DEVELOPING, VALIDATING, AND EVALUATING CLINICAL PREDICTION MODELS IN BREAST AND PROSTATE CANCER

Tom Hueting

DEVELOPING, VALIDATING, AND EVALUATING CLINICAL PREDICTION MODELS IN BREAST AND PROSTATE CANCER

Tom Hueting

Graduation Committee:

Chair / secretary:	prof.dr. T. Bondarouk
Supervisors:	prof. dr. S. Siesling, Universiteit Twente prof. dr. ir. H. Koffijberg, Universiteit Twente
Co-supervisor:	dr. M.C. van Maaren, Universiteit Twente
Committee Members:	prof. dr. D.W. Donker, Universiteit Twente dr. C.H.C. Drossaert, Universiteit Twente prof. dr. H.G. Van Der Poel, Nederlands Kankerinstituut prof. dr. J.E.A. Portielje, Leiden Universitair Medisch Centrum prof. dr. E.W. Steyerberg, Leiden Universitair Medisch Centrum prof. dr. A.L.A.J. Dekker, Maastro klinieken

DEVELOPING, VALIDATING, AND EVALUATING CLINICAL PREDICTION MODELS IN BREAST AND PROSTATE CANCER

DISSERTATION

to obtain
the degree of doctor at the Universiteit Twente,
on the authority of the rector magnificus,
prof. dr. ir. A. Veldkamp,
on account of the decision of the Doctorate Board
to be publicly defended
on Wednesday 22 June 2022 at 16.45 hours

by

Thomas Alexander Hueting
born on the 6th of June, 1992
in Hengelo, The Netherlands

This dissertation has been approved by:

Supervisors	prof. dr. ir. S. Siesling
	prof. dr. ir. H. Koffijberg
Co-supervisor	dr. M.C. van Maaren

Financial support for printing of this thesis was kindly provided by:

iKNL integraal kankercentrum Nederland

This thesis is part of the Health Sciences Series, HSS 22-43, department Health Technology and Services Research, University of Twente, Enschede, the Netherlands. ISSN: 1878-4968.

Cover design:	Herik Media
Printed by:	ProefschriftMaken.nl
Lay-out:	Dennis Hendriks || ProefschriftMaken.nl
ISBN:	978-90-365-5394-0
DOI:	10.3990/1.9789036553940

Table of contents

Chapter 1

Introduction

Prediction models

Every individual is unique. However, based on many personal characteristics individuals can be categorized in more homogenous groups, which is a way to be able to better compare people with each other to allow for the identification of potentially relevant associations between specific groups of patients and (health) outcomes. Consider, for example, characteristics such as gender, age, height, weight, and so on. The combination of all these different traits in combination with personal circumstances is what makes the individual unique. When patients are confronted with a (serious) condition, the prognosis and benefits from treatment typically depend on such a set of unique characteristics. Available treatment options and their corresponding risks and benefits are preferably tailored towards the individual as well as possible. International clinical guidelines provide recommendations regarding evidence-based treatment for specific groups of patients based on findings from randomized clinical trials. Physicians are able to combine the guideline recommendations with many individual patient and disease characteristics to provide consultation to the patient, but their judgment may be complemented by valuable insights provided by prediction models. Such models have been shown to provide more accurate estimations than physicians regarding a predicted probability for e.g. survival, benefits, and harms of specific treatment options.[1,2]

Clinical prediction models are statistical tools which can predict the probability of a certain outcome or event for an individual patient based on their (clinical) characteristics. Such predictions can support clinicians and patients in the (shared) decision-making process regarding most optimal treatment scenarios.[3] Prediction models can be used to predict the presence of a disease or condition (i.e. diagnostic model) or to predict an outcome occurring in the (near) future (i.e. prognostic model).[4] For example, a diagnostic model predicts the probability of lymph node involvement (LNI) that is currently present in a patient, and a prognostic model can predict the probability of cancer recurrence within 5-years after successful treatment. Use of prediction models can serve multiple purposes depending on the context in which the model is applied. For instance, risk-based strategies can be implemented to guide decision-making regarding the added value of a diagnostic tool or which patients may have (un)favorable balance between treatment benefits and risks (i.e. potential complications).

The number of developed and available prediction models has increased exponentially in the past decades. For example, a review on prediction models aiming to predict the probability of cardiovascular disease in the general population identified more than 360 models mainly developed in the last 10-15 years.[5] Yet, the number of models being recommended in clinical guidelines remains limited, indicating a clear gap between the development and the implementation of clinical prediction models. This gap is caused by several challenges related to the implementation of prediction models involving prediction model accessibility, transparency, generalizability, updates, impact assessment, and interpretation, which are one by one described below.

Accessibility

Prediction models are mainly described and published in (peer-reviewed) scientific papers in which the model is frequently visualized as a static nomogram. Increasingly often, they are also published on websites in the form of an online calculator. However, when online calculators are used in clinical practice, there is no guarantee that they remain available. For example, Adjuvant! Online was recommended for use in clinical practice by multiple guidelines,[6] but the online calculator unexpectedly was removed from the internet, presumably due to a committed update that includes HER2 status as a prognostic factor.[7] However, no indications for an update can be found, and the model has been inaccessible ever since. In addition, the widespread availability of online calculators on different websites with varying interfaces will hamper the implementation of all the different tools in clinical practice.

Transparency

Publications of prediction models do not always include the full details of the underlying statistical model. This hampers the reproduction of the model for external validation and subsequent use of the models in clinical practice. The lack of transparency in papers reporting on multivariable prediction models has been acknowledged by multiple researchers which provided possible solutions in the form of methodological guidelines, article series, and books that were published in the past years;

- Prognosis Research Strategy (PROGRESS)[8–10]
- Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD)[11]
- Checklist for critical Appraisal and data extraction for systematic Reviews of prediction Modelling Studies (CHARMS)[12]
- Prediction model Risk Of Bias Assessment Tool (PROBAST)[13]
- Prognosis research in healthcare[14]
- Clinical prediction models[3]

Still, despite the extensive recommendations from the abovementioned resources, quite some recently published prediction models seem to be developed using flawed development or validation methods, or at least flawed at the proper (transparent) description of the applied methods and relevant results.[15–17]

Generalizability

In case the full details of the underlying statistical model(s) are published, the model(s) can be externally validated using data of patients who were not included in the development of the model. External validation is of great importance to evaluate model robustness and generalizability to patient groups differing from the original development cohort. This especially concerns prediction models where strict inclusion criteria were applied to the development cohort. Yet, external validations are not routinely performed. For some prediction models currently being used to support clinical decision-making, a proper understanding regarding the performance of these models in the target population is lacking. For example, Adjuvant! Online[18] was recommended for use in the Dutch breast cancer guideline prior to external validation of the model in Dutch women.[19] Similarly, models predicting lymph node involvement (LNI) in prostate cancer were recommended in the Netherlands based on an external validation study performed with data of German prostate cancer patients, but no recommended model has been externally validated using Dutch patient data.[20,21]

Updates

An important step often performed during or following external validation of multivariable prediction models is updating of the models, also called recalibration. This may consist of an update of the model intercept and of the variable coefficients, and if beneficiary, a change to the included set predictors (i.e. inclusion of new variables or removal of previously included predictors). Recalibration is used to improve the fit of the model on the data while retaining valuable information from the original development of the model. Moreover, existing prediction models could become outdated over time when the clinical practice is evolving, new treatments and better outcomes might be achieved and the context changes in which the models are used. To ensure that models remain useful and accurate, external validation studies should be repeated over time, and models may need to be updated accordingly. Updating of models is not always sensible when, for instance, an existing prediction model shows good performance on external validation, and a model update would not improve the performance. In order to simplify the updating process, statistical methods have been published to give an indication of whether it makes sense to update a model during external validation.[22]

Impact assessment

After adequate external validation, it often remains unclear which impact the application of a prediction model has on clinical outcomes (e.g. treatment response), health outcomes (e.g. quality of life), and costs of care. Using prediction models in clinical practice should be regarded as an intervention that affects health outcomes by tailoring treatment to the individual. To monitor the impact of the model used on individual patient level, it is necessary to have insight into trade-offs (i.e. outcomes versus costs) that arise when a particular risk threshold value (e.g. for further diagnostic testing, therapeutic decisions, or follow-up schedules) is applied in practice. Moreover, it is needed to determine the effect of different thresholds on outcomes. Several papers have been published providing guidance for an adequate assessment of the impact of clinical prediction models.[23,24] Unfortunately, impact evaluation studies applied to clinical prediction models are rare.[25] However, with the introduction of the Medical Device Regulation (MDR) in the European Union,[26] impact studies will need to be performed increasingly frequent as part of the clinical evaluation required by the legislation.

