredging, bias, or confounding: They can all get you into the BMJ and the Friday papers*. BMJ. 2002 Dec 21;325(7378):1437-8.

Stamey TA, Yang N, Hay AR, McNeal JE, Freiha FS, et al. *Prostate-specific antigen as a serum marker for adenocarcinoma of the prostate*. N Engl J Med. 1987 Oct 8;317(15):909-16.

Sugar E, Pascoe AJ, Azad N. *Reporting of preclinical tumor-graft cancer therapeutic studies*. Cancer Biol Ther. 2012 Nov;13(13):1262-8. doi: 10.4161/cbt.21782

Suominen MI, Käkönen R, Käkönen SM, Halleen JM. *Diverging effects of doxorubicin, paclitaxel and cyclophosphamide on 4T1 mouse breast cancer primary tumor and metastases*. Poster at joint AACR-MRS meeting: Metastasis and the tumor microenvironment, September 12–15 2010, Philadelphia, USA. Available from: www.pharmatest.fi

Takahara K, Tearle H, Ghaffari M, Gleave ME, Pollak M, et al. *Human prostate cancer xenografts in lit/lit mice exhibit reduced growth and androgen-independent progression*. Prostate. 2011 Apr;71(5):525-37. doi: 10.1002/pros.21268

Tannock IF, Fizazi K, Ivanov S, Karlsson CT, Fléchon A, et al. *Aflibercept versus placebo in combination with docetaxel and prednisone for treatment of men with metastatic castration-resistant prostate cancer (VENICE): a phase 3, double-blind randomised trial.* Lancet Oncol. 2013 Jul;14(8):760-8. doi: 10.1016/S1470-2045(13)70184-0

Terada N, Shimizu Y, Kamba T, Inoue T, Maeno A, et al. *Identification of EP4 as a potential target for the treatment of castration-resistant prostate cancer using a novel xenograft model*. Cancer Res. 2010;70:1606–15. doi: 10.1158/0008-5472.CAN-09-2984

Tibshirani R. *The lasso method for variable selection in the Cox model.* Stat Med. 1997;16(4):385-395.

Tikkinen KAO, Dahm P, Lytvyn L, Heen AF, Vernooij RWM, et al. *Prostate cancer screening with prostate-specific antigen (PSA) test: a clinical practice guideline*. BMJ. 2018 Sep 5;362:k3581. doi: 10.1136/bmj.k3581

Torre LA, Siegel RL, Ward EM, Jemal A. *Global Cancer Incidence and Mortality Rates and Trends - An Update*. Cancer Epidemiol Biomarkers Prev. 2016 Jan;25(1):16-27. doi: 10.1158/1055-9965.EPI-15-0578

Trottier G, Boström PJ, Lawrentschuk N, Fleshner NE. *Nutraceuticals and prostate cancer prevention: a current review*. Nat Rev Urol. 2010 Jan;7(1):21-30. doi: 10.1038/nrurol.2009.234

Valkenburg KC, Pienta KJ. *Drug discovery in prostate cancer mouse models*. Expert Opin Drug Discov. 2015;10(9):1011-24. doi: 10.1517/17460441.2015.1052790

van der Worp HB, Howells DW, Sena ES, Porritt MJ, Rewell S, et al. *Can animal models of disease reliably inform human studies?* PLoS Med. 2010 Mar 30;7(3):e1000245. doi: 10.1371/journal.pmed.1000245

Verbeke G, Lesaffre E. *A linear mixed-effects model with heterogeneity in the random-effects population.* J Am Stat Assoc. 1996;91:217–21. doi: 10.1080/01621459.1996.10476679

Vickers AJ, Savage C, O'Brien MF, Lilja H. *Systematic review of pretreatment prostate-specific antigen velocity and doubling time as predictors for prostate cancer*. J Clin Oncol. 2009 Jan 20;27(3):398-403. doi: 10.1200/JCO.2008.18.1685

Wilkinson GF, Pritchard K. *In vitro screening for drug repositioning*. J Biomol Screen. 2015 Feb;20(2):167-79. doi: 10.1177/1087057114563024

Zappala SM, Scardino PT, Okrongly D, Linder V, Dong Y. *Clinical performance of the 4Kscore Test to predict high-grade prostate cancer at biopsy: A meta-analysis of us and European clinical validation study results*. Rev Urol. 2017;19(3):149-155. doi: 10.3909/riu0776

Zhao L, Morgan MA, Parsels LA, Maybaum J, Lawrence TS, et al. *Bayesian hierarchical changepoint methods in modeling the tumor growth profiles in xenograft experiments*. Clin Cancer Res 2011;17:1057–64. doi: 10.1158/1078-0432.CCR-10-1935

ORIGINAL PUBLICATIONS

Reprinted here with permission.

ORIGINAL PUBLICATION **I**

Laajala TD, Jumppanen M, Huhtaniemi R, Fey V, Kaur A, Knuuttila M, Aho E, Oksala R, Westermarck J, Mäkelä S, Poutanen M, Aittokallio T. *Optimized design and analysis of preclinical intervention studies in vivo.* Scientific Reports. 2016;6:30723.

SCIENTIFIC REPORTS

OPEN

Received: 14 December 2015
Accepted: 06 July 2016
Published: 02 August 2016

Optimized design and analysis of preclinical intervention studies *in vivo*

Teemu D. Laajala[1,2,3,4], Mikael Jumppanen[3,5], Riikka Huhtaniemi[3,4,6,7], Vidal Fey[3,6,8], Amanpreet Kaur[5,9,10], Matias Knuuttila[3,4,6], Eija Aho[7], Riikka Oksala[4,7], Jukka Westermarck[5,9], Sari Mäkelä[3,11], Matti Poutanen[3,6,12,*] & Tero Aittokallio[1,2,3,*]

Recent reports have called into question the reproducibility, validity and translatability of the preclinical animal studies due to limitations in their experimental design and statistical analysis. To this end, we implemented a matching-based modelling approach for optimal intervention group allocation, randomization and power calculations, which takes full account of the complex animal characteristics at baseline prior to interventions. In prostate cancer xenograft studies, the method effectively normalized the confounding baseline variability, and resulted in animal allocations which were supported by RNA-seq profiling of the individual tumours. The matching information increased the statistical power to detect true treatment effects at smaller sample sizes in two castration-resistant prostate cancer models, thereby leading to saving of both animal lives and research costs. The novel modelling approach and its open-source and web-based software implementations enable the researchers to conduct adequately-powered and fully-blinded preclinical intervention studies, with the aim to accelerate the discovery of new therapeutic interventions.

In vivo animal studies are an essential part of any drug development project. To further increase the reproducibility and translatability of preclinical studies, there is an increasing need to improve their experimental design and statistical analysis[1–6]. Recurrent concerns are especially related to lack of power calculations for sample size estimation, inadequate conduction of randomized and blinded intervention group allocations, and limited consideration of individual animal characteristics at baseline prior to interventions[2,6–10]. It has been argued that preclinical animal studies should more closely follow the established practices applied in the human clinical trials, where standardized requirements have been enforced for reporting statistical power, randomization procedures and stratification factors[1,11]. Typical sources of variation in the animal baseline characteristics include differences in gender, body weight and age, as well as in the genetic differences, cage conditions or the variability in gut microbiota[7,12–14]. Each of these experimental factors may contribute to confounding variability in the intervention responses, leading to false positive or negative findings, unless the study is carried out using adequate sample sizes or design that normalizes such confounding factors. Although these issues are widely acknowledged among the researchers, and guidelines are available for standardizing and reporting preclinical animal research[15], the implementation of the best practices is often neglected[2,8,16–19]. Accordingly, a recent survey revealed that over 85% of published animal studies did not describe any randomization or blinding, and over 95% lacked the estimation of sufficient sample size needed for detecting true effects in the intervention studies[17].

[1]Department of Mathematics and Statistics, University of Turku, Turku, Finland. [2]Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland. [3]Turku Center for Disease Modeling (TCDM), University of Turku, Turku, Finland. [4]Drug Research Doctoral Programme (DRDP), University of Turku, Turku, Finland. [5]Cancer Cell Signaling Group, Turku Centre for Biotechnology, University of Turku, Turku, Finland. [6]Department of Physiology, Institute of Biomedicine, University of Turku, Turku, Finland. [7]Orion Corporation, Orion Pharma, Department of Oncology and Critical Care Research, Turku, Finland. [8]Department of Medical Biochemistry and Genetics, Institute of Biomedicine, University of Turku, Turku, Finland. [9]Department of Pathology, University of Turku, Turku, Finland. [10]Turku Doctoral Programme of Molecular Medicine (TuDMM), University of Turku, Turku, Finland. [11]Functional Foods Forum, University of Turku, Turku, Finland. [12]Institute of Medicine, The Sahlgrenska Academy, Gothenburg University, Gothenburg, Sweden. *These authors contributed equally to this work. Correspondence and requests for materials should be addressed to M.P. (email: matti.poutanen@utu.fi) or T.A. (email: tero.aittokallio@fimm.fi)

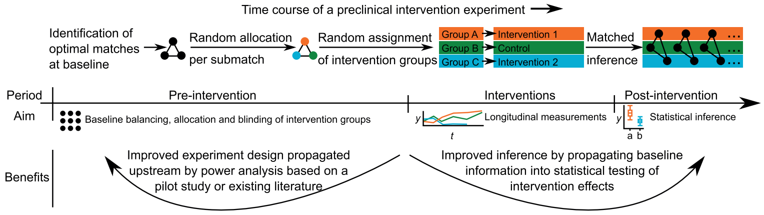

Time course of a preclinical intervention experiment
Identification of optimal matches at baseline
Random allocation per submatch
Random assignment of intervention groups
Group A
Group B
Group C
Intervention 1
Control
Intervention 2
Matched inference
Period
Pre-intervention
Interventions
Post-intervention
Aim
Baseline balancing, allocation and blinding of intervention groups
Longitudinal measurements
Statistical inference
y
t
a b
Benefits
Improved experiment design propagated upstream by power analysis based on a pilot study or existing literature
Improved inference by propagating baseline information into statistical testing of intervention effects

Figure 1. Benefits of the modelling framework over the course of the study period. The animal baseline matching improves the statistical analysis and design of preclinical animal studies in terms of power calculations, balanced allocations, and intervention blinding (pre-intervention period), as well as through the use of matching information in the statistical testing of the intervention effects (post-intervention period).

In the absence of established practices and procedures for power calculations tailored for preclinical studies, the preferred sample size is often decided through historical precedent rather than solid statistics[9,20]. Similarly, the current approaches for allocating animals to separate intervention arms are typically based on manual picking and balancing of the animal groups based on only one baseline variable[6]. However, such simple design procedures may easily miss the complex relationships between multiple baseline variables, and the subtle intervention effects. Further, it remains a challenging question how to choose among the multiple baseline markers due to inherent differences in animal experimentation. Preferably, the intervention groups should be balanced using all the available baseline factors, including information about the animal characteristics (e.g., gender, age and weight), littermates, housing conditions, and pre-treatments, among others. Otherwise, even minor uncontrolled differences between the treatment arms may cause significant variation in the response profiles[13]. Many of the experimental factors lead to complex hierarchical designs, with nested animal, host-tumor, cage, batch and litter relationships at multiple levels, thus reaching beyond the capability of the existing randomization and allocation methods available for preclinical animal studies[21,22]. The current methods often assume the independence of the baseline variables and experimental units, which may lead to over-optimistic evaluation of the effective sample size, also known as pseudo-replication[16]. This takes place, for instance, when one allocates multiple animals from a single batch or cage to a single treatment arm, or when multiple tumours are placed in the same animal.

Results

We developed and implemented a novel methodology to improve the experimental design and statistical analysis of preclinical studies carried out with experimental animals. The advances are based on a mathematical optimization framework for animal matching that improves both the unbiased allocation of the intervention groups, as well as the sensitivity and specificity of the post-intervention efficacy evaluations by making the full use of all the available baseline characteristics. To support its widespread use in various experimental settings, the modelling framework has been made available both as an open-source R-package (http://cran.r-project.org/package=hamlet) (Supplementary Note S1), and through a web-based graphical user interface (http://rvivo.tcdm.fi/) (Supplementary Note S2). To our knowledge, these implementations are the first that effectively consider the nested, hierarchical structures of preclinical animal studies across the different phases of the experiment, starting from the power analyses, to allocation of animals to the various treatment arms, and all the way to finally evaluate the intervention effects (Fig. 1). In the present work, we demonstrate the benefits of these tools over conventional analysis in two applications of orthotopic xenografts of VCaP prostate cancer cells in immune deficient mice as disease models for castration-resistant prostate cancer (CRPC) (Supplementary Fig. S1). The first study analysed the efficacy of two androgen receptor antagonists (ARN-509 and MDV3100) to suppress the growth of castration-resistant VCaP tumors[23], while the second study investigated the effect of surgical and pharmaceutic therapies on orchiectomized mice (for details; see Supplementary Methods and Supplementary Note S3).

In a given pool of animals, the matching solution provides an optimal intervention group allocation of animals (or tumors) based on several baseline characteristics (Fig. 2). Rather than considering only the optimal pairing of individual animals, the solution can be used also to identify optimal matches among a number of features, animals or tumors, e.g., triplets, quadruplets, or more (see Methods for the mathematical formulation of the matching problem). Such optimal combinations, referred here to as *submatches*, are constructed by minimizing the sum of all the pairwise distances between the members of each submatch, illustrated here by pairwise connecting edges (Fig. 2). Since the non-bipartite matching procedure does not require pre-defined group labels, the control group can be selected without any guidance from the experimenters (Supplementary Fig. S7b). Instead, the animal allocation is performed objectively within each submatch by distributing its members randomly to separate treatment arms, hence enabling fully-blinded intervention group allocation through separate matching and randomization phases (Figs 1 and 2c,d). In the present study, we demonstrate how the matching information does not only improve the pre-intervention design, such as baseline animal group balancing and allocation, but it also improves the post-intervention statistical power to detect true treatment effects.

Matching normalizes baseline variability in confounding variables. The first VCaP xenograft case study was originally conducted based on the matching procedure[23], where it showed its added value in complex

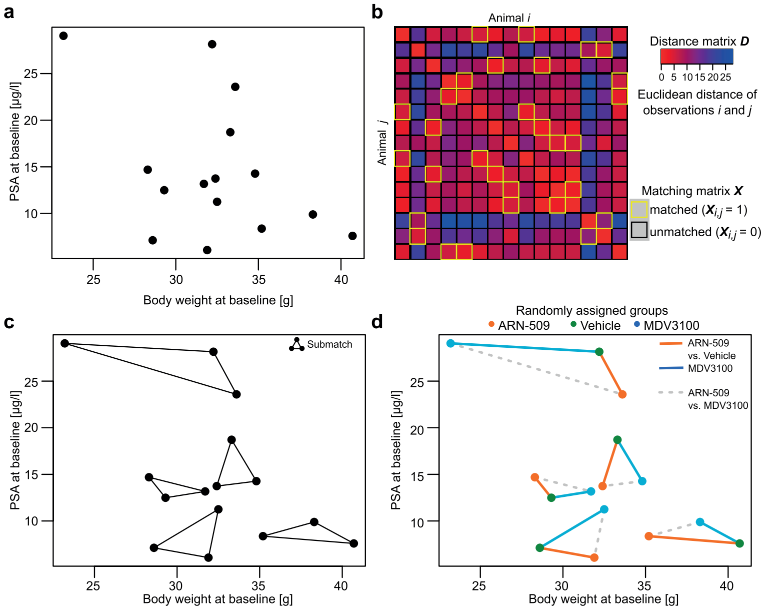


Figure 2. Optimal matching of animals in the case of orthotopic VCaP mouse xenografts. The original task was to randomly assign 75 animals into five balanced intervention groups (one control and four treatment groups, each consisting of 15 animals), but here we focus on two of the treatments only (ARN-509 and MDV3100), using a sub-sample of the complete data matrix (see Supporting Fig. S3). (**a**) Bivariate observations sampled from the VCaP study, illustrating the two selected baseline variables (body weight and PSA). (**b**) 15 × 15 dimensional distance matrix ***D*** calculated based on the baseline variables was used as an input to the matching procedure, which solves the optimal animal matching matrix ***X***. (**c**) The optimal submatches from the branch and bound algorithm, which guarantees a globally optimal solution (see Supporting Fig. S7). (**d**) The optimally matched animals were randomized into the intervention groups via blinded treatment label assignments (coloured points). The baseline matching information was used in the statistical testing of the treatment effects, mainly through paired comparisons between the treated and control animals (solid lines). Alternatively, the model also allows for direct comparisons between the two treatments (dotted lines).

designs with batch/cage effects and multiple treatment groups ($n = 15$ animals per group). While the full matching included four baseline variables, we illustrate the methodology first using two key animal characteristics (PSA and body weight at baseline; Fig. 2a–c). The optimal submatches were subsequently randomized and blinded for the experimenters to enable unbiased analysis across three intervention groups (ARN-509, MDV3100 and Vehicle) (Fig. 2d). The confounding variability from the two castration batches was normalized by treating these as two separate optimal matching problems (Supplementary Fig. S3a), which guaranteed that the two batches were allocated uniformly to the intervention groups through the use of submatches (Supplementary Fig. S3b). Notably, the matching distance matrices at baseline were also significantly correlated with the post-intervention RNA-seq profiling of a randomly chosen subset of individual tumours ($p = 0.039$, Mantel's test, $n = 4$ animals per group; Supplementary Fig. S8c), suggesting that major trends in the characteristic baseline differences used in the animal allocation were still captured by their genome-wide transcriptional responses even after the interventions (Supplementary Fig. S8a,b).

To more systematically study the degree of confounding variability and its effects on the animal allocation, we tested the frequency of statistically significant differences in all the available baseline variables between the randomized treatment groups. A total of $n = 100,000$ animal allocations were simulated either totally at random (*unmatched randomization*) or using the matching information from the optimal submatch allocations (*matched randomization*). The baseline variables considered in the optimal matching were body weight and PSA at baseline, as well as PSA fold-change from previous week prior to allocation. With the unmatched randomization, 13.8% of the treatment groups represented significant differences with respect to at least one of the baseline variables ($p < 0.05$, one-way ANOVA). In contrast, only 0.018% of the treatment groups in the matched randomizations showed any baseline differences. This indicates that matching effectively eliminates baseline differences in the

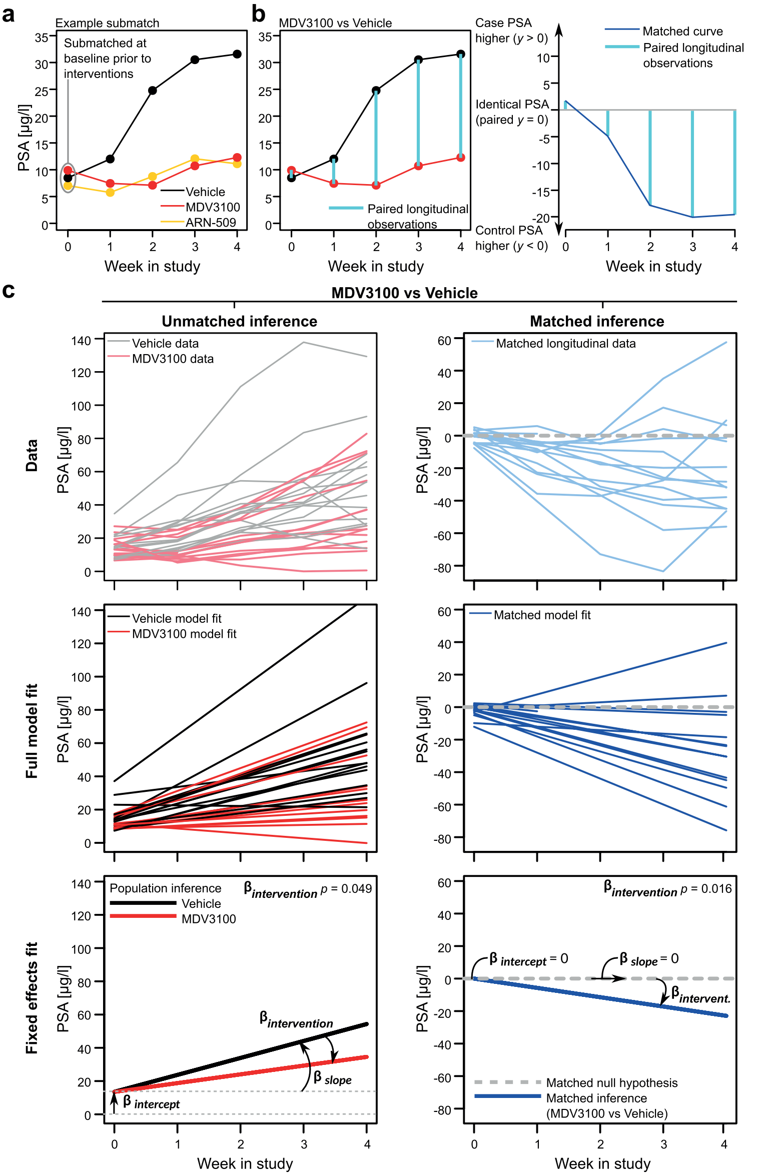


Figure 3. Statistical testing of the treatment effects using pairwise matched inference. (**a**) The matched inference makes use of the baseline matching information when testing the intervention effects by pairing the observed responses according to the optimal submatches at equal time points. (**b**) An example of the submatch-based pairing in the MDV3100 vs vehicle comparison, where the example trajectory was previously shown as a single estimate value in the original study[23]. Complex response differences are better captured when additional baseline information is incorporated into the statistical inference. The paired differences from the longitudinal

observations (left panel) construct a single treatment curve for the pairwise matched mixed-effects modelling (right panel). (**c**) Comparison of the matched and unmatched statistical inference approaches in the MDV3100 vs vehicle comparison. Even if both inference approaches yield rather similar conclusion about the possible intervention effects, the matched approach improves the sensitivity of the detection (right panel). Different aspects of the mixed-effects modelling are visualized based on the observed data (top panel): the full model fit combining both the random and fixed effects (middle panel), and the population inference depicting only the fixed effects along with their interpretation (bottom panel). In the matched inference, the population of paired differences in the intervention effects ($\boldsymbol{\beta}_{\boldsymbol{intervention}}$) is tested against a null hypothesis of no paired differences ($y = 0$ line). The statistical inference results of the intervention effects are summarized in Table 1, and the full model fits for the four treatment cases are shown in Supplementary Figs S5 and S6.

confounding variables, which unless carefully controlled during the allocation process, may contribute to the poor reproducibility of preclinical research findings[24].

Matching improves the statistical inference of treatment responses. In the post-intervention analysis, we studied the benefits of using the matching information in the mixed-effects modelling of the treatment effects (see Methods for the model formulation), focusing first on the ARN-509 and MDV3100 treatments (Fig. 3a). The matched inference approach models the paired longitudinal differences in the intervention responses (PSA in the VCaP xenografts; Fig. 3b), based on the optimal submatches of the animals at baseline (Fig. 2c; Supplementary Fig. S3). The benefits gained by such matching-based paired testing became more evident with the MDV3100 case, where we observed that the animal body weight at baseline was inversely associated with the final PSA level (correlation coefficient $\rho = -0.607$, $p = 0.021$, Supplementary Fig. S2d). Such multivariate, longitudinal relationship between the baseline variables and treatment responses cannot be captured by the conventional, unmatched model, leading to reduced statistical sensitivity (Fig. 3c, left). The MDV3100 treatment effect became clearly significant when the baseline matching information was incorporated into the mixed-effects modelling (Fig. 3c, right). The more apparent ARN-509 intervention effect was detected both with the matched and unmatched statistical models (Table 1). Of note, the non-matched approach also benefitted here from the matched randomization of the original study[23].

As another case study, we randomly allocated 100 VCaP mice using the matching algorithm into six intervention groups (Supplementary Fig. S4), out of which three are further investigated here (Control, orchiectomized (ORX) and ORX+Tx). As was expected, when compared to the intact control animals, both the matched and unmatched statistical models were able to detect the significant intervention effect from the ORX surgery (Table 1). However, the unmatched approach totally missed the additional effect from an undisclosed pharmaceutic treatment (Tx), while the ORX+Tx combination effect was found significant after using the baseline matching information in paired testing of the longitudinal intervention responses. In the combination case, the standard, non-paired analysis lacked the power to distinguish the complex response patterns between the intervention groups, in part due to the non-linear responses in the early time points (Supplementary Fig. S6). In contrast, the paired inference, enabled by the matching information, was able to capture these pairwise response differences, leading to subtle yet significant intervention-specific effect sizes (Table 1). These results support the improved statistical sensitivity gained by the baseline matching information in the detection of true treatment effects, especially when studying more complex and subtle intervention effects.

Matching increases statistical power to detect true treatment effects. Since the intervention effects in the preclinical studies are often relatively subtle, statistical power calculations are critical for estimating the sufficient number of animals needed to detect a true effect. However, preclinical experiments pose specific requirements for the power calculations, due to the complex nature of longitudinal responses, relatively high frequency of missing values originating from animal health or other exclusion criteria, complex hierarchical designs, as well as multivariate baseline characteristics, which are beyond the capacity of any standard sample size estimation procedures. We addressed the above mentioned challenges by implementing a model-based power analysis calculation. The method first samples animals with replacement from an estimated mixed-effects model, and then uses these bootstrap datasets to re-estimate the specified statistical model (see Methods for the modelling details).

When applied to the two VCaP xenograft studies, the model-based calculation enables one to estimate the study power as a function of tumors per treatment group. Although the power calculation can be done with respect to each of the terms in the mixed-effects model, we focused here on the intervention-specific term $\beta_{\text{intervention}}$ (Fig. 3c). With the more prominent intervention effects from ARN-509 and ORX, the power calculation led to similar sample size estimates between matched and unmatched models ($n < 10$; Fig. 4, left panel). However, there were notable differences in the number of animals needed when more complex or subtle interventions effects were studied; with MDV3100, the matched analysis reached the conventional power level of 0.8 at much smaller sample size compared to the unmatched model ($n = 17$ vs. $n = 26$; Fig. 4a, right panel), whereas for the intervention effect from ORX+Tx combination, the unmatched analysis remained below the sufficient power level with any practically feasible number of animals (Fig. 4b, right panel).

Although the power simulations were performed here retrospectively, these results already demonstrate that statistical inference of the intervention effects is highly dependent on the expected effect size and within-group variation, suggesting that future experimental designs should be tailored for each case individually, using e.g. data from a pilot experiment, so that the power calculations will meet the expected response patterns. Given the relatively large difference in the number of animals needed to reach sufficient power using an unmatched or matched approach, especially with the less evident cases (MDV3100 and ORX+Tx interventions), it is recommended that

Model		Fixed effects (p-value)			Random effects (SD)		
		$\beta_{intercept}$	β_{slope}	$\beta_{intervention}$	$\gamma_{intercept}$	γ_{slope}	ε_{error}
ARN-509 vs Control	*Unmatched*	14.311 (<0.001)***	10.062 (<0.001)***	**−7.627 (<0.001)*** **	8.234	5.163	5.749
	Matched	0 (−)	0 (−)	**−7.962 (0.0047)** **	7.053	8.894	8.399
MDV3100 vs Control	*Unmatched*	13.536 (<0.001)***	10.188 (<0.001)***	**−4.940 (0.0494)* **	7.635	6.259	6.395
	Matched	0 (−)	0 (−)	**−5.729 (0.0160)* **	7.013	7.401	11.247
ORX vs Intact	*Unmatched*	14.548 (<0.001)***	1.336 (<0.001)***	**−1.265 (0.0034)** **	14.578	0.997	8.518
	Matched	0 (−)	0 (−)	**−1.931 (0.0063)** **	4.251	2.157	9.522
ORX+Tx vs ORX	*Unmatched*	9.998 (<0.001)***	0.122 (0.0675)N.S.	**−0.101 (0.2704)N.S.**	10.476	0.167	9.977
	Matched	0 (−)	0 (−)	**−0.112 (0.0457)* **	2.381	0.155	4.618

Table 1. Mixed-effects model fits for the fixed effects (population inference) and random effects (individual effects and the random error term). Model estimates and their significance levels using the conventional unmatched and matching-based pairwise models are presented for each intervention comparison separately. The model term that explicitly tests for an intervention effect is highlighted in bold. N.S., not significant; *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

post-intervention statistical procedures should be defined already before the initiation of the study using tools such as proposed here[25]. However, a more fair comparison point for the matched approach would require a new study, conducted without using the baseline matching information, but this was not carried out within our preclinical studies because of ethical reasons.

Discussion

The importance of controlling for individual variation is well-acknowledged in human clinical studies, with the aim to increase the study validity and reproducibility[26]. Similarly, in preclinical studies, reproducibility of the findings is associated with transparent reporting and paying careful attention to the experimental issues, including balancing, randomization and blinding[2,3]. Even though preclinical experimental designs differ from the truly randomized treatment group testing applied in clinical trials, the preclinical studies should benefit from the best practices of human clinical trials to improve their translatability[11,15]. We demonstrated here that a more detailed animal matching and statistical modeling offers many benefits across the different phases of the preclinical intervention experiment (Fig. 1). Prior to the interventions, the baseline balancing makes the experimental and control groups as similar as possible, while the matching-based randomization ensures that all the animal groups are sufficiently representative of the underlying population. This should reduce confounding variability and false positives in the subsequent testing of the intervention effects. During interventions, blinding promotes comparable handling and treatment of the animals by experimenters, while the estimated model parameters can detect outliers and provide insights into dynamic changes in individual animals in response to the interventions, such as non-linear treatment effects in the intervention groups. This makes the outcome measurements more uniform and reduces bias when reporting the results. After the intervention period, the paired longitudinal analysis of the individuals or tumours that were similar at baseline can be utilized in more sensitive detection of treatment effects (analogous to the paired t-testing). This may reduce false negative detections, especially when testing more subtle or complex treatment relationships, such as the MDV3100 and ORX+Tx treatment responses considered in the present study. While demonstrated here in the context of orthotopic xenograft studies, the statistical analysis and design issues are widely applicable also to genetically-modified mouse models (GEMMs), and should be even more important with the use of patient-derived xenografts (PDX), where the tumor material is limited and unique to each patient case[20].

Power calculations in preclinical animal studies.

Power calculations are routinely demanded in human clinical studies, and recent reports have called for more rigorous sample size estimation also in preclinical animal studies[9,20]. Our model-based simulations enable the full use of response data from a pilot study or similar studies in the literature when estimating the sufficient sample size, rather than guessing or predicting the key model parameters and their variance. Furthermore, sampling of observations from a pre-fitted mixed-effects model offers a possibility to also incorporate indirect intervention effects, such as censoring due to death or animal exclusion, which may be difficult or even impossible to infer otherwise when determining the model parameters. Finally, the mixed-effects model requires the experimenter to specify the tested population hypotheses and the particular model structure already in the study design phase, which effectively discourages exploratory cherry picking and fishing for the 'optimal' results, a practice which severely reduces the reproducibility of the findings[27]. We note that the power simulations carried out in the present study were performed retrospectively, and hence, are applicable to designing future studies only[28]. When testing for more subtle or complex treatment effects, such as the +Tx effects in the ORX mice, sufficiently large sample sizes were required to provide statistically robust results. Even if this may lead to unexpectedly high number of test animals, it is widely acknowledged that underpowered or otherwise poorly designed studies are not only unethical but also contribute to both delays and increased costs of drug development process[4,9].

Exploratory and confirmatory study design issues.

Table 2 summarizes the experimental design issues that we feel are essential to consider while performing statistically robust preclinical intervention studies.

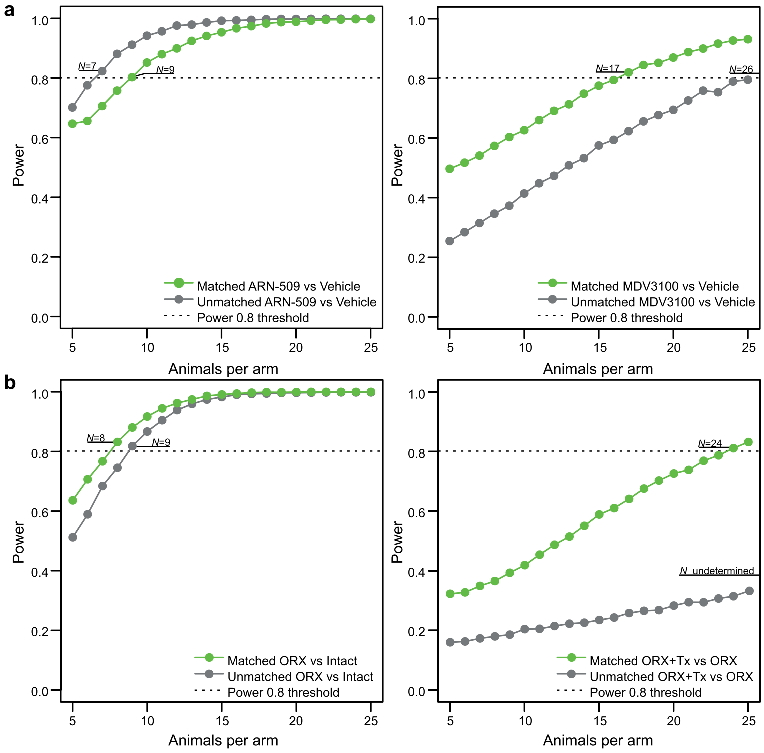


Figure 4. Model-based power calculations for sufficient sample size estimation. Statistical power (the likelihood that a true treatment effect is detected) as a function of the sample size (animals per treatment arm). Power calculations were computed by bootstrap re-sampling, either without the matching information (unmatched) or using the information from the optimal pairs of matched samples (matched). The estimated sample sizes (*N*) are defined based on the conventional threshold of 0.8 power. (**a**) ARN-509 and MDV3100 intervention effects in the VCaP mouse xenografts. (**b**) ORX and ORX+Tx intervention effects in the orchiectomized (ORX) VCaP mouse xenografts.

These issues are important both for exploratory and confirmatory preclinical studies, in order to improve their generalizability and translatability toward human diseases[29]. Exploratory studies involve preclinical screening and pathophysiological hypothesis testing, placing therefore more focus on detection sensitivity, whereas confirmatory studies are geared more toward efficacy estimation and clinical translation, where specificity of the findings is often more important. These two study classes serve as examples of the various inferential aims of the preclinical studies, and we hope our considerations will complement the current ARRIVE guidelines[15], in terms of the statistical design and analysis of intervention effects. However, there remain several other factors that are outside the scope of the statistical methods introduced here, which may have much bigger role in the generalizability and translatability of the preclinical findings. For instance, although the internal variation in the treatment response can be controlled to a large extent using the matching and randomization methodology, these cannot normalize the effects of external factors, such as the representativeness of the animal model of the actual human disease, its target population and heterogeneity[3,30]. Additionally, although the animal matching can be performed based on multiple prognostic preclinical variables, these are unlikely to directly translate into the clinical use due to differences in the preclinical and clinical experimentation and physiology. However, the success rate of the human clinical trials is likely to benefit from a more accurate modelling of the heterogeneous treatment responses already during the preclinical phase[3].

Additional simulations of the model performance. An important practical question is how many and what type of baseline covariates should be used for animal matching. To address this question, we performed

Design issue	Exploratory study	Confirmatory study	Aims and benefits
Study objective (focus on sensitivity/precision or specificity/generalizability)	Preclinical screening and pathophysiological hypothesis testing (*sensitivity*)	Estimating effect size and ensuring clinical translation (*specificity*)	Sensitivity allows effective search for intervention candidates, while specificity emphasizes translational aspects. Notably, mere statistical significance in preclinical testing does not yet guarantee clinical relevance
Example animal models[19]	Traditional cost-efficient models, e.g. subcutaneous xenografts	Translational models, e.g. orthotopic xenografts, PDX, GEMM	Seeking a balance between cost-efficiency and translatability
Number of intervention groups (Parameter *G*)	High number of candidate intervention groups (Prefer *G* over *N*)	Carefully selected interventions to be validated (Prefer *N* over *G*)	High *G* allows effective exploration of novel candidates for downstream confirmatory studies
Number of animals in each intervention arm (Parameter *N*)	Focus on testing multiple candidate intervention groups at sufficient sample size (medium *N*)	High confidence required for true positive effects as well as for effect size estimate (high *N*)	Well-characterized animals and sufficient *N* allows better translation to the target population and improved generalizability
Number of covariates *d* in matching (Data dimension *d*)	Many possible confounding covariates, with suspected effect on the primary response (flexible *d*)	Ideally only few selected confounding covariates, which affect the representative intervention outcome (low *d*)	Matched animals in separate treatment arms allows more accurate inference both in terms of sensitivity and specificity
Estimation of sample size for the study and effect sizes for the interventions	Often difficult due to lack of pilot studies for the candidate interventions	Key ingredient in ensuring sufficient statistical power[1,9]	Sufficient statistical power to identify true intervention effects and reject false effects. Accurate effect size estimation assists in evaluating clinical significance
Maximization of the consistency in handling of the individual animals and/or tumours	Relevant in all study aims	Relevant in all study aims	Prevent undesired stratification and false detections due to potential batch-effects
Taking into account potential dependence structures (e.g. tumours within the same animal)	Highly dependent on the number of *G* in relation to *N*. Some degree of compromise is acceptable to maximize sensitivity	Highly relevant, e.g. cage-effects are attributed to high attrition rates of preclinical findings [6,12–14]	Prevents over-estimation of the required sample size due to so-called pseudo-replication[15]

Table 2. Experimental design issues in exploratory and confirmatory preclinical studies. Exploratory and confirmatory study aims adopted from Kimmelman *et al.*[29]

extensive simulation study (see Supplementary Methods), which confirmed and extended the results from our real case studies, showing that the matching information and paired analysis improves the statistical inference beyond the conventional approaches; this improvement was systematically observed across the number and type of covariates in terms of both detection sensitive and specificity (Supplementary Fig. S9). The largest benefits of the matching was gained with a selected panel of predictive baseline markers (e.g. 3–10 most informative covariates), in relatively small-sized studies ($N = 5$ to 10), but even if performed using non-informative markers and in larger studies, matching did not lead to reduced sensitivity or specificity. We therefore recommend preclinical researches to use several expert-curated baseline variables to improve the animal allocation and statistical testing, with a focus on the most relevant markers for the particular inference task (Table 2).

We further performed simulation studies to assess the relative advantages of the matched regression modelling in comparison to a more standard adjusted regression modelling, where the baseline confounders are incorporated as covariates in the post-intervention testing phase (Supplementary Material). Such post-intervention adjustments in the regression modeling may suffer from confounders interacting with the intervention effect, which may be difficult to track down and control for retrospectively in the intervention effect testing, as well as from an uncertainty about which confounders should be incorporated as the regression coefficients. We noticed that a matched-based animal allocation systematically improved over the adjusted regression, while the use of the pairwise matching information in the regression modeling led to the overall best sensitivity and specificity (Supplementary Fig. S10). Taken together, these simulation results show that the matching-based design and statistical analysis generally outperforms the more conventional approaches that do not use the baseline matching information.

Current limitations and future perspectives. The presented methodology has certain limitations and potential caveats that should be understood. First of all, our specific focus here was on preclinical *in vivo* animal models, while the other forms of preclinical research are beyond the scope of this work. However, similar methods could be used also for *in vitro* experiments, where genetic and chemical perturbations and interventions are extensively modeled using dissimilarity-based methods analogous to the matching-procedure presented here[31,32]. Further, our methodology is implemented in the context of conventional preclinical study period, where animals are first selected for study inclusion, then baseline variables are measured based on which all the animals are randomly allocated into intervention groups (Fig. 1). Although the mixed-effect statistical model effectively captures the dynamic changes in the intervention responses, the baseline-based dissimilarity metrics do not typically consider time-dependent covariates; however, one can carry out also a longitudinal randomization procedure using, for instance, dynamic allocation methods that take into account dynamic cohort additions, covariate structures and intervention responses[33]. Finally, although both of our example cases were longitudinal intervention analyses of the tumor growth as a function of time modeled using linear mixed-effects models, the experimental design approach is also applicable to single end-point comparisons as a special case. We demonstrated this through the use of multivariate extension of the standard *t*-test (so-called Hotelling's T^2 test) in the VCaP xenografts, where we observed that the PSA surrogate marker correlated well with the primary outcome of tumor volume. Importantly, it was shown that the detection sensitivity of the subtle treatment effect of MDV3100 was increased when the end-point markers were coupled with the matching information through paired T^2-testing (Supplementary Fig. S11).

As a future work, it will be important to perform a more systematic review and evaluation of the practices and factors that affect treatment assessment in preclinical intervention studies *in vivo*. These include experimental factors, such as measurement frequency, structured missing information due to both lower censoring at response detection limit and right-censoring at death, extent of pseudo-replication and confounding variability due to correlated structures (e.g., multiple tumors), as well as dynamic changes in the treatment effects over time. In particular, non-random missing values pose challenges to any statistical testing approaches, including our matching-based post-intervention testing procedure, which assumes that both of the paired individuals have been observed in order to effectively model the pairwise treatment differences. Such procedure creates the caveat that highly aggressive tumor groups, which are often being censored due to ethical reasons, may fail to provide representative animal/tumor pairs with those individuals with fully-observed longitudinal response profiles. This aspect of the pairwise matching procedure may actually provide also an advantage compared to the standard statistical methods, which often treat all the missing data as missing-at-random (MAR), as censoring removes pairwise differences from both of the animals that have a matched baseline profile; therefore, right-censored missingness will not accumulate only to aggressively growing groups. Although it is possible that this allows for less-biased estimates, provided that the prognostic matching covariates can accurately predict the response, this potential advantage may come at the expense of decreased power to detect the longitudinal intervention differences as dominant right-censoring may result in insufficiently short pairwise longitudinal trajectories. Due to the complex nature of non-random missingness in the post-intervention testing, systematic evaluation of these effects warrants a separate future work in various preclinical models and experimental setups.

Methods

Optimal non-bipartite matching problem formulation.

Matching was used to allocate individual animals into homogeneous subgroups according to a pre-defined dissimilarity criterion[34] (Fig. 2; Supplementary Fig. S7a,b). Multiple baseline variables that may have either prognostic or confounding contribution to the treatment response were simultaneously used for balancing the treatment and control groups through the pre-selected dissimilarity metric (Supplementary Table 1). By incorporating such baseline information, the experimental design allows for more sensitive and specific detection of effects that are due to the interventions alone[35]. In theory, matching should not introduce loss of statistical power even when performed on irrelevant covariates[34]. Since purely deterministic allocation methods have been criticized for the risk of introducing experimental biases due to, for instance, the lack of masking[36], our constrained randomization procedure incorporated also a stochastic component, making it fully compatible with the current clinical recommendations of random allocation and balancing at baseline. The matching-based randomization approach refines all possible allocations from a single pool of individuals, and then randomly picks one of these most feasible allocation solutions. As such, the procedure greatly resembles the randomized block design, which is used in the clinical field to adjust for pre-intervention randomizations by stratifying for categorical factors (e.g. gender) or bins of numeric values (e.g. adolescent/adult/elderly), especially in studies with small or moderate sample size[37].