Interpretation

To enhance (shared) decision-making on individual patient level, it is necessary to correctly interpret the calculated probability, which decision it supports, and what the expected risks and benefits are. Moreover, threshold values should be given accompanied by advice on how to act in clinical practice.[27] (e.g. continue with further testing, therapeutic decisions, or follow-up schedules).[27] Such thresholds aim to optimize outcomes at the group level, while decision-making at patient level can deviate from the recommendation, e.g. due to personal preferences of a patient. Still, the interpretation of predicted risks from clinical prediction models used for decision-making remains challenging, in particular when a composite outcome is predicted.[28]

Combined, these challenges currently lead to the use of prediction models that have not been adequately validated in the target population, and an undesirable variation in prediction models used across hospitals and even across clinicians. At the same time, new models are constantly being published while studies externally validating and potentially updating readily available models are lacking. Consequently, the benefit of using clinical prediction models as decision support tools has (by far) not reached its full potential for both patients and healthcare professionals.[29,30]

To overcome the highlighted challenges involving clinical prediction models, it is crucial to identify currently existing models (accessibility), review the quality of the models (transparency), assess how well they perform on external validation (generalizability), and investigate the potential benefit of recalibrating the validated models (updating). Subsequently, models showing adequate performance will be ready for implementation in clinical practice after clearly defined intended model use is described (interpretation), and the intended model use is substantiated by evidence regarding added value (impact assessment). In this thesis, multiple studies aiming to overcome the challenges are described using examples on breast and prostate cancer.

Prediction models in oncology

Cancer is one of the leading causes of death worldwide. Data provided by the World Health Organization (WHO) shows that in 2019, cancer ranks as one of the four most common causes of death worldwide in 135 of 183 countries.[31] Global cancer statistics show that a total of 19.3 million new cases of cancer were diagnosed in 2020 and 9.9 million deaths were caused by cancer. In males, 10.1 million new cases were diagnosed of which the five most commonly diagnosed cancers concern lung (14.3%), prostate (14.1%), colorectal (10.6%), stomach (7.1%), and liver (6.3%) cancers. A total of 5.5 million men died due to cancer, where the five most deadly cancers are lung (21.5%), liver (10.5%), colorectal (9.3%), stomach (9.1%), and prostate (6.8%) cancer. In females, the most commonly diagnosed cancers are breast (34.5%), colorectal (9.4%), lung (8.4%), cervix uteri (6.5%), and thyroid (4.9%) cancer.[32] This thesis describes various aspects of prediction models for clinical oncology with applications in breast cancer and prostate cancer care. The prognosis and treatment options vary between different types of cancers. The availability of a large amount of data is required to accurately develop and evaluate prediction models.[33] As breast and prostate cancers rank among the most commonly diagnosed malignancies among women and men, respectively, large datasets are available. Since both the benefits and side effects of different treatment options have a major impact on patients' health outcomes, and these outcomes are largely dependent on patient-, tumor- and treatment-related characteristics, prediction models are very suitable to support decision-making for patients diagnosed with breast and prostate cancer.

Breast cancer

In the Netherlands, over 17,000 patients are diagnosed with breast cancer, of which approximately 15,000 tumors are invasive, about 2,000 concern Ductal Carcinoma in Situ (DCIS), and a little over 100 breast cancers are found in men[34]. Overall, 88% of the patients with invasive cancer are still alive 5 years after diagnosis. Non-metastatic breast cancer is treated with curative intent which consists of surgical resection of the primary tumor and sampling or removal of lymph nodes in the axillary region. Additional postoperative radiation is also common, especially when breast-conserving surgery is performed. Systemic treatment can be performed preoperatively (neoadjuvant) and postoperatively (adjuvant). The decision to administer systemic therapy relies on patient- and tumor-related characteristics. Systemic treatment options include chemotherapy, hormonal therapy, and targeted therapy. The treatment of metastatic breast cancer mainly aims to prolong life and alleviate symptoms and target for an optimal quality of life.[35]

As described before, clinical prediction models can complement cancer staging to provide an even better understanding of patient prognosis and support decisions on specific treatment options. An example of such a model recommended in international guidelines[36] is PREDICT, in which the added value of adjuvant chemotherapy, endocrine therapy, and targeted therapy is expressed as 5-, 10- and 15-year survival benefit.[37] PREDICT has been used globally for breast cancer care for several years, but has its limitations.[38] For instance, radiotherapy benefit is not taken into account, and the model does not perform equally well in all patient subgroups.[39] It is unrealistic to expect that a single tool can be used for the full spectrum of breast cancer patients, however, it is currently uncertain which other tools are available, how well they perform, and what their impact is.

Prostate cancer

Prostate cancer is the most commonly diagnosed cancer in Dutch men with approximately 13,000 newly diagnosed prostate cancers each year. Overall, 88% of the patients diagnosed with prostate cancer are still alive after 5 years.[34] The treatment of prostate cancer heavily relies on the risk classification and (initial) tumor stage. Not all patients require immediate treatment and can be included for either active surveillance (AS) or watchful waiting (WW). AS is mainly provided for low risk patients, has a curative intent, and aims to minimize overtreatment whereas WW can be applied for patients at all stages, has palliative intent, and can be considered for patients with a life expectancy below 10 years. Patients not suitable for AS or WW who are eligible for treatment with curative intent should be considered for radical treatment such as radical prostatectomy, with extended pelvic lymph node dissection if the risk of lymph node involvement exceeds 5%. Other suitable options include radiotherapeutic treatments. Depending on disease stage, androgen deprivation therapy can be considered. Multimodal treatments are also proven to be effective treatment options. For patients with metastasized or recurrent disease, hormonal therapy, radiotherapy, and chemotherapy are suitable options.[40]

Clinical prediction models are frequently used to gain insight into the prognosis of the patient. For instance, in intermediate risk patients opting for radical prostatectomy, the prostate cancer guideline from the European Association of Urology (EAU) recommends to perform an extended pelvic lymph node dissection (ePLND) if the probability of LNI exceeds 5%. Several models have been developed to predict the LNI risk in patients eligible for radical prostatectomy.[41] Even though the models have the same intended use, their predictive performance may differ, and not all recommended models have been transparently described according to reporting standards.[42] To identify the most suitable model for a target population, external validation is essential.

Aims of the thesis

For both breast cancer and prostate cancer care, several clinical prediction models are recommended and are being used to support decision-making in clinical practice. Yet, the identified challenges for proper application of prediction models in clinical practice also apply to oncological care. This thesis therefore aims to:

- Identify potentially useful clinical prediction models supporting treatment decision-making in breast cancer and prostate cancer patients.
- Externally validate identified clinical prediction models for breast cancer and prostate cancer using Dutch registry data.
- Update models of potential value when applied to Dutch patient care.
- Assess the impact of applying models to support clinical decisions in oncological patients using a cost-effectiveness analysis.