Expanding the previous formulation[35] (Supplementary Fig. S7b), the optimal non-bipartite matching problem can be formulated as follows. Let us consider binary symmetrical matching matrices $\boldsymbol{X}$ of size $N \times N$ where:

$$X_{i,j} \in \{0, 1\} \tag{1}$$

$$\textstyle\sum_i X_{i,j} = G - 1 \ \forall \ j \in \{1, 2, \ldots, N\} \tag{2}$$

$$\textstyle\sum_j X_{i,j} = G - 1 \ \forall \ i \in \{1, 2, \ldots, N\} \tag{3}$$

$$X_{i,j} = X_{j,i} \tag{4}$$

$$X_{i,i} = 0 \tag{5}$$

The two sum constraints in equations (2 and 3) guarantee that the number of edges originating from a single observation equals the number of desired groups minus 1. Here, G denotes the number of desired members per each matching structure, and is equal to the number of desired intervention groups. This means that all the rows sum to G-1 and columns to G-1 in the binary matching matrix $\boldsymbol{X}$ (Fig. 2b). Dissimilarity matrix $\boldsymbol{D}$ of size $N \times N$ is defined as:

$$D_{i,j} = D_{j,i} \tag{6}$$

$$D_{i,i} = 0 \tag{7}$$

Each element $D_{i,j}$ is computed according to the chosen dissimilarity metric with possible alternatives summarized in Supplementary Table S1. The interpretation of the constraints are follows; in equation (1): For each possible pairs of individuals *i* and *j*, the pair is either matched (value 1, connected with an edge) or not matched (value 0);

(2): Each individual *j* is connected to other *G*-1 matched individuals; (3): Each individual *i* has *G*-1 other individuals that are matched to the individual *i*; (4): If individual *i* is matched to individual *j*, then individual *j* is also matched to individual *i* (no single direction relationships allowed); (5): An individual may not be matched to itself; (6): The similarity of individual *i* to individual *j* is as great as similarity of individual *j* to individual *i* (no directionality allowed). (7): An individual is always perfectly similar to itself.

The existing optimal non-bipartite matching algorithms, for example, the one presented in the R-package '*nbpMatching*'[35], consider paired non-bipartite matching, where:

$$\sum_i X_{i,j} = 1 \ \forall \ j \in \{1, 2, \ldots, N\} \tag{8}$$

$$\sum_j X_{i,j} = 1 \ \forall \ i \in \{1, 2, \ldots, N\} \tag{9}$$

We expanded upon this problem and developed a global optimization algorithm for solving this generalized problem. In order to introduce multigroup matches, we define fully connected structures called *submatches*. Matches are considered as graphs {*V, E*}, where *V* is vertex (node) and *E* is an edge between vertices. If observations *i* and *j* have not been matched, their edge is non-existing ($X_{i,j} = 0$). Each submatch M_k is a subgraph of *V*, where the number of vertices belonging to the *k*:th submatch M_k equals to *G*, that is, the number of desired groups. The matching matrix has to have edges between all of the elements belonging to M_k, that is:

$$X_{i,j} = 1 \ \forall \ V_i, \ V_j \in M_k \tag{10}$$

Furthermore, the submatches are non-overlapping, in the sense that no edges are allowed to exist between these substructures:

$$X_{i,j} = 0 \ \forall \ V_i \in M_k, \ V_j \in M_l, \ k \neq l \tag{11}$$

The total number of these substructures is *N*/*G* in the matching solution. Supplementary Fig. S3 shows the matching problem in the ARN-509/MDV3100-experiment with the desired number of groups *G* = *5*, which illustrates the increase in computational complexity as the number of edges within a submatch increases per binomial coefficients. The optimal matching problem is:

$$\min_X \sum_i \sum_j X_{i,j} D_{i,j} \tag{12}$$

The optimization problem in equation (12) is used to identify the matching matrix ***X*** that minimizes the sums of distances that fulfill the constraints in equations (1–5) for a given distance matrix ***D*** with desired submatch size *G*. These identified submatches may then be used to allocate the intervention groups (Fig. 2), with possible additional constraints as described in Supplementary Methods.

Mixed type baseline information in the matching. We used categorical variables alongside numerical variables in the matching problem. We divided this into two options: (i) *relaxed*, where the categorical information increments distance at $D_{i,j}$ by a certain amount if the two observations *i* and *j* originated from different categorical labels, and (ii) *strict*, where observations with separate categorical labels may never be matched by setting their relative distance to infinity ($D_{i,j} = \infty$). Observations of *relaxed* type may be part of the same submatch even if they have different labels, provided that their similarity in the numerical dimensions dominates over the categorical difference. Whether or not this happens, depends on the chosen distance metric (Supplementary Table S1); for example, the Gower's dissimilarity[38] is a popular choice for combining mixed type information, but also other metrics have been proposed[39–41]. In the *strict* approach, two observations with different categorical labels may never be part of the same submatch, and therefore this option eliminates a large fraction of possible solutions by limiting the search to a smaller solution space due to infinite values in ***D***. This approach also forces each intervention group to contain an equal number of members from each sub-strata.

Branch and bound algorithm (exact optimization). The number of possible ***X*** binary matching matrix solutions that fulfill the constraints set in equations (1–5) increases exponentially as a function of the number of individuals participating in the matching. For detection of the global optimum of equation (12) in the discrete optimization task, a branch and bound algorithm relies on implicit exhaustive enumeration of all possible combinations in a tree-like structure. Within this structure, however, massive amounts of solutions are omitted based on knowledge that the omitted solutions could theoretically not be better than the current best found solution. If a branch of solutions may include a solution better than the current best found solution, it has to be searched through enumeration. The algorithm itself may be depicted as traversing a tree-like structure using alternating steps called the branching step that expands the current search tree, and the bounding steps that omit large non-optimal areas of the search tree (Supplementary Fig. S7c,d). These *branch* and *bound* steps are described in detail in our Supplementary Material, along with an alternative heuristic Genetic Algorithm (GA) that provides a faster non-exact optimization alternative for large studies.

Matched mixed-effects modeling of treatment effects. In order to evaluate the effect of interventions in a longitudinal study, we assumed that the response variable *y* (e.g. PSA concentration) for the *i*:th tumor from the intervention group g_1 or g_2 grows according to the following linear model:

$$y_{i,g,t} = \beta_0 + \beta_1 \cdot x_t + \beta_2 \cdot g_2 \cdot x_t + \gamma_{i,0} + \gamma_{i,1} \cdot x_t + e_{i,g,t} \tag{13}$$

Here, variable $x_t \in \mathrm{N}_0$ indicates the *t*:th time point in the study (e.g. day or week since starting the interventions). Fixed effects β_0 and β_1 correspond to be population based effects, where b_0 models the initial average response value at baseline ($x_t = 0$), and b_1 models the longitudinal expected linear growth pattern of the tumor, while β_2 includes a binary indicator g_2 which obtains value 1 if the *i*:th tumor belongs to the group g_2 and 0 otherwise. Random effects $\gamma_{i,0}$ and $\gamma_{i,1}$ model variation of the *i*:th individual in the initial response levels or in the growth rate patterns, respectively, and these are analogous to the fixed effects β_0 and β_1. We modeled random noise with the error term $e_{i,t}$, and the error and random effects are assumed to be normally distributed:

$$e_{i,t} \sim \mathrm{N}(0, \sigma_e), \qquad u_{i,0} \sim \mathrm{N}(0, \sigma_{\gamma,0}), \qquad u_{i,1} \sim \mathrm{N}(0, \sigma_{\gamma,1}) \tag{14}$$

The unmatched model in equation (13) does not incorporate supporting prognostic matching information beyond the baseline levels of the main response *y*, although tailored modeling approaches exist for similarly formulated models[42,43]. Therefore, we propose a matched mixed-effects model, which incorporates the matching information obtained from the matching of pairs {*i,j*} before the interventions:

$$y_{i,g1,t} – y_{j,g2,t} = y_{\{i,j\},t} \tag{15}$$

where the submatched individuals *i* and *j* have been allocated to different intervention arms g_1 and g_2 as described in (Figs 1 and 2). The resulting time point specific pairwise observations are then modeled longitudinally using a mixed-effects model:

$$y_{\{i,j\},t} = \beta_{intercept} + \beta_{slope} \cdot x_t + \beta_{intervention} \cdot x_t + \gamma_{\{i,j\},0} + \gamma_{\{i,j\},1} \cdot x_t + e_{\{i,j\},t} \tag{16}$$

where by default we propose setting $\beta_{intercept} = 0$ and $\beta_{slope} = 0$ due to their redundancy in the matched curves (see Fig. 3c bottom panel for the visual interpretation). While γ_0 effectively models the baseline ($x_t = 0$) individual level random intercept for the response variable *y*, the model term γ_1 allows pairwise variation in the growth slopes. This allows prognostic inference for the population effects, especially for the inter-group growth difference in the fixed effect $\beta_{intervention}$, since additional baseline experimental factors are incorporated through the matching {*i,j*}.

The mixed-effects model fitting was performed using the *lme4*-package[44] in the R statistical software[45]. In Table 1, the *p*-values for fixed effects β were computed using Satterthwaite's approximation for degrees of freedom using the *lmerTest*-package[46], while significances of random effects *u* can be tested using log-likelihood ratio tests as proposed in literature[47]. The concept of matching-based mixed-effects modeling is presented in Fig. 3. Example *Unmatched* equation (13) and *Matched* equation (16) model fits are shown for the Control vs. MDV3100 comparison (Single submatch visualized in Fig. 3a of the total 15 pairs). Due to incorporating prognostic submatch-information to the modeled curves (Fig. 3a,b), the matched inference resulted in an increase in sensitivity (Table 1, Fig. 4). Complete visualizations of the model fits are given in the Supplementary Figs S5 and S6. Interestingly, prognostic accuracy in the intervention testing was most likely allowed by the pairing of similar curves in ORX+Tx vs ORX testing (Supplementary Fig. S6b), where the matched curves retained approximately linear trends despite the lack of an early PSA nadir.

Power simulations from experimental datasets. Power analysis is important to ensure statistical validity of the experimental findings. So far, reliable resources have not been available for the preclinical studies in which the experiments pose a number of specific requirements, namely the complex nature of longitudinal responses, right-censoring occurs due to death of animals, limited number of individuals, batch-wise effects, and multivariate baseline characteristics. We addressed these challenges by offering a sampling based power analysis tool that samples individuals with replacement (bootstrapping) from a pre-fitted mixed-effects model, and then re-fits the specified statistical model to the sampled datasets. The method then provides a power curve as a function of *N* in respect to each of the tested population hypotheses. We draw inspiration for this simulation approach from literature[48], although we propose sampling by bootstrapping the data, rather than based on the mixed-effects model parameters.

There are a number of advantages in evaluating the power of a study through simulations: (**i**) Data based simulations do not force the experimenter to perform an expert guess on an often non-intuitive model parameter and its variance to assess required sample amounts. Instead, the experimenter may provide artificial data, e.g. data observed in literature or in pilot studies. This approach is drastically more concrete and expert curated approach to the task. (**ii**) By sampling observations from a pre-fitted mixed-effects model, our approach offers possibility to incorporate also indirect effects, such as censoring due to unexpected death of animals during the study, which may be otherwise difficult or impossible to infer directly for the model parameters. (**iii**) The sampling function relies on a readily fitted mixed-effects model for data input, automatically identifies a suitable sampling unit, and then re-fits the statistical model to the sampled datasets. This feature requires the experimenter to readily specify tested population hypotheses and the structure of the mixed-effects model already in the design phase of the experiment. By requiring such pre-experiment coordination of the tested hypotheses and pre-specified structure of the model, our method encourages well specified á priori hypotheses.

Ethics Statement. All mice were handled in accordance with the institutional animal care policies of the University of Turku (Turku, Finland). The animals were specific pathogen-free, fed with complete pelleted chow and tap water *ad libitum* in a room with controlled light (12 h light, 12 h darkness) and temperature (21 ± 1 °C). The two animal experiments were approved by the Finnish Animal Ethics Committee (licenses

ESAVI/1993/04.10.03/2011 and ESAVI/7472/04.10.03/2012). The institutional policies on animal experimentation fully meet the international requirements as defined in the NIH Guide on animal experimentation. Supplementary Methods provide further details of the intervention experiments and Supplementary Note *the ARRIVE guideline checklist* for the two animal studies.

References

1. Collins, F. S. & Tabak, L. A. Policy: NIH plans to enhance reproducibility. *Nature* **505,** 612–613, doi: 10.1038/505612a (2014).
2. Henderson, V. C. *et al.* A meta-analysis of threats to valid clinical inference in preclinical research of sunitinib. *Elife*. **4,** 1–13, doi: 10.7554/eLife.08351 (2015).
3. Begley, C. G. & Ellis, L. M. Drug development: Raise standards for preclinical cancer research. *Nature* **483,** 531–533, doi: 10.1038/483531a (2012).
4. Freedman, L. P., Cockburn, I. M. & Simcoe, T. S. The economics of reproducibility in preclinical research. *PLoS Biol* **13,** e1002165, doi: 10.1371/journal.pbio.1002165 (2015).
5. Singh, M. & Ferrara, N. Modeling and predicting clinical efficacy for drugs targeting the tumor milieu. *Nat Biotechnol* **30,** 648–657, doi: 10.1038/nbt.2286 (2012).
6. Couzin-Frankel, J. When mice mislead. *Science* **342,** 922–925, doi: 10.1126/science.342.6161.922 (2013).
7. Perrin, S. Preclinical research: make mouse studies work. *Nature* **507,** 423–425, doi: 10.1038/507423a (2014).
8. Henderson, V. C., Kimmelman, J., Fergusson, D., Grimshaw, J. M. & Hackam, D. G. Threats to validity in the design and conduct of preclinical efficacy studies: a systematic review of guidelines for *in vivo* animal experiments. *PLoS Med* **10,** e1001489, doi: 10.1371/journal.pmed.1001489 (2013).
9. Cressey, D. UK funders demand strong statistics for animal studies. *Nature* **520,** 271–272, doi: 10.1038/520271a (2015).
10. Macleod, M. Why animal research needs to improve. *Nature* **477,** 511, doi: 10.1038/477511a (2011).
11. Muhlhausler, B. S., Bloomfield, F. H. & Gillman, M. W. Whole animal experiments should be more like human randomized controlled trials. *PLoS Biol* **11,** e1001481, doi: 10.1371/journal.pbio.1001481 (2013).
12. Hildebrand, F. *et al.* Inflammation-associated enterotypes, host genotype, cage and inter-individual effects drive gut microbiota variation in common laboratory mice. *Genome Biol* **14,** R4, doi: 10.1186/gb-2013-14-1-r4 (2013).
13. Hasty, A. H. & Gutierrez, D. A. What have we really learned about macrophage recruitment to adipose tissue? *Endocrinology* **155,** 12–14, doi: 10.1210/en.2013-2027 (2014).
14. Reardon, S. A mouse's house may ruin experiments. *Nature* **530,** 264, doi: 10.1038/nature.2016.19335 (2016).
15. Kilkenny, C., Browne, W. J., Cuthill, I. C., Emerson, M. & Altman, D. G. Improving bioscience research reporting: the ARRIVE guidelines for reporting animal research. *PLoS Biol* **8,** e1000412, doi: 10.1371/journal.pbio.1000412 (2010).
16. Landis, S. C. *et al.* A call for transparent reporting to optimize the predictive value of preclinical research. *Nature* **490,** 187–191, doi: 10.1038/nature11556 (2011).
17. Baker, D., Lidster, K., Sottomayor, A. & Amor, S. Two years later: journals are not yet enforcing the ARRIVE guidelines on reporting standards for pre-clinical animal studies. *PLoS Biol* **12,** e1001756, doi: 10.1371/journal.pbio.1001756 (2014).
18. Sugar, E., Pascoe, A. J. & Azad, N. Reporting of preclinical tumor-graft cancer therapeutic studies. *Cancer Biol Ther* **13,** 1262–1268, doi: 10.4161/cbt.21782 (2012).
19. Eisen, J. A., Ganley, E. & MacCallum, C. J. Open science and reporting animal studies: who's accountable? *PLoS Biol* **12,** e1001757, doi: 10.1371/journal.pbio.1001757 (2014).
20. Day, C. P., Merlino, G. & Van Dyke, T. Preclinical mouse cancer models: a maze of opportunities and challenges. *Cell* **163,** 39–53, doi: 10.1016/j.cell.2015.08.068 (2015).
21. Su, Z. Optimal allocation of prognostic factors in randomized preclinical animal studies. *Drug Inf J* **45,** 725–729, doi: 10.1177/009286151104500508 (2011).
22. Kasturi, J. *et al.* IRINI: random group allocation of multiple prognostic factors. *Contemp Clin Trials* **32,** 372–381, doi: 10.1016/j.cct.2010.12.009 (2011).
23. Knuuttila, M. *et al.* Castration induces up-regulation of intratumoral androgen biosynthesis and androgen receptor expression in an orthotopic VCaP human prostate cancer xenograft model. *Am J Pathol* **184,** 2163–2173, doi: 10.1016/j.ajpath.2014.04.010 (2014).
24. Hutchinson, L. & Kirk, R. High drug attrition rates—where are we going wrong? *Nat Re Clin Oncol* **8,** 189–190, doi: 10.1038/nrclinonc.2011.34 (2011).
25. Laajala, T. D. *hamlet*: hierarchical optimal matching and machine learning toolbox (R-package version 0.9.5). URL http://CRAN.R-project.org/package=hamlet (2016).
26. Treasure, T. & MacRae, K. D. Minimisation: the platinum standard for trials? Randomisation doesn't guarantee similarity of groups; minimisation does. *BMJ* **317,** 362–363 (1998).
27. Smith G. D. & Ebrahim S. Data dredging, bias, or confounding : They can all get you into the BMJ and the Friday papers. *BMJ* **325,** 1437–1438 (2002).
28. Hoenig, J. M. & Heisey, D. M. The abuse of power: the pervasive fallacy of power calculations for data analysis. *Am Stat* **55,** 19–24, doi: 10.1198/000313001300339897 (2001).
29. Kimmelman, J., Mogil, J. S. & Dirnagl, U. Distinguishing between exploratory and confirmatory preclinical research will improve translation. *PLoS Biol* **12,** e1001863, doi: 10.1371/journal.pbio.1001863 (2014).
30. van der Worp, H. B. *et al.* Can animal models of disease reliably inform human studies? *PLoS Med* **7,** e1000245, doi: 10.1371/journal.pmed.1000245 (2010).
31. Seashore-Ludlow, B. *et al.* Harnessing connectivity in a large-scale small-molecule sensitivity dataset. *Cancer Discov* **5,** 1210–1223, doi: 10.1158/2159-8290.CD-15-0235 (2015).
32. Wessel, J. & Schork, N. J. Generalized genomic distance-based regression methodology for multilocus association analysis. *Am J Hum Genet* **79,** 792–806, doi: 10.1086/508346 (2006).
33. Pond, G. R., Tang, P. A., Welch, S. A. & Chen, E. X. Trends in the application of dynamic allocation methods in multi-arm cancer clinical trials. *Clin Trials* **7,** 227–234, doi: 10.1177/1740774510368301 (2010).
34. Greevy, R., Lu, B., Silver, J. H. & Rosenbaum, P. Optimal multivariate matching before randomization. *Biostatistics* **5,** 263–275, doi: 10.1093/biostatistics/5.2.263 (2004).
35. Lu, B., Greevy, R., Xu, X. & Beck, C. Optimal nonbipartite matching and its statistical applications. *Am Stat* **65,** 21–30, doi: 10.1198/tast.2011.08294 (2011).
36. Pond, G. Statistical issues in the use of dynamic allocation methods for balancing baseline covariates. *Br J Cancer* **104,** 1711–1715, doi: 10.1038/bjc.2011.157 (2011).
37. Lachin, J. M., Matts, J. P. & Wei, L. J. Randomization in clinical trials: conclusions and recommendations. *Control Clin Trials* **9,** 365–374 (1988).
38. Gower, J. C. A general coefficient of similarity and some of its properties. *Biometrics* **27,** 857–871 (1971).
39. McCane, B. & Albert, M. Distance functions for categorical and mixed variables. *Pattern Recognit Lett* **29,** 986–993, doi: 10.1016/j.patrec.2008.01.021 (2008).
40. Wilson, D. R. & Martinez, T. R. Improved heterogeneous distance functions. *J Artif Intell Res* **6,** 1–34, doi: 10.1613/jair.346 (1997).

41. Guojun, G., Chaoqun, M. & Jianhong, W. Similarity and dissimilarity measures in *Data clustering: theory, algorithms, and applications* (eds. Wells, M. T. *et al.*) Ch. 6, 67–106 (ASA-SIAM, 2007).
42. Zhao, L. *et al.* Bayesian hierarchical changepoint methods in modeling the tumor growth profiles in xenograft experiments. *Clin Cancer Res* **17,** 1057–1064, doi: 10.1158/1078-0432.CCR-10-1935 (2011).
43. Laajala, T. D. *et al.* Improved statistical modeling of tumor growth and treatment effect in preclinical animal studies with highly heterogeneous responses *in vivo*. *Clin Cancer Res* **18,** 4385–4396, doi: 10.1158/1078-0432.CCR-11-3215 (2012).
44. Bates, D. M., Maechler, M. & Bolker, B. *lme4*: Linear mixed-effects models using S4 classes (R-package version 1.1-6). URL http://CRAN.R-project.org/package=lme4 (2014).
45. R Development Core Team. *R*: a language and environment for statistical computing (version 3.2.1). R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org (2015).
46. Kuznetsova, A. *lmerTest*: Tests for random and fixed effects for linear mixed effect models (R-package version 2.0-6). URL http://CRAN.R-project.org/package=lmerTest (2014).
47. Pinheiro, J. C. & Bates, D. M. Hypothesis tests and confidence intervals in *Mixed effects models in S and S-PLUS* (eds. Chambers, J. *et al.*) Ch. 2.4, 82–96 (Springer-Verlag, 2000).
48. Gelman, A. & Hill, J. Sample size and power calculations in *Data analysis using regression and multilevel/hierarchical models* (eds. Alvarez, R. M. *et al.*) Ch. 20, 437–454 (Cambridge University Press, 2007).

Acknowledgements

The authors thank Nina Messner, Heli Niittymäki, Katri Hovirinta, Paulina Chrusciel and Taina Kirjonen for their expert technical assistance in the ARN-509 and MDV3100 animal experiments, and Emmy Verschuren and Jennifer Devlin for their useful comments on the manuscript. This work was funded by Turku Drug Research Doctoral Programme (T.D.L., R.H., M.K., R.O.), Turku Doctoral Programme of Molecular Medicine (A.K.), the Finnish Cultural Foundation (T.D.L., M.K.), Academy of Finland (grants 8217676 and 122546 to J.W.; 269862, 272437, 292611, 279163 and 295504 to T.A.; 264153 to S.M.; 13856 to M.P.), the Sigrid Juselius Foundation (J.W., M.P.), National Cancer Institute (16X064), the Foundation for the Finnish Cancer Institute (J.W.), and the Finnish Cancer Organisations (M.P., T.A.). The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Author Contributions

T.D.L., R.H., A.K., M.K., E.A., R.O., J.W., S.M. and M.P.: conceived and designed the animal experiments; T.D.L., R.H., A.K. and M.K.: performed the animal experiments; T.D.L. and T.A.: conceived and designed the statistical analyses; T.D.L.: implemented the open source R-package; T.D.L., M.J. and V.F.: implemented the web-based GUI; T.D.L., R.H., A.K., M.K., M.P. and T.A.: analysed the data; T.D.L., J.W., M.P. and T.A.: wrote the manuscript. All authors read and approved the final manuscript.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: M.P. is the director of Turku Center for Disease Modeling (TCDM), providing preclinical mouse models including statistical analyses of drug interventions. All other authors have declared no competing interests.

How to cite this article: Laajala, T. D. *et al.* Optimized design and analysis of preclinical intervention studies *in vivo*. *Sci. Rep.* **6**, 30723; doi: 10.1038/srep30723 (2016).

Supplementary Material (Publication I):

Supplementary Figures S1 - S11

Supplementary Table S1

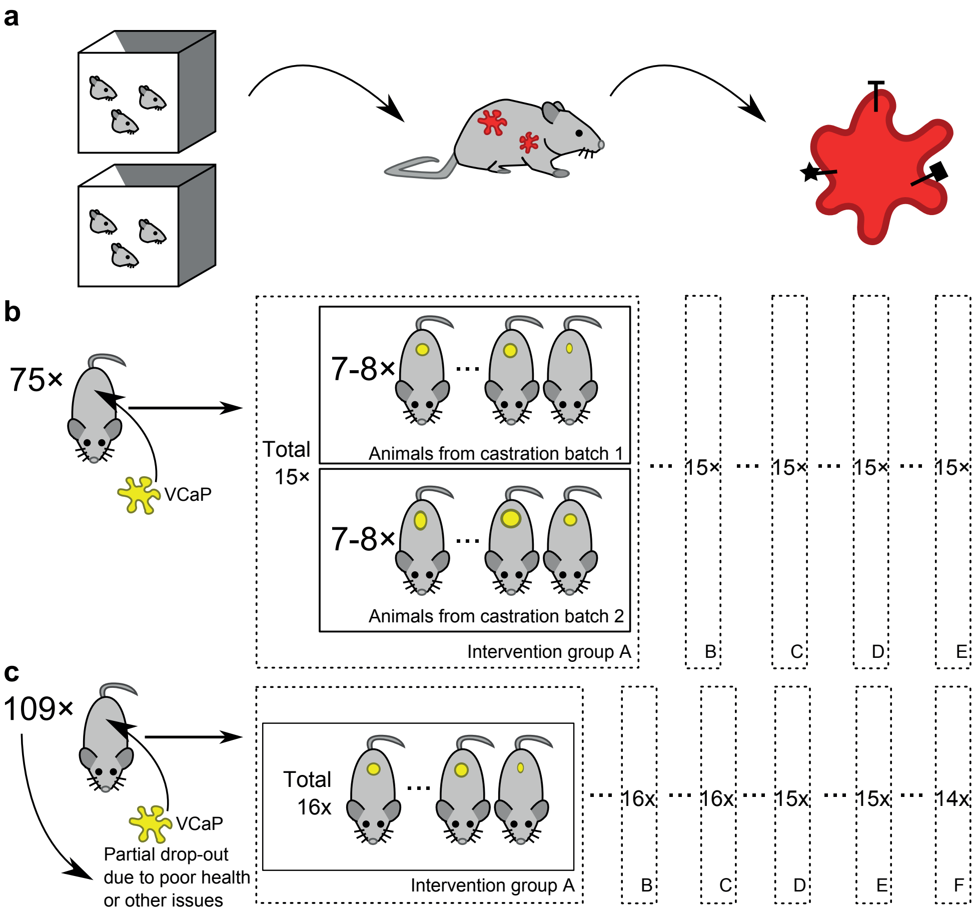
Supplementary Methods: Additional details for the preclinical experiments and algorithms. (*Available online*)*

Supplementary Note S1: Hamlet R-package: step-by-step user instructions. (*Available online*)*

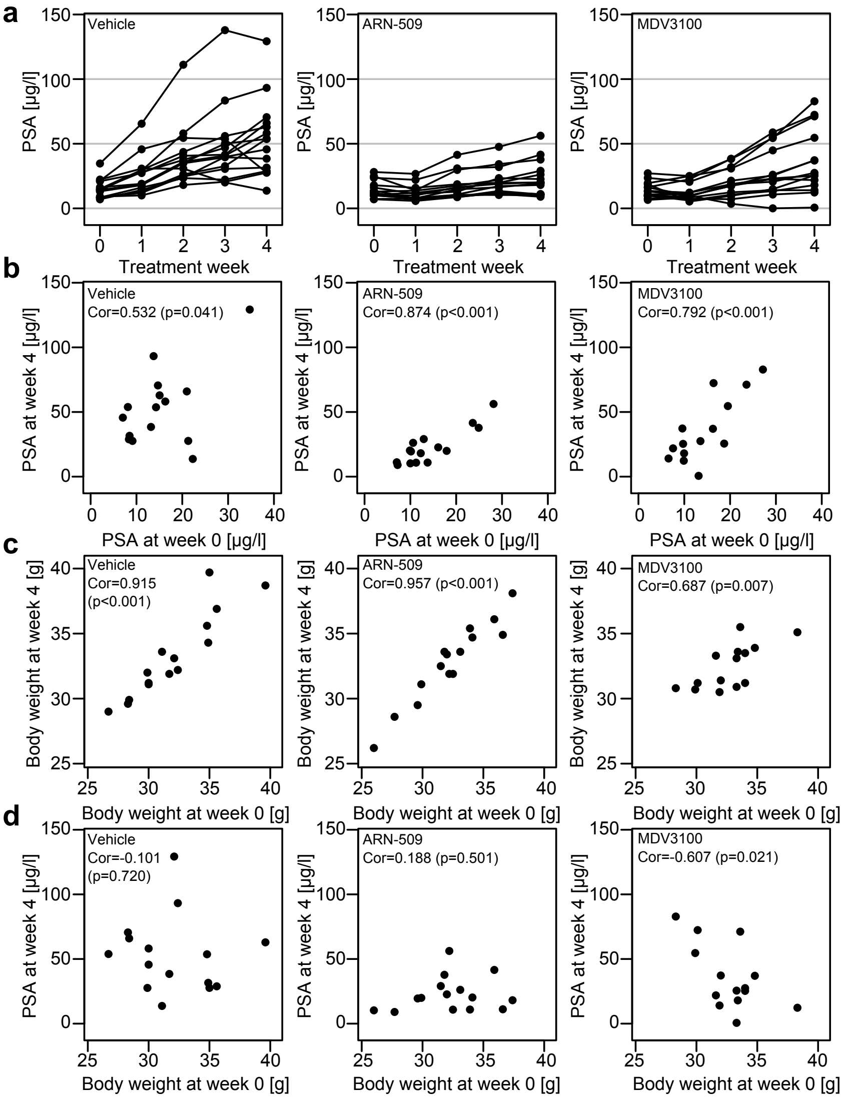
Supplementary Note S2: R-vivo user instructions. (*Available online*)*

Supplementary Note S3: The ARRIVE checklist for the two animal treatment studies. (*Available online*)*

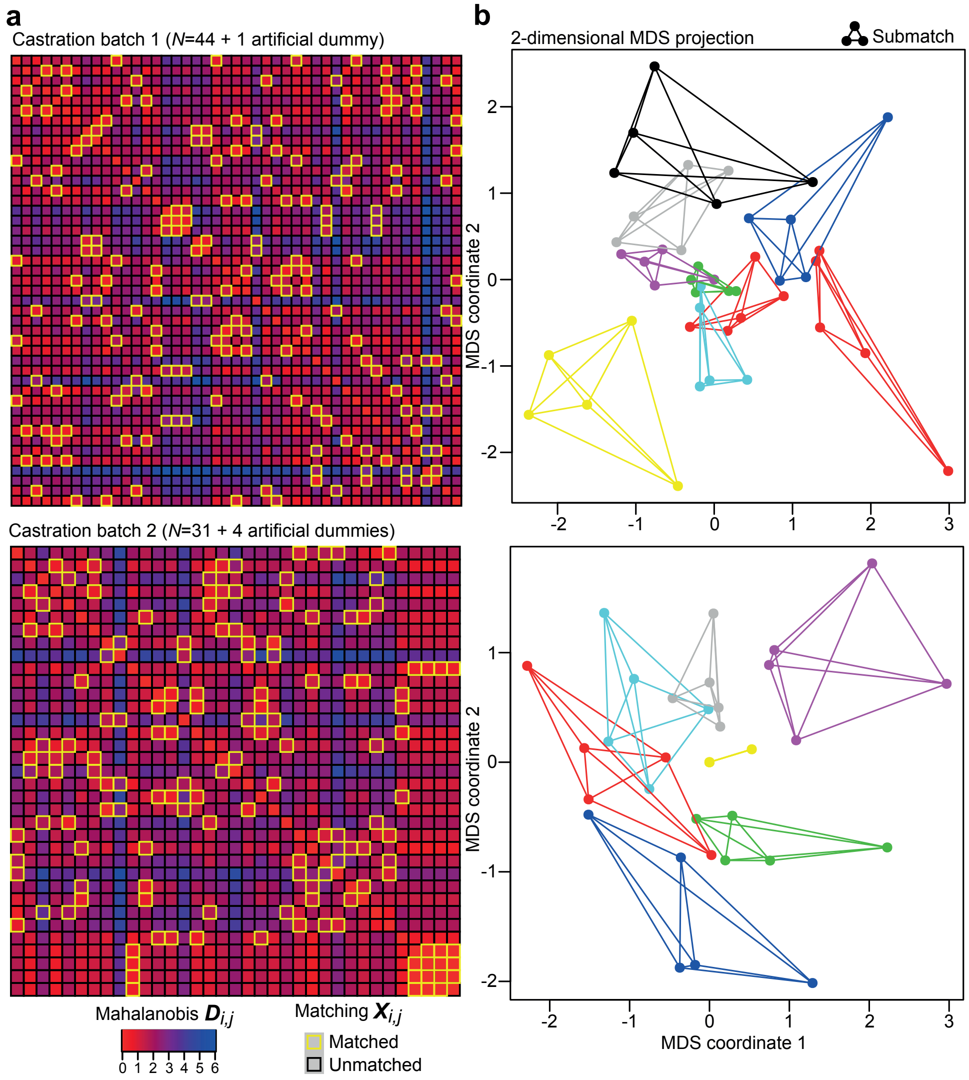
* : `https://www.nature.com/articles/srep30723#supplementary-information`



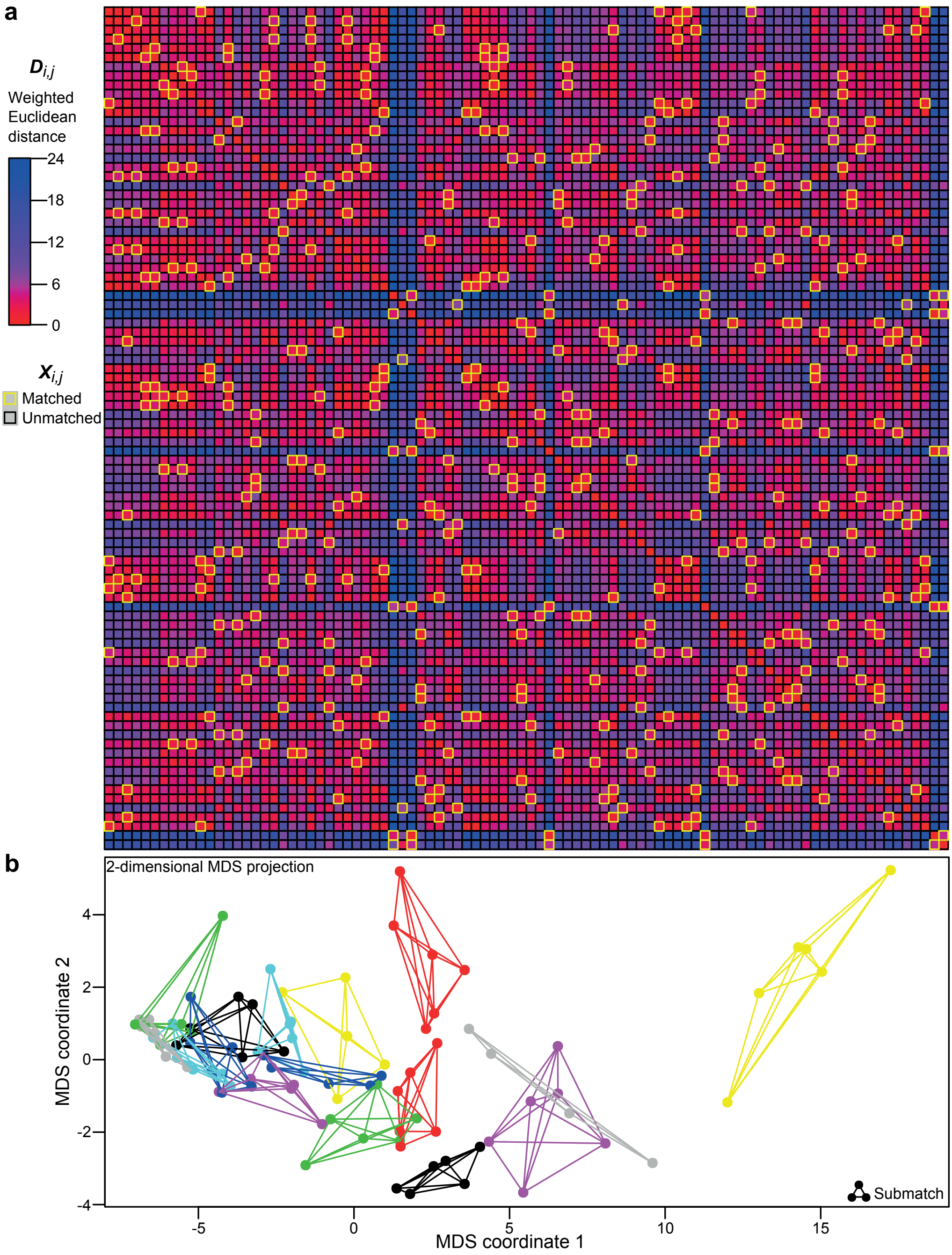
Supplementary Figure S1: Schematic illustrations of the common hierarchical structures that need to be taken into account in animal allocation and the experiment designs of the two case studies. (**a**) Preclinical cancer studies commonly incorporate a layered hierarchical design, where multiple nested animals may originate from a single batch or a cage, while multiple tumors may be located in a single animal. (**b**) ARN-509/MDV3100 -intervention study with orthotopic VCaP prostate cancer cells in male immunodeficient mice (HSD: Athymic Nude Foxn 1nu). According to the experimental procedure, orthotopic tumors were generated by injecting the cancer cells into the prostate of each animal. The growth of the tumors was followed by weekly measurements of the serum PSA indicating the tumor burden. The animals were castrated in two separate batches on subsequent weeks, resulting in two substrata with different tumor growth characteristics. The mice were followed by serum PSA measurements and after the re-appearance of the tumors the mice were allocated into several CRPC treatment arms . Hierarchical allocation procedure based on the global matching algorithm ensures that the substrata are evenly distributed among the intervention groups. (**c**) ORX/ORX+Tx -intervention study with analogous subcutaneous VCaP xenografts. A single substrata of animals was allocated into several intervention groups (out of which Control, ORX and ORX+Tx are presented in this paper), while some animals had to be dropped out due to ethical reasons.



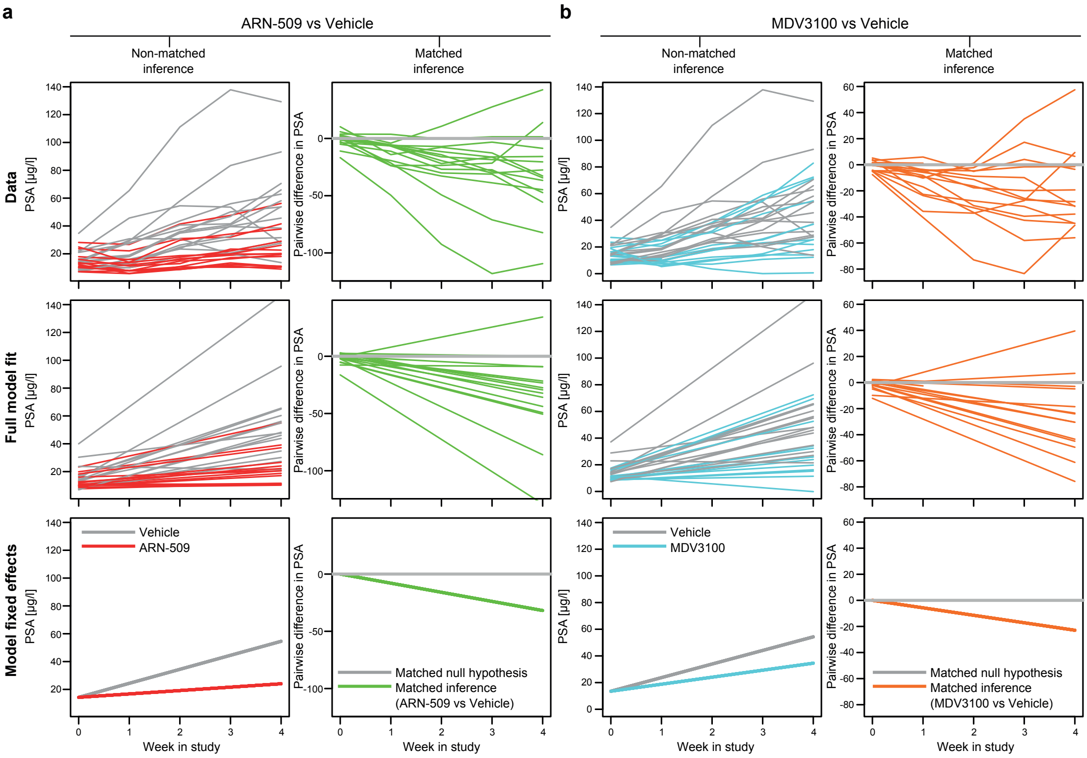
Supplementary Figure S2: Experimental data of the VCaP study. (**a**) Two selected treatment alternatives, ARN-509 ($n = 15$) and MDV3100 ($n = 15$), were compared to the vehicle group ($n = 15$). (**b**) As expected, the initial PSA level at baseline was predictive of the PSA value measured after 4 weeks of treatment. (**c**) Similarly, body weights of the animals at baseline were correlated with the body weights after 4 weeks of treatment. (**d**) Interestingly, the initial body weight showed a borderline inverse correlation with the PSA level after 4 weeks of treatment in the MDV3100 group (p=0.021), while this relationship was not seen in the other groups. Such a multivariate association between the treatment response (final PSA) and an initial animal characteristic (body weight at baseline) would be missed with simple univariate animal matching procedures.



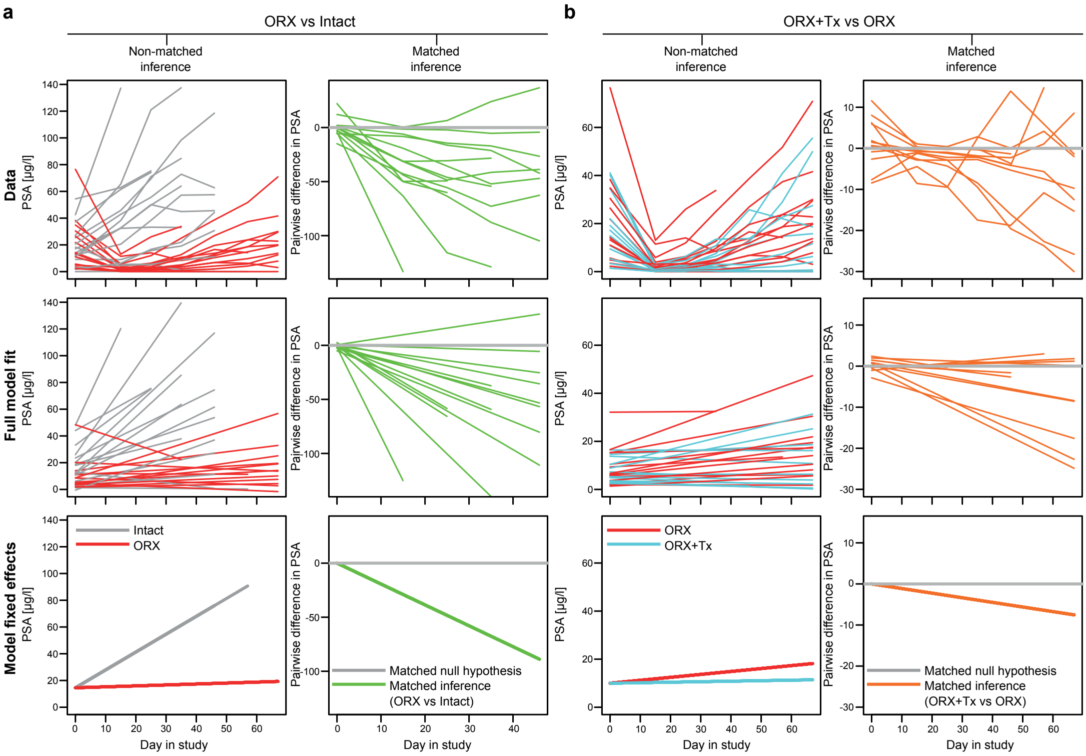
Supplementary Figure S3: Solving the non-bipartite submatching problem in the MDV3100/ARN-509 intervention study. (**a**) The animals ($n = 75$) were divided to two different castration sub-strata, which were separately submatched only within a strata and subsequently allocated evenly to the intervention arms (see **Supplementary Fig. S1b**). The matrix colors indicate dissimilarities in the baseline characteristics, and the box color indicates animals being part of same submatch. (**b**) Multidimensional Scaling (MDS) 2-dimensional projection of the complex baseline characteristics, with each submatch indicated with connecting edges and different coloring.



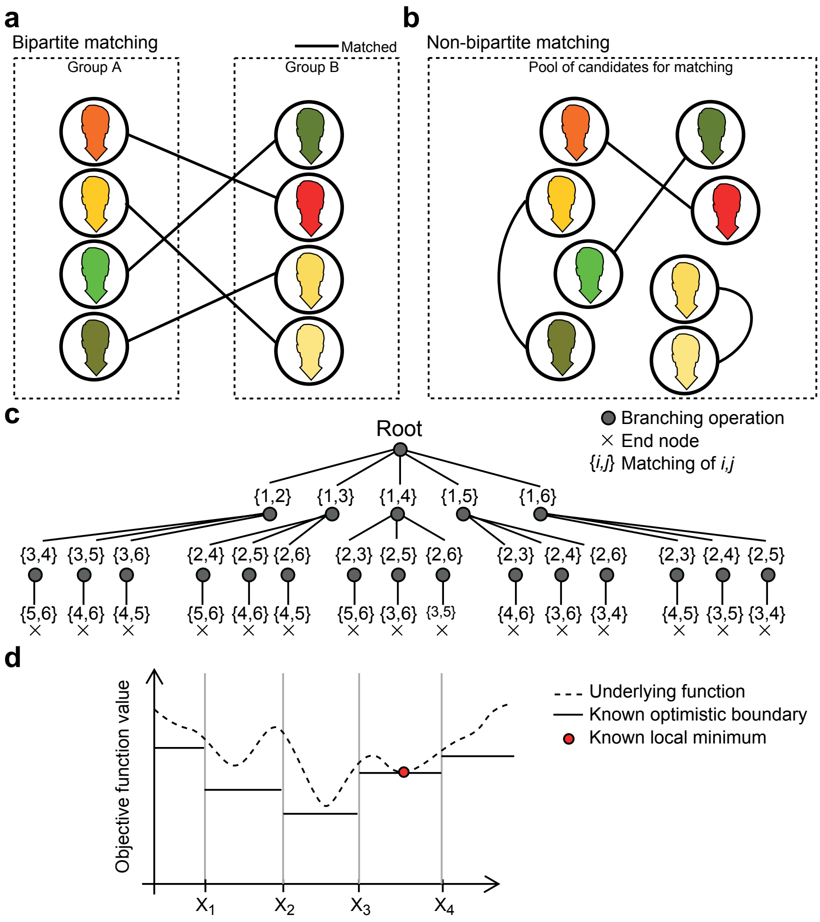
Supplementary Figure S4: Solving the non-bipartite submatching problem in the ORX/ORX+Tx intervention study. (**a**) The animals ($n = 109$) were matched to submatches of size 6, and subsequently allocated to different intervention arms within each submatch. Only three of the intervention groups are analyzed here (Control, ORX, ORX+Tx). The matrix colors indicate dissimilarities in the baseline characteristics, and the box color indicates two animals being part of same submatch. (**b**) Multidimensional Scaling (MDS) 2-dimensional projection of the complex baseline characteristics, with each submatch indicated with connecting edges and different coloring.



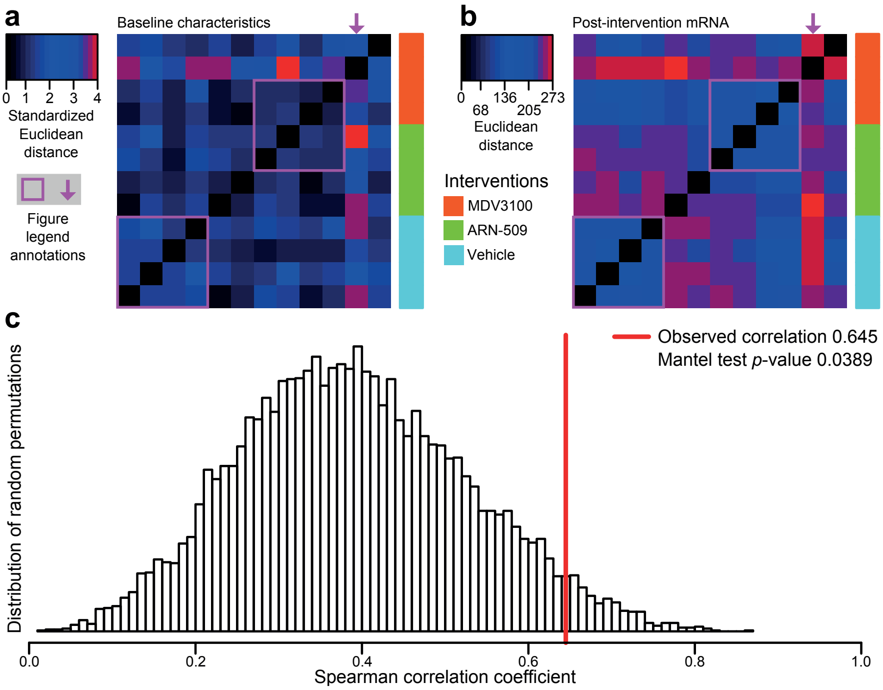
Supplementary Figure S5: Mixed-effects model fits in the ARN-509/MDV3100 intervention study. Top panel: response data; middle panel: full model fit; bottom panel: fixed effects fit. (**a**) ARN-509 versus Vehicle. Left panel: unmatched inference; right panel: matched inference. (**b**) MDV3100 versus Vehicle. Left panel: unmatched inference; right panel: matched inference. Model coefficient estimates, standard deviations and *p*-values are presented in **Table 1**.



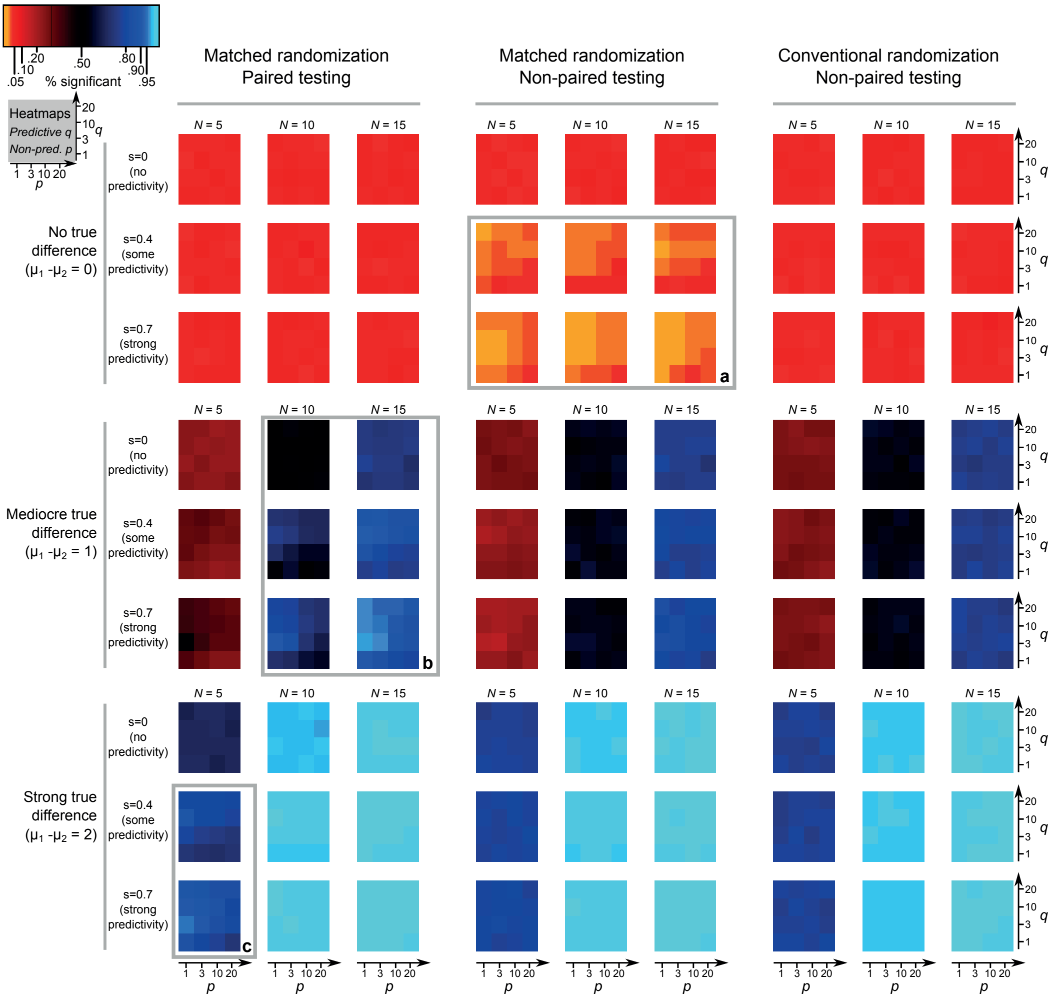
Supplementary Figure S6: Mixed-effects model fits in the ORX/ORX+Tx intervention study. Top panel: response data; middle panel: full model fit; bottom panel: fixed effects fit. (**a**) ORX versus Control. Left panel: unmatched inference; right panel: matched inference. (**b**) ORX+Tx versus ORX. Left panel: unmatched inference; right panel: matched inference. Model coefficient estimates, standard deviations and p-values are presented in **Table 1**.



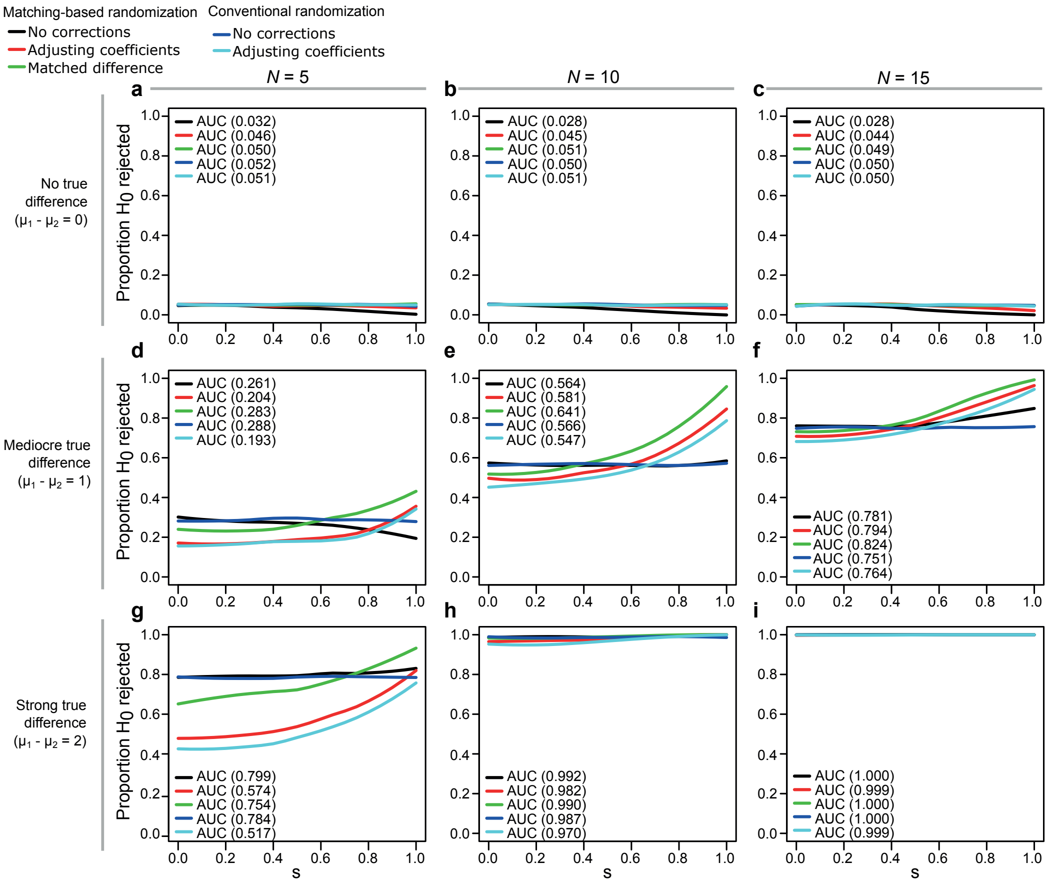
Supplementary Figure S7: The difference between bipartite and non-bipartite matching, and a graphical representation of the steps in the branch and bound algorithm for solving the non-bipartite problem. (**a**) A bipartite matching problem, where the matching is identified between two pre-defined groups. (**b**) In the preclinical cancer context, the non-bipartite matching enables detection of comparable individuals from a single pool of animals, based on similarities in their baseline characteristics. (**c**) Branching implicitly enumerates all possible combinations of matches in the solution. In this particular example, the branching structure is presented for matching of pairs ($G = 2$) for 6 individuals. (**d**) Concept of the bounding function in a continuous minimization task (lower objective function values are preferred). A bounding function is utilized to discard branches in the tree-like structure (panel **c**), by concluding that a certain range (branch) of solutions cannot improve the current best solution. In this example, ranges $x \leq X_1$ and $X_4 \leq x$ do not have to be searched, as the bounding function hints that the current best solution (indicated in red) cannot be improved in this solution range. However, solutions in $X_1 \leq x \leq X_2$ and $X_2 \leq x \leq X_3$ have to be tested, since the bounding function suggests a possible lower theoretical boundary in these solution ranges.



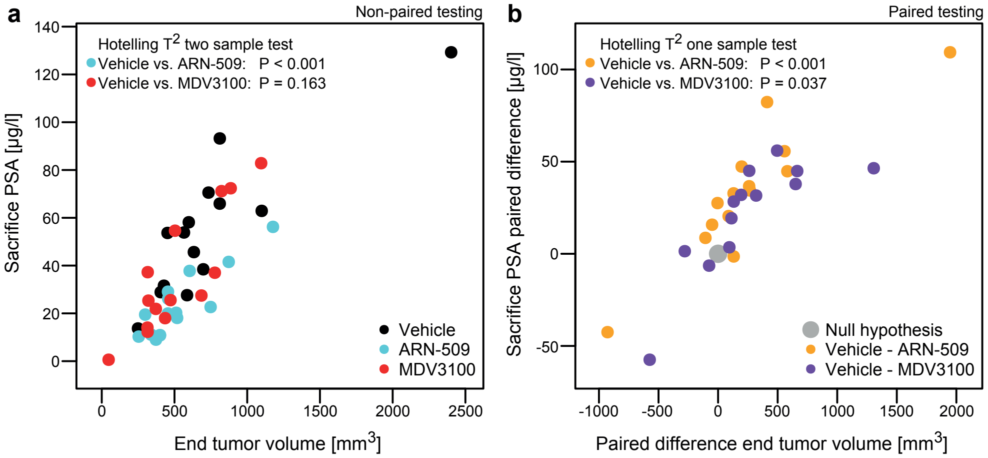
Supplementary Figure S8: Evaluation of the animal allocations in the ARN-509 / MDV3100 VCaP xenograft study using Mantel's test that compares the pre-intervention dissimilarity matrices of the baseline animal characteristics to the post-intervention mRNA gene expression profiles of the treated tumors. By visual inspection, two interesting dissimilarity sub-groups were identified (pink boxes). Further, one exceptional baseline animal remained an outlier also at the tumor mRNA-level (pink arrow). (**a**) Dissimilarity matrix of the baseline characteristics for the sequenced animals ($n = 12$) was calculated using standardized Euclidean distance. (**b**) Dissimilarity matrix of the RNA-seq expression profiles (fragments per kilobase of exon per million mapped reads, FPKM) was calculated using Euclidean distance. (**c**) Distribution of the permutated correlation statistic. Statistically significant Spearman correlation was observed between the baseline characteristics and post-intervention mRNA expression (red line), by conducting $n = 10,000$ permutations of the dissimilarity matrices (Mantel's test).



Supplementary Figure S9: A simulation run of 1,000 matched 2-group datasets were generated for each combination in the parameter grid, resulting in a total of 432,000 datasets for which matching was conducted and data drawn from mulvariate normal distributions with given parameters. The matching procedure was used as in the manuscript, and conventional randomization randomly allocated groups of equal size ignoring baseline information to both experiment groups. Paired or non-paired t-test was used to determine whether there was a difference with $\alpha = 0.05$ significance threshold. The following parameters were varied: Magnitude of true group difference $\mu_1 - \mu_2 \in \{0, 1, 2\}$; Sample size per group $N \in \{5, 10, 15\}$; Magnitude of informativeness in (parameter q) predictive baseline variables $s \in \{0, 0.4, 0.7\}$; Count of predictive baseline variables $q \in \{1, 3, 10, 20\}$; Count of non-predictive baseline variables $p \in \{1, 3, 10, 20\}$. Few interesting key results were annotated in the simulation results: (**a**) Interestingly, when matched allocation was used, the specificity in testing was highly increased in the case when no true group difference was present. This phenomenon persistent in the non-paired testing, highlighting that matching-based allocation also serves to improve specificity and that non-paired testing can be benefit even if the matching information is not utilized in the post-intervention testing. (**b**) A benefit in sensitivity was observed in the small group-wise different $(\mu_1 - \mu_2 = 1)$ in comparison to the non-matched testing as long as the number of predictive markers was greater than non-informative baseline markers $(q \geq p)$. As expected, this advantage was lost if no informative markers were present $(s = 0)$, but no loss of accuracy was observed in comparison to the conventional methods. (**c**) In small explorative studies $(N = 5)$, a slight advantage in sensitivity was observed especially if the baseline markers were highly predictive $(s = 0.7)$, highlighting that predictive markers may help narrow down candidates more effectively in such explorative studies with typically smaller sample sizes N.



Supplementary Figure S10: Simulation results as a function of the predictive s parameter, with default R loess smoothing applied for the visualization of the curves. Positive detection was defined using the conventional significance threshold of $p < 0.05$ for the multiple regression term to test differences between the two simulated groups. The overall performance of each modeling strategy was assessed with the area under curve (AUC) over the whole range of correlation of the covariate with the outcome (s), which summarizes the findings over the whole correlation spectrum both where there was no predictive baseline information (low s) or where the single covariate had strong predictive power (high s), but was confounded by the three additional random confounder-covariates. The three columns indicated the different sample sizes $N \in \{5, 10, 15\}$. (**a-c**) Simulations when no group difference was present. (**d-f**) Mediocre group difference. (**g-i**) Strong group difference.



Supplementary Figure S11: A single end-point testing example from the VCaP ARN-509 / MDV3100 -study. Hotelling's T^2 multivariate extension of the t-test was used to illustrate how two end-point markers can be tested with or without the matching information. In this case the two end-point markers were highly correlated, illustrating that the PSA was a feasible surrogate marker to serve as a proxy for the actual tumor size in the orthotopic VCaP animal model. (**a**) In the non-paired case, MDV3100 was to some extent overlapping with the sacrifice measurements from the Vehicle group. (**b**) Pairing the end-point markers and comparing to the null hypotheis that the multivariate normal distribution $\mu = \{0, 0\}$. The paired adjustment revealed difference between Vehicle and MDV3100, which was consistent with the results observed in the longitudinal PSA analysis (Table 1).

Supplementary Table S1. Distance/dissimilarity measures for capturing similarities between two *d*-dimensional variable vectors $\boldsymbol{x}$ and $\boldsymbol{y}$. The symbol s_i denotes the standard deviation of the associated *i*:th variable; $\boldsymbol{S}$ denotes the $d \times d$ -dimensional covariance-variance matrix computed between the variables, thus incorporating also inter-variable correlations; *R* denotes the range of the variable. Some of the measures can be obtained as special cases of Minkowski or Mahalanobis (listed as footnotes).

Distance measure	**Formula**
Minkowski †	$\left(\sum_{i=1}^{d} \lvert x_i - y_i \rvert^r\right)^{\frac{1}{r}}, r \geq 1$
Mahalanobis	$\sqrt{(\boldsymbol{x}-\boldsymbol{y})\boldsymbol{S}^{-1}(\boldsymbol{x}-\boldsymbol{y})^T}$
Euclidean [a,b,]†	$\sqrt{\sum_{i=1}^{d} (x_i - y_i)^2}$
Standardized Euclidean [c]	$\sqrt{\sum_{i=1}^{d} \left(\frac{x_i}{s_i} - \frac{y_i}{s_i}\right)^2}$
Manhattan [d,]†	$\sum_{i=1}^{d} \lvert x_i - y_i \rvert$
Maximum [e,]†	$\max_{1 \leq i \leq d} \lvert x_i - y_i \rvert$
Gower dissimilarity *	continuous: $\lvert x_i - y_i \rvert / R_i$ binary/categorical: 0 if $x_i = y_i$, 1 otherwise

[a] obtained as a special case of Minkowski when $r = 2$; [b] obtained as a special case of Mahalanobis when *S* is a unit diagonal matrix; [c] obtained as a special case of Mahalanobis when *S* is a diagonal matrix; [d] obtained as a special case of Minkowski when $r = 1$; [e] obtained as a special case of Minkowski when $r \rightarrow \infty$; † is not scale-invariant, thus data normalization should be considered; * Suitable for mixed-type data. Gower's dissimilarity coefficient is obtained by summarizing over all the available variables *i*=1,2,..., *d*.

ORIGINAL PUBLICATION **II**

Laajala TD, Corander J, Saarinen NM, Mäkelä K, Savolainen S, Suominen MI, Alhoniemi E, Mäkelä S, Poutanen M, Aittokallio T. *Improved statistical modeling of tumor growth and treatment effect in preclinical animal studies with highly heterogeneous responses in vivo.* Clinical Cancer Research. 2012;18(16):4385-96.

Clinical Cancer Research

Cancer Therapy: Preclinical

Improved Statistical Modeling of Tumor Growth and Treatment Effect in Preclinical Animal Studies with Highly Heterogeneous Responses *In Vivo*

Teemu D. Laajala[1,6], Jukka Corander[8], Niina M. Saarinen[4,5], Katja Mäkelä[2,4,5], Saija Savolainen[2,5], Mari I. Suominen[7], Esa Alhoniemi[7], Sari Mäkelä[4,5,3], Matti Poutanen[2,5,10], and Tero Aittokallio[1,6,9]

Abstract

Purpose: Preclinical tumor growth experiments often result in heterogeneous datasets that include growing, regressing, or stable growth profiles in the treatment and control groups. Such confounding intertumor variability may mask the true treatment effects especially when less aggressive treatment alternatives are being evaluated.

Experimental design: We developed a statistical modeling approach in which the growing and poorly growing tumor categories were automatically detected by means of an expectation-maximization algorithm coupled within a mixed-effects modeling framework. The framework is implemented and distributed as an R package, which enables model estimation and statistical inference, as well as statistical power and precision analyses.

Results: When applied to four tumor growth experiments, the modeling framework was shown to (i) improve the detection of subtle treatment effects in the presence of high within-group tumor variability; (ii) reveal hidden tumor subgroups associated with established or novel biomarkers, such as ERβ expression in a MCF-7 breast cancer model, which remained undetected with standard statistical analysis; (iii) provide guidance on the selection of sufficient sample sizes and most informative treatment periods; and (iv) offer flexibility to various cancer models, experimental designs, and treatment options. Model-based testing of treatment effect on the tumor growth rate (or slope) was shown as particularly informative in the preclinical assessment of treatment alternatives based on dietary interventions.

Conclusions: In general, the modeling framework enables identification of such biologically significant differences in tumor growth profiles that would have gone undetected or had required considerably higher number of animals when using traditional statistical methods. *Clin Cancer Res; 18(16); 4385–96. ©2012 AACR.*

Introduction

Preclinical tumor growth studies using animal models have a fundamental role in anticancer drug development. Experimental cancer models in mice and rats include, among others, implanting human tumor cells into immunocompromised animals (xenograft models) or inducing tumor-promoting mutations in rodents using carcinogens such as 7,12-dimethylbenz(*a*)anthracene (DMBA). Regardless of the model type, the typical experimental design involves dividing the animals into the treatment groups (representing different doses or treatment combinations), and monitoring the relative effects of the treatments on tumor growth, in comparison with the control group (no treatment). The tumor growth is typically measured at a number of time intervals until the animals die, become moribund, or reach a planned time of sacrifice (1).

Despite careful control of the experiments, the longitudinal tumor growth measurements reflect multiple sources of both biologic and experimental variation that may severely confound the actual treatment responses. Along with measurement noise, additional experimental challenges include missing data points due to animal morbidity, mortality, and quantitation limits, as well as very aggressively growing outlying profiles. Such experimental variation can be compensated to some degree by increasing the number of animals and tumors analyzed. However, due to economical and practical reasons, most

Authors' Affiliations: [1]Department of Mathematics, [2]Departments of Physiology and [3]Cell Biology and Anatomy, Institute of Biomedicine, [4]Functional Foods Forum, [5]Turku Center for Disease Modeling, University of Turku; [6]Turku Centre for Biotechnology; [7]Pharmatest Services Ltd, Turku; [8]Department of Mathematics and Statistics, [9]Institute for Molecular Medicine (FIMM), University of Helsinki, Finland; and [10]Institute of Medicine, The Sahlgrenska Academy, Gothenburg University, Gothenburg, Sweden

Note: Supplementary data for this article are available at Clinical Cancer Research Online (clincancerres.aacrjournals.org).

Current address for S. Savolainen: Orion Pharma Ltd, Turku, Finland.

Corresponding Author: Tero Aittokallio, Institute for Molecular Medicine Finland, University of Helsinki, FI-00014, Finland. Phone: 358-50-318-2426; Fax: 358-9-191-25737; E-mail: tero.aittokallio@fimm.fi

doi: 10.1158/1078-0432.CCR-11-3215

Translational Relevance

Heterogeneous responses observed in many preclinical models of cancer treatment may lead to frequent false-negative results and therefore to ineffective translation of *in vivo* results to clinical trial designs. Using various preclinical animal models, cancer cell lines, and *in silico* simulations, we show here how modeling and exploring of different categories of tumor growth profiles can improve statistical testing and biologic understanding of treatment effects, especially when less aggressive treatment alternatives are being evaluated. Statistical power and precision analyses offer possibilities for further improving the design of the experimental protocols for preclinical assessment of cancer treatments. Taking into account the individual characteristics already in the preclinical stages should also help to propagate information on the intertumor variability to the subsequent clinical studies.

experiments are still being carried out on relatively small sample sizes including less than 10 tumors per group (1). Moreover, even when using genetically standardized and well-characterized animal strains, the experiments often represent substantial between-animal variability, which cannot be controlled simply by increasing the number of animals. Such confounding factors often result in hidden subgroups, which are not predefined but may associate with divergent treatment outcomes in terms of the growth profiles observed over the treatment period. Some tumors may grow aggressively in a treatment group, even if the same treatment inhibits the growth of other tumors, or some untreated tumors do not grow well or even completely regress in the control group (2–8).

The heterogeneous nature of the tumor growth profiles pose severe challenges to the statistical models that typically rely on the assumption that the groups being compared are relatively homogeneous. Many studies have used single end points, such as tumor volume at a prespecified time point or tumor doubling time, together with traditional statistical tests, such as *t* test and ANOVA, or their nonparametric counterparts (5–11). However, such univariate approaches often lead to suboptimal statistical power because of their ineffective use of the longitudinal growth patterns (1, 12). In contrast, repeated measures and regression models use the entire growth profiles and enable more systematic between-group comparisons through model parameters (1). In particular, mixed-effects models have become a convenient approach to model various experimental factors, such as treatment effects or base levels (fixed effects) while accounting for variation expressed by individual animals or tumors (random effects). This model family has successfully been used to analyze specific types of xenograft experiments or study questions (12–17). However, further challenges remain. In particular, the conventional model cannot detect subtle treatment effects in the presence of heterogeneous responses, due to unfeasible model estimation, resulting in skewness or multimodality in the random effects (18).

The present work introduces a novel modeling framework for in-depth statistical analysis of tumor growth experiments in which the underlying tumor heterogeneity is modeled by dividing the longitudinal growth profiles into growing and poorly growing categories within the treatment and control groups. The framework is based on well-established linear mixed-effects models enabling robust estimation and statistical inference of treatment effects through parameters such as tumor growth rates (slopes) or average tumor levels (offset). By means of such elemental parameters that are descriptive of both strong and more subtle modes of tumor growth inhibition, the modeling framework enables the investigator to address a range of questions relevant in many practical settings, such as the degree of dynamic treatment effect on the growth rates, the amount of tumor heterogeneity present in the given data, and how the experimental design should be modified to find significant treatment effects. To promote its widespread application in the future studies, we provide an easy-to-use R implementation with accompanying tools for model visualization and diagnostics. Using 4 tumor growth experiments as application use cases, we show here how the categorizing mixed-effects model enables the extraction of full information from these longitudinal profile datasets.

Materials and Methods

The model was applied to 4 tumor growth experiments, including prostate and breast cancer mouse xenograft models, a syngeneic mammary cancer model with 4T1 mouse mammary tumor cells, and a DMBA-induced mammary carcinoma in the rat. These experiments represent with a wide range of properties encountered in many treatment settings, including various treatment options and dosages (Table 1). Moreover, the experiments included designs with and without a designated target size that the tumors need to reach before treatment initiation. The designs differed also in the number of tumors per treatment group, a parameter, which is directly related to the power of detecting statistically significant treatment effects. Other experimental design parameters included diverse setups for treatment periods and sampling frequencies as well as different response readouts such as tumor volume or area. Importantly, 3 of the 4 experiments showed different degrees of intertumor heterogeneity in terms of evidence for within-group growing and poorly growing categories (Supplementary Fig. S1).

DMBA-induced mammary cancer model

Anticarcinogenic activity of the diet-derived lignan metaboline, enterolactone (ENL), was studied by applying a mammary cancer model in the rat (6) in which the mammary tumors were induced by the use of DMBA. The induction caused a varying number of tumors per animal (1–5 measurable tumors) and thus the total tumor

Table 1. Summary of the experimental datasets used in the present work

Experiment	DMBA case	MCF-7 case	LNCaP case	4T1 case
Strain	Female Sprague-Dawley rat	Female athymic nude mouse	Male athymic nude mouse	Female immunocompetent balb/c mouse
Cell line (source)		MCF-7 (human)	LNCaP (human)	4T1 (mouse)
Cancer model	Breast cancer, carcinogen	Breast cancer, xenograft	Prostate cancer, xenograft	Breast cancer, syngeneic
Treatment	ENL	LAR	DPN or ENL	Doxorubicin or cyclophosphamide
Dosage and route of administration	Daily 1 or 10 mg/kg *per os*	Daily 20 or 100 mg/kg *per os*	DPN 4.5 mg/60 days s.c. or ENL 100 mg/kg in feed	Doxorubicin: weekly 7.5 mg/kg; cyclophosphamide: 100 mg/kg at days 0, 2, and 4
Measurement frequency	Once a week	Once a week	Twice a week	Twice a week
Number of time points	9	6	11	6
Sample sizes	13 animals per group	Control (15), lower dose (20), higher dose (20) tumors	Control (12), DPN (10), ENL (8) tumors	8 tumors per group
Target size	No	20 mm^{2,a}	200 mm^{3,b}	No
Response readout	Total tumor volume per animal	Tumor area	Tumor volume	Tumor volume
Missing value proportions	Control 4% Low dose 4% High dose 0%	All groups 0%	Control 14% DPN 10% ENL 9%	Control 0% Doxorubicin 2% Cyclophosphamide 0%
Additional cell markers	Tumor histologic types	ERα, ERβ	PSA	Metastases in the lung and liver
Reference	6	19	Unpublished	21

[a]Treatment starting time defined by average tumor area

[b]Treatment starting time defined by individual tumor volume

volume per animal was used as the response readout (Table 1). Two different dosages of ENL (1 and 10 mg/kg *per os* by gavage) were introduced 9 weeks after the DMBA induction. Each of the treatment groups included both growing (growth profiles with positive slope) and poorly growing (horizontal profiles near zero volume) tumors; the lower dosage group also contained 2 outlier profiles (Supplementary Fig. S1A). All the profiles were used here in the statistical modeling.

Histologic classification of the tumors was carried out as described earlier (6). Briefly, the tumor contributing most to the total volume per animal was considered, as it was most often histologically analyzed and could be considered as most representative for the animal. Some of the tumors could not be analyzed due to issues related to tumor suppression, volume below detection accuracy, or quality of the sample. The histologic types of "poorly differentiated", "well differentiated", and "atrophic" included more than one tumor and these were used in the analyses.

MCF-7 breast cancer xenograft model

MCF-7 breast cancer xenografts were grown in ovariectomized athymic mice in the presence of estradiol (19). The antitumor activity of the dietary lignan, lariciresinol (LAR), was studied by applying 2 different dosages (20 or 100 mg/kg *per os* by gavage) of the compound, and the tumor growth was compared with mice treated with the vehicle only (Table 1). The tumor growth profiles were analyzed along with biomarkers, such as estrogen receptors α (ERα, ESR1) and β (ERβ, ESR2), to identify explanatory factors for the observed heterogeneous growth profiles (Supplementary Fig. S1B).

LNCaP prostate cancer xenograft model

This experiment studied the effects of a synthetic ERβ-selective agonist [DPN; 2,3-bis(4-hydroxyphenol)-propionitrile] and of a tissue-specific ER activator, diet-derived lignan metabolite (ENL) on the growth of the LNCaP prostate cancer xenografts in immunocompromised mice (Athymic Nude-Fown 1 nu, Harlan). The cells (2×10^6 cells/200 µL medium/Matrigel) were subcutaneously inoculated into 5- to 6-week-old male mice. DPN was administered as pellets (4.5 mg for 60 days, Innovative Research of America). The mice in both control and treatment groups were fed purified control diet (AIN-93G; ref. 20). ENL was provided within a special diet including 100 ppm of the compound. The tumors were palpated twice a week, and the treatment was commenced once a tumor reached the target volume of 200 mm^3. To maximize the number of tumors, the growth period was allowed to reach the target volume level within 4 to 6 weeks. Because the number of tumors in the experiment remained relatively small, a maximal number of the short and outlier profiles, which often are filtered out in standard analyses, were included in the statistical analysis of the heterogeneous dataset (Supplementary Fig. S1C). In addition to the tumor size, serum prostate-specific antigen (PSA), a known prostate cancer biomarker, was measured at sacrifice (Table 1).

4T1 syngeneic mammary cancer model

Mouse mammary adenocarcinoma 4T1-cells (American Type Culture Collection) were inoculated into the thoracic mammary fat-pads of 6-week-old female immunocompetent Balb/c mice (Harlan Laboratories Inc.; ref. 21). Two established drugs were used for the treatments (Table 1): doxorubicin (22) and cyclophosphamide (23). The drug treatments were started 6 days after the inoculation of the cells. Doxorubicin (Doxorubicin Ebewe; Ebewe Pharma GmbH) was administered 7.5 mg/kg once a week and cyclophosphamide (Sendoxan, Baxter) 100 mg/kg was administered at days 0, 2, and 4 since the beginning of the treatment. The tumor growth profiles showed very homogeneous patterns within each of the treatment groups (Supplementary Fig. S1D), possibly due to the host environment being native to the 4T1 cancer cell line (24).

The categorizing mixed-effects model

The mixed-effects models have a number of advantages in the statistical analysis of tumor growth profiles. First, the whole longitudinal growth profile, with possible missing data points, can be used in the model estimation and parametric inference thereby avoiding the need for selecting predefined endpoints or *ad hoc* imputation of missing values. Second, the random effects give flexibility for the model to take into account individual tumor- and animal-specific variation that originates from the given experimental setup and data. We extended the standard model and developed a novel, hierarchical mixed-effects model, which learns the growing and poorly growing tumor categories in a given set of longitudinal tumor growth profiles. The categorizing mixed-effects model is conceptually formulated as:

$$\begin{aligned}\text{Tumor response} = {} & b_1 + b_2 \times \text{Treatment} + b_3 \\ & \times \text{Time point} \times \text{Growth} + b_4 \\ & \times \text{Treatment} \times \text{Time point} \times \text{Growth} \\ & + u_{1,T} + u_{2,T} \times \text{Time point} \quad (\text{Model 1})\end{aligned}$$

Here, the binary treatment covariate indicates the control and treatment groups and time point indicates the discrete measurement time points (Supplementary Table S1). The binary growth covariate is used to distinguish between the growing and poorly growing tumor categories. The terms b_i represent the model's fixed effects accounting for factors such as the base level tumor size (b_1), treatment-induced shift in the average tumor levels over the timepoints (offset, b_2), overall growth rate of those tumors categorized as growing (b_3), and treatment-induced difference in the growth rate of the growing tumors (slope effect, b_4). The random effects $u_{1,T}$ and $u_{2,T}$ represent variation specific to an individual tumor T. The full mathematical model formulation and details of its estimation, inference, and validation are given in Supplementary Methods.

Testing for the treatment-effects is done through the parameter estimates from the fitted categorizing model (Fig. 1A). The slope effect term b_4 evaluates time-dependent changes in the relative tumor growth rate per time

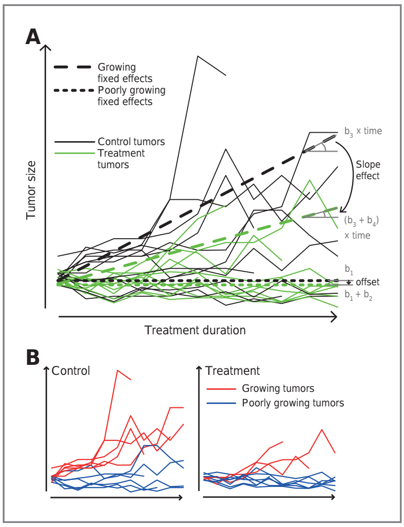


Figure 1. Schematic illustration of the treatment effect assessment in the LNCaP DPN experiment. A, fixed effects of the categorizing mixed-effects model are estimated from the data (b_1-b_4). The slope effect evaluates a treatment-induced and time point–dependent decrease in the growth rates of the growing tumors, whereas the offset term evaluates a treatment-induced shift in the horizontal tumor levels over all the time points and tumors. B, once the growing and poorly growing categories have been found by the model, the category labels are tested against the treatment labels, hence enabling evaluation of potentially more complex growth inhibiting treatment effects that may not be directly reflected in the offset or slope effects (here $P = 0.415$, Fisher exact test; Table 3).

unit in tumors categorized as growing. The slope effect therefore captures also a subtle suppressive treatment-effect relative to the overall growth rate b_3. An effective growth inhibition rate was defined as $|b_4/b_3|$. The offset term b_2 in turn evaluates more dramatic changes in the horizontal base level profiles of the tumors in those studies with a designated target size; otherwise, the terms b_1 and b_2 are set equal to zero. Because these terms do not account for the dynamic changes in the treated or control profiles, the offset term effectively captures the average treatment response in the poorly growing tumors over the entire treatment duration.

We implemented a novel clustering method based on the expectation-maximization (EM) algorithm for categorizing the tumor profiles into the growing and poorly growing subgroups (Supplementary Fig. S2). The model fitting was done using the restricted maximum likelihood (REML) estimation in the lme4 package (25) within the R statistical software (26). The statistical significance of the treatment-specific fixed effects was assessed through Markov-Chain Monte Carlo (MCMC) simulation (27). The full details of the implementation of the modeling framework are given in Supplementary Methods. The source code of the implemented R package, named XenoCat, is freely available (28).

Post hoc statistical analyses

After the growth categories were detected from the fitted model, 2-sided Fisher exact test was used to assess whether the found categorization into the growing and poorly growing subcategories can be explained by the proportion of tumors from the control and treatment groups (Fig. 1B). Significant overrepresentation of the treated tumors in the poorly growing category is indicative of such treatment effect that inhibits the tumor growth but may not be directly reflected in the fixed effect terms of the model. Hence, the offset and slope effect terms, together with the *post hoc* analysis of the detected growth categories using the Fisher exact test, can be used to draw conclusions on the treatment effects and underlying mechanisms of action.

In addition to the treatment labels, other external biologic and experimental explanatory factors for the growing and poorly growing categories were subsequently tested. For discrete explanatory factors, such as the histologic tumor classification, the Fisher exact test was used to assess the association between the tumor growth labels and the histologic classes. For normally distributed continuous factors, such as the ERβ positivity, the Welch 2-sample unpaired t test was used to evaluate the difference in the ERβ expression between the 2 growth categories. In case the Shapiro–Wilk normality test null hypothesis was rejected, the Wilcoxon rank-sum test was used instead as a nonparametric alternative.