Thesis outline

The thesis focuses on different aspects of prediction models for clinical oncology with chapters on breast cancer and prostate cancer. The thesis can be divided into two parts. The first part revolves around applications of clinical prediction models for breast cancer care and the second part focuses on prostate cancer care. Both parts consist of three chapters.

Part 1: Breast cancer

Multiple prediction models have been used for years to support decision-making in the treatment of breast cancer patients. In addition to these models, further decisions could be supported with the use of previously published clinical prediction models. In addition, there are some known limitations to the models that are currently being used for which useful alternative models may be available. With these premises in mind, we conducted a systematic literature review to identify potentially valuable models that have been published in recent years. This systematic review tackles challenges regarding accessibility and transparency of clinical prediction models and is described in **Chapter 2.** Subsequently, all identified models in the systematic review were considered for external validation using data from the Netherlands Cancer Registry to assess their generalizability. When sufficient data were available, and the models were not previously developed or validated on the NCR data, these models were externally validated. **Chapter 3** describes these external validation studies. In the systematic review and external validation study, we focused on clinical prediction models that can be used for a wide variety of decisions for patients previously diagnosed with breast cancer. One of the decisions involves the intensity of surveillance for patients who have been treated with curative intent. Patients with a low risk may require less frequent follow-up visits aimed at the detection of recurrent disease. A model predicting locoregional recurrence over 5-years called "INFLUENCE" has previously been developed for this purpose.[43] However, this model predicted solely locoregional recurrence as outcome variable and not the probability of contralateral breast cancer, which is also important in surveillance. Also, some relevant predictors such as the HER2 status were not incorporated in the model. **Chapter 4** describes an update of the INFLUENCE model to version 2.0, incorporating the desired improvements. The newly updated model was developed by comparing different modelling techniques, including a cox regression, parametric spline, and random survival forest approach.

Part 2: Prostate cancer

Important prognostic information for patients with prostate cancer regards the risk of metastasis, which are most commonly diagnosed in bone and lymph nodes. Patients with a high probability of LNI may require an ePLND. The probability of LNI can be estimated using clinical prediction models. The use of such models has been recommended in international guidelines on the management of prostate cancer for several years. However, different models are recommended to estimate the risk of LNI, and the recommended threshold to perform an ePLND varies between guidelines. As a first step to deal with some of these challenges, such as the generalizability of the models in the Dutch setting, **Chapter 5** evaluates the performance of a set of popular models predicting the risk of LNI in prostate cancer patients using a sample of Dutch prostate cancer patients who underwent radical prostatectomy and concomitant ePLND. **Chapter 6** describes an external validation study that aimed to assess the effect of using imaging methods to measure predictor information instead of conventional methods. For example, clinical tumor stage is assessed using digital rectal examination (DRE), but can also be assessed with multiparametric magnetic resonance imaging (mpMRI). This way, it is assessed whether existing models can also be generalized to patients who have been staged with mpMRI, or whether the models are required to be updated in this patient group. The two models with the best performance in the previous chapter (i.e. Memorial Sloan Kettering Cancer Center (MSKCC) and Briganti 2012 models), were compared head-to-head with predictor information measured either with DRE or with mpMRI. Finally, the impact of using the models with an adequate performance in Dutch patients has been assessed in a health economic evaluation described in **Chapter 7**. Here, a decision analytic model was constructed in which the impact of applying a set of reasonable thresholds was compared to a scenario in which no patient would undergo an ePLND.

References

1. Hoffmann, T. C. & Del Mar, C. Clinicians' expectations of the benefits and harms of treatments, screening, and tests: A systematic review. *JAMA Intern. Med.* **177**, 407–419 (2017).
2. Buchan, T. A. *et al.* Physician Judgement vs Model-Predicted Prognosis in Patients With Heart Failure. *Can. J. Cardiol.* **36**, 84–91 (2020).
3. Steyerberg, E. Clinical prediction models. (2019).
4. Hendriksen, J. M. T., Geersing, G. J., Moons, K. G. M. & de Groot, J. A. H. Diagnostic and prognostic prediction models. *J. Thromb. Haemost.* **11**, 129–141 (2013).
5. Damen, J. A. A. G. *et al.* Prediction models for cardiovascular disease risk in the general population: systematic review. *BMJ* **353**, i2416 (2016).
6. Senkus, E. *et al.* Primary breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann. Oncol.* **26**, v8–v30 (2015).
7. Borstkanker - Risicoprofilering - Richtlijn - Richtlijnendatabase. Available at: https://richtlijnendatabase.nl/richtlijn/borstkanker/risicoprofilering.html. (Accessed: 18th February 2022)
8. Hemingway, H. *et al.* Prognosis research strategy (PROGRESS) 1: A framework for researching clinical outcomes. *BMJ* **346**, (2013).
9. Riley, R. D. *et al.* Prognosis Research Strategy (PROGRESS) 2: Prognostic Factor Research. *PLoS Med.* **10**, (2013).
10. Steyerberg, E. W. *et al.* Prognosis Research Strategy (PROGRESS) 3: Prognostic Model Research. *PLoS Med.* **10**, (2013).
11. Collins, G. S., Reitsma, J. B., Altman, D. G. & Moons, K. G. M. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement. *Ann. Intern. Med.* **162**, 55 (2015).
12. Moons, K. G. M. M. *et al.* Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies: The CHARMS Checklist. *PLoS Med.* **11**, e1001744 (2014).
13. Wolff, R. F. *et al.* PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. **170**, (2019).
14. Riley, R., Windt, D. van der, Croft, P. & Moons, K. Prognosis research in healthcare: concepts, methods, and impact. (2019).
15. Steyerberg, E. W. *et al.* Poor performance of clinical prediction models: the harm of commonly applied methods. **98**, 133–143 (2018).
16. Collins, G. S. & Le Manach, Y. Nomograms need to be presented in full. *Cancer* **123**, 177–178 (2017).
17. Wynants, L. *et al.* Prediction models for diagnosis and prognosis of covid-19: Systematic review and critical appraisal. *BMJ* **369**, (2020).
18. Ravdin, P. M. *et al.* Computer Program to Assist in Making Decisions About Adjuvant Therapy for Women With Early Breast Cancer. *J. Clin. Oncol.* **19**, 980–991 (2001).

19. de Glas, N. A. *et al.* Validity of the online PREDICT tool in older patients with breast cancer: a population-based study. *Br. J. Cancer* **114**, 395–400 (2016).

20. Prostaatcarcinoom - Algemeen - Richtlijn - Richtlijnendatabase. Available at: https://richtlijnendatabase.nl/richtlijn/prostaatcarcinoom/algemeen.html. (Accessed: 18th February 2022)

21. Hansen, J. *et al.* External validation of the updated briganti nomogram to predict lymph node invasion in prostate cancer patients undergoing extended lymph node dissection. *Prostate* **73**, 211–218 (2013).

22. Vergouwe, Y. *et al.* A closed testing procedure to select an appropriate method for updating prediction models. *Stat. Med.* **36**, 4529–4539 (2017).

23. Moons, K. G. M. *et al.* Risk prediction models: II. External validation, model updating, and impact assessment. *heart.bmj.com* doi:10.1136/heartjnl-2011-301247

24. van Giessen, A., de Wit, G. A., Moons, K. G. M., Dorresteijn, J. A. N. & Koffijberg, H. An alternative approach identified optimal risk thresholds for treatment indication: an illustration in coronary heart disease. *J. Clin. Epidemiol.* **94**, 122–131 (2018).

25. van Giessen, A. *et al.* Systematic Review of Health Economic Impact Evaluations of Risk Prediction Models: Stop Developing, Start Evaluating. *Value Heal.* **20**, 718–726 (2017).