Power, precision, and sample size estimation

Comparisons between different experimental designs and modeling setups were carried out to provide further model-guided information on their operation and suggestions for future improvements. The comparisons were based on parameters, such as the number of tumors and/or timepoints, which were investigated in relation to the calculated statistical study power, defined as the probability of detecting a statistically significant treatment effect, provided that the effect is truly present and that the model is correct. Estimation of the sample size N that is needed to achieve a given statistical power was based on simulated data generated according to the model fit (29). Furthermore, a precision analysis was implemented using the modeling framework to give guidance on the most informative time periods. Precision here means the reciprocal of the variance of the test statistic, given the estimated model and the experimental design (30). A general overview of the modeling workflow is available in Supplementary Fig. S3.

Results

An efficient implementation of the statistical modeling framework was developed and distributed as an open-source R package, named XenoCat, with accompanying user instructions (28). Here, the framework was applied to 4 case studies and the results from the categorizing mixed-effects model were compared with those obtained using the conventional mixed-effects model in terms of statistical inference, power, precision, and suggested sample size. The conventional noncategorizing mixed-effects model is a special case of the Model 1, in which the growth covariate is omitted (i.e., set to unity).

DMBA case

Estimation of the categorizing mixed-effects model in the DMBA experiment illustrates how the model can effectively describe the growing and poorly growing tumor subcategories within the treatment and control groups (Fig. 2). By taking into account such tumor growth heterogeneity, the categorizing model gave highly significant treatment effect on the slope effect term consistently both in the in low-dose and the high-dose groups ($P = 7.4 \times 10^{-5}$, $|b_4/b_3| = 40\%$ and $P = 3.8 \times 10^{-5}$, $|b_4/b_3| = 49\%$; Table 2). The subtle suppressive effect of the dietary intervention (ENL treatment) on the growth rate was missed by the conventional mixed-effects model even in the high dosage treatment group ($P > 0.05$). The increased sensitivity of the categorizing model is due to improved model fit, as indicated by the loss of skewness and multimodality in the distribution of the random slopes (Supplementary Fig. S4).

Because the identified growing and poorly growing categories could not be explained by the ENL treatment groups (Fig. 2D; Table 3), we searched for explanatory factors from the histologic analysis of the tumors. According to expectations, the tumors classified as "well differentiated" or "atrophic" were decreased in proportion in the growing tumor category consistently under both dosage levels, whereas the tumors classified as "poorly differentiated" were more abundant in the growing category (Table 3). Even if showing only a borderline statistical

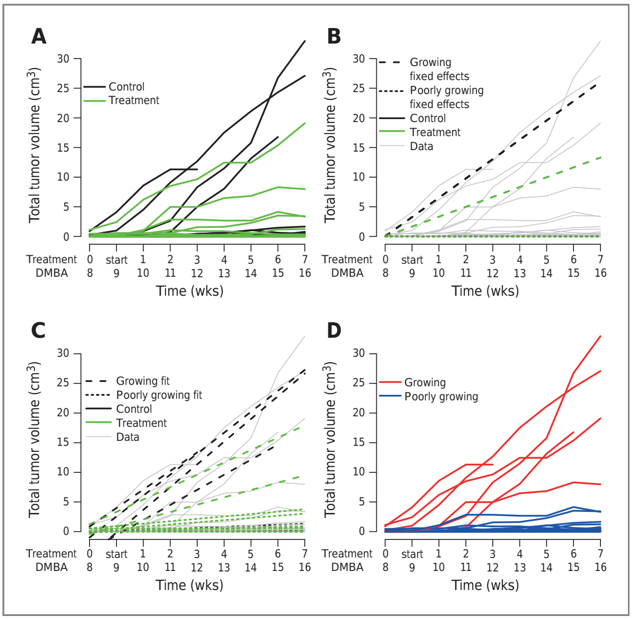


Figure 2. Operation of the categorizing modeling procedure in the DMBA ENL high-dose experiment. A, distinct growth patterns are evident in individual rats within both groups where some tumors grow aggressively, whereas others remain completely stabilized during the treatment period. B, the fixed effect fit for the population growth profiles show treatment-induced slope difference in the growing tumors ($P < 0.001$). C, the full model fit, where the individual variation is modeled using random-effects. D, The growing and poorly growing tumor categories found by the EM algorithm. In each panel, the 'DMBA' time scale depicts the time as tumor induction by the carcinogen DMBA, and the 'treatment' time scale depicts the time since the treatment initiation (start).

association ($P = 0.069$), the relative proportion of tumors in the histologic classes supported the existence and relevance of the 2 growth categories. In contrast, the association between the treatment groups and the histologic types was highly insignificant ($P = 0.955$) indicating that the treatment *per se* did not influence the differentiation process.

MCF-7 case

In the MCF-7 xenograft experiment, the effects of the dietary lignan LAR treatment were found insignificant both on the offset and slope terms (Table 2). However, even in the absence of statistically significant treatment effects, the growing and poorly growing tumor categories could be explained by the treatment groups under the high dosage LAR treatment (Fisher exact test, $P = 0.022$; Table 3), suggesting that the dietary lignan treatment successfully blocks a significant portion of tumors into the poorly growing category. Interestingly, the tumors in the growing and poorly growing categories were also different in terms of their measured ERβ levels in the high dosage group ($P = 0.008$; Table 3) indicating that ERβ inhibits tumor growth, as has been previously suggested on the basis of results obtained from other experimental breast cancer models (31).

LNCaP case

A xenograft study with LNCaP cells was analyzed in terms of possible treatment effects, and to provide guidance for a sufficient sample size and the most informative time periods to be used in further studies. The categorizing model showed, already in the present data, a statistically highly significant slope effect in response to the ENL treatment ($P = 0.001$, $|b_4/b_3| = 80\%$), and a slightly significant slope effect in response to the DPN treatment ($P = 0.037$, $|b_4/b_3| = 48\%$). Both of these effects were undetected by the conventional mixed-effects model ($P > 0.05$; Table 2). However, both model types captured well the target tumor volume of 200 mm^3 in their base level terms under both treatments ($P < 10^{-5}$), whereas the categorization emphasized the overall growth terms ($P < 10^{-5}$).

The measured PSA concentrations at sacrifice were significantly different between the tumors classified into the growing or poorly growing categories. According to expectations, the PSA levels were consistently higher in the growing category than in the poorly growing category both in response to the DPN ($P = 0.005$) and ENL ($P = 0.001$) treatments (Table 3). Interestingly, the PSA levels at sacrifice were similar in the control and treatment groups both on DPN and ENL ($P > 0.05$) indicating that factors other than

Table 2. Fixed effect estimates, confidence intervals and statistical significance

DMBA case	**Categorizing model**			**Noncategorizing model**		
ENL low dose	Estimate	HPD interval	P	Estimate	HPD interval	P
Overall growth b_3	3.12	[2.56 to 3.39]	<0.001	0.989	[0.34 to 1.54]	<0.01
Slope effect b_4	−1.25	[−1.74 to −0.703]	<0.001	−0.174	[−1.04 to 0.66]	0.659
ENL high dose						
Overall growth b_3	3.26	[2.78 to 3.47]	<0.001	1.02	[0.50 to 1.44]	<0.001
Slope effect b_4	−1.60	[−2.01 to −0.871]	<0.001	−0.681	[−1.28 to 0.04]	0.066
MCF-7 case						
LAR low dose						
Base level b_1	21.4	[16.9 to 25.8]	<0.001	21.0	[16.5 to 25.8]	<0.001
Offset b_2	−2.66	[−8.68 to 3.01]	0.359	−4.18	[−10.5 to 1.97]	0.183
Overall growth b_3	8.71	[6.92 to 10.6]	<0.001	8.37	[6.55 to 10.2]	<0.001
Slope effect b_4	−0.599	[−3.10 to 1.94]	0.619	−1.04	[−3.44 to 1.44]	0.398
LAR high dose						
Base level b_1	21.4	[17.0 to 25.7]	<0.001	21.0	[16.5 to 25.7]	<0.001
Offset b_2	1.05	[−5.00 to 6.28]	0.802	−0.945	[−7.12 to 5.08]	0.761
Overall growth b_3	8.71	[6.98 to 10.5]	<0.001	8.37	[6.46 to 10.3]	<0.001
Slope effect b_4	−1.72	[−4.08 to 1.07]	0.237	−3.07	[−5.63 to −0.53]	<0.05
LNCaP case						
DPN						
Base level b_1	234	[196 to 272]	<0.001	233	[192 to 273]	<0.001
Offset b_2	−22.7	[−78.5 to 33.9]	0.421	−19.5	[−78.5 to 40.5]	0.540
Overall growth b_3	101	[75.4 to 130]	<0.001	52.8	[22.8 to 81.5]	<0.01
Slope effect b_4	−48.9	[−95.6 to −2.91]	<0.05	−41.0	[−83.2 to 2.99]	0.072
ENL						
Base level b_1	234	[194 to 275]	<0.001	233	[191 to 276]	<0.001
Offset b_2	−8.31	[−72.6 to 53.3]	0.784	−5.19	[−72.1 to 60.9]	0.862
Overall growth b_3	101	[74.5 to 130]	<0.001	52.7	[22.6 to 81.7]	<0.01
Slope effect b_4	−81.1	[−125 to –36.1]	<0.01	−45.1	[−92.2 to 1.02]	0.058
4T1 case						
Doxorubicin						
Overall growth b_3	68.4	[57.6 to 79.3]	<0.001	68.4	[57.6 to 79.3]	<0.001
Slope effect b_4	−16.8	[−32.4 to −1.40]	<0.05	−16.8	[−32.0 to −1.12]	<0.05
Cyclophosphamide						
Overall growth b_3	68.4	[60.5 to 76.5]	<0.001	68.4	[60.2 to 76.4]	<0.001
Slope effect b_4	−66.5	[−78.3 to −54.7]	<0.001	−66.8	[−78.2 to −55.3]	<0.001

NOTE: The highest posterior density (HPD), 95% confidence intervals, and P values were estimated using 100,000 MCMC simulations. Negative estimates for the treatment specific terms (b_2, b_4) indicate potential treatment effects. Model terms b_1 and b_2 were set to zero in studies without a designated tumor target size (DMBA, 4T1). The fixed effects presented in this table are visualized in a parallel fashion in Supplementary Fig. S10.

the treatment contribute to the identified between-tumor differences in terms of their growth profiles and PSA levels.

We further used the modeling framework to predict that a significant slope effect ($P < 0.05$ at 0.8 power) in response to the DPN treatment could be obtained when the number of tumors is 19 per group (Supplementary Fig. S5A). Notably, with the noncategorizing model, the same sample size estimate would be 25, showing the benefits of the categorizing model already in the initial power analysis. The power analysis also predicted that significant offset effect will not be obtained within reasonable animal numbers. The precision analysis showed differences in the model types and treatment periods when assessing treatment effects (Supplementary Fig. S5B); in particular, the relative importance of the initial time points for the statistical precision (Supplementary Fig. S5C).

4T1 case

In cases such as 4T1, where there is no evident within-group tumor heterogeneity, the EM algorithm classifies all the tumors into the growing category, and therefore the categorizing and noncategorizing models gave the same results (Table 2). More specifically, after adjustment to quadratic growth using residual plots (Supplementary

Table 3. *Post hoc* association analysis of the detected tumor growth categories

	Treatment classes (% within category)			**Histologic classes[a] (% within category)**			
DMBA case	**Control**	**Treatment**	***P***	**Poorly differentiated**	**Well differentiated**	**Atrophic**	***P***
ENL low dose							
Growing	4 (44%)	5 (56%)		4 (67%)	2 (33%)	0 (0%)	
Poorly growing	9 (53%)	8 (47%)	1.000	2 (17%)	7 (58%)	3 (25%)	0.156
ENL high dose							
Growing	4 (67%)	2 (33%)		3 (60%)	1 (20%)	1 (20%)	
Poorly growing	9 (45%)	11 (55%)	0.645	2 (13%)	10 (67%)	3 (20%)	0.069
MCF-7 case				**ERβ expression[b] (per 1,000 cells)**			
LAR low dose							
Growing	14 (48%)	15 (52%)		248.1 ± 238.7			
Poorly growing	1 (17%)	5 (83%)	0.207	82.0 ± 56.6			0.115
LAR high dose							
Growing	14 (56%)	11 (44%)		213.0 ± 127.0			
Poorly growing	1 (10%)	9 (90%)	0.022	329.7 ± 32.7			0.008
LNCaP case				**PSA concentration[b] (at sacrifice, μg/L)**			
DPN							
Growing	6 (67%)	3 (33%)		97.3 ± 48.3			
Poorly growing	6 (46%)	7 (54%)	0.415	29.3 ± 17.7			0.005
ENL							
Growing	6 (60%)	4 (40%)		99.1 ± 45.5			
Poorly growing	6 (60%)	4 (40%)	1.000	29.1 ± 15.4			0.001

NOTE: Underlining indicates statistical significance ($P \leq 0.05$).

[a] Some of the tumors could not be histologically typed

[b] Values expressed as mean ± SD

Fig. S6), it was confirmed that doxorubicin resulted in regressed tumor growth profiles ($P < 0.05$), whereas cyclophosphamide completely stabilized the growth of each treated tumor ($P < 10^{-6}$).

To test the relative benefits of the categorizing model in a setting where the underlying growth categories and true treatment effect were predefined, we constructed a simulated dataset by combining the doxorubicin- and cyclophosphamide-treated tumors into a single treatment group. The EM algorithm separated the sources of these growth profiles with 100% accuracy within the control and treatment groups (Fig. 3A). The categorizing model also enabled detection of a significant treatment slope effect ($P = 0.002$), which remained undetected by the noncategorizing model (Fig. 3B). This is due to the inability of the noncategorizing model to adjust for the distinct sources of intertumor variation, leading to poor model fit and multimodality in the slope estimates, which could be corrected by taking into account the tumor heterogeneity with the categorizing model (Fig. 3C).

Finally, we also conducted simulations under the null hypothesis of no true treatment effect (Supplementary Material). As expected, an increase in the type-I error appeared under such situation if the categorization approach was applied to homogeneous data or if the noncategorizing approach was applied to heterogeneous data (Supplementary Table S2). The model diagnostic tools should therefore be used to make informed decisions about the model type and structure that is most preferred for the dataset under analysis.

Discussion

This study showed (i) the benefits of modeling the growing and poorly growing categories in terms of improved statistical inference (e.g., DMBA and 4T1 cases); (ii) how the detected categories may be associated with interesting biologic factors, such as endogenous ERβ levels in the MCF-7 case, which provide insights into the underlying tumor heterogeneity; and (iii) how the framework can provide informed suggestions on designing more effective tumor growth experiments in terms of sufficient sample sizes and most informative treatment periods (LNCaP case). The generic modeling framework can also be extended to include additional covariates, such as quadratic growth profiles (Supplementary Fig. S6) or probabilistic tumor categorization (Supplementary Fig. S2). For instance, as the heterogeneity in the growth profiles in the MCF-7 and LNCaP studies was not so clear-cut, continuous growth

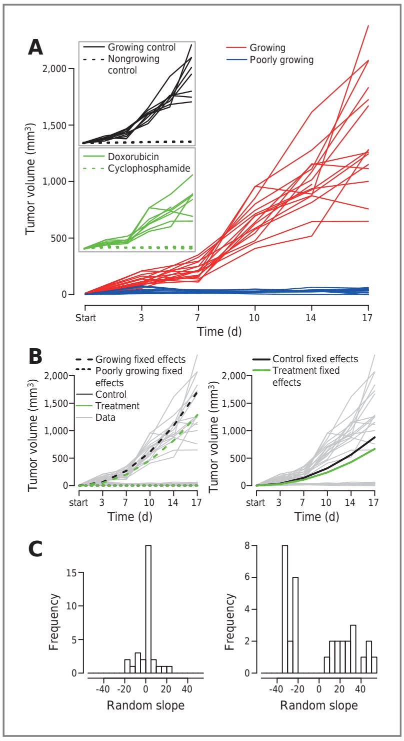


Figure 3. Modeling the simulated 4T1 dataset with and without categorization. A, the EM algorithm and the categorizing model correctly identified the growing and poorly growing subgroups both in the combined control group (original controls and square root of their response traces) and in the combined treatment group (doxorubicin- and cyclophosphamide-treated tumors). B, the categorizing approach detects the doxorubicin-specific treatment effect (left, $P = 0.002$), which is missed by the noncategorizing approach (right, $P = 0.441$). The fixed effect estimates of the noncategorizing model are not feasible due to the assumption of homogeneous growth profiles. C, the random-effects of the categorizing approach show reasonable model fit (left), whereas the random slopes of the noncategorizing approach exhibit severe multimodality (right), suggesting that the growth profiles indeed originate from 2 distinct distributions (i.e., tumor subcategories).

covariates were further used to show that such probabilistic categorization resulted in similar conclusions as obtained from the binary categorization (Supplementary Table S3).

Existing statistical approaches and their limitations

Tumor growth profiles have traditionally been analyzed using univariate statistical approaches that do not fully take into account the tumor heterogeneity within the treatment and control groups. These approaches are typically based on the comparison of tumor sizes at a prespecified time point, using statistical methods such as *t* tests and ANOVA, or their rank-based alternatives such as Wilcoxon–Mann–Whitney and Kruskal–Wallis test (1, 4–12). Another commonly used end point is the time until tumor size doubling, which is analyzed using statistical methods from survival analysis such as the log-rank test (1). There are, however, some potential pitfalls in the use of such single end point approaches. First, an invalid choice of the evaluation time point or the target tumor size may lead to substantial loss of information, in case a large fraction of tumors have not reached the predefined endpoint (32, 33). Second, any single end point is unpowered to detect treatment mechanisms behind dynamic patterns of tumor growth (12). This was exemplified in the DMBA case, where only 2 of the 9 time points showed a significant treatment effect in the original ANOVA-based analysis (6), making the inference upon the efficacy of the dietary intervention more difficult.

Longitudinal statistical modeling methods have also been developed for tumor growth experiments, but these are often restricted to rather specific study designs or questions, and lack effective modeling of intertumor heterogeneity (1). Related approaches that share similar methodologies include, for instance, a standard *t* test together with an EM algorithm as well as Bayesian modeling approaches for testing differences in treatment regimens (13–15). Other authors have developed a nonlinear method for summing 2 exponential functions (16), or a nonparametric approach for estimating tumor growth profiles using penalized spline functions (17). However, even if these models can deal, for instance, with missing and censored data values, other important characteristics of the growth profiles, such as tumor regression or growth rates, cannot be estimated using such approaches. Finally, many of the more advanced statistical models introduced for analyzing tumor growth experiments are not implemented as user-friendly software packages, which hinders their routine use in data analysis.

Recently, an interesting Bayesian hierarchical changepoint (BHC) model was proposed for analyzing long treatment experiments (12). The model assumes that the treated tumors will first suppress in response to the treatment, then reach a minimum, and later, rebound with both the decline and the regrowth curves assumed being linear on the log scale. The main difference between our framework and the BHC models is that the latter categorizes the growth profile of each individual tumor into these specific growth periods (i.e., it models intratumor variability), whereas our model categorizes the given set of tumor profiles into growing and poorly growing classes (i.e., it models intertumor variability). The BHC model is especially useful for estimating regression period and nadir tumor volume for such tumors that contain measurements below the limit of quantitation leading to missing values and censored data (1). This is often the case when assessing more aggressive treatment options, which can totally regress the tumor growth and the main focus lies on testing rebound effects and possible side effects. On the other hand, when experimenting with less aggressive treatment alternatives, more subtle treatment effects are easily missed in case the intertumor heterogeneity is not properly taken into account.

Benefits of the categorizing mixed-effects model

To our knowledge, there are no existing approaches towards modeling the growing and poorly growing tumor categories, even if the presence of such categories in the tumor growth experiments has been long evident (2–8). While there are various approaches to reduce model fit heteroscedasticity, such as the Box-Cox or logarithmic transformations, these cannot model the intrinsic heterogeneity encountered within control and treatment groups. This study showed that when the observed within- and between-group variation is effectively modeled, it is possible to improve the sensitivity of the treatment evaluation through relevant model parameters. In particular, the slope parameter was shown informative when evaluating the efficacy of dietary plant lignans. Our modeling framework also enables comparison of different experimental designs in terms of their associated study power, precision, and sample size estimates, something that is rarely available from other modeling works. However, it should be appreciated that the operation of these modeling tools depends on the data under analysis. Therefore, data visualization and model diagnostics should always be used to confirm that the model assumptions are fulfilled and the model results are valid (see Supplementary Methods for details).

To promote its widespread application in tumor growth studies, we have made publicly available the modeling framework in the form of an R package, named XenoCat, with implementation, source code, user-instructions, and step-by-step example available (28). In contrast to most existing models, our framework can be robustly applied to various tumor growth experiments without making strong assumptions about the type or amount of data under analysis. For instance, the 4 case studies analyzed here were conducted using different tumor models, representing a wide range of experimental setups, such as different number of tumors and various response readouts and their distributional characteristics, which can drastically affect the performance of the traditional statistical methods. The model can deal with short or even outlier profiles, which may be present in the data due to various filtering criteria or very aggressively growing tumors, respectively, and which are frequently excluded from the standard statistical analysis. Therefore, the model can use the full information captured in the entire longitudinal profiles to maximize the output of the tumor growth studies.

The novel tumor categorizing algorithm does not only enable calculating interesting growth parameters, but it also allows for detection of hidden subgroups of differentially growing tumors within treatment and control groups that may associate with the underlying tumor biology. In particular, differences observed in the ERβ expression between the growing and poorly growing categories in the MCF-7 breast cancer model are highly intriguing. Previous studies on genetically modified breast cancer cell lines with high constitutive or inducible expression of ERβ show that tumor growth is significantly reduced when the transgene is turned on (31). Our study is the first, to our knowledge, to show the inverse association between tumor growth and endogenous ERβ expression and suggests that endogenous ERβ levels may be regulated by interventions (here, dietary lignans). This phenomenon may be linked to underlying differences in tumor progression mechanisms (34) and can even give insights into treatment resistance (35). Besides providing additional explanations for the detected tumor growth categories, biologic correlates behind the model-captured tumor heterogeneity could thus open up new possibilities for identifying novel targets and treatment opportunities for cancer.

Limitations of the model and its future extensions

A number of simplifying assumptions were made here to make the implemented model as robust and flexible as possible. The methodology proposed here is based on linear mixed-effects models with dichotomous categorization and assumption that the poorly growing profiles are approximately horizontal. However, in cases where deemed appropriate, the generic model can be extended to more complex settings, including nonlinear growth patterns or several growth categories with non-zero slope parameters or probabilistic tumor categorization, allowing, for instance, partially overlapping groups such as growing, regressing, and stabilizing profiles (6). Another interesting future question we intend to tackle is that whether combining multiple phenotypic readouts for treatment response, such as tumor sizes and PSA levels, would improve statistical power in the case of the prostate cancer model. The current implementation of the power analysis also assumes complete data, but missing values, either informatively censored or missing-at-random (36), could be incorporated in the future work. Finally, the computationally, rather intensive, power calculations could easily be split into parallel processes for maximal computational efficiency (Supplementary Fig. S7).

Disclosure of Potential Conflicts of Interest

M.I. Suominen and E. Alhoniemi were the employees of Pharmatest Services Ltd (at the time of the study), a company that sells services for preclinical cancer treatment testing. M.I. Suominen is also a shareholder of Pharmatest Services Ltd and has ownership interest (including patents) in the same. No potential conflicts of interest were disclosed by other authors.

Authors' Contributions

Conception and design: T.D. Laajala, J. Corander, S.I. Mäkelä, M. Poutanen, T. Aittokallio
Development of methodology: T.D. Laajala, J. Corander, T. Aittokallio
Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): N.M. Saarinen, S. Savolainen, M.I. Suominen, S.I. Mäkelä, M. Poutanen
Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): T.D. Laajala, M. Poutanen, T. Aittokallio
Writing, review, and/or revision of the manuscript: T.D. Laajala, J. Corander, S. Savolainen, M.I. Suominen, E. Alhoniemi, S.I. Mäkelä, M. Poutanen, T. Aittokallio
Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): T.D. Laajala, N.M. Saarinen, K. Mäkelä, E. Alhoniemi, M. Poutanen
Study supervision: J. Corander, M. Poutanen, T. Aittokallio

Interpretation of the biological relevance of the obtained results: N.M. Saarinen

Acknowledgments

The authors thank Dr. Emmy Verschuren (FIMM) for critical reading and useful suggestions for the manuscript and Pharmatest Services Ltd for providing the data from the 4T1 syngeneic mammary cancer model.

Grant Support

This work was supported by the Finnish Funding Agency for Technology and Innovation (Tekes grant 40106/10 to M. Poutanen and T. Aittokallio and Tekes grant 40226/08 to S. Mäkelä); the Academy of Finland (grants 120569, 133227, and 140880 to T. Aittokallio, grant 251170 to J. Corander, grant 138561 to M. Poutanen, grant 123382 to S. Mäkelä; and grant 115459 to N. M. Saarinen); and the European Research Council (ERC grant 239784 to J. Corander).

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received December 13, 2011; revised May 25, 2012; accepted June 8, 2012; published OnlineFirst June 27, 2012.

References

1. Heitjan DF. Biology, models, and the analysis of tumor xenograft experiments. Clin Cancer Res 2011;17:949–52.
2. Enmon R, Yang WH, Ballangrud AM, Solit DB, Heller G, Rosen N, et al. Combination treatment with 17-N-allylamino-17-demethoxy geldanamycin and acute irradiation produces supra-additive growth suppression in human prostate carcinoma spheroids. Cancer Res 2003;63: 8393–9.
3. Bedogni B, Welford SM, Kwan AC, Ranger-Moore J, Saboda K, Broome Powell M. Inhibition of phosphatidylinositol-3-kinase and mitogen-activated protein kinase 1/2 prevents melanoma development and promotes melanoma regression in the transgenic TPRas mouse model. Mol Cancer Ther 2006;5:3071–7.
4. Gutman M, Couillard S, Labrie F, Candas B, Labrie C. Effects of the antiestrogen EM-800 (SCH 57050) and cyclophosphamide alone and in combination on growth of human ZR-75-1 breast cancer xenografts in nude mice. Cancer Res 1999;59:5176–80.
5. Saarinen NM, Power K, Chen J, Thompson LU. Flaxseed attenuates the tumor growth stimulating effect of soy protein in ovariectomized athymic mice with MCF-7 human breast cancer xenografts. Int J Cancer 2006;119:925–31.
6. Saarinen NM, Huovinen R, Wärri A, Mäkelä SI, Valentín-Blasini L, Sjöholm R, et al. Enterolactone inhibits the growth of 7,12-dimethylbenz(a)anthracene-induced mammary carcinomas in the rat. Mol Cancer Ther 2002;1:869–76.
7. Galaup A, Opolon P, Bouquet C, Li H, Opolon D, Bissery MC, et al. Combined effects of docetaxel and angiostatin gene therapy in prostate tumor model. Mol Ther 2003;7:731–40.
8. Ribonson SP, Jordan VC. Reversal of the antitumor effects of tamoxifen by progesterone in the 7,12-dimethylbenzanthracene-induced rat mammary carcinoma model. Cancer Res 1987;47:5386–90.
9. Terada N, Shimizu Y, Kamba T, Inoue T, Maeno A, Kobayashi T, et al. Identification of EP4 as a potential target for the treatment of castration-resistant prostate cancer using a novel xenograft model. Cancer Res 2010;70:1606–15.
10. Takahara K, Tearle H, Ghaffari M, Gleave ME, Pollak M, Cox ME. Human prostate cancer xenografts in lit/lit mice exhibit reduced growth and androgen-independent progression. Prostate 2011;71: 525–37.
11. Shusterman S, Grupp SA, Barr R, Carpentieri D, Zhao H, Maris JM. Angiogenesis inhibitor TNP-470 effectively inhibits human neuroblastoma xenograft growth, especially in the setting of subclinical disease. Clin Cancer Res 2001;7:977–84.
12. Zhao L, Morgan MA, Parsels LA, Maybaum J, Lawrence TS, Normolle D. Bayesian hierarchical changepoint methods in modeling the tumor growth profiles in xenograft experiments. Clin Cancer Res 2011;17: 1057–64.
13. Tan M, Fang HB, Tian GL, Houghton PJ. Small sample inference for incomplete longitudinal data with truncation and censoring in tumour xenograft models. Biometrics 2002;58:612–20.
14. Fang HB, Tian GL, Tan M. Hierarchical models for tumour xenograft experiments in drug development. J Biopharm Stat 2004;14:931–45.
15. Tan M, Fang HB, Tian GL, Houghton PJ. Repeated-measures models with constrained parameters for incomplete data in tumour xenograft experiments. Stat Med 2005;24:109–19.
16. Liang H, Sha N. Modeling antitumor activity by using a non-linear mixed-effects model. Math Biosci 2004;189:61–73.
17. Liang H. Modeling antitumor activities in xenograft tumor treatment. Biometrical J 2005;3:358–68.
18. Verbeke G, Lesaffre E. A linear mixed-effects model with heterogeneity in the random-effects population. J Am Stat Assoc 1996;91: 217–21.
19. Saarinen NM, Wärri A, Dings RP, Airio M, Smeds AI, Mäkelä S. Dietary lariciresinol attenuates mammary tumor growth and reduces blood vessel density in human MCF-7 breast cancer xenografts and carcinogen-induced mammary tumors in rats. Int J Cancer 2008;123: 1196–204.
20. Reeves PG, Nielsen FH, Fahey GC Jr. AIN-93 purified diets for laboratory rodents: final report of the American Institute of Nutrition ad hoc writing committee on the reformulation of the AIN-76A rodent diet. J Nutr 1993;123:1939–51.
21. Suominen MI, Käkönen R, Käkönen SM, Halleen JM. Diverging effects of doxorubicin, paclitaxel and cyclophosphamide on 4T1 mouse breast cancer primary tumor and metastases. Poster at joint AACR-MRS meeting: Metastasis and the tumor microenvironment, September 12–15 2010, Philadelphia, USA. Available from: www.pharmatest.fi
22. Du G, Lin H, Wang M, Zhang S, Wu X, Lu L, et al. Quercetin greatly improved therapeutic index of doxorubicin against 4T1 breast cancer by its opposing effects on HIF-1α in tumor and normal cells. Cancer Chemother Pharmacol 2010;65:277–87.
23. Viola RJ, Provenzale JM, Li F, Li CY, Yuan H, Tashjian J, et al. *In vivo* bioluminescence imaging monitoring of hypoxia-inducible factor 1alpha, a promoter that protects cells, in response to chemotherapy. Am J Roentgenol 2008;191:1779–84.
24. Kamb A. What's wrong with our cancer models? Nat Rev Drug Discov 2005;4:161–5.
25. Bates DM, Maechler M, Bolker B. lme4: Linear mixed-effects models using S4 classes. R package version 0.999375-39; 2011 [cited 2012 May 25]. Available from: http://CRAN.R-project.org/package=lme4
26. R Development Core Team. R: A language and environment for statistical computing. R statistical software version 2.14; 2011 [cited 2012 May 25]. Available from: http://www.R-project.org
27. Baayen RH. Modeling data with fixed and random effects. In: Analyzing linguistic data, a practical introduction to statistics using R, Cambridge University Press; 2008. p. 242–58.
28. XenoCat-project, R package version 1.0.3; 2012 [cited 2012 May 25]. Available from: http://code.google.com/p/r-xenocat/.
29. Gelman A, Hill J. Multilevel power calculation using fake-data simulation. In:Alvarez RM, Beck NL, Wu LL editors. Data analysis using regression and multilevel/hierarchical models. New York: Cambridge University Press; 2007. p. 449–54.
30. Stroup WW. Mixed model procedures to assess power, precision and sample size in the design of experiments. In: ASA Proceedings of the biopharmaceutical section. Alexandria, VA: American Statistical Association; 1999. p. 15–24.
31. Hartman J, Lindberg K, Morani A, Inzunza J, Ström A, Gustafsson JA. Estrogen receptor beta inhibits angiogenesis and growth of T47D breast cancer xenografts. Cancer Res 2006;66: 11207–13.

32. Begg AC. Analysis of growth delay data: potential pitfalls. Br J Cancer Suppl 1980;4:93–97.

33. Heitjan DF, Manni A, Santen RJ. Statistical analysis of *in vivo* tumor growth experiments. Cancer Res 1993;53:6042–50.

34. Horimoto Y, Hartman J, Millour J, Pollock S, Olmos Y, Ho KK, et al. ERβ1 represses FOXM1 expression through targeting ERα to control cell proliferation in breast cancer. Am J Pathol 2011;179:1148–56.

35. Hopp TA, Weiss HL, Parra IS, Cui Y, Osborne CK, Fuqua SA. Low levels of estrogen receptor β protein predict resistance to Tamoxifen therapy in breast cancer. Clin Cancer Res 2004;10:7490–9.

36. Aittokallio T. Dealing with missing values in large-scale studies: microarray data imputation and beyond. Brief Bioinform 2010;11:253–64.

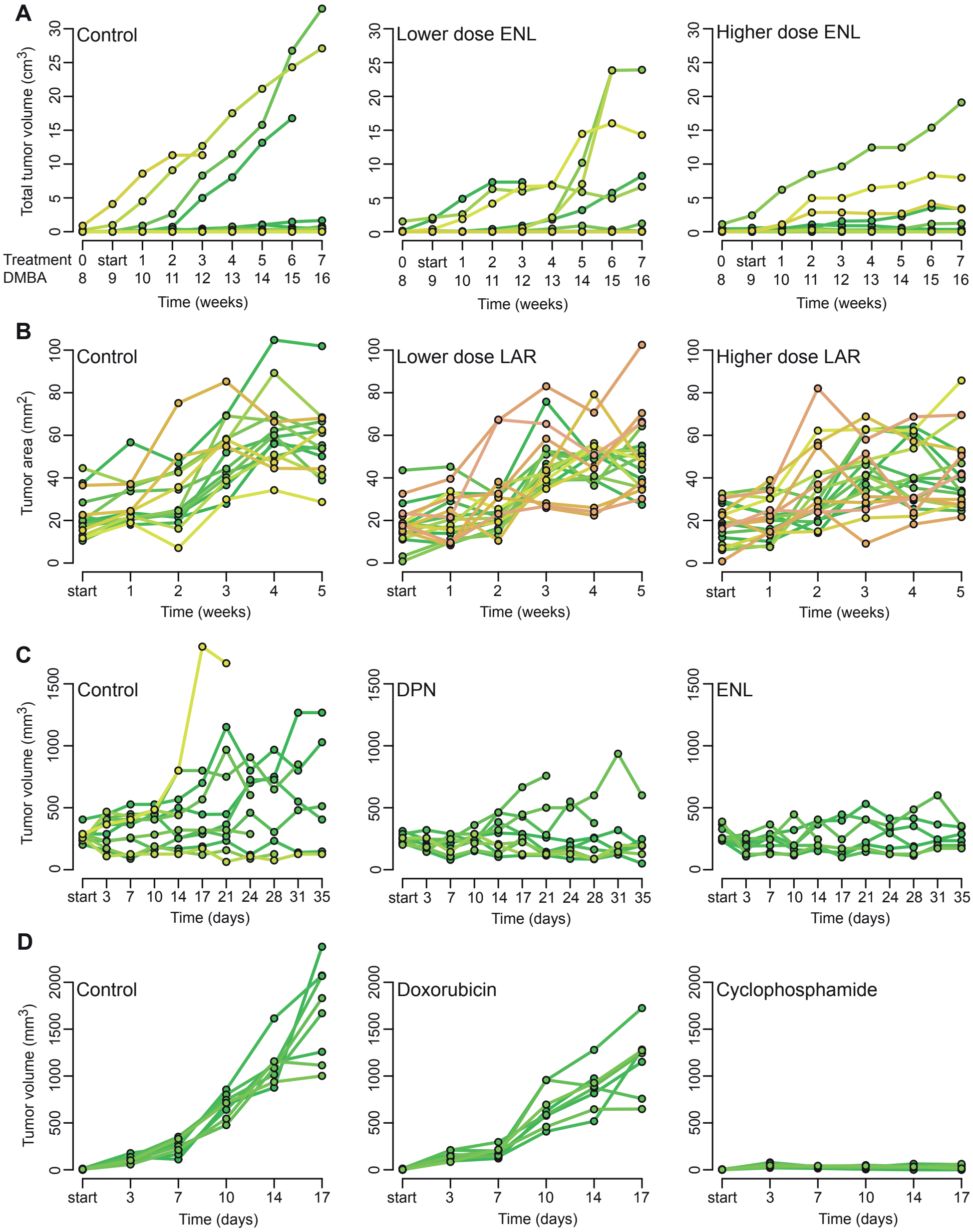
Supplementary Material (Publication II):

Supplementary Figures S1 - S10

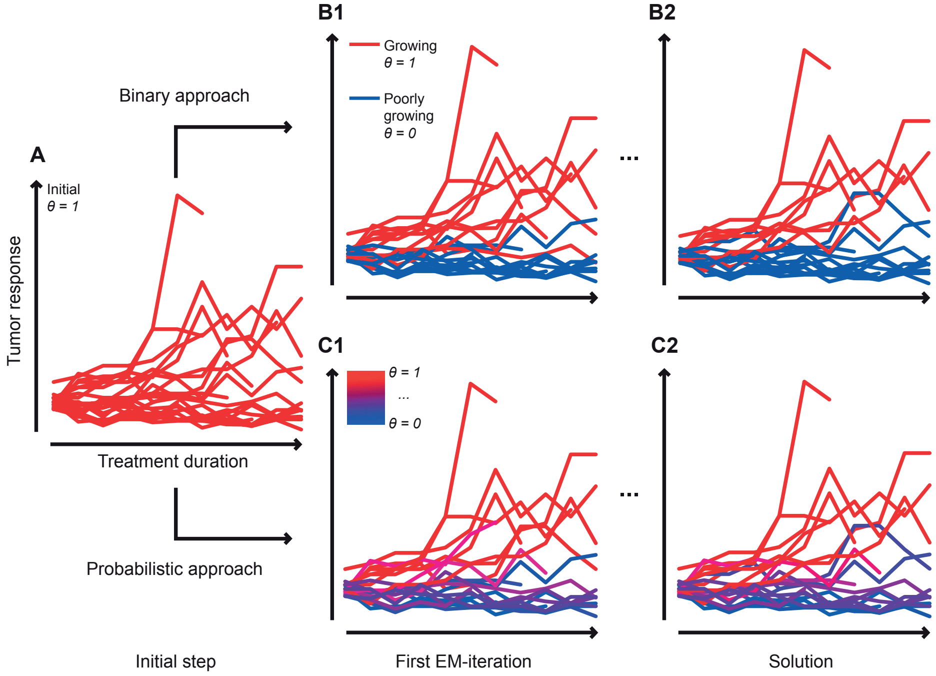
Supplementary Tables S1 - S3

Supplementary Methods (*Available online*)*

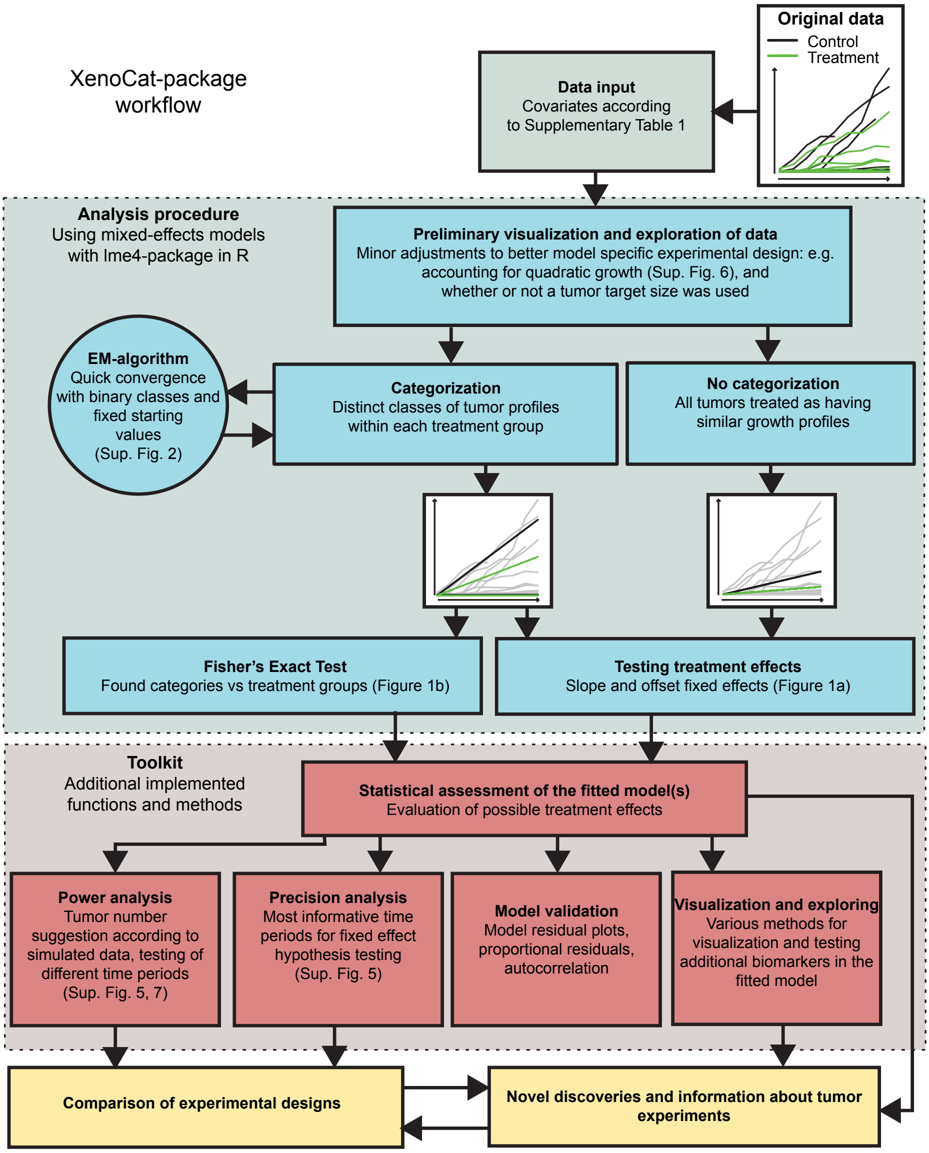
*: https://clincancerres.aacrjournals.org/content/clincanres/suppl/2012/07/02/1078-0432.CCR-11-3215.DC1/Supplementary_material_Methods.pdf



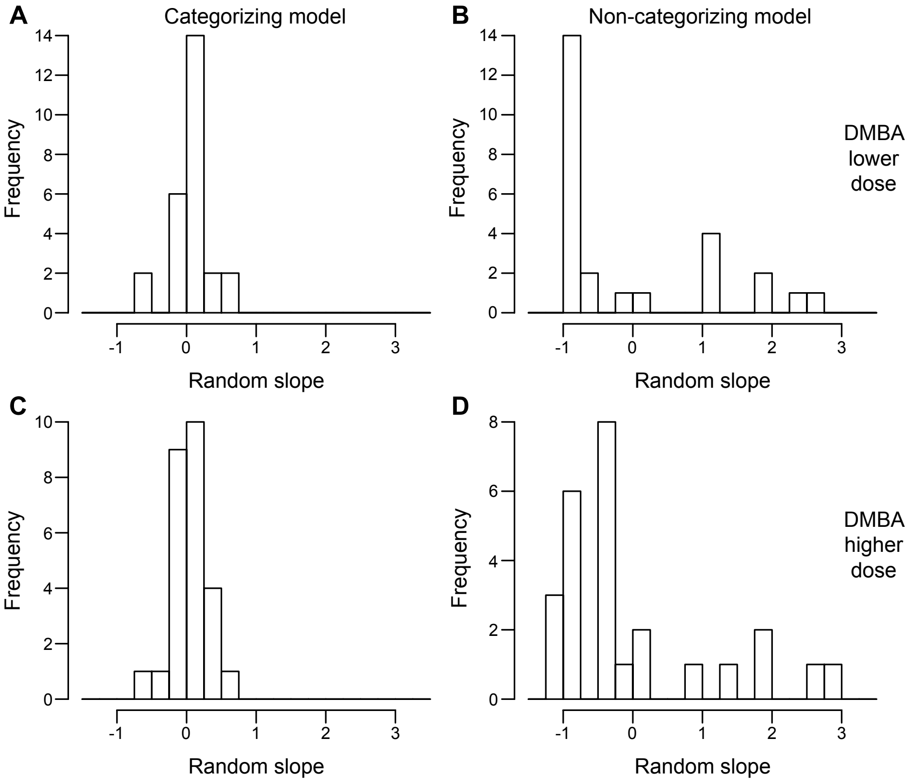
Supplementary Figure S1: Individual tumor profiles within the control and treatment groups. (A) DMBA dataset, (B) MCF-7 dataset, (C) LNCaP dataset, (D) 4T1 dataset.



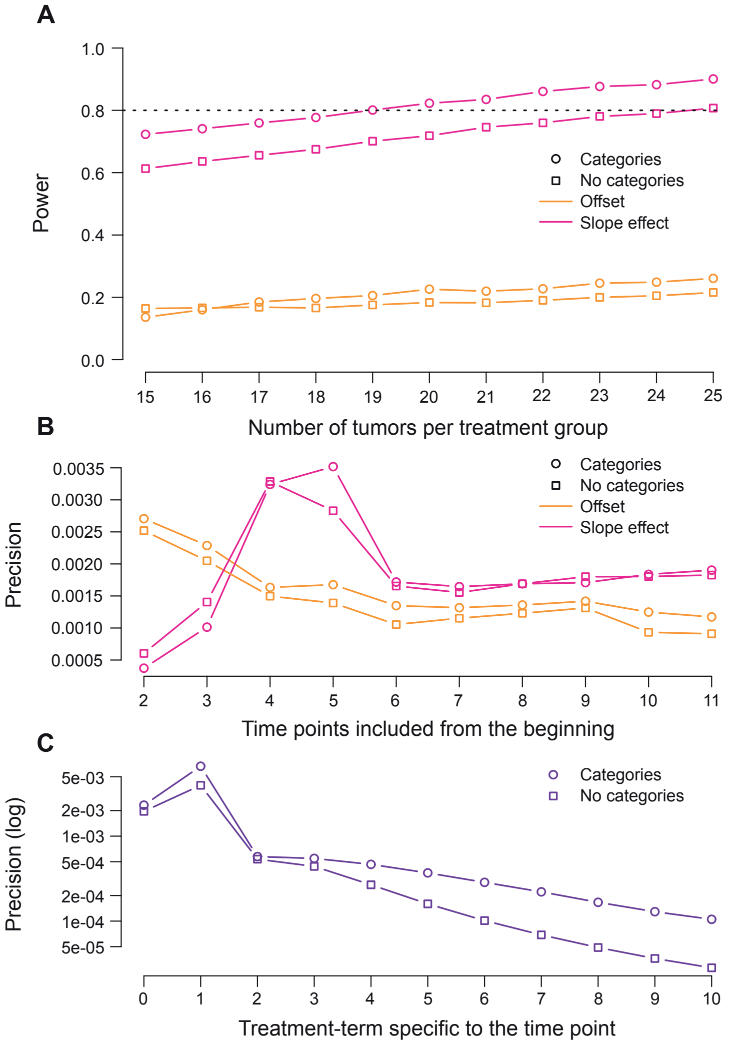
Supplementary Figure S2: Convergence of the EM-algorithm categories in the LNCaP DPN experiment. Likelihoods for the observations within a single tumor are combined, thus they share the same category at each step. (A) When the initial starting point is fixed, the process is deterministic and the same stable solution is found over different runs. (B) In the default approach, the identified categories are discriminated at each EM-iteration to yield a binary discrimination into the growing and poorly growing categories. (C) The probabilistic approach utilizes the probability of a tumor to belong to the growing category yielded by the likelihood ratio test. This provides a continuous version for the latent growth covariate for model fitting. (1) First EM-iteration, (2) Final EM-iteration.


XenoCat-package workflow
Original data
Control
Treatment
Data input
Covariates according to Supplementary Table 1
Analysis procedure
Using mixed-effects models with lme4-package in R
Preliminary visualization and exploration of data
Minor adjustments to better model specific experimental design: e.g. accounting for quadratic growth (Sup. Fig. 6), and whether or not a tumor target size was used
EM-algorithm
Quick convergence with binary classes and fixed starting values (Sup. Fig. 2)
Categorization
Distinct classes of tumor profiles within each treatment group
No categorization
All tumors treated as having similar growth profiles
Fisher's Exact Test
Found categories vs treatment groups (Figure 1b)
Testing treatment effects
Slope and offset fixed effects (Figure 1a)
Toolkit
Additional implemented functions and methods
Statistical assessment of the fitted model(s)
Evaluation of possible treatment effects
Power analysis
Tumor number suggestion according to simulated data, testing of different time periods (Sup. Fig. 5, 7)
Precision analysis
Most informative time periods for fixed effect hypothesis testing (Sup. Fig. 5)
Model validation
Model residual plots, proportional residuals, autocorrelation
Visualization and exploring
Various methods for visualization and testing additional biomarkers in the fitted model
Comparison of experimental designs
Novel discoveries and information about tumor experiments

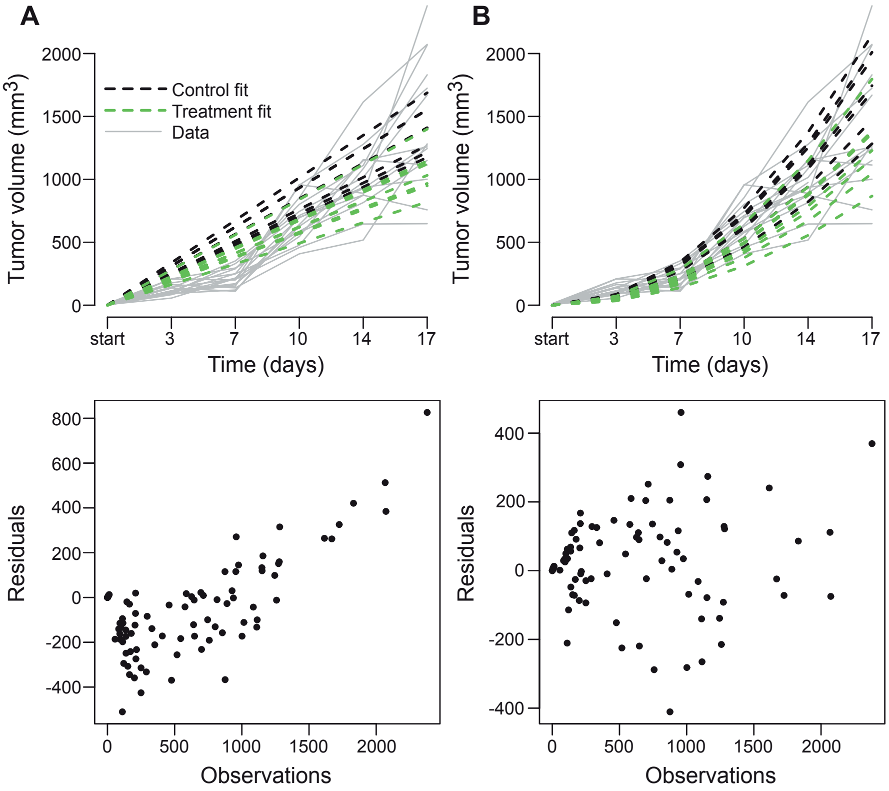
Supplementary Figure S3: Workflow of the tumor growth analysis using the presented methodology in R. The workflow is split into two phases: the first phase includes the model fitting and testing of the model. The second phase includes a kit of various tools for practical tasks in tumor growth studies, such as power and precision analysis, validation of the model and further exploration of the model in terms of testing additional biomarkers related to the study question.



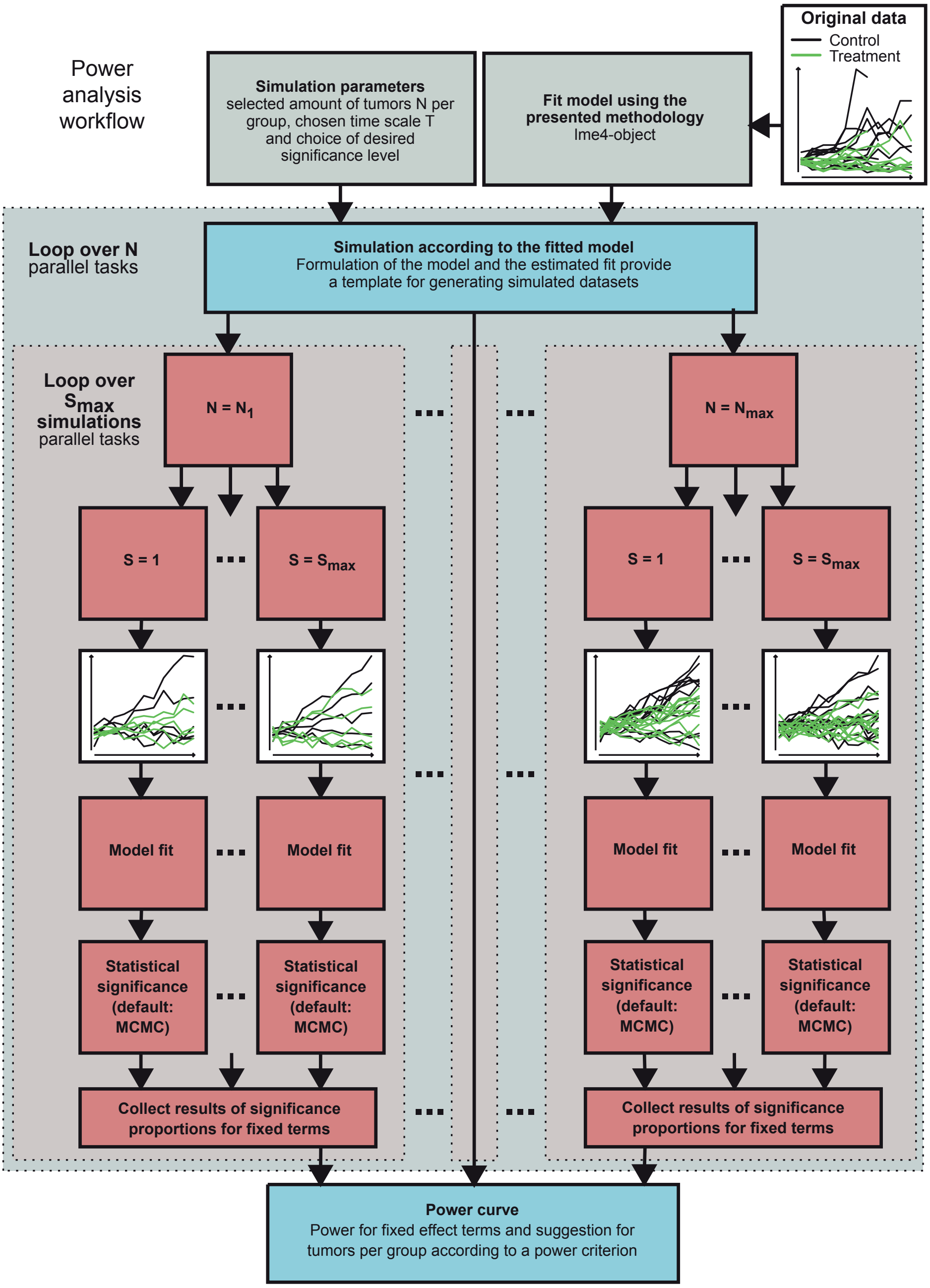
Supplementary Figure S4: Random slopes in the DMBA lower and higher dosage datasets. Since the components for each individual were equivalent in the random effects' model matrices histogram plots of the estimated values were used to assess if heterogeneity was observed also in the fitted random effects (9). Upper panels A,B: DMBA lower dose, Lower panels C,D: DMBA higher dose. (A),(C): With the categorizing model, the distributions for random slopes do not exhibit multimodality or skewness. (B),(D): The random slopes fitted without categorization are highly skewed and multimodal, with the mode of the distribution residing below zero random slope.



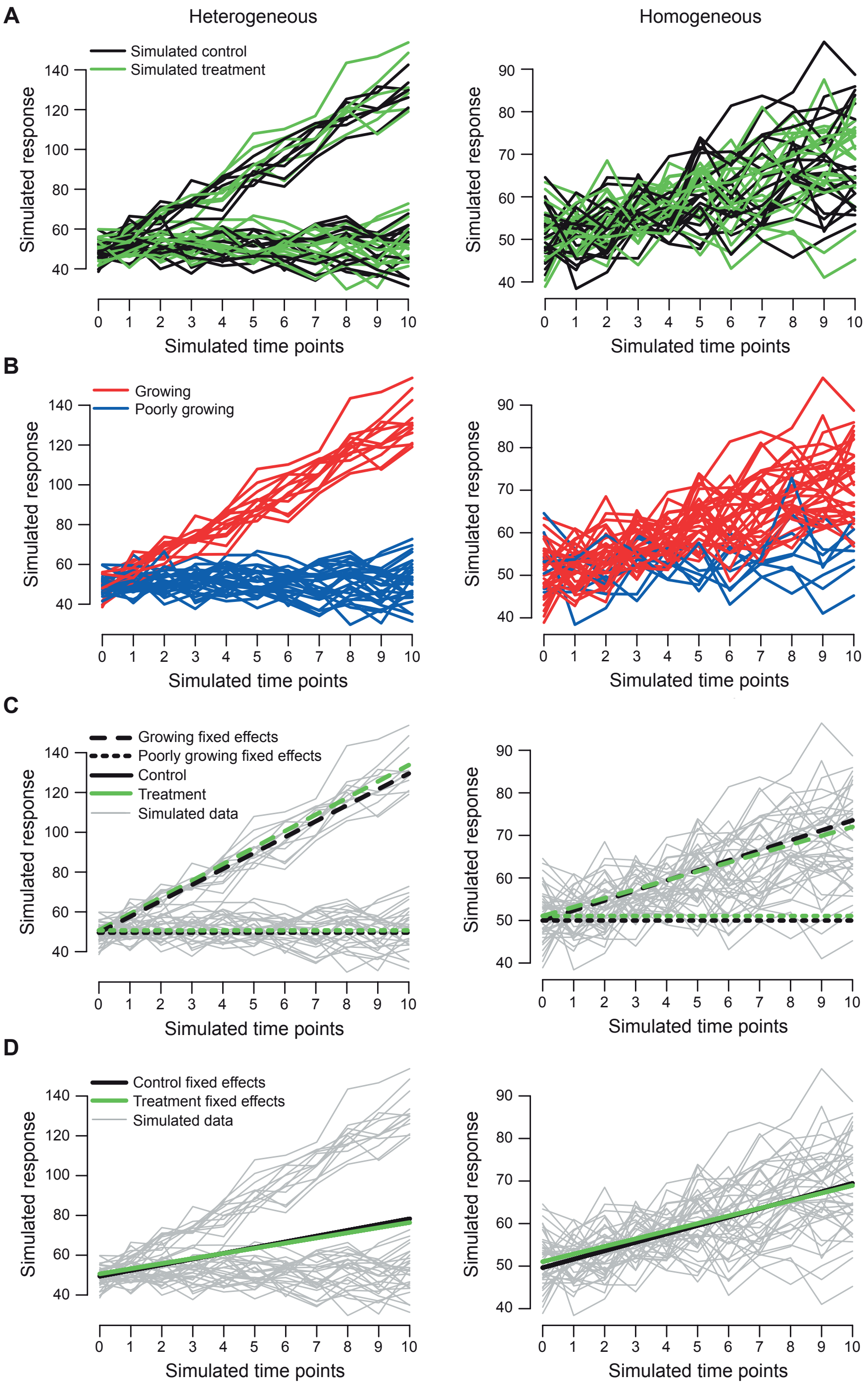
Supplementary Figure S5: Power and precision analysis in the LNCaP DPN experiment. (A) Statistical power with and without categorization. Using the 0.8 power (black dashed line), the suggestion for tumor number per group is 19 with and 25 without categorization to achieve a significant slope treatment effect. The offset effect did not reach the set threshold. (B) Statistical precision for the slope and offset effects when exploring the model accuracy for shorter treatment durations by including varying number of time points from the beginning. (C) Time point -specific statistical precision for the treatment-effect.



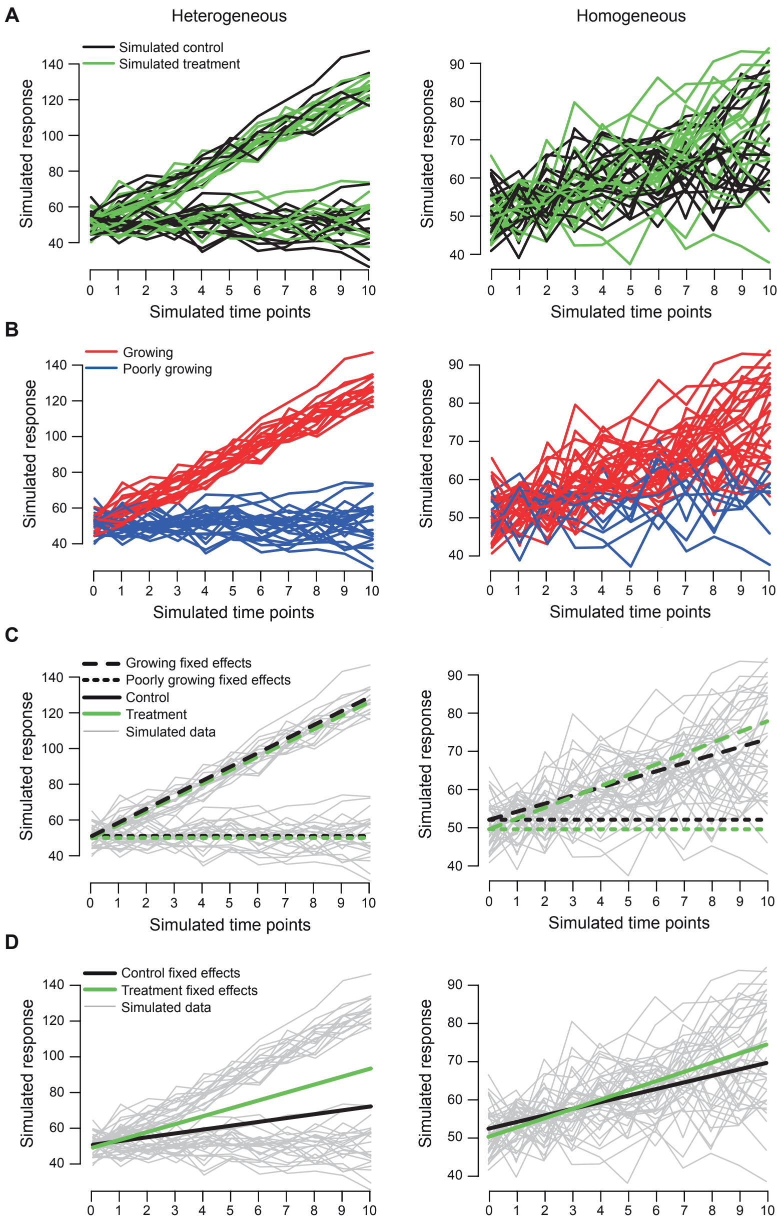
Supplementary Figure S6: Adjusting the model fit according to the validation tools in the 4T1 Doxorubicin experiment. (A) Model fit with the default model covariates does not account for the quadratic growth in the data, which is evident from the visual inspection of the residuals. (B) The adjusted model provides an adequate fit for the quadratic growth profiles.


Power analysis workflow
Simulation parameters
selected amount of tumors N per group, chosen time scale T and choice of desired significance level
Fit model using the presented methodology
lme4-object
Original data
Control
Treatment
Loop over N parallel tasks
Simulation according to the fitted model
Formulation of the model and the estimated fit provide a template for generating simulated datasets
Loop over S_{max} simulations parallel tasks
$N = N_1$
$N = N_{max}$
$S = 1$
$S = S_{max}$
$S = 1$
$S = S_{max}$
Model fit
Model fit
Model fit
Model fit
Statistical significance (default: MCMC)
Statistical significance (default: MCMC)
Statistical significance (default: MCMC)
Statistical significance (default: MCMC)
Collect results of significance proportions for fixed terms
Collect results of significance proportions for fixed terms
Power curve
Power for fixed effect terms and suggestion for tumors per group according to a power criterion

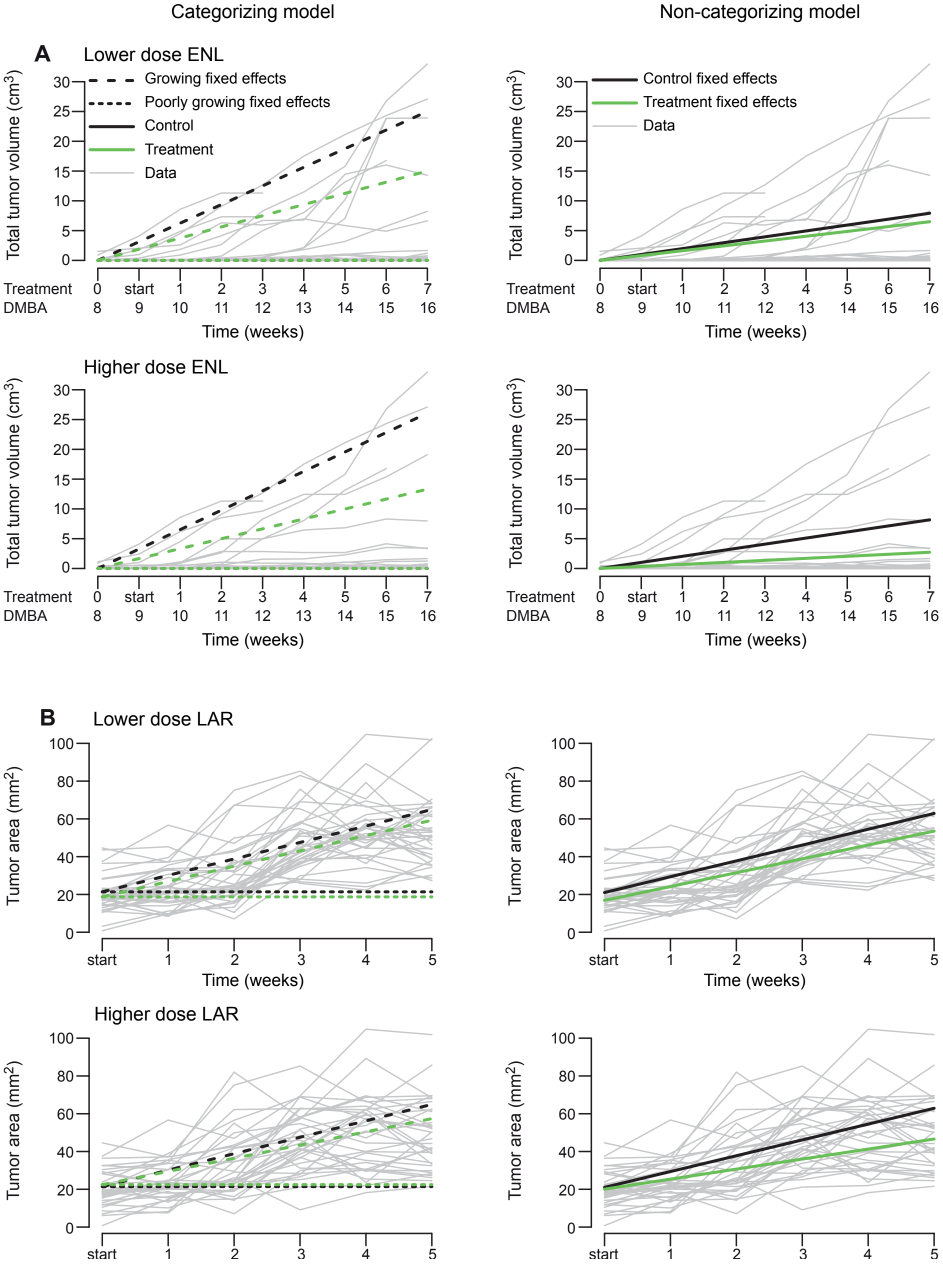
Supplementary Figure S7: Workflow of the power analysis. The fitted model is used as a base for simulation of artificial datasets for sample size estimation. The computational tasks are independent and can be run as separate processes in a more efficient parallel implementation.



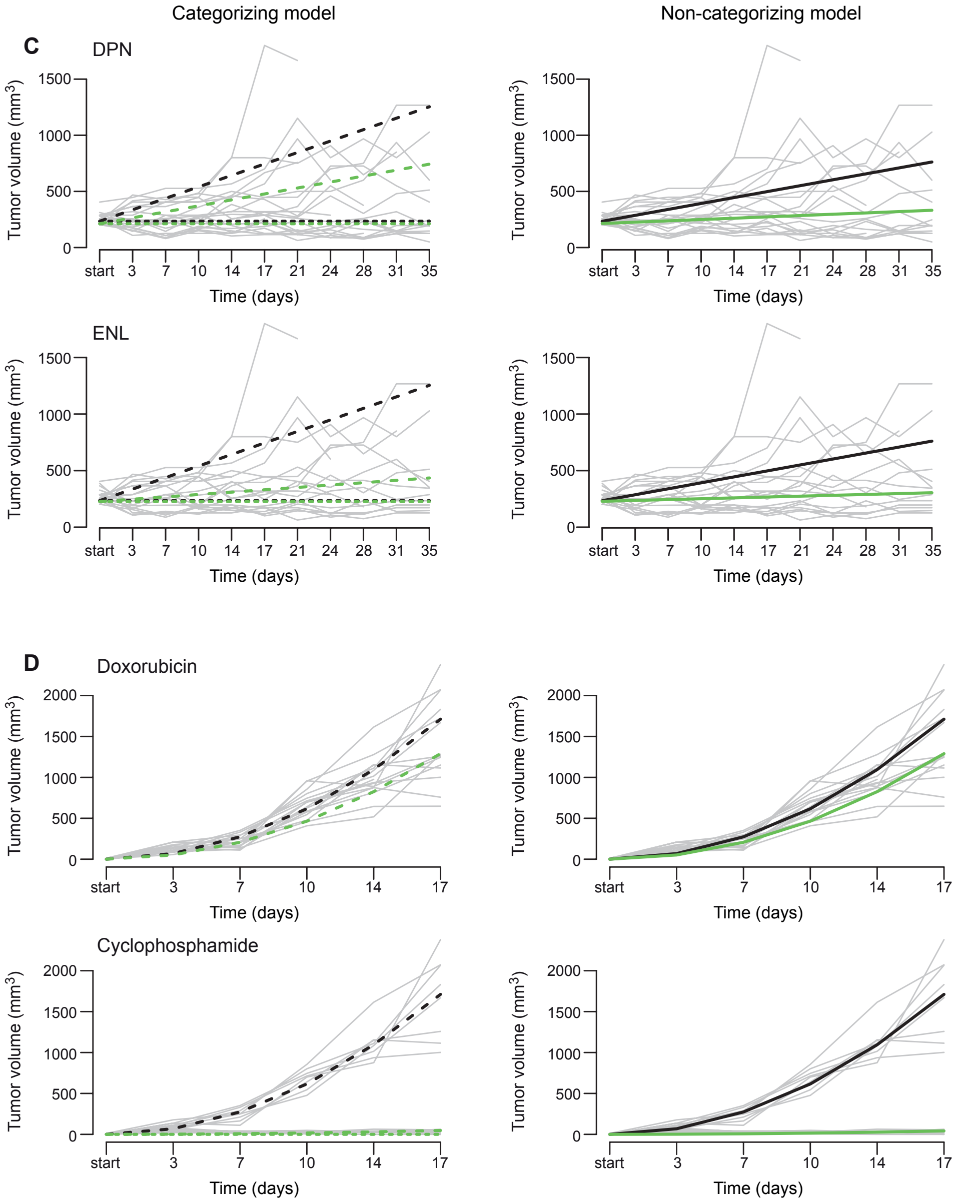
Supplementary Figure S8: Visualization of the null simulation study for heterogeneous (left panel) and homogeneous (right panel) cases. (A) An example of the simulated datasets, (B) Identified categorization in the categorizing approach, (C) Fixed effects inference in the categorizing approach (here $p > 0.05$ for treatment effects), (D) Fixed effects inference in the conventional non-categorizing approach (here $p > 0.05$ for treatment effects).



Supplementary Figure S9: Examples of such null model simulations, where use of model diagnostics would avoid spurious findings. (A) Heterogeneous and homogeneous cases of simulated datasets, (B) Identified tumor sub-categories in the categorizing approach, (C) Fixed effects inference in the categorizing approach, (D) Fixed effects inference in the conventional non-categorizing approach. Here, the conventional model detects a statistically significant slope effect in the example heterogeneous dataset ($p < 0.05$, panel D, left column). However, the fixed effects fit does not capture well the distinct populations of the tumor growth profiles, which resulted in severe bimodality in the random slopes (terms $u_{2,i}$). This indicates that the non-categorizing model fit was not feasible, and that the categorizing approach should have been used instead (panel C, left column). Similarly, the categorizing model detects a statistically significant slope effect in the example homogeneous case ($p < 0.05$, panel C, right column). However, the model fit has resulted here in false convergence and the random base levels (terms $u_{1,i}$) were all estimated as zero. Additionally, this solution led to a significant autocorrelation between the first and last observation of the growth profiles ($p < 0.001$). This indicates that the categorizing model formulation is possibly too complex for this particular dataset, and that an alternative formulation should be considered, such as the conventional model (panel D, right column).



Categorizing model
Non-categorizing model
A
Lower dose ENL
Growing fixed effects
Poorly growing fixed effects
Control
Treatment
Data
Control fixed effects
Treatment fixed effects
Data
Total tumor volume (cm^3)
30
25
20
15
10
5
0
Treatment
DMBA
0
start
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
Time (weeks)
Higher dose ENL
B
Lower dose LAR
Higher dose LAR
Tumor area (mm^2)
100
80
60
40
20
0
start
1
2
3
4
5
Time (weeks)



Supplementary Figure S10: Fixed effect fits from the categorizing (left column) and non-categorizing (right column) models, shown in a parallel fashion to Table 2. (A) DMBA case, (B) MCF-7 case, (C) LNCaP case, (D) 4T1 case. In those studies without a designated tumor target size (panels A, D), the poorly growing category is depicted using the horizontal population profile at zero response, to visualize the poorly growing fixed effect fit even in the absence of the base level and offset terms in the corresponding models (i.e. $b_1 = b_2 = 0$)

ORIGINAL PUBLICATION **III**

Laajala TD*, Seikkula H*, Seyednasrollah F, Mirtti T, Boström PJ, Elo LL. *Longitudinal modeling of ultrasensitive and traditional prostate-specific antigen and prediction of biochemical recurrence after radical prostatectomy.* Scientific Reports. 2016;6:36161.

*: Equal contribution

www.nature.com/scientificreports

SCIENTIFIC REPORTS

OPEN

Received: 26 May 2016

Accepted: 03 October 2016

Published: 02 November 2016

Longitudinal modeling of ultrasensitive and traditional prostate-specific antigen and prediction of biochemical recurrence after radical prostatectomy

Teemu D. Laajala[1,2,*], Heikki Seikkula[3,4,*], Fatemeh Seyednasrollah sadat[1,2], Tuomas Mirtti[5,6], Peter J. Boström[4] & Laura L. Elo[1]

Ultrasensitive prostate-specific antigen (u-PSA) remains controversial for follow-up after radical prostatectomy (RP). The aim of this study was to model PSA doubling times (PSADT) for predicting biochemical recurrence (BCR) and to capture possible discrepancies between u-PSA and traditional PSA (t-PSA) by utilizing advanced statistical modeling. 555 RP patients without neoadjuvant/adjuvant androgen deprivation from the Turku University Hospital were included in the study. BCR was defined as two consecutive PSA values >0.2 ng/mL and the PSA measurements were log_2-transformed. One third of the data was reserved for independent validation. Models were first fitted to the post-surgery PSA measurements using cross-validation. Major trends were then captured using linear mixed-effect models and a predictive generalized linear model effectively identified early trends connected to BCR. The model generalized for BCR prediction to the validation set with ROC-AUC of 83.6% and 95.1% for the 1 and 3 year follow-up censoring, respectively. A web-based tool was developed to facilitate its use. Longitudinal trends of u-PSA did not display major discrepancies from those of t-PSA. The results support that u-PSA provides useful information for predicting BCR after RP. This can be beneficial to avoid unnecessary adjuvant treatments or to start them earlier for selected patients.

Prostate-specific antigen (PSA) is the most widely used tool to detect and monitor prostate cancer (PCa)[1]. PSA detection methods with detection levels under 0.1 ng/mL are considered ultrasensitive and some assays are capable of detecting levels approaching 0.001 ng/mL[2]. The use of ultrasensitive PSA assays (u-PSA) remains controversial due to questions regarding reliability and usefulness of u-PSA[3]. However, u-PSA could potentially detect biochemical recurrence (BCR) after radical prostatectomy (RP) significantly earlier than traditional PSA (t-PSA) assays[4].

Early detection of BCR is important because salvage radiation therapy (RT) is most efficient when given shortly after BCR[5]. BCR is defined here as two or more consecutive PSA values over 0.2 ng/mL concordant to the EAU consensus[6]. Currently, there is no evidence that salvage RT prompted by elevated u-PSA values after RP would improve patient survival. Nevertheless, it could save high-risk patients from unnecessary adjuvant RT and favor more selective salvage RT[7].

[1]Computational Biomedicine Group, Turku Centre for Biotechnology, University of Turku, Turku, Finland. [2]Department of Mathematics and Statistics, University of Turku, Turku, Finland. [3]Department of Surgery, Central Hospital of Central Ostrobothnia, Kokkola, Finland. [4]Department of Urology, Turku University Hospital, Turku, Finland. [5]Department of Pathology (HUSLAB), Helsinki University Hospital, Helsinki, Finland. [6]Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland. *These authors contributed equally to this work. Correspondence and requests for materials should be addressed to L.L.E. (email: laliel@utu.fi)

PSA doubling time (PSADT) has been used to estimate the risk of disease progression after radical surgery; PSADT of nine months or less is an independent risk factor for prostate cancer specific mortality[8]. Detectable u-PSA levels after RP can predict PCa recurrence[9]. Patients with undetectable u-PSA two years after surgery are unlikely to develop rapid clinical progression of PCa (PSADT <9 months if experiencing BCR later)[10]. Based on current literature, the correlation between general PSADT and ultrasensitive doubling times (u-DT) is poor[11]. False positive findings from u-PSA may also originate from laboratory measurement errors[12,13].

The aim of this study was to develop novel tools that reduce the unreliability related to u-PSA. Furthermore, we assessed the potential prognostic significance of u-DT for predicting BCR after RP and applied comprehensive mathematical modeling of u-PSA and t-PSA, in order to establish an accurate predictive link between early measurements of PSA and the risk of BCR.

Methods

Patient material.

Patients undergoing open RP and limited pelvic lymphadenectomy at Turku University Hospital during 2004–2008 were included (n = 604). The follow-up period was a minimum of 6 years. Open RP was performed as initially described by Walsh *et al.*[14]. Patients who received neoadjuvant or adjuvant androgen deprivations (ADT) were excluded; this also meant the exclusion of node positive patients, resulting in 555 patients. For practical reasons all ADTs during the follow-up period were called as adjuvant ADT, resulting population including no patients with hormonal treatment. From the 555 patients, full follow-up information was unavailable for 33 patients and 19 patients died of causes unrelated to PCa, resulting in a final set of 503 patients. Patients who died from other causes than PCa were excluded to avoid bias, as also very early follow-up information from some of these patients was lacking. Patients with adjuvant RT (ART) were not excluded based on our earlier findings demonstrating no differences between DTs in RT patients with or without adjuvant[15]. This study from Seikkula *et al.* composed of almost identical study population, and assessed optimal u-PSA threshold for upcoming BCR. In the conducted multivariate analysis there were no significant differences between patients with or without ART. According to the Finnish national rules and regulations for medical registry studies with retrospective nature no patient consents are required. Study protocol was approved by the IRB of the hospital district of South West Finland and the study was carried out taking into account all the study guidelines and national laws in Finland.

The patients were followed every 3 months for the first year after the surgery and semiannually thereafter. The follow-up included a physical examination and u-PSA measurements. Data was collected retrospectively from Turku University Hospital's medical records and PSA data was obtained from Turku University Hospital laboratory data sources. All the PSA-analyses were done with electrochemiluminescence-immunoassay (ECLIA, Roche Diagnostics GmbH), which has a lowest limit of detection (LLD) 0.003 ng/mL. The collected data included essential clinicopathological variables, neoadjuvant and adjuvant therapies, and follow-up information, which were later used also in multivariate analysis of a potentially more accurate LASSO penalized prediction model by expanding beyond just PSA-derived information.

Processing of PSA measurements.

PSA measurements with non-detected quantities were imputed using the smallest non-zero measurement. Of all the eligible post-surgery measurements, 4502 (79.6%) were u-PSA ($\leq$0.1 ng/mL) and 1151 (20.4%) t-PSA ($>$0.1 ng/mL). Post-surgery PSA nadir was defined as the lowest PSA measurement within a 3 month window after surgery. The 3 month period was chosen because 8 weeks is ample time to allow PSA levels to clear after RP and detectable u-PSA values in 1–3 months after RP are suggested as a marker for BCR progression[9,16]. The mathematical modeling was based only on post-nadir measurements prior to possible salvage treatments.

To evaluate the generalization ability of the modeling, the data was randomized into 3 subgroups of subjects prior to model development, where factors such as age, BCR status, and Gleason score (GS) were balanced. 2 of the subgroups were randomly chosen as the exploratory data and fully utilized in model development. Within this exploratory data, generalization ability was maintained through cross-validation. The remaining third of the data was utilized as a validation set, to retain an objective view to the robustness of the final model (Table 1).

Mathematical modeling.

Cubic penalized splines were used in the exploratory set with a wide range of values for the spline smoothing parameter λ. The optimal smoothing parameter was identified by minimizing the cross-validation Median Squared Error (MSE) of the spline fits. Penalized splines provided a flexible approach to explore whether the log_2-transformed PSA would display complex non-linear patterns (low λ) or linear patterns (high λ).

Based on the observed highly linear patterns of the log_2-transformed PSA, a linear mixed-effects model was built. The parameter estimates of the model for the log_2-PSA nadir and PSADT were used for detecting differences between the BCR and non-BCR patients. A clinical risk assessment tool was further derived using generalized linear mixed-effects models as a binary classifier for BCR using parameter derivatives from the patient-wise nadir and PSADT. Furthermore, we then subjected the binary classification task of nadir and PSADT along with clinical parameters from Table 1 to penalized LASSO regression, where the multivariate regression model is optimized to maximal generalizability by penalizing the inclusion of non-zero coefficients, i.e. non-informative, overlapping or correlated variables are eliminated.

The mathematical modeling was conducted using the R statistical software (version 3.2)[17], along with the R-packages *psplines*[18], *lme4*[19] and *glmnet*[20] for the penalized cubic splines, linear mixed-effects models and penalized LASSO regression, respectively. See the Supplementary Methods for a more detailed description of the mathematical modeling process, including splines, linear mixed-effects models and the LASSO multivariate regression.

		Dataset			
Variable	*Instance*	*Exploratory 2/3*		*Validation 1/3*	
pT	**2**	180 (53.3%)		92 (55.8%)	
	3	156 (46.2%)		73 (44.2%)	
	4	1 (0.3%)			
	Missing	1 (0.3%)			
Gleason score (GS)	**≤6**	157 (46.4%)		80 (49.1%)	
	7 (3+4)	101 (29.9%)		49 (30.1%)	
	7 (4+3)	50 (14.8%)		18 (11.0%)	
	≥8	28 (8.3%)		16 (9.8%)	
	Missing	2 (0.6%)			
Margins	**Negative**	200 (59.2%)		100 (60.6%)	
	Positive	137 (40.5%)		65 (39.4%)	
	Missing	1 (0.3%)			
Adjuvant RT	**No**	295 (87.3%)		147 (89.1%)	
	Yes	42 (12.4%)		18 (10.9%)	
	Missing	1 (0.3%)			
Salvage RT	**No**	275 (81.4%)		136 (82.4%)	
	Yes	63 (18.6%)		29 (17.6%)	
PSA at surgery	**<10**	251 (74.3%)		121 (73.3%)	
	10–20	67 (19.8%)		36 (21.8%)	
	≥20	19 (5.6%)		8 (4.8%)	
	Missing	1 (0.3%)			
Age	**<60**	123 (36.4%)		61 (37.0%)	
	60–70	193 (57.1%)		96 (58.2%)	
	>70	21 (6.2%)		8 (4.8%)	
	Missing	1 (0.3%)			
Total counts of PSA measurements in different time windows	**Time post-surgery**	*t-PSA*	*u-PSA*	*t-PSA*	*u-PSA*
	<1y	166	875	161	466
	1y–3y	120	788	78	413
	>3y	236	1000	164	649
Patient status	**No recurrence**	279 (82.5%)		140 (84.8%)	
	Recurrence (BCR)	52 (15.4%)		22 (13.3%)	
	Metastasis/other	7 (2.1%)		3 (1.8%)	

Table 1. Patient characteristics, PSA measurement counts, and patient counts in the exploratory and validation datasets (proportions in parentheses).