26. Medical Devices Regulation. (2017). Available at: https://eur-lex.europa.eu/eli/reg/2017/745/2017-05-05. (Accessed: 8th February 2021)

27. Kappen, T. H. *et al.* Barriers and facilitators perceived by physicians when using prediction models in practice. *J. Clin. Epidemiol.* **70**, 136–145 (2016).

28. Lagerweij, G. R., Moons, K. G. M., De Wit, G. A. & Koffijberg, H. Interpretation of CVD risk predictions in clinical practice: Mission impossible? *PLoS One* **14**, (2019).

29. Yang, C., Kors, J., Ioannou, S. & … L. J.-J. of the. Trends in the conduct and reporting of clinical prediction model development and validation: a systematic review. *academic.oup.com*

30. Shah, N. D., Steyerberg, E. W. & Kent, D. M. Big data and predictive analytics: Recalibrating expectations. *JAMA - J. Am. Med. Assoc.* **320**, 27–28 (2018).

31. Global health estimates: Leading causes of death. Available at: https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates/ghe-leading-causes-of-death. (Accessed: 18th February 2022)

32. Sung, H. *et al.* Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA. Cancer J. Clin.* **71**, 209–249 (2021).

33. Van Der Ploeg, T., Austin, P. C. & Steyerberg, E. W. Modern modelling techniques are data hungry: A simulation study for predicting dichotomous endpoints. *BMC Med. Res. Methodol.* **14**, 1–13 (2014).

34. NKR Cijfers. Available at: https://iknl.nl/nkr-cijfers?fs%7Cepidemiologie_id=526&fs%7Ctumor_id=292%2C295%2C297&fs%7Cregio_id=550&fs%7Cperiode_id=564%2C565%2C566%2C567%2C568%2C569%2C570%2C571%2C572%2C573%2C574%2C575%2C576%2C577%2C578%2C579%2C580%2C581%2C582%2C583%2C584%2C585%2C586%2C587%-2C588%2C589%2C590%2C591%2C592%2C593%2C563%2C562%2C561&fs%7Cgeslacht_

id=644&fs%7Cleeftijdsgroep_id=677&fs%7Cjaren_na_diagnose_id=687&fs%7Ceenheid_id=703&cs%7Ctype=line&cs%7CxAxis=periode_id&cs%7Cseries=tumor_id&ts%7CrowDimensions=periode_id&ts%7CcolumnDimensions=tumor_id&lang%7Clanguage=nl. (Accessed: 9th February 2022)

35. Waks, A. G. & Winer, E. P. Breast Cancer Treatment: A Review. *JAMA - J. Am. Med. Assoc.* **321**, 288–300 (2019).
36. Cardoso, F. *et al.* Early breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann. Oncol.* **30**, 1194–1220 (2019).
37. Wishart, G. C. *et al.* PREDICT: a new UK prognostic model that predicts survival following surgery for invasive breast cancer. *Breast Cancer Res.* **12**, R1 (2010).
38. Candido dos Reis, F. J. *et al.* An updated PREDICT breast cancer prognostication and treatment benefit prediction model with independent validation. *Breast Cancer Res.* **19**, 58 (2017).
39. van Maaren, M. C. C. *et al.* Validation of the online prediction tool PREDICT v. 2.0 in the Dutch breast cancer population. *Eur. J. Cancer* **86**, 364–372 (2017).
40. Cornford, P. *et al.* EAU-EANM-ESTRO-ESUR-SIOG Guidelines on Prostate Cancer. Part II—2020 Update: Treatment of Relapsing and Metastatic Prostate Cancer[Formula presented]. *Eur. Urol.* **79**, 263–282 (2021).
41. Mottet, N. *et al.* EAU-ESTRO-SIOG Guidelines on Prostate Cancer. Part 1: Screening, Diagnosis, and Local Treatment with Curative Intent. *Eur. Urol.* **71**, 618–629 (2017).
42. Briganti, A. *et al.* Updated Nomogram Predicting Lymph Node Invasion in Patients with Prostate Cancer Undergoing Extended Pelvic Lymph Node Dissection: The Essential Importance of Percentage of Positive Cores. *Eur. Urol.* **61**, 480–487 (2012).
43. Witteveen, A. *et al.* Personalisation of breast cancer follow-up: a time-dependent prognostic nomogram for the estimation of annual risk of locoregional recurrence in early breast cancer patients. *Breast Cancer Res. Treat.* **152**, 627–636 (2015).

PART I

Breast Cancer

Chapter 2

Clinical prediction models to support treatment decisions in breast cancer patients: a systematic review

Tom A. Hueting, Marissa C. van Maaren, Mathijs P. Hendriks,
Hendrik Koffijberg, Sabine Siesling

Submitted for peer-review

The supplementary data to this chapter are available upon request

Abstract

Background: In breast cancer the number of existing prediction models that may support treatment decision-making, the necessary predictors, predicted outcomes, modeling methods, quality, and validity are currently unknown.

Methods: Literature was systematically searched to identify studies reporting on development of prediction models aiming to support breast cancer treatment decision-making, published between January 2010 and December 2020. Data extraction was performed according to the Checklist for critical Appraisal and data extraction for systematic Reviews of prediction Modelling Studies (CHARMS). Quality and potential risk of bias was assessed using the Prediction model Risk Of Bias (ROB) Assessment Tool (PROBAST).

Results: After screening 20460 studies, 534 studies were included, reporting on 922 models. Most common predictors were age (n=426 46%), tumor size (n=373, 40%), lymph node involvement (n=337, 37%). In the 922 identified models the following outcomes were predicted; mortality (n=417 45%), recurrence (n=217, 24%), lymph node involvement (n=141, 15%), adverse events (n=58, 6%), treatment response (n=56, 6%), or other outcomes (n=33, 4%). Much models (n=285, 31%) lacked a complete description of the final model and could not be applied to new patients. Most models (n=878, 95%) were considered to contain high ROB.

Conclusion: A substantial overlap in predictor variables and outcomes between the models was observed. A large number of models were not reported according to established reporting guidelines or showed methodological flaws during the development and/or validation of the model. Further development of prediction models with thorough quality and validity assessment is an essential first step for future clinical application.

Introduction

Breast cancer is the most commonly diagnosed cancer in women worldwide. Disease severity and treatment options for breast cancer depend on various factors such as subtype, tumor stage, personal context, and genetic characteristics.[1] The heterogeneity of breast cancer challenges clinicians to optimize treatment for each individual patient. Pros and cons of different treatment options (i.e. improvement of prognosis versus (late) adverse events) should be considered before treatment initiation on an individual patient level. Clinical prediction models can support clinical decision-making by estimating individual predictions on certain outcomes using combinations of different relevant patient and disease characteristics.

Multiple prediction models have been available to guide treatment decision-making for breast cancer patients in the past years. For example, Predict[2] is a prediction model that has been available as an online model to support decision-making on adjuvant treatment strategies. The use of Predict or other similar tools such as Cancermath[3] or the Nottingham Prognostic Index[4] have been recommended in international guidelines[5]. Yet there are more treatment decisions for breast cancer patients that could be well supported by prediction models. There may be potentially valuable models already available that are not currently used because their quality and reliability is unclear.

Before prediction models may be implemented in clinical practice, multiple steps should be performed. These methods include the steps for development, internal validation, external validation, updating, and impact assessment of prediction models.[6–9] Ideally, the development and validation of a model should be described according to the guideline for transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD).[10]

However, with regard to the application of prediction models aimed at supporting decision-making in breast cancer care, it is currently unknown how many different models have been developed, which outcomes can be (accurately) predicted, and with which variables the outcomes can be predicted. We therefore aimed to systematically review prediction models that may be used to support treatment decision-making in breast cancer patients and to assess the quality of studies reporting on the development and (internal) validation of prediction models.