Results

Detailed patient characteristics are reported in Table 1. Majority of the post-surgery PSA measurements were detectable only in the u-PSA range: 83.6% and 79.1% in the exploratory and validation sets, respectively. There were 156 (46.2%) and 73 (44.2%) patients with pT3, and nearly half of the patients had Gleason $\leq$ 6. Rate of positive margins was approximately 40%. Only 15.4% and 13.3% of the patients reached BCR during follow-up. Representative longitudinal curves of 30 randomly chosen patients are shown prior and post to the log_2-transformation in Fig. 1a,b, respectively. Due to the log_2-transformation a unit change corresponded to PSA doubling in the original PSA scale. For detailed modeling results, see the Supplementary Results within the Supplementary Material.

Splines and linear parametric models. The major PSA trends were effectively captured readily by linear components in the model based on the optimality of high values of the smoothing parameter λ (Fig. 1c) as well as upon visual inspection (Fig. 1d–f; Fig. 2a,b). Interestingly, the first order derivatives that capture longitudinal changes in PSADT clearly distinguished between the BCR and non-BCR, suggesting that longitudinal follow-up of PSADT could provide an accurate predictor of BCR (Fig. 2c). The u-PSA and t-PSA did not exhibit markedly different patterns in the splines (Fig. 2b,c).

Since splines suggested that linear model families were suitable for modeling the log_2-PSA patterns, we fitted linear regression models to perform parametric inference for the population effects. The focus was on the log_2-PSA nadir and PSADT. Patient-wise estimates for these coefficients are shown in Fig. 3a,b with 1 or 3 year follow-up, respectively.

Finally, generalized linear models were used as binary classifiers to connect the patient-wise characteristics from Fig. 3a,b to the known BCR statuses. The prediction accuracy using 1 year or 3 year post-nadir follow-up was 85.3% or 88.8%, respectively, using the prediction surfaces provided in Fig. 3c,d. Overall, only minor variation was detected between the u-PSA and t-PSA in model diagnostics, exemplified by the slight decrease of

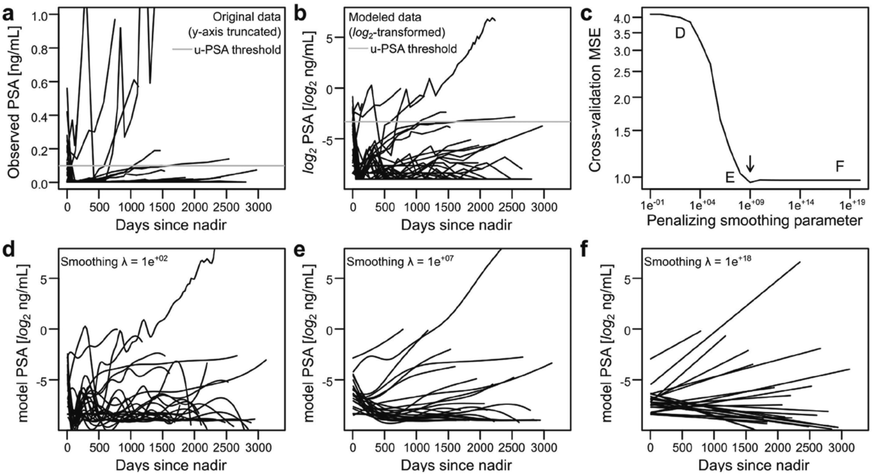


Figure 1. Longitudinal PSA profiles for 30 randomly chosen patients using penalized cubic splines. (**a**) The raw PSA-profiles exhibited varying patterns as a function of time since post-surgery nadir. (**b**) After *log*$_2$-transformation, unit increase in the response corresponds to doubling in the original scale. (**c**) Model complexity was chosen according to Cross-Validation (CV) Median Squared Error (MSE). Optimal model ($\lambda = 10^9$) is indicated with the arrow. (**d**–**f**) Example model fits for varying λ are shown for the *log*$_2$-scale data from panel **b**.

heteroscedasticity over the threshold for model residuals (Fig. 3e). A computational example for predicting future patient risks with our given model estimates is provided for the mathematically inclined readers through the conventional theoretical connection to simple linear regression in the Supplementary Table S1, which generalizes to any standard spreadsheet software.

Clinical parameters, such as pT-classification or GS, in connection to the patient-wise estimates of PSA nadir and DT are reported in Supplementary Table S2. GS classes (≤ 6, 7 or ≥ 8), positive subsequent salvage treatment status, and a pre-surgery PSA >10 ng/mL were associated to differences in post-nadir PSADT estimated by the model. For the *log*$_2$-PSA nadir model parameter, multiple associations were identified, excluding GS, indicating that multiple clinical parameters may be associated with the sensitive detections possible only in the u-PSA range (Supplementary Table S2). Their interpretation remains to be further studied, thus the prediction model was based solely on PSA trends.

When LASSO regression was cross-validated (CV) and the optimal model was fitted using penalization parameter within a single standard error of minimal CV error (Supplementary Fig. S2a), the multivariate regression model proposed utilizing only the estimated PSA nadir and PSADT as variables for BCR prediction. While multiple clinical variables were informative and almost included (Supplementary Fig. 2b), the generalized multivariate model highlights usefulness of the nadir and PSADT over conventional clinical parameters.

Validation. One representative third of the data was left for objective validation of the modeling procedure in a wider context (Table 1 right panel). The validation predictions resulted in high sensitivity and specificity both for the 1 and 3 year models (Fig. 3f) with the Area Under the ROC-curve (ROC-AUC) of 0.836 (95% CI 0.72–0.96) and 0.951 (95% CI 0.91–0.99), respectively.

Graphical user-interface pipeline for future predictions. In order to provide the analysis pipeline widely accessible to clinicians, a graphical user interface (GUI) was implemented using the R Shiny (RStudio Inc) platform with the underlying mathematical methodology outlined in the Supplementary Methods. The GUI is freely available at the Shinyapps.io (RStudio Inc) service-platform (http://compbiomed.shinyapps.io/u-pa/). The tool allows automated analysis of novel measurements with the existing methodology, and is provided with the exploratory dataset for illustrative purposes. Its design allows clinicians to conveniently run the pipeline and generate PDF-based risk reports for new patients. A typical workflow of the GUI is presented in Fig. 4.

Discussion

In the current study we applied mathematical modeling to investigate the role of u-PSA as means of follow-up after RP. Based on our results, u-PSA provides useful information for predicting BCR after RP and we developed an easily applicable prediction platform (Fig. 4), which to our knowledge is the first clinically relevant predictive tool focused on u-PSA. Our results show highly linear trends in PSADT (Fig. 1). This offers a clinically convenient

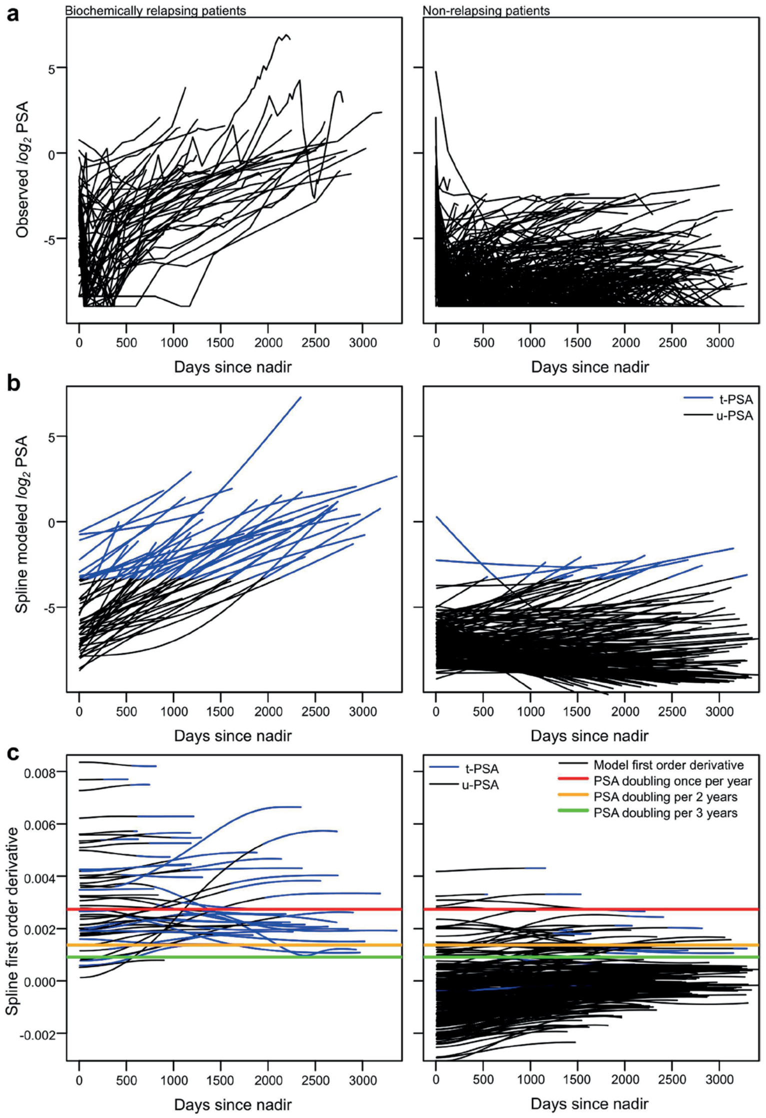


Figure 2. All the modeled exploratory data, model fits and the first order derivatives of the penalized splines for the relapsing (left column; $N = 52$) and non-relapsing patients (right column; $N = 279$). (**a**) Modeled log_2-transformed data. (**b**) Corresponding penalized cubic spline fits. (**c**) The first order derivatives. With few exceptions, derivatives maintained relatively constant levels over the follow-up period. Once per year or once per two years PSA doubling criteria were good indicators of relapse or non-relapse of patients. Noticeable differences between u-PSA (black) and t-PSA (blue) were not present.

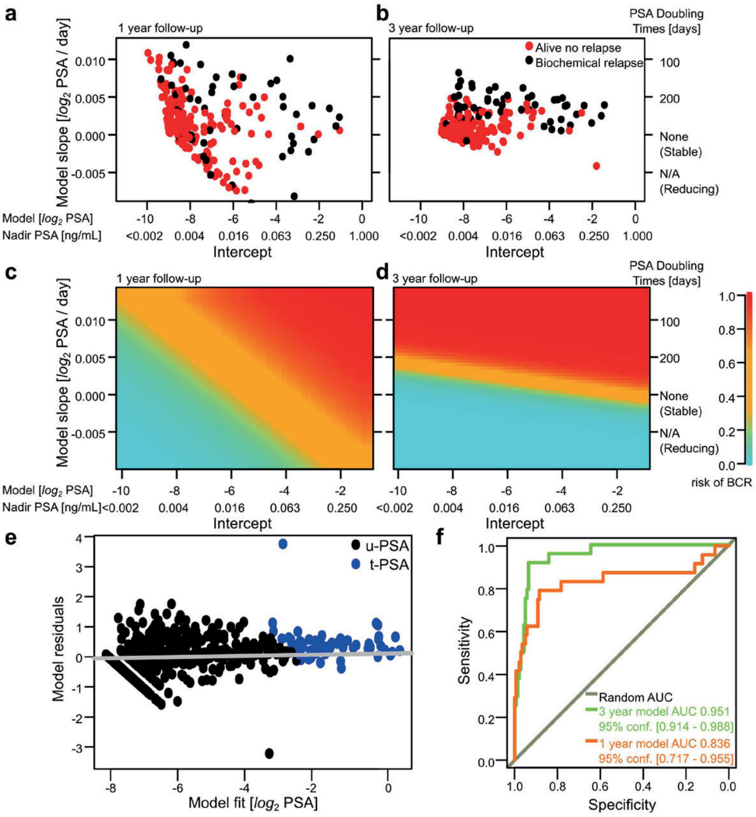
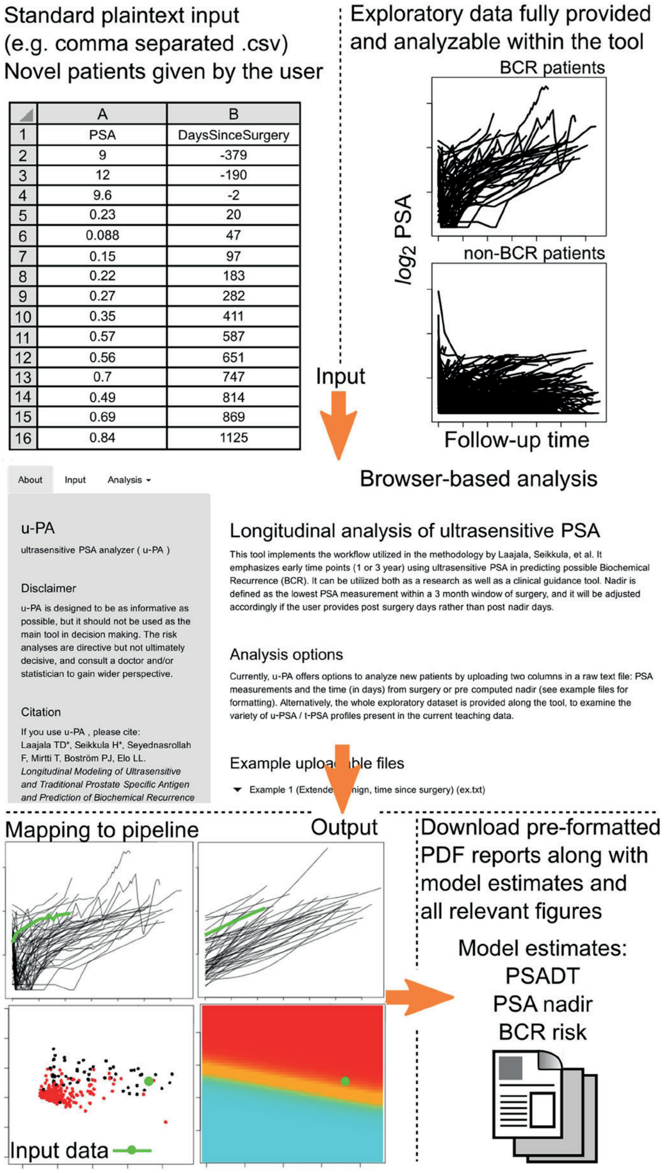


Figure 3. (**a**,**b**) Linear mixed-effects models yielded estimates for patient-specific nadir intercept and doubling coefficient using a 1 year (panel **a**) or a 3 year post-nadir censoring window (panel **b**). (**c**,**d**) Using generalized regression, we identified prediction surfaces for the risk of BCR using the 1 year (panel **c**) or 3 year post-nadir time window (panel **d**). Logistic regression predictions for the generalized linear models for the generalized linear models were annotated using the color key on the right. (**e**) Regression residuals for the 1 year post-nadir window using linear-mixed effects models display slight decrease in residual variance as a function of u-PSA versus t-PSA, though no systematic increasing or decreasing trends were detected. (**f**) The validation dataset suggested high predictive accuracy for BCR using the fitted models from the exploratory portion of data.

analysis approach, as raw PSA measurements may be transformed to PSADT through the log_2-transformation, after which the linear trends may be captured using conventional tools widely available in any statistical or spreadsheet software. According to previous studies the specificity of u-PSA is poor[7], but in our study we show that by using sophisticated computational techniques the sensitivity and specificity are high.

Based on our analysis of the second order spline derivatives, major trends in u-PSA curvature were established by the end of the first year and only slight individual variation occurred thereafter (Supplementary Fig. 1). Stabilization in the curvature of the splines suggested consistent changes in log_2-PSA after the first follow-up year. Motivated by the spline analysis supporting linear trends (high optimal smoothing parameter λ), we tested parametric linear inference using both a 1 year and a 3 year post-nadir window. A 3 year time window was also utilized by Malik *et al.*, in a study where they assessed non-detectable vs. detectable u-PSA with the threshold of 0.05 ng/mL[21]. Detectable u-PSA 2–2.5 years after RP was independent prognostic factor for PSA progression also according to Chang *et al.,* but in their study the presence of a detectable u-PSA level earlier than 2 years from surgery did not reliably predict the subsequent clinical course of BCR[11]. Although our model can predict BCR accurately already within an early follow-up window after surgery, it may still suffer from over-diagnosis related issues and



Standard plaintext input
(e.g. comma separated .csv)
Novel patients given by the user
Exploratory data fully provided
and analyzable within the tool
BCR patients
non-BCR patients
log_2 PSA
Follow-up time
Input
Browser-based analysis

	A	B
1	PSA	DaysSinceSurgery
2	9	-379
3	12	-190
4	9.6	-2
5	0.23	20
6	0.088	47
7	0.15	97
8	0.22	183
9	0.27	282
10	0.35	411
11	0.57	587
12	0.56	651
13	0.7	747
14	0.49	814
15	0.69	869
16	0.84	1125

u-PA

ultrasensitive PSA analyzer (u-PA)

Disclaimer

u-PA is designed to be as informative as possible, but it should not be used as the main tool in decision making. The risk analyses are directive but not ultimately decisive, and consult a doctor and/or statistician to gain wider perspective.

Citation

If you use u-PA , please cite:
Laajala TD*, Seikkula H*, Seyednasrollah F, Mirtti T, Boström PJ, Elo LL.
Longitudinal Modeling of Ultrasensitive and Traditional Prostate Specific Antigen and Prediction of Biochemical Recurrence

Longitudinal analysis of ultrasensitive PSA

This tool implements the workflow utilized in the methodology by Laajala, Seikkula, et al. It emphasizes early time points (1 or 3 year) using ultrasensitive PSA in predicting possible Biochemical Recurrence (BCR). It can be utilized both as a research as well as a clinical guidance tool. Nadir is defined as the lowest PSA measurement within a 3 month window of surgery, and it will be adjusted accordingly if the user provides post surgery days rather than post nadir days.

Analysis options

Currently, u-PA offers options to analyze new patients by uploading two columns in a raw text file: PSA measurements and the time (in days) from surgery or pre computed nadir (see example files for formatting). Alternatively, the whole exploratory dataset is provided along the tool, to examine the variety of u-PSA / t-PSA profiles present in the current teaching data.

Example uploa ble files

▾ Example 1 (Extende nign, time since surgery) (ex.txt)

Figure 4. Graphical user interface workflow for predicting future patients or for analyzing the provided exploratory dataset of the current study.

patient-specific risk evaluation is recommended[12,22]. Earlier studies assessing the risk of PCa-progression from u-PSA have detected one year average lead time from detectable u-PSA threshold to BCR[7]. According to our analyses, the 2 year follow-up period utilized by Chang *et al.* and 3 year utilized by Malik *et al.* may already be well established by the end of the first year post-nadir[10,21]. In our analyses the ability to distinguish between the non-BCR (>80%) and BCR (<20%) patients was only marginally improved if 3 years of post-nadir follow-up was allowed instead of 1 year (Fig. 3f).

Generalized linear regression model can be used as a binary classifier to evaluate BCR risks for future patients, for example by mapping the potential patients to the prediction surfaces provided by the current modeling process (Fig. 3c,d). The validation dataset in this study highly supported the hypothesis that predictive accuracy could be obtained readily by the end of the first follow-up year (Fig. 3f). We provide a practical computational example for the validated generalized linear models for predicting the BCR risk of a patient (Supplementary Table S1) and an easy-to-use graphical user interface that is freely available at http://compbiomed.shinyapps.io/u-pa/ (Fig. 4).

When t-PSA levels are used, nadir after surgery is usually undetectable (<0.1 ng/mL)[23]. In u-PSA range the undetectable level has a wider spectrum[24]. Currently the significant threshold level of u-PSA relapse is unknown. Recently we suggested a threshold between 0.03–0.05 ng/mL[15]. Malik *et al.* showed a clear association for delayed BCR with u-PSA values of <0.05 to >0.05 ng/mL 3 years after RP[21]. Previously, clear survival benefit was shown among men with low u-PSA nadir after RP[9]. In this study, the nadir intercept and PSADT estimates were found to be highly statistically significantly associated with BCR. Our definition of PSA nadir was the lowest PSA measurement within a 3 month window from RP. More sophisticated parametric methods to determine nadir include piece-wise change-point models, which can incorporate knots that are inter-connected with linear curves[25]. In order to assess the true cutoff-point for reliable u-PSAs LLD, modeling the exact time to nadir would be an interesting future research question.

In our exploratory dataset, the median time between two subsequent post-surgery PSA measurements was 152 days. Therefore, the first year of follow-up mostly consisted of 3 measurements. This amount is the minimal number of observations required to fit a linear regression model. The 3 year follow-up period was less sensitive to the nadir point and more likely holds a more realistic amount of observations for reliable PSADT estimation. However, it remains to be validated to what extent the doubling trends are established by the end of the first year, and for this purpose more intensive coverage of PSA trends would be already required for the early follow-up.

Accurate methods to determine the clinical risk represented by a rising PSA value are critical to develop rational treatment strategies. So far no studies have demonstrated that u-PSA triggered therapy will improve outcome. On the other hand, u-PSA kinetics may provide predictive information. Only few studies have compared DTs in traditional and ultrasensitive ranges[10,11,15,26]. It is possible that past negative findings for u-PSA have been susceptible to utilizing single measurements as predictors[10,11,26]. When multiple measurements are not considered, variation in single measurements may dominate instead of averaging more coherent trends through regression curves. This highlights the need for feasibly chosen mathematical models that capture all patient-specific variation in a more effective manner. Some authors claim that u-PSA measurements are helpful to determinate early BCR after RP[4,27,28]. Others claim that it will offer no benefit and mainly cause unnecessary anxiety for patients[29]. Previously Reese *et al.* demonstrated a poor correlation between PSADTs, possibly due to inconsistency of u-PSA measurements[11]. Some authors have reported unreliability of u-PSA measurements[13,30]. Also according to literature, specificity of the u-PSA is relatively poor[7]. In our study, when utilizing sophisticated mathematical modeling over time we identified no major discrepancies between the u-PSA and t-PSA. In contrast, large portion of our data at 1 year window post-nadir consisted of u-PSA (Table 1; median 3 measurements by the end of the first follow-up year), while retaining a good prediction and generalization ability. Most importantly, because u-PSA may improve the time to detection as a supplement to t-PSA of BCR by months or years, this advantage of earlier prediction for relapse has potential to improve the patients' chance of durable progression-free survival with salvage therapy given at a lower cancer burden and a wider time window for cure[4,31]. Furthermore, by presenting mathematically extensive approach with both univariate and multivariate modeling of BCR, we highlight the need of accurate prediction tools that outperform and raise awareness that arbitrary chosen simple thresholds (e.g. in t-PSA range) are likely to be subpar.

Major strength of this study is the extensive mathematical modeling of both the u-PSA and t-PSA measurements, all of which is offered as an easy to use web-based graphical user interface (GUI) platform. All the PSA measurements were done with the same PSA assay, reducing error caused by varying assays. A limitation of the study is that all the patients were from the same hospital district, and thus a larger sample size and longer follow-up is needed for more accurate validation of these findings in order to guarantee generalizability.

Conclusions

Our results indicate that u-PSA provides useful information for predicting BCR after RP. The utilized approach of considering PSADT was easily achieved using log_2-transformation of the data, and makes our conclusions and estimates comparable to any study utilizing the well-established PSADT as an end-point[32,33]. Using this convenient approach, we developed a novel mathematical modeling a mathematical modeling pipeline and implementation utilizing only PSADT and PSA nadir for predicting BCR mainly based on u-PSA measurements in early follow-up of PSA response. To our knowledge this is the first clinically relevant predictive tool focused on systematically complementing t-PSA with u-PSA and displaying the coherence between the two; however, for future studies to be clinically widely applicable, albeit thresholds have been suggested[15], a more extensive exploration of u-PSA key thresholds is imperative. We believe that such threshold most likely would be a combination of the estimated nadir (in u-PSA range), the readily established PSADT potentially expanding both u-PSA and t-PSA ranges, and would most likely also include other clinically relevant variables (Supplementary Table S2).

We believe that in salvage RT policy early risk evaluation is beneficial and we are optimistic about the predictive use of u-PSA in supplement to the more established t-PSA measurements. Our easily accessible mathematical pipeline established a novel baseline for future validation studies of u-PSA importance and method development.

References

1. Welch, H. G. & Albertsen, P. C. Prostate cancer diagnosis and treatment after the introduction of prostate-specific antigen screening: 1986–2005. *J Natl Cancer Inst* **101,** 1325–1329, doi: 10.1093/jnci/djp278 (2009).
2. Ferguson, R. A., Yu, H., Kalyvas, M., Zammit, S. & Diamandis, E. P. Ultrasensitive detection of prostate-specific antigen by a time-resolved immunofluorometric assay and the Immulite immunochemiluminescent third-generation assay: potential applications in prostate and breast cancers. *Clin Chem.* **42,** 675–684 (1996).
3. Heidenreich, A. *et al.* EAU guidelines on prostate cancer. Part II: Treatment of advanced, relapsing, and castration-resistant prostate cancer. *Eur Urol.* **65,** 467–479, doi: 10.1016/j.eururo.2013.11.002 (2014).
4. Shen, S., Lepor, H., Yaffee, R. & Taneja, S. S. Ultrasensitive serum prostate specific antigen nadir accurately predicts the risk of early relapse after radical prostatectomy. *J Urol.* **173,** 777–780 (2005).
5. Eggener, S. E. *et al.* Predicting 15-year prostate cancer specific mortality after radical prostatectomy. *J Urol.* **185,** 869–875, doi: 10.1016/j.juro.2010.10.057 (2011).
6. Boccon-Gibod, L. *et al.* Management of prostate-specific antigen relapse in prostate cancer: a European Consensus. *Int J Clin Pract.* **58,** 382–390 (2004)
7. Tilki, D., Kim, S. I., Hu, B., Dall'Era, M. A. & Evans, C. P. Ultrasensitive prostate specific antigen and its role after radical prostatectomy: a systematic review. *J Urol.* **193,** 1525–1531, doi: 10.1016/j.juro.2014.10.087 (2015).
8. Freedland, S. J. *et al.* Risk of prostate cancer-specific mortality following biochemical recurrence after radical prostatectomy. *JAMA* **294,** 433–439 (2005).
9. Eisenberg, M. L., Davies, B. J., Cooperberg, M. R., Cowan, J. E. & Carroll, P. R. Prognostic implications of an undetectable ultrasensitive prostate-specific antigen level after radical prostatectomy. *Eur Urol.* **57,** 622–629, doi: 10.1016/j.eururo.2009.03.077 (2010).
10. Chang, S. L. *et al.* Freedom from a detectable ultrasensitive prostate-specific antigen at two years after radical prostatectomy predicts a favorable clinical outcome: analysis of the SEARCH database. *Urology* **75,** 439–444 (2010).
11. Reese, A. C., Fradet, V., Whitson, J. M., Davis, C. B. & Carroll, P. R. Poor agreement of prostate specific antigen doubling times calculated using ultrasensitive versus standard prostate specific antigen values: important impact on risk assessment. *J Urol.* **186,** 2228–2232 (2011).
12. Ellis, W. J. *et al.* Early detection of recurrent prostate cancer with an ultrasensitive chemiluminescent prostate-specific antigen assay. *Urology* **50,** 573–579, doi: 10.1016/s0090-4295(97)00251-3 (1997).
13. Yu, H. & Diamandis, E. P. Measurement of serum prostate specific antigen levels in women and in prostatectomized men with an ultrasensitive immunoassay technique. *J Urol* **153,** 1004–1008 (1995).
14. Walsh, P. C., Lepor, H. & Eggleston, J. C. Radical prostatectomy with preservation of sexual function: anatomical and pathological considerations. *Prostate* **4,** 473–485 (1983).
15. Seikkula, H. *et al.* Role of ultrasensitive prostate-specific antigen in the follow-up of prostate cancer after radical prostatectomy. *Urol Oncol.*, doi: 10.1016/j.urolonc.2014.10.010 (2014).
16. Oesterling, J. E. Prostate specific antigen: a critical assessment of the most useful tumor marker for adenocarcinoma of the prostate. *J Urol.* **145,** 907–923 (1991).
17. R Core Team, R: A Language and environment for statistical computing, R software version 3.3.1. http://www.r-project.org/ (Date of access: 21/06/2016) (2016).
18. Ripley, B. *psplines: Smoothing splines with penalties on order m derivatives, R-package version 1.0–16*. http://cran.r-project.org/package=pspline (2013) (Date of access: 21/06/2015).
19. Bates, D. *et al. lme4:Linear mixed-effects models using Eigen and S4, R-package version 1.1–7*. http://cran.r-project.org/package=lme4 (2014) (Date of access: 21/06/2016).
20. Friedman, J., Hastie T., Simon N. & Tibshirani R. *glmnet: Lasso and elastic-net regularized generalized linear models, R-package version 2.0–5*. http://cran.r-project.org/package=glmnet (2016) (Date of access: 02/08/2016).
21. Malik, R. D., Goldberg, J. D., Hochman, T. & Lepor, H. Three-year postoperative ultrasensitive prostate-specific antigen following open radical retropubic prostatectomy is a predictor for delayed biochemical recurrence. *Eur Urol.* **60,** 548–553, doi: 10.1016/j.eururo.2011.05.036 (2011).
22. Witherspoon, L. R. Early detection of cancer relapse after prostatectomy using very sensitive prostate-specific antigen measurements. *Br J Urol.* **79** Suppl 1, 82–86 (1997).
23. Stamey, T. A. *et al.* Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate. II. Radical prostatectomy treated patients. *J Urol.* **141,** 1076–1083 (1989).
24. Sokoll, L. J. *et al.* Do Ultrasensitive PSA Measurements Have a Role in Predicting Long-Term Biochemical Recurrence-Free Survival in Men Following Radical Prostatectomy? *J Urol.*, doi: 10.1016/j.juro.2015.08.080 (2015).
25. Zhao, L. *et al.* Bayesian hierarchical changepoint methods in modeling the tumor growth profiles in xenograft experiments. *Clin Cancer Res.* **17,** 1057–1064, doi: 10.1158/1078-0432.ccr-10-1935 (2011).
26. Teeter, A. E. *et al.* Does early prostate-specific antigen doubling time (ePSADT) after radical prostatectomy, calculated using PSA values from the first detectable until the first recurrence value, correlate with standard PSADT? A report from the Shared Equal Access Regional Cancer Hospital Database Group. *BJU Int.* **104,** 1604–1609 (2009).
27. Doherty, A. P. *et al.* Undetectable ultrasensitive PSA after radical prostatectomy for prostate cancer predicts relapse-free survival. *Br J Cancer* **83,** 1432–1436, doi: 10.1054/bjoc.2000.1474 (2000).
28. Haese, A. *et al.* Ultrasensitive detection of prostate specific antigen in the followup of 422 patients after radical prostatectomy. *J Urol.* **161,** 1206–1211 (1999).
29. Taylor, J. A. 3rd, Koff, S. G., Dauser, D. A. & McLeod, D. G. The relationship of ultrasensitive measurements of prostate-specific antigen levels to prostate cancer recurrence after radical prostatectomy. *BJU Int.* **98,** 540–543, doi: 10.1111/j.1464-410X.2006.06294.x (2006).
30. Khosravi, M. J., Papanastasiou-Diamandi, A. & Mistry, J. An ultrasensitive immunoassay for prostate-specific antigen based on conventional colorimetric detection. *Clin Biochem.* **28,** 407–414 (1995).
31. Stephenson, A. J. *et al.* Predicting the outcome of salvage radiation therapy for recurrent prostate cancer after radical prostatectomy. *J Clin Oncol.* **25,** 2035–2041, doi: 10.1200/JCO.2006.08.9607 (2007).
32. Patel, A., Dorey, F., Franklin, J. & deKernion, J. B. Recurrence patterns after radical retropubic prostatectomy: clinical usefulness of prostate specific antigen doubling times and log slope prostate specific antigen. *J Urol.* **158,** 1441–1445 (1997).
33. Pound, C. R. *et al.* Natural history of progression after PSA elevation following radical prostatectomy. *JAMA* **281,** 1591–1597 (1999).

Acknowledgements

The authors would like to thank PhD Johanna Mäkelä for helpful comments regarding the manuscript. T.D.L. was funded by the Drug Research Doctoral Programme (DRDP, University of Turku) and the Finnish Cultural Foundation. T.M. was funded by the Cancer Society of Finland and the Academy of Finland (grant 268531). L.L.E. was funded by the Sigrid Juselius Foundation and the Academy of Finland (grant 296801).

Author Contributions

T.D.L. and L.L.E.: Designed the statistical methodology; T.D.L.: Implemented the statistical methodology; T.D.L: Implemented the web-based tool; H.S. and T.M.: Provided and processed the clinical material; H.S., T.M. and P.J.B.: Provided the clinical interpretation. L.L.E.: Supervised the statistical development. P.J.B.: Supervised the clinical aspects of the study; F.S.: Performed the independent model evaluation in the validation dataset. T.D.L., H.S., P.J.B. and L.L.E.: Wrote the manuscript draft. All authors read the final manuscript and approved of its content.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Laajala, T. D. *et al.* Longitudinal modeling of ultrasensitive and traditional prostate-specific antigen and prediction of biochemical recurrence after radical prostatectomy. *Sci. Rep.* **6**, 36161; doi: 10.1038/srep36161 (2016).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

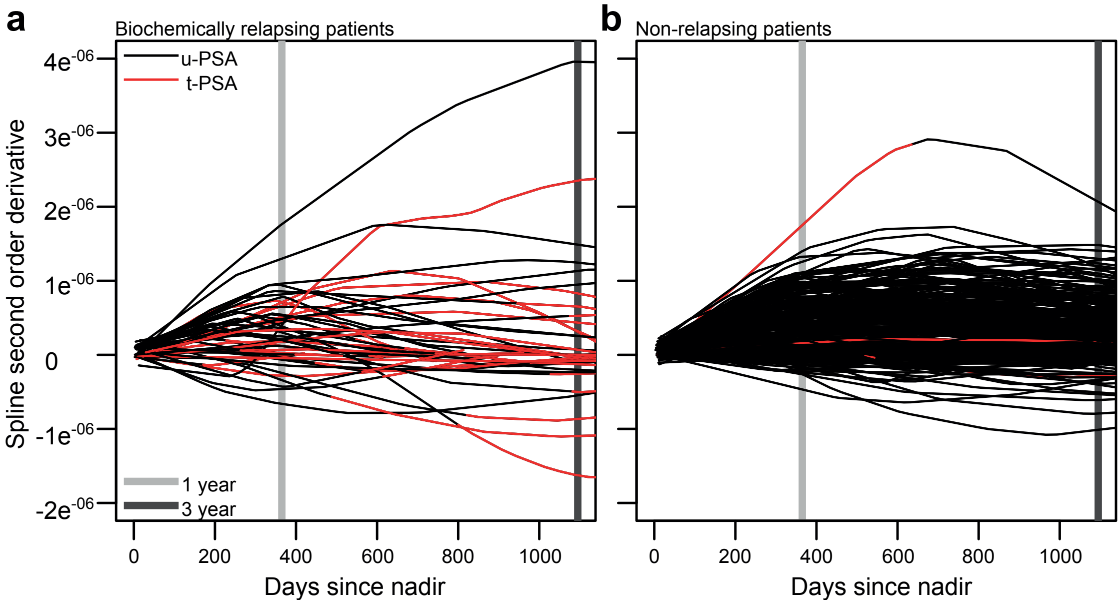
Supplementary Material (Publication III):

Supplementary Figures S1 - S2

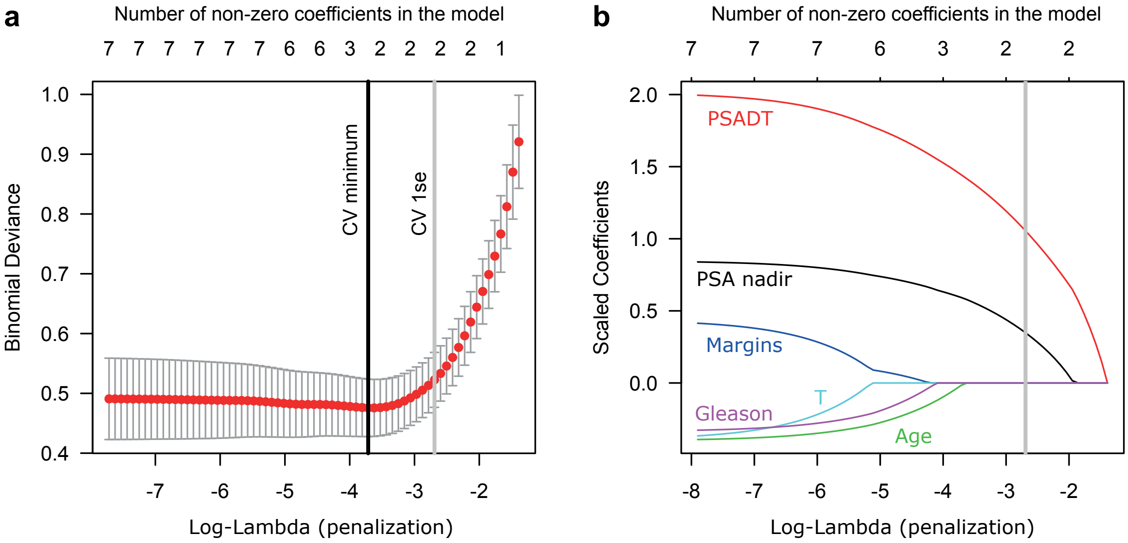
Supplementary Tables S1 - S2

Supplementary Methods and Results (*Available online*)[*]

[*]: `https://www.nature.com/articles/srep36161#supplementary-information`



Supplementary Figure S1: Second order derivatives of the spline fits. (**a**) Biochemically relapsing patients (BCR, $N = 52$), (**b**) non-BCR patients ($N = 279$). The vertical light and dark grey lines indicate one year and three year time points, respectively. Corresponding fitted measurements for the u-PSA and t-PSA are annotated using black and red colors, respectively.



Supplementary Figure S2: LASSO model cross-validation and penalization curve. (**a**) 10-fold cross-validation (CV), based on which optimal penalization was chosen to be first penalization parameter within a standard error of the CV minimum. (**b**) Penalization curves, which display that the log_2 nadir and PSADT were selected by the final model, albeit some traditional clinical parameters were almost included.

Supplementary Table S1: A computational spreadsheet example of simple linear regression in predicting BCR risk from the proposed generalized linear model

Col →	**A**	**B**	**C**	**D**	**E**	**F**
Row ↓	**PSA**	log_2**-PSA (= y)**	**DaysSinceSurgery**	**DaysSinceNadir (= x)**	x^2	$x \cdot y$
1	10	3.321928	-139	-184		
2	7.6	2.925999	-1	-46		
3	0.091	-3.45799	21	-24		
4	0.006	-7.381	45	0	0	0
5	0.022	-5.506	78	33	1089	-181.7
6	0.033	-4.921	100	55	3025	-270.7
7	0.006	-7.381	225	180	32400	-1329
8	0.004	-7.966	335	290	84100	-2310
9	0.003	-8.381	710	665		
10	0.003	-8.381	1092	1047		
11	0.003	-8.381	1289	1244		
12	0.008	-6.966	1429	1384		
13	0.006	-7.381	1584	1539		

Each row corresponds to a single PSA measurement. In our current study, we defined nadir to be the lowest point in PSA within a 3-month window post-surgery, thus limiting observations for our model to observations $Row \geq 4$ in this example. Similarly, in order to evaluate model parameters in a 1-year window post-nadir, the lower limit for utilized observations is set at $Row \leq 8$. These constraints were obtained by observing the days since nadir column at **D**. The simple regression coefficients can be computed in closed form:

$$\begin{array}{lll}
\overline{x^2} & = \text{AVERAGE(E4:E8)} & = 24123 \\
\overline{xy} & = \text{AVERAGE(F4:F8)} & = \text{-818.2} \\
\bar{x} & = \text{AVERAGE(D4:D8)} & = 111.6 \\
\bar{y} & = \text{AVERAGE(B4:B8)} & = \text{-6.63103} \\
\bar{x}^2 & = \text{POWER(AVERAGE(D4:D8), 2)} & = 111.6^2 = 12454.56
\end{array}$$

Thus, for this particular individual, the simple regression estimates are:

$$\begin{array}{lllll}
\hat{\beta_2} & = \dfrac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2} & = \dfrac{-818.2 - (111.6 \cdot -6.63103)}{24123 - 12454.56} & = -0.00669987 & \text{(PSADT)} \\
\hat{\beta_1} & = \bar{y} - \hat{\beta_1}\bar{x} & = -6.63103 - (-0.00669987 \cdot 111.6) & = -5.883325 & (log_2\text{-PSA nadir})
\end{array}$$

where $\hat{\beta_2}$ corresponds to the PSADT and $\hat{\beta_1}$ to the log_2-PSA nadir. Above estimates may be inspected in **Figure 3 D** to evaluate the individual's risk for BCR. In this particular case, the risk for BCR is very low, which is expected when the PSADT coefficient is negative (no doubling occurs). 1 year follow-up was used as a criterion for including observations in estimating $\hat{\beta_1}$ and $\hat{\beta_1}$. The coefficients $\{\beta_{base}$, β_{nadir} and $\beta_{doubling}\}$ reported in our study for 1-year follow up were {2.736, 0.640, 218.488}. Thus, the risk for BCR for this individual may be computed as provided in the **Supplementary Methods**:

$$\frac{1}{1 + e^{-(\beta_{base} + \beta_{nadir} \times x_1 + \beta_{doubling} \times x_2)}} = \frac{1}{1 + e^{-(2.736 + 0.640 \cdot -5.883325 + 218.488 \cdot -0.00669987)}}$$

$$= 0.07633843...$$

which would indicate a very low risk of BCR, as was later observed in follow-up. Similarly, a BCR risk for a hypothetical patient undergoing PSADT every 150 days ($1/150 \approx 0.00667$) and an estimated log_2-PSA nadir of -5 (or $2^{-5} = 0.03125$ in the original PSA scale) would yield ≥ 0.5 risk:

$$\frac{1}{1 + e^{-(2.736 + 0.640 \cdot -5 + 218.488 \cdot 0.00667)}} = 0.7297422...$$

Supplementary Table S2: Estimated 3 year follow-up patient-wise *log*$_2$-PSA nadir levels (intercepts) and PSADT (doubling slopes) in the exploratory dataset in connection to the patients' clinical parameters

		log$_2$-PSA nadir (intercepts $\beta_0 + \gamma_{0,i}$)						PSADT (slopes $\beta_1 + \gamma_{1,i}$)						
		Min.	***1st Qu.***	***Median***	***Mean***	***3rd Qu.***	***Max.***	***Min.***	***1st Qu.***	***Median***	***Mean***	***3rd Qu.***	***Max.***	***N***
pT	**2**	-10.448	-8.542	-8.266	-7.877	-7.729	-2.557	-0.007475	-0.000039	0.000322	0.000804	0.000797	0.006146	*179*
	3	-9.603	-8.374	-7.815	-7.131	-6.045	-1.432	-0.001640	-0.000058	0.000474	0.001190	0.002202	0.006671	*154*
Gleason score	**<=6**	-9.115	-8.588	-8.065	-7.904	-7.699	-5.418	-0.000488	-0.000064	0.000195	0.000558	0.000569	0.005971	*59*
	7	-9.100	-8.620	-7.990	-7.488	-7.031	-2.568	-0.000371	-0.000044	0.000297	0.001097	0.002097	0.006633	*58*
	>=8	-10.448	-8.903	-7.864	-7.791	-7.216	-3.815	-0.000565	0.000384	0.002190	0.002002	0.003372	0.005052	*12*
Margins	**Neg.**	-10.448	-8.526	-8.226	-7.772	-7.633	-1.432	-0.001640	-0.000062	0.000322	0.000858	0.000853	0.006671	*199*
	Pos.	-9.115	-8.405	-7.815	-7.176	-6.045	-1.651	-0.000551	-0.000049	0.000483	0.001169	0.002148	0.006633	*134*
Adjuvant RT	**No**	-10.448	-8.494	-8.161	-7.656	-7.408	-1.432	-0.001640	-0.000044	0.000345	0.000947	0.001137	0.006671	*293*
	Yes	-9.112	-8.386	-7.010	-6.627	-5.352	-1.651	-0.000551	-0.000056	0.000356	0.001246	0.002266	0.005044	*40*
Salvage RT	**No**	-9.282	-8.537	-8.236	-7.885	-7.650	-2.852	-0.001640	-0.000118	0.000194	0.000425	0.000586	0.005677	*273*
	Yes	-10.448	-7.901	-5.916	-5.929	-4.246	-1.432	0.000072	0.002534	0.003813	0.003519	0.004555	0.006671	*60*
PSA at surgery	**<10**	-10.448	-8.531	-8.177	-7.738	-7.506	-2.557	-0.000748	-0.000096	0.000276	0.000829	0.000929	0.006671	*248*
	10-20	-9.282	-8.409	-8.065	-7.104	-6.128	-1.432	-0.000342	0.000201	0.000704	0.001412	0.002175	0.005971	*67*
	>20	-9.603	-8.304	-7.273	-6.288	-4.362	-1.651	-0.001640	-0.000015	0.000506	0.001503	0.003216	0.004756	*18*

Holm-method multiple testing corrected *p*-values according to one-way ANOVA: white N.S.; orange p<0.05.

ORIGINAL PUBLICATION **IV**

Guinney J*, Wang T*, **Laajala TD***, Winner KK, Bare JC, Neto EC, Khan SA, Peddinti G, Airola A, Pahikkala T, Mirtti T, Yu T, Bot BM, Shen L, Abdallah K, Norman T, Friend S, Stolovitzky G, Soule H, Sweeney CJ, Ryan CJ, Scher HI, Sartor O, Xie Y, Aittokallio T, Zhou FL, Costello JC, Prostate Cancer Challenge DREAM Community. *Prediction of overall survival for patients with metastatic castration-resistant prostate cancer: development of a prognostic model through a crowdsourced challenge with open clinical trial data.* Lancet Oncology. 2017;18(1):132-142.

*: Equal contribution

Prediction of overall survival for patients with metastatic castration-resistant prostate cancer: development of a prognostic model through a crowdsourced challenge with open clinical trial data

Justin Guinney, Tao Wang*, Teemu D Laajala*, Kimberly Kanigel Winner, J Christopher Bare, Elias Chaibub Neto, Suleiman A Khan, Gopal Peddinti, Antti Airola, Tapio Pahikkala, Tuomas Mirtti, Thomas Yu, Brian M Bot, Liji Shen, Kald Abdallah, Thea Norman, Stephen Friend, Gustavo Stolovitzky, Howard Soule, Christopher J Sweeney, Charles J Ryan, Howard I Scher, Oliver Sartor, Yang Xie†, Tero Aittokallio†, Fang Liz Zhou†, James C Costello†, and the Prostate Cancer Challenge DREAM Community‡*

Summary

Background Improvements to prognostic models in metastatic castration-resistant prostate cancer have the potential to augment clinical trial design and guide treatment strategies. In partnership with Project Data Sphere, a not-for-profit initiative allowing data from cancer clinical trials to be shared broadly with researchers, we designed an open-data, crowdsourced, DREAM (Dialogue for Reverse Engineering Assessments and Methods) challenge to not only identify a better prognostic model for prediction of survival in patients with metastatic castration-resistant prostate cancer but also engage a community of international data scientists to study this disease.

Methods Data from the comparator arms of four phase 3 clinical trials in first-line metastatic castration-resistant prostate cancer were obtained from Project Data Sphere, comprising 476 patients treated with docetaxel and prednisone from the ASCENT2 trial, 526 patients treated with docetaxel, prednisone, and placebo in the MAINSAIL trial, 598 patients treated with docetaxel, prednisone or prednisolone, and placebo in the VENICE trial, and 470 patients treated with docetaxel and placebo in the ENTHUSE 33 trial. Datasets consisting of more than 150 clinical variables were curated centrally, including demographics, laboratory values, medical history, lesion sites, and previous treatments. Data from ASCENT2, MAINSAIL, and VENICE were released publicly to be used as training data to predict the outcome of interest—namely, overall survival. Clinical data were also released for ENTHUSE 33, but data for outcome variables (overall survival and event status) were hidden from the challenge participants so that ENTHUSE 33 could be used for independent validation. Methods were evaluated using the integrated time-dependent area under the curve (iAUC). The reference model, based on eight clinical variables and a penalised Cox proportional-hazards model, was used to compare method performance. Further validation was done using data from a fifth trial—ENTHUSE M1—in which 266 patients with metastatic castration-resistant prostate cancer were treated with placebo alone.

Findings 50 independent methods were developed to predict overall survival and were evaluated through the DREAM challenge. The top performer was based on an ensemble of penalised Cox regression models (ePCR), which uniquely identified predictive interaction effects with immune biomarkers and markers of hepatic and renal function. Overall, ePCR outperformed all other methods (iAUC 0·791; Bayes factor >5) and surpassed the reference model (iAUC 0·743; Bayes factor >20). Both the ePCR model and reference models stratified patients in the ENTHUSE 33 trial into high-risk and low-risk groups with significantly different overall survival (ePCR: hazard ratio 3·32, 95% CI 2·39–4·62, p<0·0001; reference model: 2·56, 1·85–3·53, p<0·0001). The new model was validated further on the ENTHUSE M1 cohort with similarly high performance (iAUC 0·768). Meta-analysis across all methods confirmed previously identified predictive clinical variables and revealed aspartate aminotransferase as an important, albeit previously under-reported, prognostic biomarker.

Interpretation Novel prognostic factors were delineated, and the assessment of 50 methods developed by independent international teams establishes a benchmark for development of methods in the future. The results of this effort show that data-sharing, when combined with a crowdsourced challenge, is a robust and powerful framework to develop new prognostic models in advanced prostate cancer.

Funding Sanofi US Services, Project Data Sphere.

Lancet Oncol 2016

Published **Online**
November 15, 2016
http://dx.doi.org/10.1016/S1470-2045(16)30560-5

See **Online/Comment**
http://dx.doi.org/10.1016/S1470-2045(16)30564-2

*Contributed equally as first authors

†Contributed equally as senior authors

‡Members listed in the appendix

Sage Bionetworks, Seattle, WA, USA (J Guinney PhD, J C Bare BS, E C Neto PhD, T Yu BS, B M Bot MS, T Norman PhD, S Friend MD); **Quantitative Biomedical Research Center, Department of Clinical Sciences** (T Wang PhD, Y Xie PhD), **Center for the Genetics of Host Defense** (T Wang), **The Simmons Comprehensive Cancer Center,** (Y Xie), **and Lyda Hill Department of Bioinformatics** (Y Xie), **University of Texas Southwestern Medical Center, Dallas, TX, USA; Department of Mathematics and Statistics** (T D Laajala MSc, Prof T Aittokallio PhD), **and Department of Information Technology** (A Airola PhD, T Pahikkala PhD), **University of Turku, Turku, Finland; Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland** (T D Laajala, S A Khan PhD, G Peddinti PhD, T Mirtti MD, Prof T Aittokallio); **Department of Pathology (HUSLAB), Helsinki University Hospital, Helsinki, Finland** (T Mirtti); **Department of Pharmacology and Computational Biosciences Program** (K Kanigel Winner PhD, J C Costello PhD), **and University**

Introduction

Prostate cancer is the most common cancer among men in high-income countries and ranks third in terms of mortality after lung cancer and colorectal cancer.[1] Of more than two million men diagnosed with prostate cancer in the USA over the past 10 years, roughly 10%

of Colorado Comprehensive Cancer Center (J C Costello), University of Colorado, Anschutz Medical Campus, Aurora, CO, USA; AstraZeneca, Gaithersburg, MD, USA (K Abdallah MD); IBM T J Watson Research Center, IBM, Yorktown Heights, NY, USA (G Stolovitzky PhD); Prostate Cancer Foundation, Santa Monica, CA, USA (H Soule PhD); Department of Medical Oncology, Dana-Farber Cancer Institute and Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA (C J Sweeney MBBS); Genitourinary Medical Oncology Program, Division of Hematology and Oncology, University of California, San Francisco, CA, USA (Prof C J Ryan MD); Genitourinary Oncology Services, Department of Medicine, Sidney Kimmel Center for Prostate and Urologic Cancers, Memorial Sloan-Kettering Cancer Center and Weill Cornell Medical College, New York, NY, USA (Prof H I Scher MD); Tulane Cancer Center, Tulane University, New Orleans, LA, USA (Prof O Sartor MD); and Sanofi, Bridgewater, NJ, USA (L Shen PhD, F L Zhou MD)

Correspondence to:
Dr James C Costello, University of Colorado Anschutz Medical Campus, Aurora, CO 80045, USA
james.costello@ucdenver.edu

See Online for appendix

For more on DREAM challenges see http://dreamchallenges.org

For more on Project Data Sphere see http://www.projectdatasphere.org

For more on the CEO Roundtable on Cancer's Life Sciences Consortium see http://ceo-lsc.org

To access data via the Synapse platform see https://www.synapse.org/ProstateCancerChallenge

Research in context

Evidence before this study

We searched PubMed between January, 2012, and July, 2015, with the terms "prognosis", "overall survival", "mCRPC", and "docetaxel". Our search yielded a 2014 study in which an updated prognostic model was described for metastatic castration-resistant prostate cancer that had been developed from the CALGB-90401 study (a randomised, double-blind, phase 3 clinical trial) and validated with data from the phase 3 ENTHUSE 33 trial. The study focused on a subset of clinical variables using datasets that were not in the public domain. Leveraging the wealth of data already generated from clinical trials is challenging on several fronts, but is complicated in particular by data access.

Added value of this study

Project Data Sphere is an independent not-for-profit initiative that aims to provide open access to historical patient-level data. The prostate cancer DREAM (Dialogue for Reverse Engineering Assessments and Methods) challenge is an open-data, crowdsourced competition to develop and assess prognostic models in metastatic castration-resistant prostate cancer. Using data from the comparator arms of four phase 3 clinical trials of chemotherapy-naive patients with metastatic castration-resistant prostate cancer, 50 independent teams—a diverse group of experts including biostatisticians, computer scientists, and clinical experts—developed prognostic models for the DREAM challenge, representing, to the best of our knowledge, the most comprehensive set of benchmarked models to date. The best-performing model was based on an ensemble of penalised Cox regression models that judged the prognostic value of interactions between predictor covariates and substantially outperformed the 2014 model. Strong support was provided for previously identified prognostic variables in the 50 models, and additional important variables were identified along with novel interactions between covariates. Data are available publicly through the Project Data Sphere initiative, and all method predictions and code are available for download through the Sage Bionetworks Synapse platform.

Implications of all the available evidence

Clinical trial data-sharing is both feasible and useful, and the DREAM challenge is an appropriate vehicle on which to build and rigorously assess prognostic or predictive models quickly, openly, and robustly. We established a new prognostic benchmark in metastatic castration-resistant prostate cancer, with applications in trial design and guidance for clinicians and patients. Robust and accurate prognostic predictors can be used to homogenise risk in clinical trials of metastatic castration-resistant prostate cancer and enable smaller trials for assessment of treatment effects.

presented with metastatic disease. For these men, the mainstay of treatment is androgen deprivation therapy, with a high proportion of response. However, responses are not durable, and nearly all tumours eventually progress to the lethal metastatic castration-resistant state. Although substantial improvements in outcome for men with metastatic castration-resistant prostate cancer have been achieved after approval of next-generation hormonal agents, an immunotherapeutic drug, a radiopharmaceutical agent, and a cytotoxic drug,[2–10] how best to deploy these treatments has not been ascertained. Elucidation of variables associated with patients' outcomes independent of treatment will facilitate the design of future trials by homogenising risk, thus enabling clinical trial questions to be answered more rapidly because smaller sample sizes will be needed.

Prognostic models in metastatic castration-resistant prostate cancer have been described[11–13] using baseline variables from independent cohort studies. A 2014 prognostic model for metastatic castration-resistant prostate cancer [14] included eight clinical factors predictive of overall survival: Eastern Cooperative Oncology Group (ECOG) performance status; disease site; use of opioid analgesics; lactate dehydrogenase; albumin; haemoglobin; prostate-specific antigen; and alkaline phosphatase. Can innovative models with improved performance be developed through a systematic search using data-driven approaches while providing insights into biological aspects of the disease that affect patients' outcomes? An example of a novel clinical factor that is underexplored in contemporary prognostic model development is interaction effects between clinical variables, even though interactions between genetic variants are used widely and known to improve genetic-based risk prediction and patients' stratification.[15,16]

Here, we present results from the prostate cancer DREAM (Dialogue for Reverse Engineering Assessments and Methods) challenge—an open-data, crowdsourced challenge in metastatic castration-resistant prostate cancer. A major contribution to this effort was removal of privacy and legal barriers associated with open access to phase 3 clinical trial data[17] by Project Data Sphere—a not-for-profit initiative of the CEO Roundtable on Cancer's Life Sciences Consortium that broadly shares oncology clinical trial data with researchers. The challenge was designed to accomplish two goals. First, we aimed to leverage open clinical trial data, enabling a community-based approach to identify the best-performing prognostic model in a rigorous and unbiased manner. Second, participating teams aimed to develop predictive models to both validate previously characterised predictive clinical variables and discover new prognostic features. Consistent with the mission of DREAM, all challenge data, results, and method descriptions from participating teams are available publicly through the open-access Synapse platform.

Methods

Trial selection

In April 2014, the DREAM challenge organising team reviewed all existing and incoming prostate cancer trial datasets (comparator arm only) in Project Data Sphere and selected four trials, which were the source of training and validation datasets for the DREAM challenge—ASCENT2,[18] MAINSAIL,[19] VENICE,[20] and ENTHUSE 33.[21] All four trials were randomised phase 3 clinical trials in which the comparator arm consisted of a docetaxel regimen and overall survival was the primary endpoint. These four trials also had similar inclusion and exclusion criteria: eligible patients were aged 18 years and older, had progressive metastatic castration-resistant prostate cancer, were chemotherapy-naive, and had an ECOG performance status of 0–2. Further details of inclusion and exclusion criteria for each trial are provided in the appendix (p 3). The patient-level trial datasets were deidentified by data providers and made available for the DREAM challenge through Project Data Sphere. No institutional review board approval was needed to access data.

Patient populations

We compiled training datasets from the comparator arms of ASCENT2, MAINSAIL, and VENICE. ASCENT2[18] is a randomised open-label study assessing DN-101 in combination with docetaxel. Patients with metastatic castration-resistant prostate cancer were randomly assigned either docetaxel and prednisone (comparator arm) or docetaxel and DN-101, stratified by geographical region and ECOG performance status. MAINSAIL[19] is a randomised double-blind study to assess efficacy and safety of docetaxel and prednisone with or without lenalidomide in patients with metastatic castration-resistant prostate cancer. Participants were randomly assigned to either docetaxel, prednisone, and placebo (comparator arm) or lenalidomide, docetaxel, and prednisone. Stratification of patients in MAINSAIL was done based on ECOG performance status (0–1 *vs* 2), geographical region (USA and Canada *vs* Europe and Australia *vs* rest of world), and type of disease progression after hormonal treatment (rising prostate-specific antigen only *vs* tumour progression). VENICE[20] is a randomised double-blind study comparing the efficacy and safety of aflibercept versus placebo, in which patients with metastatic castration-resistant prostate cancer were randomly assigned either docetaxel, prednisone or prednisolone, and placebo (comparator arm) or docetaxel, prednisone or prednisolone, and aflibercept. Participants were stratified by baseline ECOG performance status (0–1 *vs* 2). The validation dataset was from the ENTHUSE 33 trial,[21] a double-blind study in which patients with metastatic castration-resistant prostate cancer were randomly allocated (1:1) either docetaxel and placebo (comparator arm) or docetaxel with zibotentan, stratified by centre.

Data curation

The original datasets from Project Data Sphere contained patient-level raw tables that conformed to either Study Data Tabulation Model (SDTM) standards or company-specific clinical database standards. To optimise use of these data for the DREAM challenge, we compiled the four sets of trial data into a set of five standardised raw event-level tables, meaning all four clinical trials were combined into the same tables based on laboratory values, medical history, lesion sites, previous treatments, and vital signs. Including patients' demographic information, these tables presented most measurements made for the patient in that category. To summarise these data on a per-patient level, we created a core table, distilling the raw event-level tables and patients' demographics into 129 clinically defined baseline and outcome variables. Full details of the data curation process are provided in the appendix (pp 3, 4).

We supplied participating teams with the full set of baseline and raw variables from the core and raw event-level tables. We encouraged challenge participants to derive additional baseline clinical variables from the five standardised raw event-level tables for modeling. We also provided teams with outcome variables for the ASCENT2, MAINSAIL, and VENICE trials, but we did not release the outcome variables for the ENTHUSE 33 trial because they would serve to independently evaluate the performance of models. The primary endpoint used for model development was overall survival, defined as the time from date of randomisation to the date of death from any cause.

We did principal component analysis to investigate systematic similarities or differences between the four clinical trials, using either all available variables or binary variables only. We visualised the principal component analysis by plotting the first principal component against the second principal component for all patients.

Further validation

After the DREAM challenge was completed using data from ENTHUSE 33 for method evaluation, we further validated the top-performing and reference models with data from a fifth trial, ENTHUSE M1,[22] to assess whether the top-performing model could be used to stratify risk for patients with metastatic castration-resistant prostate cancer who received placebo alone and no docetaxel. ENTHUSE M1 is a randomised double-blind study to assess the efficacy and safety of 10 mg zibotentan in patients with metastatic castration-resistant prostate cancer (specifically, bone metastasis). By contrast with ENTHUSE 33, the ENTHUSE M1 trial included a comparator arm of placebo alone. Patients were randomly allocated (1:1) either zibotentan or placebo and were stratified by centre. The inclusion and exclusion criteria were similar to those used for ENTHUSE 33 except that patients in ENTHUSE M1 were pain free or mildly symptomatic. To be consistent for validation, curation of

ENTHUSE M1 data followed the same process as was done for ASCENT2, MAINSAIL, VENICE, and ENTHUSE 33, resulting in a core table and five raw event-level tables.

Challenge procedures

The DREAM challenge was hosted and fully managed on Synapse, a cloud-based platform for collaborative scientific data analysis, through which all model predictions were submitted. The challenge was run in two phases (appendix pp 4, 17). First, teams were allowed to train and test their models in an open testing leaderboard phase. Second, teams were permitted one last submission to the final scoring phase, after which teams were scored and ranked. Accordingly, we split data from ENTHUSE 33 into two separate sets, consisting of 157 patients and 313 patients. The smaller dataset was used for the open testing phase and the larger dataset was used for the final scoring phase. Moreover, all reported performance values for the evaluated methods and all comparisons between the top-performing model and reference model used the larger set of data from the ENTHUSE 33 trial. The reference prognostic model for prediction of overall survival was a penalised Cox proportional-hazards model using the adaptive least absolute shrinkage and selection operator (LASSO) penalty.[14]

For method evaluation, we used the integrated AUC (iAUC)[23] calculated from 6–30 months as our primary scoring metric. For robust determination of the best performing team or teams, we used Bayes factor analysis and randomisation test based on iAUC (appendix pp 4, 5). For each team, we calculated the Bayes factor to directly compare the performance of a model with the reference model; coefficients for the reference model were obtained from reported hazard ratios (HRs).[14] Furthermore, we evaluated model predictions by plotting Kaplan-Meier curves, after dichotomising patients for each team separately by median risk score. We used the log-rank test to compare the two groups using the *coxph* function in the *survival R* package. We calculated CIs by inverting the Wald test statistic. The risk scores generated by each model have their own dynamic range; thus, we used the rankings of patients for scoring by iAUC or Kaplan-Meier analysis. Accordingly, we selected the median risk score as a means to compare different methods in a fair manner. A major goal of the challenge was to encourage teams to develop and test novel methods outside of standard survival analysis approaches; thus, risk score predictions across all teams varied in their range and distribution. A standard threshold could not be established fairly for all teams; therefore, we relied on rank-based scoring methods, including the iAUC, and stratifying risk scores based on the median. We also calculated other statistics, including median survival and 1-year and 2-year survival for the dichotomised high-risk and low-risk groups. We did hierarchical clustering on rank-normalised risk score predictions from all models in the challenge, using Euclidean distance and average linkage.

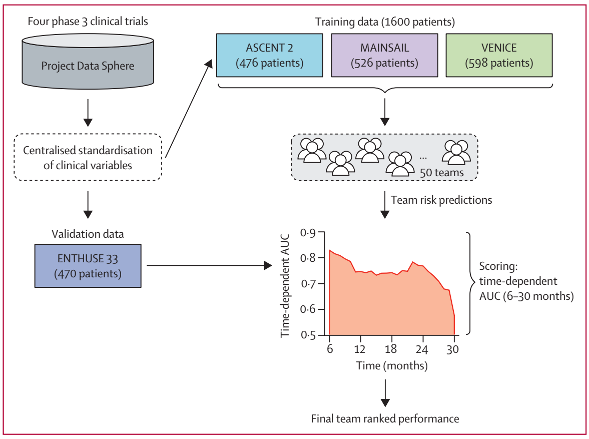

Four phase 3 clinical trials
Project Data Sphere
Centralised standardisation of clinical variables
Validation data
ENTHUSE 33 (470 patients)
Training data (1600 patients)
ASCENT 2 (476 patients)
MAINSAIL (526 patients)
VENICE (598 patients)
... 50 teams
Team risk predictions
Time-dependent AUC
0·9
0·8
0·7
0·6
0·5
6
12
18
24
30
Time (months)
Scoring: time-dependent AUC (6–30 months)
Final team ranked performance

Figure 1: **Study design**
Data were acquired from Project Data Sphere and curated centrally by the organising team to provide a harmonised dataset across the four studies. Three studies were provided as training data (ASCENT2, MAINSAIL, and VENICE) and the fourth (ENTHUSE 33) was the validation dataset. Teams submitted risk scores for ENTHUSE 33, then their predictions were scored and ranked using an integrated time-dependent area under the curve (AUC) metric.

We used the ENTHUSE 33 dataset to assess the calibration of the top-performing model. We plotted the predicted survival probability based on the top-performing model against the observed survival proportions at 18, 24, 30, and 36 months. For each time cutoff, we divided the population into seven equally spaced categories based on the ranked predicted risk by the top-performing model. We then calculated the true survival proportion within each category and plotted it as a point estimate and 95% CI. A 45° line on the plots indicated perfect calibration.

The organisers of the DREAM challenge used SAS version 9.3 for data curation and *R* version 3.2.4 for statistical analyses. *R* packages used for challenge evaluation included *survival* version 2.38-3, *ROCR* version 1.0-7, *timeROC* version 0.3, and *Bolstad2* version 1.0-28. The top-performing model also used *glmnet* version 2.0-5 and *hamlet* version 0.9.4-2.

Clinical trial data used in the prostate cancer DREAM challenge can be accessed online.[24] Write-ups, model code, and predictions for all teams are reported in the appendix (pp 7, 8). Challenge documentation, including a detailed description of its design, overall results, scoring scripts, and the clinical trials data dictionary can be accessed via the Synapse platform.

	ASCENT2 (n=476)	MAINSAIL (n=526)	VENICE (n=598)	ENTHUSE 33 (n=470)	ENTHUSE M1 (n=266)
Age (years)					
18–64	111 (23%)	171 (33%)	219 (37%)	160 (34%)	58 (22%)
65–74	211 (44%)	246 (47%)	254 (42%)	217 (46%)	111 (42%)
≥75	154 (32%)	109 (21%)	125 (21%)	93 (20%)	97 (36%)
ECOG performance score					
0	220 (46%)	257 (49%)	280 (47%)	247 (53%)	196 (74%)
1	234 (49%)	247 (47%)	291 (49%)	223 (47%)	70 (26%)
2	22 (5%)	20 (4%)	27 (5%)	0 (0%)	0 (0%)
3	0 (0%)	1 (<1%)	0 (0%)	0 (0%)	0 (0%)
Missing	0 (0%)	1 (<1%)	0 (0%)	0 (0%)	0 (0%)
Metastasis					
Liver	5 (1%)	58 (11%)	60 (10%)	64 (14%)	12 (5%)
Bone	345 (72%)	439 (83%)	529 (88%)	470 (100%)	266 (100%)
Lungs	8 (2%)	74 (14%)	88 (15%)	56 (12%)	13 (5%)
Lymph nodes	163 (34%)	298 (57%)	323 (54%)	208 (44%)	80 (30%)
Analgesic use					
No	338 (71%)	347 (66%)	419 (70%)	339 (72%)	256 (96%)
Yes	138 (29%)	179 (34%)	179 (30%)	131 (28%)	10 (4%)
Lactate dehydrogenase (U/L)	202 (176–250)	210 (174–267)	NA	213 (181–287)	188 (170–219)
Missing	13 (3%)	1 (<1%)	596 (100%)	5 (1%)	7 (3%)
Prostate-specific antigen (ng/mL)	68·8 (24·2–188·4)	84·9 (32·2–271·2)	90·8 (30·8–260·6)	99·6 (33·6–236·8)	52·3 (17·3–153·0)
Missing	1 (<1%)	4 (1%)	6 (1%)	12 (3%)	4 (2%)
Haemoglobin (g/dL)	12·6 (11·6–13·6)	12·7 (11·5–13·7)	12·7 (11·7–13·5)	12·5 (11·3–13·5)	12·9 (12·2–13·7)
Missing	3 (1%)	10 (2%)	0 (0%)	4 (1%)	2 (1%)
Albumin (g/L)	NA	43 (41–45)	42 (38–45)	43 (40–46)	43 (41–45)
Missing	476 (100%)	1 (<1%)	16 (3%)	2 (<1%)	1 (<1%)
Alkaline phosphatase (U/L)	113 (80–213)	124 (81–265)	135 (85–270)	155 (98–328)	130 (83–222)
Aspartate aminotransferase (U/L)	24 (20–31)	24 (19–31)	25 (20–33)	25 (20–33)	24 (19–29)
Missing	4 (1%)	1 (<1%)	8 (1%)	3 (1%)	3 (1%)

Data are median (IQR) or number of patients (%). NA=not available. ECOG=Eastern Cooperative Oncology Group.

Table: **Patients' baseline characteristics**

Role of the funding source

Project Data Sphere had a collaborative role in design and logistics of the DREAM challenge but played no part in data collection, data analysis, and data interpretation or in the writing of this report. Sanofi US Services provided an in-kind contribution of human resources for curation of the raw datasets for the DREAM challenge and for clinical and scientific support of the challenge organisation, at the request of Project Data Sphere. Sanofi personnel participated in design of the DREAM challenge, in data analysis and data interpretation, and in writing of the report, but had no role in data collection. Raw clinical trial datasets for ASCENT2, MAINSAIL, and VENICE were available on the Project Data Sphere platform and were accessible by all registered users of Project Data Sphere, including all DREAM challenge participants and organisers, throughout the challenge. JG, TW, KKW, BMB, LS, KA, YX, FLZ, and JCC had access to raw data for ENTHUSE 33. JG, TW, KKW, LS, KA, FLZ, and JCC had access to raw data for ENTHUSE M1, during the post-challenge analysis. Data for ENTHUSE 33 and ENTHUSE M1 have been made freely accessible through the Project Data Sphere platform with publication of this report. The corresponding author had full access to all data in the study and had final responsibility for the decision to submit for publication.

Results

The overall DREAM challenge design is shown in figure 1, with full details in the appendix (p 4). The table presents baseline characteristics of patients in the five clinical trials included in this analysis. The training dataset included: 476 individuals from ASCENT2; 526 participants in MAINSAIL; and 598 men from VENICE. The validation dataset consisted of 470 patients from the ENTHUSE 33 trial; 528 men were initially enrolled to that trial but, because of regulatory restrictions in one country, data for 58 individuals were not made public through the challenge. The second validation dataset comprised 266 patients from ENTHUSE M1. Because of the same regulation restriction mentioned for

Panel: Top-performing model construction in training datasets

The top-performing model was based on an ensemble of penalised Cox regression models (ePCR), as shown in the equation. For each trial-specific ensemble component, the model estimation procedure identified an optimum penalisation parameter (λ), which controls for the number of non-zero coefficients in the prediction model, and simultaneously the regularisation parameter (α) with respect to the objective function:

$$argmax_{\beta}[\frac{2}{n}\sum_{i=1}^{n}(x_{j(i)}^{T}\beta - \ln(\sum_{j\in R_i} e^{x_j^T\beta})) - \lambda(\alpha\sum_{i=1}^{p}|\beta_i| + \frac{1}{2}(1-\alpha)\sum_{i=1}^{p}\beta_i{}^2)]$$

Here, *x* are the predictors (clinical variables or their pairwise interactions), *β* are the model coefficients subjected to the absolute error and squared error penalisations (|*β*| and *β*², respectively), *p* is the number of predictors, *n* is the number of observations, *j(i)* is the index of the observation event at time T_i, and R_i is the set of indices *j* for which $y_j \geq T_i$ (patients at risk at time T_i), where y_j is the observed death or right-censoring time. The set of indices R_i is redefined for each patient *i* using the above risk criterion incorporating *y* and *T*. With suitable regularisation, the penalised regression identifies an optimum balance between the model fit and top predictors, effectively generalising the Cox model for future predictions. To reduce the risk of overfitting and to avoid randomness bias in the binning, the final ensemble models were optimised using ten-fold cross-validation of the iAUC, averaged over multiple cross-validation runs. By modelling each trial individually as a separate ensemble component with different optima in the equation, we are able to account effectively for trial-specific variation (appendix p 12). The optimum parameters (penalisation λ and norm α) for each trial were first identified using cross-validation, after which the model coefficients (*β*) are estimated by optimising the above objective function.

Data processing entailed missing value imputation with a penalised Gaussian regression variant of the equation, with cross-validation when variables with non-missing values were used as predictors. Variables with missing values were inferred by training an optimum model with the non-missing variables and then imputing the missing values. Laboratory values were modelled as continuous variables. Data curation entailed unsupervised explorative analyses (appendix pp 5, 6, 12). ASCENT2 trial data were used in the imputation and unsupervised learning phases but were omitted from construction of the final supervised ensemble predictor, which was based on three components: MAINSAIL alone, VENICE alone, and their combination (appendix p 12). The final ensemble prediction was done by averaging over the ranks of the component-predicted risks for the ENTHUSE 33 dataset (appendix p 12). Averaging of risk score ranks was selected to be more robust to trial-specific variation and potential outliers. Full details of the model and its network visualisation are in the appendix (pp 5, 6, 12) with a list of chosen predictors (appendix pp 10, 11).

ENTHUSE 33, some data were not provided to Project Data Sphere.

129 clinical baseline variables were measured for laboratory values, lesion site, previous medicines, medical history, and vital signs. When combined and assessed, the clinical variables for each trial were similar (appendix p 13), although when binary variables—mainly representing lesion sites—were judged separately, differences in clinical trials were recorded (appendix p 13). ASCENT2 had a lower frequency of patients with visceral metastases (1·1% liver and 1·7% lung) compared with individuals in the other three trials (10–14% liver, 11–15% lung). By contrast, the proportion of patients with bone metastases was high across the four trials (72–100%). Median follow-up differed among the four studies: 11·7 months (IQR 8·6–15·8) in ASCENT2; 9·2 months (6·4–13·1) in MAINSAIL; 21·1 months (12·9–29·6) in VENICE; and 15·3 months (10·9–20·8) in ENTHUSE 33. Risk profiles for each of the trials—specifically, mortality—were similar among the four trials (proportionality of hazards, p>0·5; appendix p 14). The proportion of patients who died in each trial was 138 (29%) of 476 in ASCENT2, 92 (17%) of 526 in MAINSAIL, 433 (72%) of 598 in VENICE, and 255 (54%) of 470 in ENTHUSE 33.

50 international teams—comprising 163 individuals—submitted predictions from their models to the challenge; with the reference model, the total number of models is 51. The distribution of all team scores by iAUC is shown in the appendix (p 15). The top-performing model was developed by a collaborative team from the Institute for Molecular Medicine Finland and the University of Turku. The method was based on an ensemble of penalised Cox regression (ePCR) models. The ePCR model extended beyond the LASSO-based reference model by using an elastic net to select additional correlated groups of clinical variables and their interactions, modelled as interaction terms (panel). The risk predictions from the trial-specific ensemble components were rank-averaged to produce the final ensemble risk score predictions and to avoid trial-specific variation.

The top-scoring ePCR model reported an iAUC of 0·791 and outscored all other teams, with a Bayes factor greater than 5, surpassing the threshold that defines significantly different performances (Bayes factor >3). The reference model achieved an iAUC of 0·743, with a significant difference in scores between the ePCR model and the reference model (Bayes factor >20). With a time-dependent AUC metric, the ePCR model outperformed the reference model at every timepoint, with the biggest difference in performance at later timepoints between 18 and 30 months (figure 2A). A median split of patients into low-risk and high-risk groups for the ePCR model resulted in a low-risk group comprising 156 patients and 56 deaths (median follow-up 27·6 months [IQR 18·2–31·9]) and a high-risk group containing 157 patients and 107 deaths (15·1 [8·5–20·1]). Similarly for the reference group, a low-risk group including 156 patients and 59 deaths (median follow-up 26·5 months [IQR 17·2–31·9]) and a high-risk group with 157 patients and 104 deaths (15·6 [8·6–21·8]) were generated. Kaplan-Meier analysis showed that low-risk and high-risk groups had significantly different overall survival in each model (ePCR, HR 3·32, 95% CI 2·39–4·62, p<0·0001; reference, 2·56, 1·85–3·53, p<0·0001; figure 2B, 2C). A full comparison is provided in the appendix (p 9). We assessed the calibration of the ePCR model by comparing predicted probabilities versus actual probabilities at multiple timepoints (appendix p 16).

Figure 3 shows a network visualisation of the significant groups of variables identified in the ePCR model and their predictive relations, based on the importance of the model covariates and their interactions. Although many of the variables used in the reference model were also included in the ePCR model, aspartate aminotransferase was identified as a new important predictor. We also recorded a number of factors that were included as interaction terms, and of particular note were those reflecting the immunological or renal function of the patient. Prostate-specific antigen was an independent but weak prognostic factor that interacted strongly with lactate dehydrogenase and aspartate aminotransferase.

In addition to identifying the top-performing model, the challenge also tested the other independent models, with 30 of 50 outperforming the reference model (Bayes factor >3; appendix p 15). We performed hierarchical clustering of risk scores from the 51 models to identify three distinct risk groups (figure 4A), with 98 patients (77 deaths) in group A (high risk), 131 patients (61 deaths) in group B (moderate risk), and 84 patients (25 deaths) in group C (low risk). Differences in overall survival among these three groups were significant (log-rank p<0·0001), with median overall survival of 12·9 months (95% CI 10·7–15·3) for group A, 20·8 months (18·3–25·6) for group B, and 27·7 months (26·6–not available) for group C (figure 4B).

40 of 50 teams provided a list of common clinical factors that were incorporated into their final models; the frequencies with which a feature was reported as being important or significant in a team's model are summarised in the appendix (p 18). The results not only confirmed the variables identified previously in the reference model but also highlighted several factors that were not. Of note, aspartate aminotransferase was included in more than half the team models. Other novel variables that were included in at least 15% of the models are total white blood cell count, absolute neutrophil count, red blood cell count, region of the world, body-mass index, and creatinine.

Application of the ePCR and reference models to the ENTHUSE M1 dataset showed model performances comparable with the primary challenge, with an iAUC of 0·768 for the ePCR model and 0·727 for the reference model (figure 5A). A median split of risk scores in the ePCR model led to a high-risk group of 133 patients, of which 45 were right-censored, and a low-risk group of 133 patients, of which 88 were right-censored. Kaplan-Meier analysis of the ENTHUSE M1

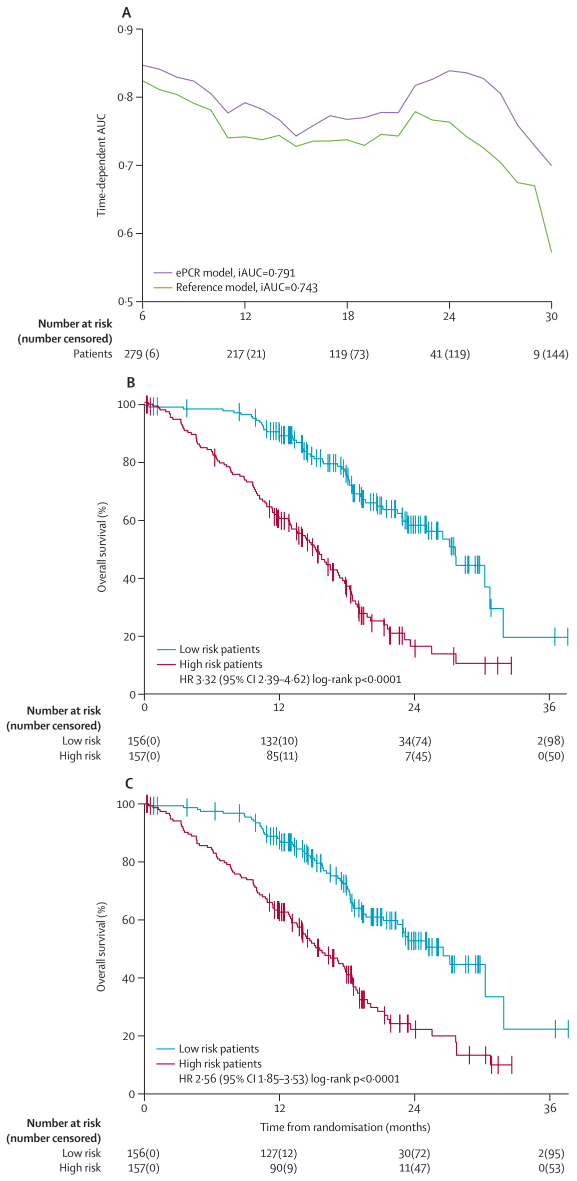


***Figure 2*: Performance of ePCR model, using data from ENTHUSE 33**
(A) Time-dependent AUC was measured from 6 months to 30 months at 1-month intervals, reflecting the performance of predicting overall survival at different timepoints. (B, C) Overall survival was assessed by the Kaplan-Meier method, stratified by the median in the top-performing ePCR model (B) and the reference model (C). The log-rank test was used to compare risk groups. ePCR=ensemble of penalised Cox regression models. iAUC=integrated time-dependent area under the curve. HR=hazard ratio.

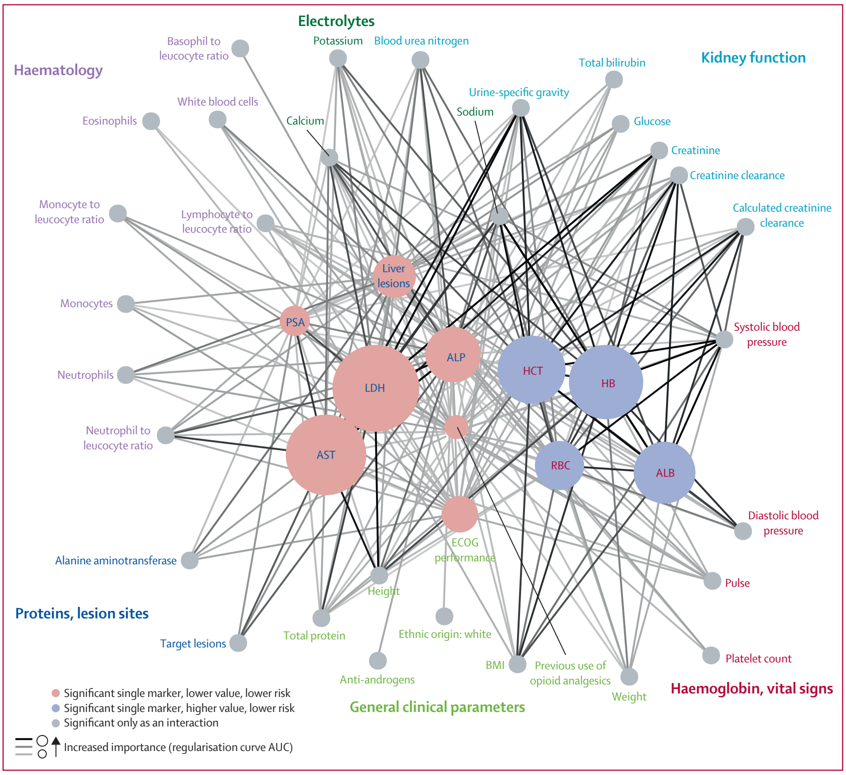


Figure 3: **Projection of the most important variables and interactions in the ePCR model**

Automated data-driven network layout of the most significant model variables, according to their interconnections with other model variables. Node size and colour indicate the importance of the variable alone for prediction of overall survival and its coefficient sign, respectively. This importance was calculated as the area under the curve (AUC) of the penalised model predictors, as a function of penalisation parameter λ. Edge colour indicates the importance of an interaction between two model variables, with a darker colour corresponding to a stronger interaction effect. Coloured subnetwork modules annotate the variables based on expert curated categories. Variable and interaction statistics can be found in the appendix (pp 10, 11). ALB=albumin. ALP=alkaline phosphatase. AST=aspartate aminotransferase. BMI=body-mass index. ECOG=Eastern Cooperative Oncology Group. ePCR=ensemble of penalised Cox regression models. HB=haemoglobin. HCT=haematocrit. LDH=lactate dehydrogenase. PSA=prostate-specific antigen. RBC=red blood cell count.

data showed significant separation of the high-risk and low-risk predicted patients (p<0·0001), with median survival of 15·8 months (95% CI 12·8–18·7) for high-risk patients and 27·1 months (23·2–not available) for low-risk patients (figure 5B).