Methods

The systematic review study protocol has been registered in PROSPERO (registration number: CRD42020134826). The PRISMA checklist for transparent reporting of systematic reviews and meta-analysis was followed for reporting the results. (Supplementary data S2)[11]

Search strategy

Medline and Embase were searched for studies published between January 1st 2010 and December 31st 2020. The search strategy was constructed using validated search filters to find prognostic prediction studies (supplementary data S1).[12] In addition, the references listed in the studies selected for full-text assessment were screened for potentially useful studies. One reviewer (TAH) screened the title of all identified studies and screened the abstract of all studies that could not immediately be excluded based on the title. To validate this first selection, another reviewer (MvM) also screened the title and abstract of a random sample of 1600 of the studies for full-text inclusion. Discrepancies were resolved after discussions between the two reviewers.

Study selection

Studies were included if they reported the development of a prediction model intended to be used for treatment decision-making in patients (both men and women) who have been diagnosed with breast cancer. Such outcomes include; survival (overall or disease-specific), recurrence ((loco)-regional or second primary breast cancer), metastasis (including contralateral lymph nodes), adverse events, quality of life, and treatment response. Included studies must be aimed at providing predictions for breast cancer patients using a combination of two or more predictor variables, possibly demonstrated by providing a calculation method (i.e. logistic regression, neural network). Studies can report the development of multiple prediction models. We defined separate models when either the predictor-outcome association was different, or when the predictor-outcome association was the same, but a different baseline hazard or intercept was reported. Two types of prediction models were distinguished, diagnostic and prognostic models. Diagnostic models aim to estimate the likelihood for currently having the outcome, whereas prognostic models aim to estimate the probability of the outcome at a specified future time.

Data extraction

Data extraction was performed using the checklist for critical appraisal and data extraction for systematic reviews of prediction modelling studies (CHARMS).[13]

The definition of the same predictors sometimes varied between different studies. For instance, HER2-status could be defined as negative or positive, or could be incorporated in a subtype variable including both HER2 and hormonal receptor status. We decided to report the definition of the predictor as reported in the study.

To assess whether the sample size was likely sufficient to develop the model, the events per variable (EPV) were estimated for each model. The EPV is a traditional criterion that is used to estimate how many predictors could be included in the multivariable analysis. Even though the EPV has its limitations, it provides a rough indication of whether the sample size was sufficient.[14] The sample size is likely to be insufficient with an EPV <10. An EPV between 10 and 20 could be sufficient, but is still fairly low, and an EPV >20 is likely to be sufficient.

Risk of Bias

To assess whether reviewed studies are at low or high risk of bias (ROB), the prediction model risk of bias assessment tool (PROBAST) was used.[15] The PROBAST tool includes 20 signaling questions in four domains; participants, predictors, outcome, and analysis. In addition, an overall conclusion regarding low, or high ROB for the reviewed prediction models was determined. The participants domain covered the ROB related to study data and the methods used to enroll study participants. The predictors domain covered ROB caused by the measurement and definition of predictors. The outcome domain assessed the ROB caused by the estimation and definition of the outcome. The analysis domain covered the ROB related to the statistical methods used to develop and validate the model (Box 1). The ROB assessment was performed for all prediction models by one reviewer (TAH), and for a subset of 20 models also by a second reviewer (MvM), to identify potential discrepancies. Based on the similarities in ROB assessment between the two reviewers, the subset of 20 models assessed by the second reviewer was deemed sufficiently large to ensure high quality ROB assessment.

Box 1: Explanation of the prediction model risk of bias assessment tool (PROBAST).

The Prediction model Risk Of Bias ASsessment Tool (PROBAST).
The PROBAST was developed to critically appraise the development and validation of prediction models. Even though the PROBAST can be used to assess both risk of bias (ROB) and concerns regarding applicability, it was mainly used in this review to assess the ROB. The PROBAST aims to judge the risk of bias in four domains. Each domain has a set of signaling questions that needs to be answered with either "(Probably) Yes", "(Probably) No", or "No information". The ROB is subsequently judged as low, high, or unclear. The following domains are identified by the PROBAST:

1. **Participants**

The first domain has two signaling questions regarding the appropriateness of used data sources and the applied inclusion and exclusion criteria.

2. **Predictors**

This domain includes three signaling questions regarding uniformly described predictors, predictor assessment, and availability of predictors at the time the model is intended to be used.

3. **Outcome**

The Outcome domain has six signaling questions regarding its determination, definition, and time interval between predictor measurement and outcome occurrence.

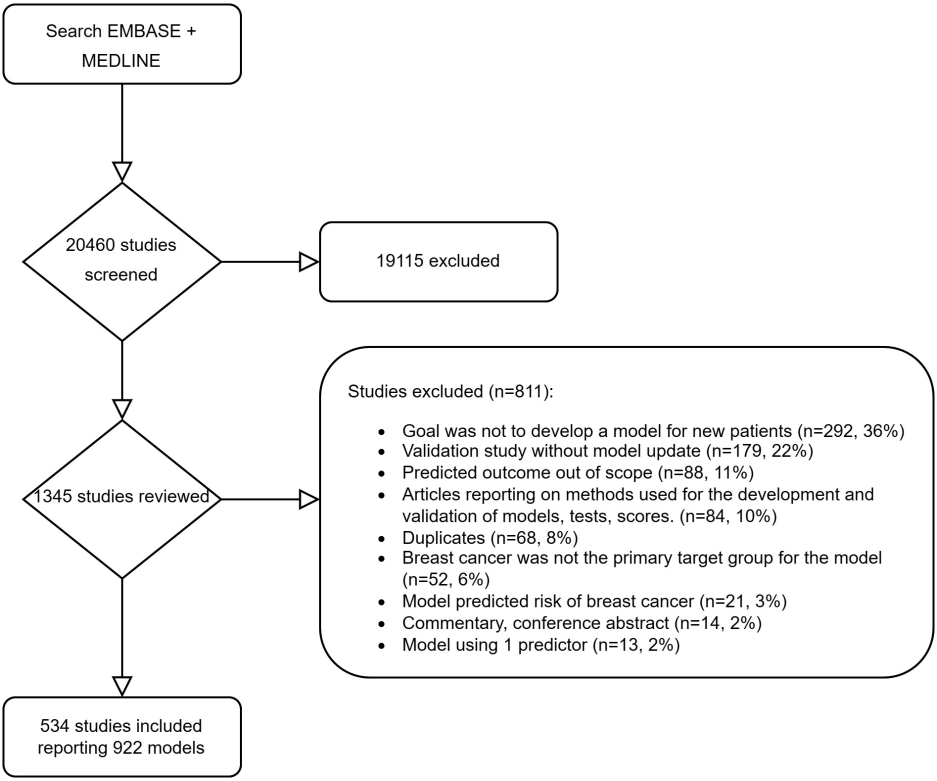
4. **Analysis**

The analysis domain has nine signaling questions regarding the statistical methods used to develop and validate the model. Topics include the sample size, handling of continuous predictors, inclusion of patients in the analysis, dealing with missing data, avoidance of univariable analysis, dealing with complexities in the data, appropriateness of performance measures, dealing with overfitting, underfitting, and optimism in the model, and whether the weights in the final model correspond with the results from the analysis
All signaling questions should be answered with "(Probably) Yes" for a low ROB rating. At least one "(Probably) No" results in a high ROB, and at least one "No information" (and no "(Probably) No" results in an unclear ROB rating.

Results

The search strategy identified 20,460 studies, of which the titles were screened. The abstract was screened for studies that could not be excluded based on the title alone. Subsequently, 1345 studies were selected for full text screening. Finally, a total of 534 studies were included, reporting on 922 models. The inclusion and exclusion criteria of the different studies and the reasons for excluding studies are shown in figure 1.

Figure 1. Flowchart of study selection.


Search EMBASE + MEDLINE
20460 studies screened
19115 excluded
1345 studies reviewed
Studies excluded (n=811):
• Goal was not to develop a model for new patients (n=292, 36%)
• Validation study without model update (n=179, 22%)
• Predicted outcome out of scope (n=88, 11%)
• Articles reporting on methods used for the development and validation of models, tests, scores. (n=84, 10%)
• Duplicates (n=68, 8%)
• Breast cancer was not the primary target group for the model (n=52, 6%)
• Model predicted risk of breast cancer (n=21, 3%)
• Commentary, conference abstract (n=14, 2%)
• Model using 1 predictor (n=13, 2%)
534 studies included reporting 922 models

Predictors

A total of 228 different model predictors were identified in the included 922 models. A total of 14 predictors were used in more than 100 different models: age (n=426, 48%), tumor size (n=373, 40%), lymph node involvement (n=337, 37%), tumor grade (n=297, 32%), ER-status (n=187, 20%), HER2-status (n=158, 17%), surgery (n=149, 16%), radiotherapy (n=141, 15%), chemotherapy (n=141, 15%), subtype (n=132, 14%), PR-status (n=130, 14%), metastasis (n=123, 13%), and genetic risk score (n=115, 12%). In the supplementary materials (S4, sheet "predictors"), an overview of all predictors per outcome is displayed. The five most common predictors per outcome are shown in table 1.

Table 1. Overview of models by outcome

Outcome	Specified outcome	Models (n = 922)	Average C-index	Most common predictors (n (X%))
Mortality	Overall survival	316	0.740	Age (184 (58%)), Tumor size (146 (46%)), Lymph node involvement (126 (40%)), Tumor grade (126 (40%)), Metastasis (82 (26%))
	Disease specific survival	94	0.763	Tumor grade (64 (68%)), Tumor size (62 (66%)), Age (54 (57%), Lymph node involvement (50 (53%)), ER status (39 (41%))
	Other cause specific survival	7	0.746	Age (7 (100%)), Tumor size (7 (100%), Surgery (4 (57%)), Chemotherapy (3 (43%)), Marital status (3 (43%))
Recurrence	Recurrence (free survival)	132	0.769	Lymph node involvement (54 (41%)), Age (42 (32%)), Tumor size (38 (29%)), Grade (35 (27%)), Genetic risk score (27 (21%))
	Locoregional recurrence	33	0.728	Age (27 (82%)), Tumor grade (24 (73%)), Tumor size (19 (58%)), Lymph node involvement (18 (55%)), Hormonal therapy (17 (52%))
	Metastasis	45	0.760	Lymph node involvement (30 (67%)), Tumor size (21 (47%)), Age (16 (36%)), Genetic risk score (13 (29%)), Subtype (7 (16%)), Lymphovascular invasion (7 (16%))
	Contralateral recurrence	7	0.589	Age (6 (86%)), Histology (5 (71%)), Radiotherapy (5 (71%)), Tumor size (4 (57%)), ER status (3 (43%)), Grade (3 (43%)), Surgery (3 (43%)), Hormonal therapy (3 (43%)), Family History (3 (43%))
Lymph nodes	Lymph node involvement	83	0.791	Tumor Size (31 (37%)), Age (27 (33%)), Lymph node status (21 (25%)), Grade (21 (25%)), Lymphovascular invasion (19 (23%))
	Sentinel lymph node involvement	13	0.763	Age (5 (38%)), Lymphovascular invasion (5 (38%)), Tumor size (3 (23%)), ER status (3 (23%)), PR status (3 (23%)), HER2 status (3 (23%)), Tumor location (3 (23%)), Multifocality (3 (23%))
	Non-sentinel lymph node involvement	45	0.758	Lymph node involvement (21 (47%)), Lymphovascular invasion (20 (44%)), Diameter largest lymph node (17 (38%)), Tumor size (13 (29%)), Lymph node ratio (11 (24%))
Treatment response	Pathologic complete response	56	0.812	ER status (21 (38%)), HER2 status (20 (36%)), Tumor size (16 (29%)), KI67 (15 (27%)), Age (9 (16%)), PR status (9 (16%)), Grade (9 (16%))
Adverse events	Lymphedema	15	0.775	BMI (10 (67%)), Radiotherapy (10 (67%)), chemotherapy (10 (67%)), Surgery (6 (40%)), Age (5 (33%)), Lymph nodes dissected (5 (33%)), Lymph node surgery (5 (33%))
	Cardiovascular complications	9	0.780	Age (8 (89%)), Chemotherapy (3 (33%)), BMI (2 (22%)), Tumor size (1 (11%)), Surgery (1 (11%))
	Pain	7	0.703	Preoperative pain (5 (71%)), BMI (4 (57%)), Age (3 (43%)), Lymph nodes dissected (3 (43%)), Postoperative pain (2 (29%))
	Other adverse events*	27	0.711	Age (10 (36%)), BMI (9 (32%)), Smoking status (8 (29%)), Comorbidities (8 (29%)), Radiotherapy (7 (25%))

Table 1. Continued

Outcome	Specified outcome	Models (n = 922)	Average C-index	Most common predictors (n (X%))
Other outcomes	Menopausal status	8	0.814	Age (8 (100%)), BMI (4 (50%)), Chemotherapy (3 (38%)), Hormonal therapy (3 (38%)), FSH (3 (38%))
	Quality of life	8	Not applicable	Age (6 (75%)), Chemotherapy (5 (63%)), Hormonal therapy (5 (63%)), Radiation therapy (5 (63%)), Stage (5 (63%)), Surgery (5 (63%)), Complications (5 (63%)), menopausal status (5 (63%)), Ambulatory (5 (63%)), Charlson Deyo score (5 (63%)), Education (5 (63%)), Postoperative length of stay (5 (63%))
	Surgical margin	6	0.726	Tumor grade (2 (33%)), lymph node involvement (2 (33%)), ER-status (2 (33%)), Tumor size (2 (33%)), Her2-status (2 (33%)), PR-status (2 (33%)), Metastasis (2 (33%)), Histology (2 (33%)), Multifocality (2 (33%))
	Treatment as the outcome	7	0.705	Age (4 (57%)), Tumor size (3 (43%)), ER status (2 (29%)), Race (2 (29%)), Radiotherapy (2 (29%))
	Good cosmetic outcome	2	0.790	Lymphovascular invasion (2), Multifocality (1 (50%)), total tumor load (1 (50%)), Tumor size (1 (50%)), ER status (1 (50%)), Lymph node involvement (1 (50%)), Grade (1 (50%))
	Nipple-Areola complex invasion	2	0.879	Distance from nipple (2 (100%)), Lymph node involvement (1 (50%)), Tumor size (1 (50%)), Location (1 (50%)), Imaging outcome (1 (50%)), Multicentricity (1 (50%))

The details of all included models were added as an additional spreadsheet in supplementary material S4.

*Other adverse events include pneumonitis, necrosis, seroma, infection, exposure, explantation, overall complication, falls, symptomatic skeletal events, fibrosis, rash, fatigue, neutropenia, and cognitive impairment

Outcome

The included studies described models that were developed to predict the following outcomes; mortality (n=417, 45%), recurrence(-free survival) (n=217, 24%), lymph node involvement (n=141, 15%), adverse events (n=58, 6%), treatment response (n=56, 6%), and other outcomes (n=33, 4%) such as menopausal status, quality of life, surgical margin, receiving treatment, cosmetic outcome, nipple-areola complex invasion. The number of models per outcome is displayed in Table 1.

The majority of the models predicted similar outcomes, although the models often differed in the specific definition of the outcome (i.e. lymph node involvement could include both sentinel and non-sentinel lymph node involvement), or the models used different in- and exclusion criteria to develop the model. Out of the 922 models, 693 (75%) were prognostic, and 229 (25%) were diagnostic models. The details of all included models were added as an additional spreadsheet in supplementary material S4.