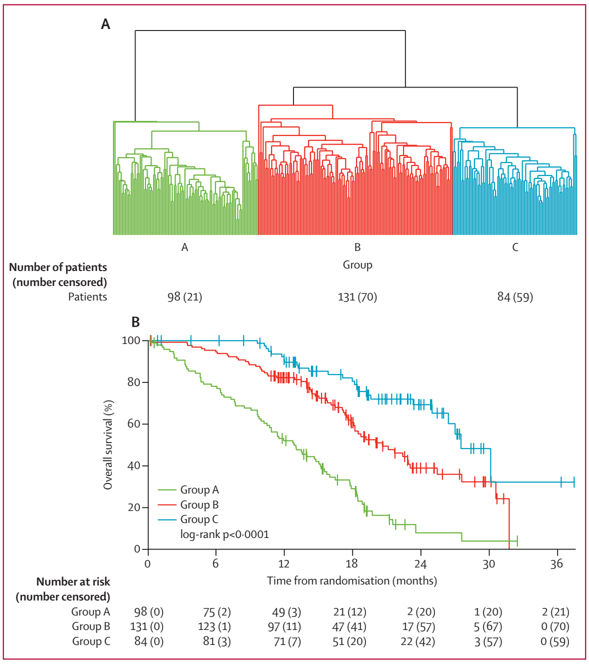
Discussion

The prostate cancer DREAM challenge resulted in one prognostic model to predict overall survival significantly outperforming all other methods, including a reference model reported by Halabi and colleagues,[14] and led to a network perspective of predictive biological variables and their interactions. The results from the top-performing team's model pointed to important interaction effects with immune biomarkers and markers of hepatic function (potentially reflected in the increased amounts of aspartate aminotransferase) and renal function. The network visualisation of the prediction model suggests a complex relation and dependency structure among many of the predictive clinical variables. Many of these noted interactions, although not significant as independent variables, might be important modulators of key clinical traits—eg, haematology-related measurements such as haemoglobin and haematocrit.

Although further investigation is necessary to determine the clinical implication of these associations and provide new insights into tumour–host interaction, these findings shed light on the complex and interwoven nature of prognostic factors on patients' survival.

Open-data, crowdsourced, scientific challenges have been highly effective at drawing together large cross-disciplinary teams of experts to solve complex problems.[25–30] To our knowledge, this DREAM challenge represented the first public collaborative competition[31] to use open-access registration trial datasets in cancer with the intention of improving outcome predictions. In total, 163 individuals comprising 50 teams participated in the challenge, applying state-of-the-art machine learning and statistical modelling methods. The contribution of five clinical trial datasets from industry and academic institutions to Project Data Sphere, and their subsequent use in an open challenge, enabled the advancement of prognostic models in metastatic castration-resistant prostate cancer that up to now was not possible. Modellers had access to several independent clinical trial cohorts with subtle differences in eligibility that increased the diversity (heterogeneity) of the total patient population considered for model development. Access was also provided to data for 150 independent and standardised variables over the trials; by contrast, only 22 variables were considered by the reference model.[14] The challenge resulted in creative data-mining approaches that used standardised raw event-level tables, which are rarely leveraged for prognostic model development, and enabled innovative clinical features to be derived for modelling. Several teams—including the top-performing team—made use of these event-level tables. Finally, evaluation of the 50 methods (validated by an independent and neutral party) provided the most comprehensive assessment of prognostic models in metastatic castration-resistant prostate cancer. These results are both a benchmark for future prognostic model development and a rich source of information that can be mined for additional insights into both patients' stratification and the robustness of clinical predictive factors.

This study has shown the benefits of open data access at a time when clinicians, researchers, and the public are advocating for improved platforms and policies that encourage sharing of clinical trial data.[32,33] Project Data Sphere has overcome major barriers to data sharing with support of data providers, to allow broad access to cancer clinical trial data. To researchers who are interested in leveraging open-access cancer trial data, this study represents a novel research approach that encompassed scientific rigor and a deep understanding of clinical data through effective collaboration of multidisciplinary teams of experts. The top-performing ePCR model was free of any a-priori clinical assumptions, with the exception of exclusion of non-relevant variables in early data curation. The data-driven modelling process identified automatically the best combination of predictors through cross-validation. Furthermore, the ePCR modelling process was fully agnostic to the variables used in the previous reference model; however, many of the same predictors were identified, in addition to novel ones. Such data-driven, unbiased modelling approaches can mine effectively the predictive variables and their combinations from large-scale and open clinical trial data.



Number of patients (number censored)	A	B	C
Patients	98 (21)	131 (70)	84 (59)

Number at risk (number censored)	0	6	12	18	24	30	36
Group A	98 (0)	75 (2)	49 (3)	21 (12)	2 (20)	1 (20)	2 (21)
Group B	131 (0)	123 (1)	97 (11)	47 (41)	17 (57)	5 (67)	0 (70)
Group C	84 (0)	81 (3)	71 (7)	51 (20)	22 (42)	3 (57)	0 (59)

Figure 4: **Challenge meta-analysis**

(A) Hierarchical clustering of patients (Euclidean distance, average linkage) by rank-normalised prediction scores from all 51 models using the ENTHUSE 33 data. (B) Kaplan-Meier plot of survival probability for the three patient clusters from (A). Group A=high risk. Group B=moderate risk. Group C=low risk.

The trials used here represent the standard of care at the time when the trials were done, which is a limitation of this study. Since 2010, several treatments have become available, for use both before and after first-line chemotherapy, and new trials have changed the way clinicians approach this disease.[34] Abiraterone and enzalutamide—both approved for first-line treatment of metastatic castration-resistant prostate cancer—are not included within the scope of this challenge because of a limitation of control arm data; both COU-AA-302[5] and PREVAIL[10] have placebo or prednisone controls, and comparative trials using these agents as control have not been done. Accordingly, trial sponsors should be encouraged to contribute data from the experimental

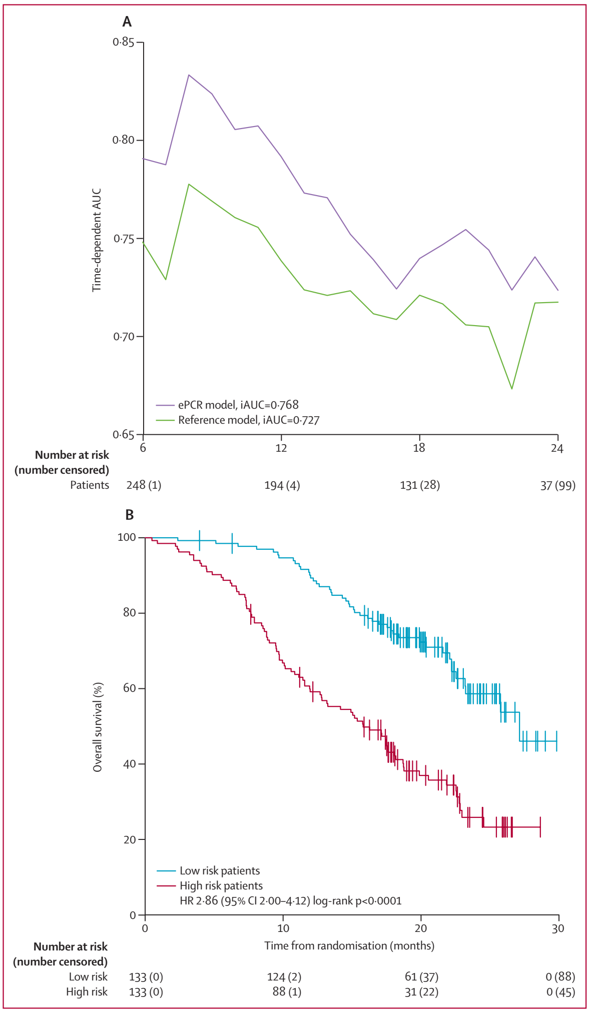


Figure 5: **Performance of ePCR model, using data from ENTHUSE M1**
(A) Time-dependent AUC was measured from 6 months to 24 months at 1-month intervals, reflecting the performance of predicting overall survival at different timepoints. The top-performing model (ePCR) is shown compared with the reference model. (B) Overall survival was assessed by the Kaplan-Meier method, stratified by median risk score. The log rank test was used to compare risk groups. ePCR=ensemble of penalised Cox regression models. iAUC=integrated time-dependent area under the curve. HR=hazard ratio.

arm (particularly for approved drugs) to an active and engaged research community. Although sponsors are concerned that virtual comparisons might be made between treatments in experimental arms of different trials, there is far more benefit in leveraging these data to validate prognostic factors and models and to investigate intermediate clinical endpoints predictive of survival.

The DREAM challenge described here has shown that there is opportunity to further optimise prognostic models in metastatic castration-resistant prostate cancer using baseline clinical variables. For substantial advances beyond the work presented here, clinical trial data must be made available that reflects current advancements in treatment paradigms, including new data-capture techniques such as genomics, immunogenomics, and metabolomics that might more accurately describe the malignant state of the tumour and its microenvironment. Vital to either of these will be the need to share patient-level oncology data with the research community for the development of the next generation of prognostic and predictive models in cancer.

Contributors

JG, TW, JCB, ECN, TY, BMB, KA, TN, SF, GS, HS, CJS, CJR, HIS, OS, YX, FLZ, and JCC designed the DREAM Challenge. FLZ and LS led the efforts by Project Data Sphere to obtain and process the clinical trial data. TDL, SAK, GP, AA, TP, TM, and TA designed the top-performing method. JG, TW, TDL, KKW, JCB, ECN, GS, TA, FLZ, and JCC did the post-challenge data analysis and interpretation. HS, CJS, CJR, HIS, and OS assisted in clinical variable interpretation. All members of the Prostate Cancer Challenge DREAM Consortium (appendix pp 20–22) submitted prognostic models to the Challenge and provided method write-ups and the code to reproduce their predictions. JG, TW, TDL, HS, CJS, CJR, HIS, OS, TA, FLZ, and JCC wrote the report.

Declaration of interests

CJS reports personal fees from Sanofi, Janssen, Astellas, and Bayer, outside the submitted work. FLZ and LS are employed by and have stock in Sanofi. KA is employed by and has stock in AstraZeneca, outside the submitted work. HIS reports non-financial support from AstraZeneca, Bristol-Myers Squibb, Ferring Pharmaceuticals, Medivation, and Takeda Millennium, outside the submitted work; consultancy fees from Astellas, BIND Pharmaceuticals, Blue Earth Diagnostics, Clovis Oncology, Elsevier (PracticeUpdate website), Genentech, Med IQ, Merck, Roche, Sanofi Aventis, WCG Oncology, and Asterias Biotherapeutics, outside the submitted work; and grants from Illumia, Innocrin Pharma, Janssen, and Medivation, outside the submitted work. GP reports grants from the Academy of Finland (decision number 265966), during the conduct of the study and outside the submitted work. SAK reports grants from the Academy of Finland (decision number 296516), during the conduct of the study and outside the submitted work. TDL reports research grants from the Finnish Cultural Foundation and Drug Research Doctoral Programme, during the conduct of the study; and a research contract from the National Cancer Institution, during the conduct of the study. TA reports grants from the Academy of Finland and the Cancer Society of Finland, during the conduct of the study; and a research contract from the National Cancer Institute, during the conduct of the study. OS reports grants and personal fees from Sanofi, outside the submitted work. AA, JCB, BMB, JCC, SF, JG, TM, ECN, TN, TP, CJR, HS, GS, TW, KKW, TY, and YX declare no competing interests.

Acknowledgments

This report is based on research using information obtained from Project Data Sphere. Neither Project Data Sphere nor the owners of any information from the website have contributed to, approved, or are in any way responsible for the contents of this report. We thank the Sage Bionetworks Synapse team for the development and design of the DREAM challenge website. This work is supported in part by: the National Institutes of Health, National Library of Medicine (2T15-LM009451), National Cancer Institute (16X064; 5R01CA152301), Boettcher Foundation, Academy of Finland (grants 265966, 296516, 292611, 269862, 272437, 279163, 295504), Cancer Society of Finland, Drug Research Doctoral Programme (DRDP) at the University of Turku, and Finnish Cultural Foundation. Sanofi US Services provided an in-kind contribution of human resources for curation of raw datasets for the challenge and for clinical and scientific support of challenge organisation, at the request of Project Data Sphere.

References

1 Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D. Global cancer statistics. *CA Cancer J Clin* 2011; **61:** 69–90.

2 Tanimoto T, Hori A, Kami M. Sipuleucel-T immunotherapy for castration-resistant prostate cancer. *N Engl J Med* 2010; **363:** 1966.

3 Berruti A, Pia A, Terzolo M. Abiraterone and increased survival in metastatic prostate cancer. *N Engl J Med* 2011; **365:** 766.

4 Fizazi K, Carducci M, Smith M, et al. Denosumab versus zoledronic acid for treatment of bone metastases in men with castration-resistant prostate cancer: a randomised, double-blind study. *Lancet* 2011; **377:** 813–22.

5 Ryan CJ, Smith MR, de Bono JS, et al. Abiraterone in metastatic prostate cancer without previous chemotherapy. *N Engl J Med* 2013; **368:** 138–48.

6 Scher HI, Fizazi K, Saad F, et al. Increased survival with enzalutamide in prostate cancer after chemotherapy. *N Engl J Med* 2012; **367:** 1187–97.

7 de Bono JS, Oudard S, Ozguroglu M, et al, for the TROPIC investigators. Prednisone plus cabazitaxel or mitoxantrone for metastatic castration-resistant prostate cancer progressing after docetaxel treatment: a randomised open-label trial. *Lancet* 2010; **376:** 1147–54.

8 Parker C, Nilsson S, Heinrich D, et al. Alpha emitter radium-223 and survival in metastatic prostate cancer. *N Engl J Med* 2013; **369:** 213–23.

9 Kantoff PW, Higano CS, Shore ND, et al. Sipuleucel-T immunotherapy for castration-resistant prostate cancer. *N Engl J Med* 2010; **363:** 411–22.

10 Beer TM, Armstrong AJ, Rathkopf DE, et al. Enzalutamide in metastatic prostate cancer before chemotherapy. *N Engl J Med* 2014; **371:** 424–33.

11 Halabi S, Small EJ, Kantoff PW, et al. Prognostic model for predicting survival in men with hormone-refractory metastatic prostate cancer. *J Clin Oncol* 2003; **21:** 1232–37.

12 Smaletz O, Scher HI, Small EJ, et al. Nomogram for overall survival of patients with progressive metastatic prostate cancer after castration. *J Clin Oncol* 2002; **20:** 3972–82.

13 Armstrong AJ, Garrett-Mayer ES, Yang Y-CO, de Wit R, Tannock IF, Eisenberger M. A contemporary prognostic nomogram for men with hormone-refractory metastatic prostate cancer: a TAX327 study analysis. *Clin Cancer Res* 2007; **13:** 6396–403.

14 Halabi S, Lin C-Y, Kelly WK, et al. Updated prognostic model for predicting overall survival in first-line chemotherapy for patients with metastatic castration-resistant prostate cancer. *J Clin Oncol* 2014; **32:** 671–77.

15 Okser S, Pahikkala T, Airola A, Salakoski T, Ripatti S, Aittokallio T. Regularized machine learning in the genetic prediction of complex traits. *PLoS Genet* 2014; **10:** e1004754.

16 Papaemmanuil E, Gerstung M, Bullinger L, et al. Genomic classification and prognosis in acute myeloid leukemia. *N Engl J Med* 2016; **374:** 2209–21.

17 Longo DL, Drazen JM. Data sharing. *N Engl J Med* 2016; **374:** 276–77.

18 Scher HI, Jia X, Chi K, et al. Randomized, open-label phase III trial of docetaxel plus high-dose calcitriol versus docetaxel plus prednisone for patients with castration-resistant prostate cancer. *J Clin Oncol* 2011; **29:** 2191–98.

19 Petrylak DP, Vogelzang NJ, Budnik N, et al. Docetaxel and prednisone with or without lenalidomide in chemotherapy-naive patients with metastatic castration-resistant prostate cancer (MAINSAIL): a randomised, double-blind, placebo-controlled phase 3 trial. *Lancet Oncol* 2015; **16:** 417–25.

20 Tannock IF, Fizazi K, Ivanov S, et al, on behalf of the VENICE investigators. Aflibercept versus placebo in combination with docetaxel and prednisone for treatment of men with metastatic castration-resistant prostate cancer (VENICE): a phase 3, double-blind randomised trial. *Lancet Oncol* 2013; **14:** 760–68.

21 Fizazi K, Fizazi KS, Higano CS, et al. Phase III, randomized, placebo-controlled study of docetaxel in combination with zibotentan in patients with metastatic castration-resistant prostate cancer. *J Clin Oncol* 2013; **31:** 1740–47.

22 Nelson JB, Fizazi K, Miller K, et al. Phase III study of the efficacy and safety of zibotentan (ZD4054) in patients with bone metastatic castration-resistant prostate cancer (CRPC). *Proc Am Soc Clin Oncol* 2011; **29:** abstr 117.

23 Hung H, Chiang C-T. Estimation methods for time-dependent AUC models with survival data. *Can J Stat* 2010; **38:** 8–26.

24 Project Data Sphere. Prostate cancer DREAM challenge. https://www.projectdatasphere.org/projectdatasphere/html/pcdc (accessed Oct 21, 2016).

25 Costello JC, Stolovitzky G. Seeking the wisdom of crowds through challenge-based competitions in biomedical research. *Clin Pharmacol Ther* 2013; **93:** 396–98.

26 Bender E. Challenges: crowdsourced solutions. *Nature* 2016; **533:** S62–64.

27 Bansal M, Yang J, Karan C, et al. A community computational challenge to predict the activity of pairs of compounds. *Nat Biotechnol* 2014; **32:** 1213–22.

28 Costello JC, Heiser LM, Georgii E, et al. A community effort to assess and improve drug sensitivity prediction algorithms. *Nat Biotechnol* 2014; **32:** 1202–12.

29 Margolin AA, Bilal E, Huang E, et al. Systematic analysis of challenge-driven improvements in molecular prognostic models for breast cancer. *Sci Transl Med* 2013; **5:** 181re1.

30 Ewing AD, Houlahan KE, Hu Y, et al. Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. *Nat Methods* 2015; **12:** 623–30.

31 Saez-Rodriguez J, Costello JC, Friend SH, et al. Crowdsourcing biomedical research: leveraging communities as innovation engines. *Nat Rev Genet* 2016; **17:** 470–86.

32 Merson L, Gaye O, Guerin PJ. Avoiding data dumpsters: toward equitable and useful data sharing. *N Engl J Med* 2016; **374:** 2414–15.

33 Bierer BE, Li R, Barnes M, Sim I. A global, neutral platform for sharing trial data. *N Engl J Med* 2016; **374:** 2411–13.

34 Lewis B, Sartor O. The changing landscape of metastatic prostate cancer. *Am J Hematol* 2015; **11:** 11–20.

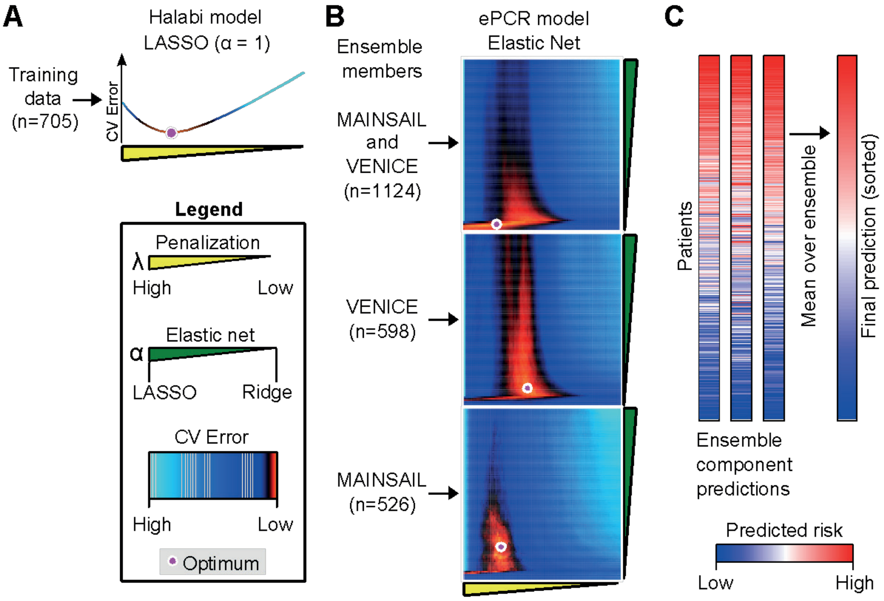
Supplementary Material (Publication IV):

Supplementary Figures S1 - S7

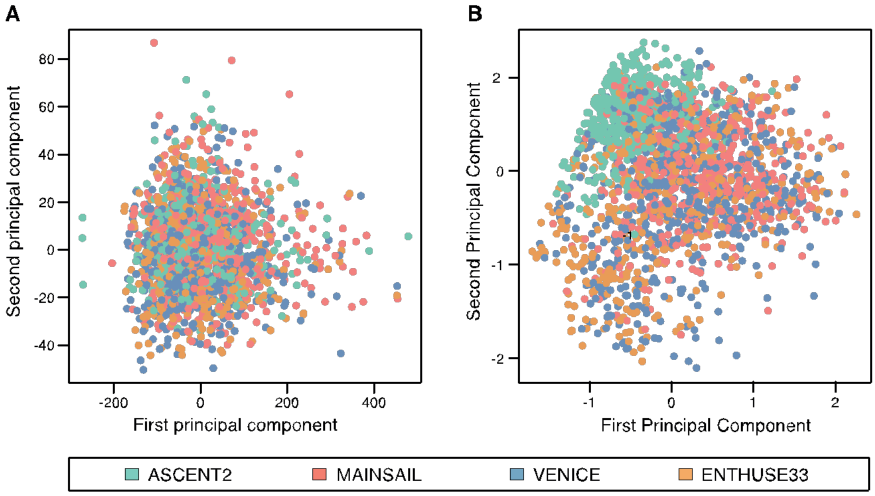
Supplementary Tables S1 - S3

Supplementary Methods (*Available online*)*

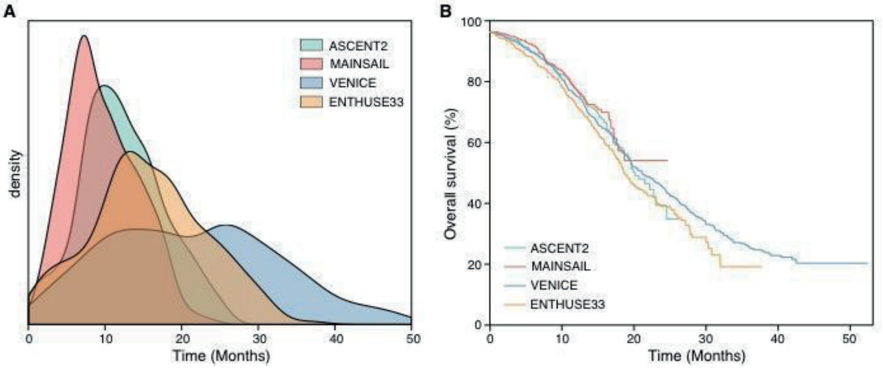
*: `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5217180/#SD1`



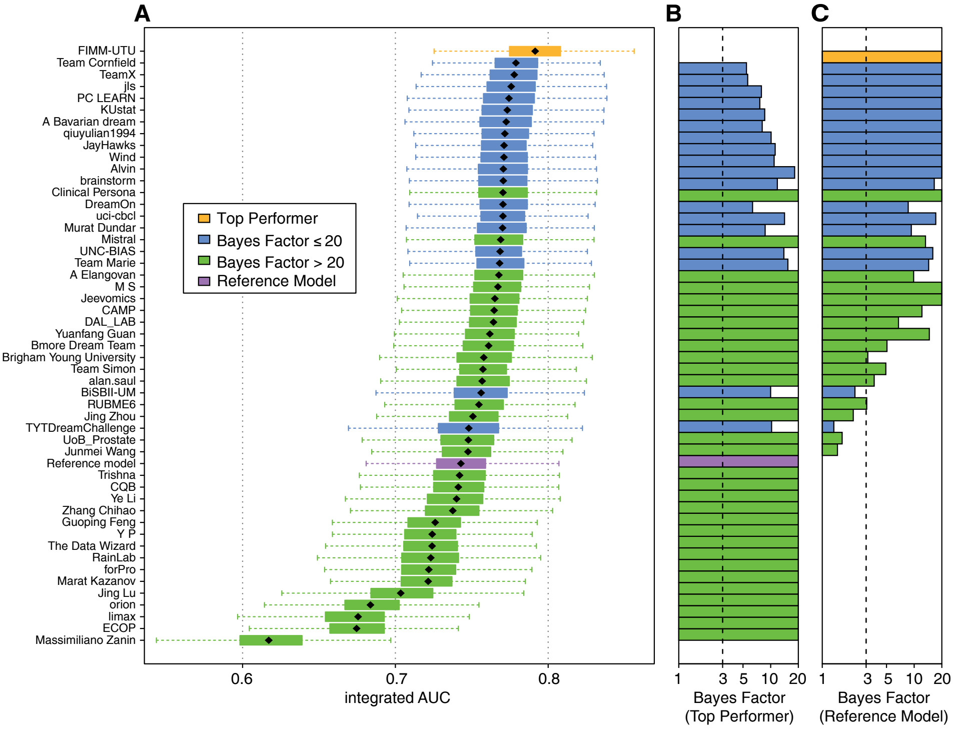
Supplementary Figure 1. Overview of the top-performing ePCR method in comparison to the Reference model (Halabi model). (A) The benchmarking Reference model explored the LASSO model (α = 1) in a training data cohort with respect to the regularization parameter (λ) using cross-validation (CV). (B) The top-performing ePCR approach is based on an ensemble of Penalized Cox Regression models (ePCR), which are optimized separately for each cohort or a combination of cohorts in terms of the regularization parameter (λ) as well as the full range of the L1/L2 regularization parameter (0 <= α <= 1). The optimal model was identified with low values of α, indicating that the Ridge Regression (α = 0)-like models performed better for modeling the complex interactions than the benchmarking Reference LASSO-model (α = 0). (C) Ensemble predictions were generated by averaging over the predicted risk ranks from each ensemble component.



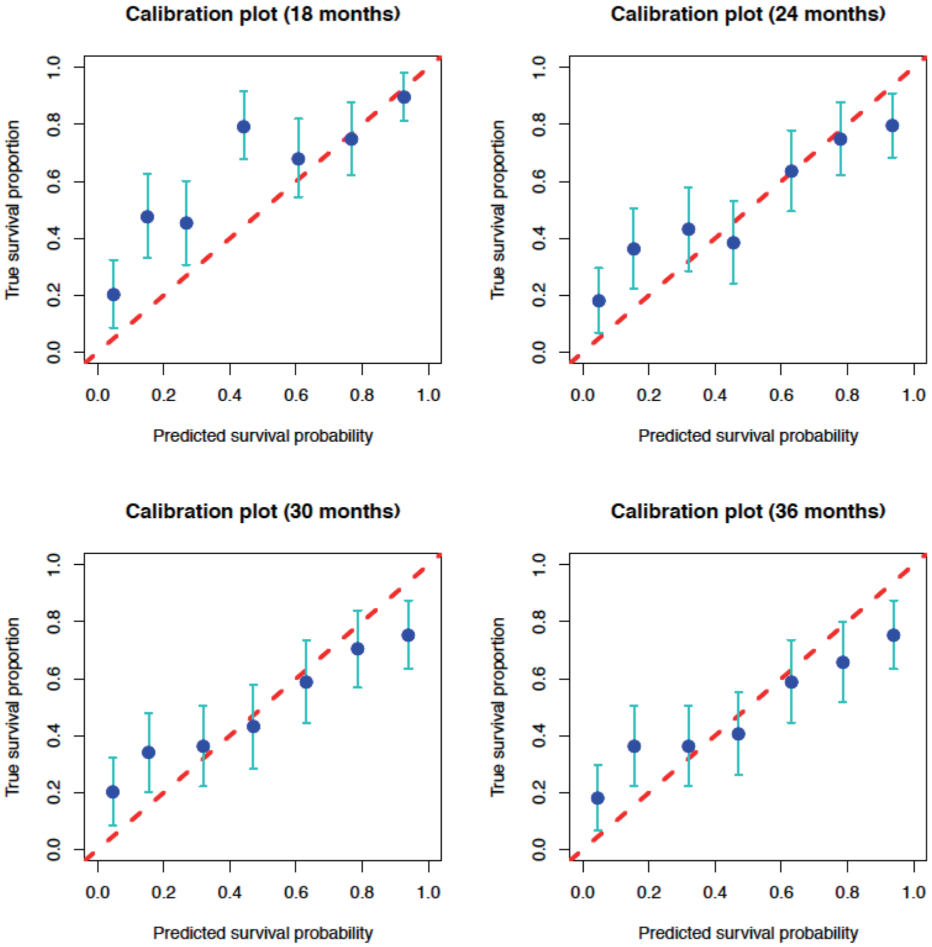
Supplementary Figure 2. (A) All data across ASCENT2, MAINSAIL, VENICE, and ENTHUSE 33– both binary and continuous data – were used in a PCA. (B) All data across the 4 studies – only binary variables – were used in PCA.



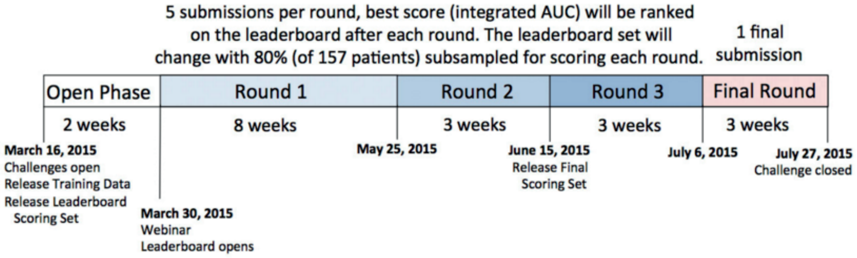
Supplementary Figure 3. (A) Density plot of follow-up times per study for the ASCENT2, MAINSAIL, VENICE, and ENTHUSE 33 trials. (B) Survival profile for each of the trials.



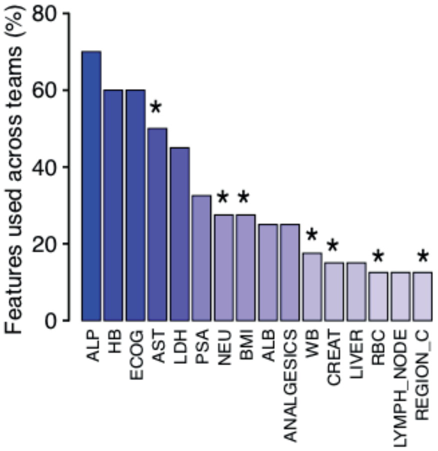
Supplementary Figure 4. Summary of Challenge results across all 50 teams plus the Reference model evaluated using the ENTHUSE 33 dataset. (A) Performance of submissions. Each submission underwent 1,000 paired bootstrap of final scoring patient set to calculate a Bayes factor against the top-performer a Bayes factor against the Reference model. A p value was calculated from randomization test of 1000 permutations. X-axis is iAUC and y-axis is submissions ranked by iAUC from high to low. Each team's bootstrapped iAUC scores are shown as horizontal boxplot with the black diamonds representing the point estimate of a team's performance. The colored boxes show the inter-quartile ranges and the whiskers extend to 1.5 times the corresponding interquartile ranges. Top-performer is colored in orange, other teams within Bayes factor of 20 were labeled in blue, and the rest of the teams were labeled in green. The Reference model is labeled in purple. (B) Bayes factors of all submissions against the top-performer are shown. Bayes factors greater than 20 were truncated to 20. (C) Bayes factors of all submissions against the Reference model. Bayes factors greater than 20 were truncated to 20.



Supplementary Figure 5. Calibration plots for the ePCR model of predicted survival probability versus true survival proportion for the ENTHUSE 33 dataset at 18, 24, 30, and 36 months.


5 submissions per round, best score (integrated AUC) will be ranked on the leaderboard after each round. The leaderboard set will change with 80% (of 157 patients) subsampled for scoring each round.
1 final submission
Open Phase
Round 1
Round 2
Round 3
Final Round
2 weeks
8 weeks
3 weeks
3 weeks
3 weeks
March 16, 2015
Challenges open
Release Training Data
Release Leaderboard Scoring Set
March 30, 2015
Webinar
Leaderboard opens
May 25, 2015
June 15, 2015
Release Final Scoring Set
July 6, 2015
July 27, 2015
Challenge closed

Supplementary Figure 6. Timeline for the Challenge. Five submissions were allowed per round, and only a single submission for the final validation round.



Supplementary Figure 7. Most frequently utilized variables by teams to build their final models using the ASCENT2, MAINSAIL, and VENICE trials. Results are self-reported from a post-Challenge survey over 40 teams. * variables are not used in the Reference model.

Supplementary Table 1. Full results from all 50 teams plus the Reference model across several scoring metrics from the Challenge. Performance measures were evaluated using the ENTHUSE 33 trial. Teams are listed with the links to their predictions, methods write-up, and code.

Team	Risk score predictions	Method write-up & code	iAUC	c-index	AUC12	AUC18	AUC24
FIMM-UTU (ePCR)	syn4732198	syn4227610	0.7915	0.7307	0.7918	0.7674	0.8388
Team Cornfield	syn4732339	syn4732274	0.7789	0.7263	0.7708	0.7663	0.8147
TeamX	syn4732955	syn4732218	0.7778	0.7157	0.7492	0.7645	0.8369
jls	syn4732934	syn4732827	0.7758	0.7212	0.7713	0.7553	0.8085
PC LEARN	syn4733119	syn3822697	0.7743	0.7205	0.7577	0.762	0.8258
KUstat	syn4741808	syn4260742	0.7732	0.7126	0.7436	0.7533	0.8376
A Bavarian dream	syn4732177	syn5592405	0.7725	0.7237	0.7721	0.7664	0.8019
qiuyulian1994	syn4732213	syn4732205	0.7716	0.711	0.7423	0.7506	0.8297
JayHawks	syn4731663	syn4214500	0.7711	0.7193	0.7717	0.7607	0.8124
Wind	syn4731647	syn4731645	0.771	0.7181	0.7625	0.7688	0.8124
Alvin	syn4732814	syn4229406	0.7707	0.7136	0.7586	0.7568	0.7927
brainstorm	syn4730818	syn3821841	0.7706	0.718	0.7617	0.7614	0.8175
uci-cbcl	syn4731657	syn4227279	0.7704	0.717	0.76	0.7716	0.8206
DreamOn	syn4731710	syn4731708	0.7704	0.712	0.7559	0.7582	0.8245
Clinical Persona	syn4681602	syn4681529	0.7704	0.7149	0.7533	0.7545	0.8328
Murat Dundar	syn4595033	syn4595029	0.7701	0.7305	0.7763	0.7773	0.773
Mistral	syn4622079	syn4622016	0.7689	0.7073	0.7382	0.7624	0.8268
UNC-BIAS	syn4731768	syn4731674	0.7685	0.717	0.7559	0.7568	0.8293
Team Marie	syn4731882	syn4485029	0.7682	0.7142	0.7519	0.7705	0.8151
A Elangovan	syn4643159	syn4212102	0.7677	0.7135	0.7655	0.7461	0.7977
M S	syn4730601	syn4229266	0.7671	0.707	0.7372	0.7652	0.8256
Jeevomics	syn4733845	syn4074987	0.7651	0.719	0.7733	0.7526	0.7917
CAMP	syn4731373	syn3647478	0.7646	0.7077	0.7331	0.758	0.8143
DAL_LAB	syn4731755	syn4731746	0.7642	0.7103	0.7521	0.7486	0.8305
Yuanfang Guan	syn7152471	syn7152438	0.7618	0.7143	0.7545	0.7631	0.8005
Bmore Dream Team	syn4733165	syn3616830	0.761	0.7121	0.7464	0.766	0.7948
Brigham Young University	syn4733391	syn4382527	0.7578	0.7048	0.7381	0.7685	0.7599
Team Simon	syn4733651	syn4732901	0.7573	0.7033	0.7278	0.7611	0.827
alan.saul	syn4731492	syn4587469	0.7568	0.7078	0.7464	0.7606	0.7961
BiSBII-UM	syn4733056	syn4229636	0.7561	0.6992	0.7394	0.7397	0.8007
RUBME6	syn4733262	syn4590933	0.7547	0.6994	0.7419	0.7198	0.7866
Jing Zhou	syn4646618	syn3685423	0.7507	0.6994	0.7361	0.7491	0.803
TYTDreamChallenge	syn4733257	syn4228911	0.748	0.7002	0.7343	0.7402	0.7657

UoB_Prostate	syn4733441	syn4591879	0.7478	0.7057	0.7468	0.7367	0.7699
Junmei Wang	syn4732891	syn4225820	0.7475	0.694	0.7319	0.7332	0.7955
Halabi Model	syn4770841	syn3324780	0.7429	0.6985	0.7418	0.7375	0.7634
Trishna	syn4730580	syn4730570	0.742	0.6922	0.7285	0.7383	0.774
CQB	syn4732202	syn3566822	0.7412	0.6914	0.7185	0.7293	0.7686
Ye Li	syn4731357	syn4731355	0.74	0.6907	0.7258	0.7249	0.806
Zhang Chihao	syn4748861	syn4259433	0.7376	0.7063	0.7561	0.7426	0.745
Guoping Feng	syn4730823	syn4730561	0.7261	0.6781	0.7073	0.707	0.7504
Y P	syn4732913	syn4732909	0.7241	0.6799	0.732	0.7057	0.7594
RainLab	syn4730829	syn4238316	0.7232	0.6708	0.7141	0.7394	0.7821
forPro	syn4707761	syn4707464	0.7219	0.6839	0.7267	0.7249	0.739
Marat Kazanov	syn4731369	syn4730567	0.7215	0.6675	0.7089	0.7112	0.7524
Jing Lu	syn4732498	syn4556277	0.7035	0.6689	0.6931	0.7073	0.7154
orion	syn4733693	syn4732963	0.6837	0.6457	0.717	0.7359	0.7952
limax	syn4732094	syn4721051	0.6756	0.6484	0.7033	0.6685	0.689
ECOP	syn4647266	syn4647259	0.6746	0.6554	0.6774	0.6881	0.6949
Massimiliano Zanin	syn4732241	syn4732239	0.6171	0.6081	0.6206	0.432	0.3852
The Data Wizard	syn4229053	syn4228992	0.5945	0.5815	0.6039	0.5824	0.6085
Compiled set of all predictions		syn7071669					

Supplementary Table 2. Comparison of risk stratification of patients in the ENTHUSE 33 trial by the ePCR and Reference models. Patients were dichotomized at median risk scores. All intervals reported are 95% confidence intervals. PPV = positive predictive value, NPV = negative predictive value. Values for Cases, Survivors, and Censored are cumulative.

ePCR model	**Patient count**	**Event count**	**Median survival time, month (CI)**	**1 year survival rate (CI)**	**2 year survival rate (CI)**	
Low risk group	156	56	27.6 (23.4-NA)	90.20% (85.5%-95.00%)	58.60% (49.7%- 69.00%)	
High risk group	157	107	15.1 (13.0-17.2)	59.90% (52.55%-68.20%)	15.70% (9.28%- 26.70%)	
Reference model	**Patient count**	**Event count**	**Median survival time, month (CI)**	**1 year survival rate (CI)**	**2 year survival rate (CI)**	
Low risk group	156	59	26.5 (22.5-NA)	87.40% (82.30%-92.90%)	52.80% (43.90%-63.50%)	
High risk group	157	104	15.6 (14.0-18.4)	62.70% (55.50%-70.80%)	22.20% (15.00%-32.90%)	
	Time (months)	t=6	t=12	t=18	t=24	t=30
	Cases	28	75	121	153	160
	Survivors	279	214	118	41	9
	Censored	6	24	74	119	144
Sensitivity (%)	ePCR	92.89	81.32	72.63	65.86	60.67
	Reference	85.73	75.94	67.43	61.19	61.21
Specificity (%)	ePCR	54.48	60.28	68.64	82.93	66.67
	Reference	53.76	57.94	64.41	73.17	44.44
PPV (%)	ePCR	16.96	40.15	64.2	86.31	82.41
	Reference	15.65	37.17	59.46	78.85	73.93
NPV (%)	ePCR	98.71	90.78	76.41	59.78	39.7
	Reference	97.41	88.02	71.86	53.57	30.8

Supplementary Table 3. Top 15 single and interacting variables from the final ePCR model built from the MAINSAIL and VENICE trials. Comprehensive list of evaluated variables is available at: https://www.synapse.org/#!Synapse:syn7113819

Top 15 single variables in the ePCR model	Ensemble p value	Ensemble effect size
Lactate dehydrogenase (LDH)	< 0.0001	3405.667
Aspartate aminotransferase (AST)	< 0.0001	3376.667
Hemoglobin (HB)	< 0.0001	3369.667
Hematocrit (HCT)	< 0.0001	3354.333
Albumin (ALB)	0.0004	3316.667
Alkaline phosphatase (ALP)	< 0.0001	3291.333
Red blood cell count (RBC)	< 0.0001	3237.333
Systolic blood pressure (SYSTOLICBP)	0.0012	3192.000
Lesions at liver (LIVER)	< 0.0001	3184.000
Sodium (NA)	0.0205	3032.000
Lesions at target site (TARGET)	0.0118	3001.000
ECOG performance status (ECOG_C)	0.0003	2923.000
Medical history: cardiac disorders (MHCARD)	0.1100	2827.667
Lymphocyte/Leukocyte ratio (LYMperLEU)	0.0143	2684.333
Body mass index (BMI)	0.0214	2679.333

Top 15 interactions in the ePCR model		Ensemble p value	Ensemble effect size
AST	LDH	< 0.0001	3408.333
ALP	LDH	< 0.0001	3406.667
ALP	AST	< 0.0001	3404.333
HB	SYSTOLICBP	< 0.0001	3402.333
LDH	Urine Specific Gravity	< 0.0001	3400.667
SYSTOLICBP	HCT	< 0.0001	3400.333
Creatinine	LDH	< 0.0001	3397.333
LDH	LDH	< 0.0001	3392.000
HB	ALB	< 0.0001	3387.333

AST	AST	< 0.0001	3384.333
HB	NA	< 0.0001	3382.667
Height	LDH	< 0.0001	3381.667
ALB	SYSTOLICBP	< 0.0001	3379.333
HB	Creatinine clearance	< 0.0001	3378.000
ALB	HCT	< 0.0001	3377.333

Annales Universitatis Turkuensis

UNIVERSITY
OF TURKU

Grano Oy, Turku , Finland 2018

ISBN 978-951-29-7443-6 (PRINT)
ISBN 978-951-29-7444-3 (PDF)
ISSN 0082-7002 (Print) ISSN 2343-3175 (Online)