Modelling methods

Relevant findings related to methods used to develop and validate the prediction models were rated (Table 2). To develop diagnostic models, logistic regression was mostly used (n = 197 (86%)). For prognostic models, Cox regression was used in 510 (74%) of the models. The majority of models were developed using data from patients in Asian (n=319, 35%), North-American (n=262, 28%), or European (n=183, 20%), countries. A total of 429 (47%) models were developed with patient data from multiple centers, and 386 (42%) models were developed with data from a single institution.

The median number of participants used to develop a model was 699 (IQR 272– 2970), with a median number of events of 130 (IQR 58–416). Regarding the sample sizes used to develop the models, the EPV could not be determined for 269 (29%) models, 162 (18%) models were developed with an EPV <10, and 159 (17%) models with an EPV between 10 and 20. The remaining 332 (36%) models were developed with an EPV ≥20.

For 525 (57%) of the developed models, it was unclear how the developers dealt with missing data in the derivation dataset, 297 (32%) of the models were developed using complete-case analysis, and only 80 (9%) of the models were developed using an imputation (i.e. multiple or single) method to deal with missing data as recommended by the TRIPOD statement.[10] A total of 285 (31%) models were not reported with sufficient information to apply the model in practice. This was mostly caused by the absence of either the predictor coefficients (n=119, 13%), the baseline hazard, (n=96, 10%), or the intercept (n=51, 6%).

Table 2. Summary of extracted items for all included models.

		Diagnostic models (N = 229)	**Prognostic models (N = 693)**	**Total included models (N = 922)**
Modelling method	Cox regression	0 (0%)	510 (74%)	510 (55%)
	Fine and Gray model	0 (0%)	25 (4%)	25 (3%)
	Logistic regression	197 (86%)	93 (13%)	290 (31%)
	Linear regression	2 (1%)	9 (1%)	11 (1%)
	Machine learning	25 (11%)	41 (6%)	66 (7%)
	Other*	4 (2%)	13 (2%)	17 (2%)
	Unclear	1 (0.4%)	2 (0.3%)	3 (0.3%)
Location of participants used to develop the model	Asian	121 (53%)	199 (29%)	319 (35%)
	North-American	35 (15%)	227 (33%)	262 (28%)
	European	61 (27%)	121 (17%)	183 (20%)
	South-American	1 (0.4%)	4 (1%)	5 (1%)
	African	1 (0.4%)	1 (0.1%)	2 (0.2%)
	Oceania	0 (0%)	3 (0.4%)	3 (0.3%)
	Multiple continents	4 (2%)	16 (2%)	20 (2%)
	Unknown	6 (3%)	122 (18%)	128 (14%)
Database used to develop the model	Single center	149 (65%)	237 (34%)	386 (42%)
	Multicenter	52 (23%)	105 (15%)	157 (17%)
	Registry	23 (10%)	249 (36%)	272 (30%)
	Unclear	5 (2%)	102 (15%)	107 (12%)
Participants in derivation cohort (n)	< 100	24 (10%)	16 (2%)	40 (4%)
	100 – 200	50 (22%)	63 (9%)	113 (12%)
	200 – 500	67 (29%)	143 (21%)	210 (23%)
	500 – 1000	39 (17%)	130 (19%)	169 (18%)
	1000 – 10000	40 (17%)	196 (28%)	236 (26%)
	≥ 10000	9 (4%)	119 (17%)	128 (14%)
	Unclear	0 (0%)	26 (4%)	26 (3%)
Events per variable	< 10	55 (24%)	107 (15%)	162 (18%)
	10 – 20	45 (20%)	114 (16%)	159 (17%)
	20 – 50	62 (27%)	84 (12%)	146 (16%)
	≥ 50	40 (17%)	146 (21%)	186 (20%)
	Unclear	27 (12%)	242 (35%)	269 (29%)
Dealing with missing data	Excluded patients with missing data	61 (27%)	236 (34%)	297 (32%)
	Imputation	9 (4%)	71 (10%)	80 (9%)
	Unknown modelled as covariate	4 (2%)	8 (1%)	12 (1%)
	No Missing data	1 (0%)	7 (1%)	8 (1%)
	Unclear	154 (67%)	371 (54%)	525 (57%)
Model performance (discrimination)	Quantified	215 (94%)	599 (86%)	814 (88%)
	Not quantified	14 (6%)	94 (14%)	108 (12%)
Model performance (calibration)	Plot (observed vs. expected)	89 (39%)	419 (60%)	508 (55%)
	Hosmer-Lemeshow goodness of fit test	11 (5%)	22 (3%)	33 (4%)
	Other**	3 (1%)	44 (6%)	47 (5%)
	Unclear	126 (55%)	208 (30%)	334 (36%)

Table 2. Continued.

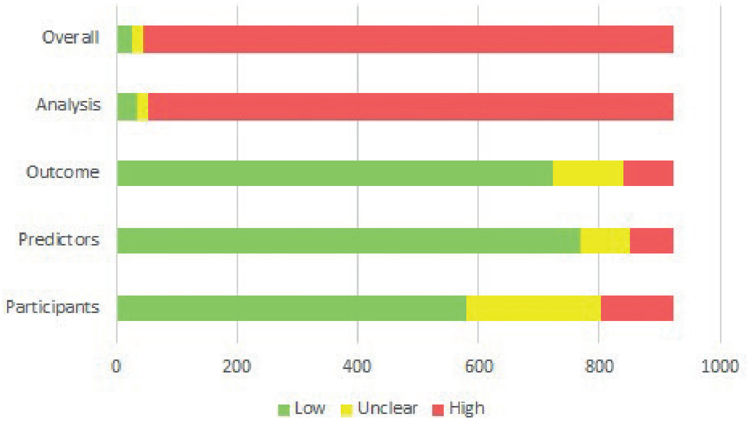
		Diagnostic models (N = 229)	Prognostic models (N = 693)	Total included models (N = 922)
Validation method	Apparent	41 (18%)	73 (11%)	114 (12%)
	x-fold cross validation	20 (9%)	27 (4%)	47 (5%)
	Bootstrap	30 (13%)	108 (16%)	138 (15%)
	External validation cohort	29 (13%)	147 (21%)	176 (19%)
	Temporal validation cohort	12 (5%)	29 (4%)	41 (4%)
	Split sample	61 (27%)	171 (25%)	232 (25%)
	Combination of multiple methods	22 (10%)	85 (12%)	107 (12%)
	Unclear	14 (6%)	53 (8%)	67 (7%)
Model is reproducible	No	79 (34%)	206 (30%)	285 (31%)
	Yes	150 (66%)	487 (70%)	637 (69%)

The details of all included models were added as an additional spreadsheet in supplementary material S4. Percentages added together may not be equal to 100% due to rounding
* Other modelling methods include classification and regression trees (CART), parametric survival regression, principal component analysis, and structural equation modelling.
** Other calibration methods include the use of a table, description of observed vs expected, or a bar chart.

Risk of bias (ROB)

The models were rated as either low (n=27, 3%), high (n=878, 95%) or unclear (n=17, 2%) ROB. The majority of the models were considered at high ROB, mainly due to the assessment of the domain 'analysis' in the PROBAST tool. Figure 2 shows the general assessment of the ROB and supplementary table 1 displays the ROB assessment per model. Discrepancies in ROB assessments performed by the two reviewers were sometimes found between answers to signaling questions, but the assessment for each PROBAST domain was similar for all studies that were assessed by both reviewers. The studies with a low ROB were added in supplementary table S3

Figure 2. Risk of bias by PROBAST domains.



A rating of high was given for a subdomain when at least one signaling question was answered with a "No". A low risk of bias rating was given if all signaling questions were answered with "Yes".
An unclear risk of bias is assigned if at least one signaling question could not be answered and if the remaining signaling questions were answered with "yes".

Reasons for defining PROBAST domains to be unclear or high risk were often similar for different models. Figure 3 represents the risk of bias stratified per outcome. An unclear or high ROB in the outcome domain occurred more often in models predicting recurrence or adverse events compared to the other models. Only 50% and 47% of the models predicting recurrence and adverse events were deemed at low ROB in the outcome domain where this percentage was 93%, 87%, 96%, and 73% for mortality, LNI, response, and other outcomes, respectively. The reason for this difference is mostly due to differences in methods to define the outcome (i.e. assessment via telephone follow-up) or to lack of description on the intensity and method of the follow-up. Notably, the ROB for the 'analysis' domain was defined as high for the majority of the models (95%). Common reasons for the high ROB concerned inadequate dealing with missing data, using univariable analysis to select candidate predictors, or not dealing with overfitting or optimism in the model. Out of all the models that were high ROB in the analysis domain, 82% showed concerns on two or more of the signaling questions, whereas the PROBAST tool advises to assign high ROB already if one of the signaling questions is not appropriately addressed.

Discussion

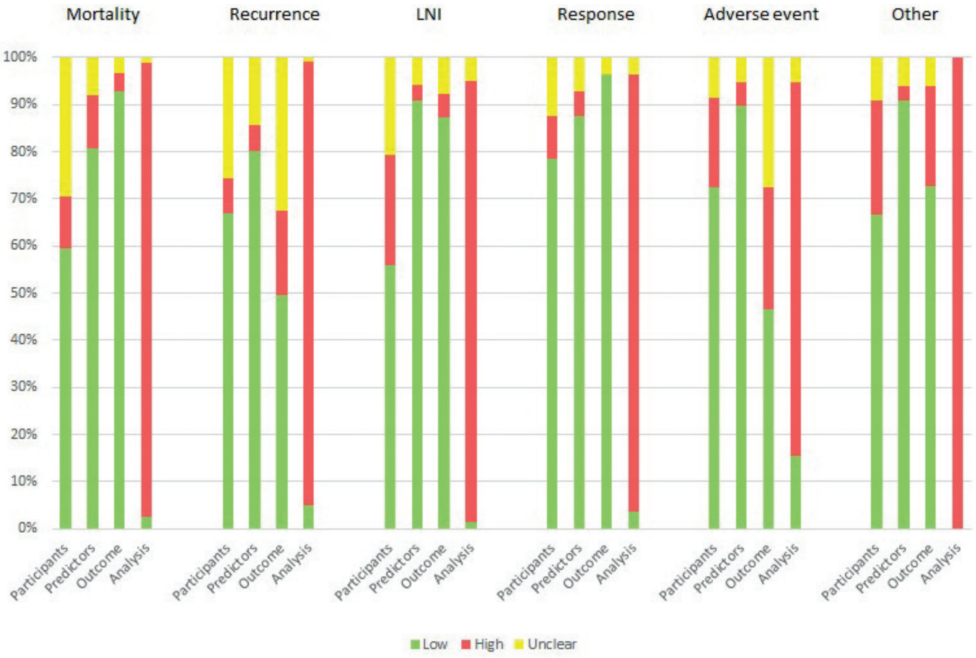
This systematic review identified a total of 534 studies published between 2010-2020, reporting the development of 922 different models. The patient's age, tumor size, and lymph node involvement were the most common predictors and were used in more than a third of the models. Models were categorized as either predicting a prognostic (n=693, 75%) or a diagnostic (n=229, 25%) outcome, The quality of the identified models was poor as only 35 models (4%) were developed with appropriate statistical methods according to the PROBAST tool, and only 27 models (3%) were deemed at low ROB overall.

Predictors used in the identified models were overlapping to a large extent. This makes sense as these predictors were proven to provide significant prognostic information regarding relevant health outcomes. ER status is an example of a predictor that was often used to predict different outcomes. ER status was mostly entered in the model as a dichotomous variable (i.e. negative, or positive). Even though the registration of such predictors as dichotomous variables is commonly applied and accepted, the dichotomization of continuous variables is regarded as bad practice.[16] As multiple predictors are commonly accepted as dichotomous variables in clinical practice, the use of these variables was no reason for a high ROB rating as suggested in the PROBAST tool. Accepted dichotomous variables were ER status, PR status, HER2-status, KI67 status, and tumor stage. Even though the use of the EPV criterion is regarded as sub-optimal,[14] the EPV could not be determined for 269 models (29%) mainly due to the lack of reporting on the number of events.

This review identified a disproportionate number of models predicting the same outcome. The majority of identified prediction models in breast cancer were developed using suboptimal or inappropriate methodology. This result aligns with previous findings in other disease areas. The review by Damen et al. assessed 363 prediction models for cardiovascular disease and concluded that most models were reported inadequately according to the CHARMS checklist.[17] A more recent systematic review of prediction models for diagnosis and prognosis of covid-19 found all models to be at high risk of bias.[18] A systematic review by Phung et al on prognostic models for breast cancer focused more on their performance.[19] Even though only 58 models were identified, the authors concluded that the performance for most models was suboptimal in independent cohorts, which could have been expected as the methods for development of most models were also suboptimal. Clearly, the lack of adequate reporting and the methodological shortcomings for the development of prediction models is not specific for breast cancer, but seems to be a common problem within clinical research. Adherence to reporting guidelines such as the TRIPOD is necessary to improve the quality of developed prediction models. A limitation of this review concerns the fact that a subset of 1600 of the identified studies in the search strategy was assessed by a second reviewer. From this subset of 1600 studies only three studies were additionally included for full-text analysis. For this reason, we do not expect that the review protocol led to exclusion of relevant studies. In addition, referenced models in studies reporting on the validation of prediction models, as well as references in previous systematic reviews on breast cancer prediction models were assessed to minimize the risk of missing relevant studies. The same limitation is applicable to the ROB assessment. In our study, a second reviewer assessed 20 prediction models using the PROBAST, and no model would have been rated differently even though some differences were found in the signaling questions.

One of the most important findings in this review concerns the high proportion of models regarded at high ROB. The majority of the models were at high ROB due to the 'analysis' domain, in which the ROB due to statistical methods is assessed. Still, a high ROB rating does not necessarily mean that the model has no or limited clinical value and a low ROB rating does not automatically constitute a valuable model. For instance, studies reporting on the update of the Predict model were rated as low ROB (Supplementary data S3), but an external validation study demonstrated suboptimal performance of the model in different patient groups.[20,21] Besides, each model only predicts a single outcome, whereas clinical decision-making also requires individual estimates of other relevant outcomes such as adverse events. Before clinical use of a model can be justified, different steps have to be taken for the development, internal and external validation, update, and impact assessment. Even then, the model needs to be trusted and understood by clinicians or adopted in clinical guidelines, and both the preferences and context of the patient should be taken into account before widespread implementation of a model is accepted in daily clinical practice. Nevertheless, the development of a valuable model starts with a good performance on internal validation, carried out with the appropriate statistical methods. Further (external) validation of the models may ultimately conclude whether the models may be generalized to different patient cohorts and perhaps different health care settings.[22] Even when models were proven to perform sufficiently well in external populations, additional (clinical) evaluations should be performed to assess the clinical and health impact of a prediction model.[23] Besides, with changing regulations in the European Union, the majority of prediction models in the current review are very likely to require certification as a medical device according to the Medical Devices Regulation before clinical use is enabled.[24] The fact that such a low number of models (n=27, 3%) were considered to be reported adequately based upon the model development stage underpin the need for improved reporting of prediction model development, perhaps now more than ever.

Figure 3. Risk of Bias assessment per outcome.



LNI = Lymph node involvement. A rating of high was given for a subdomain when at least one signaling question was answered with a "No". A low risk of bias rating was given if all signaling questions were answered with "Yes". An unclear risk of bias is assigned if at least one signaling question could not be answered and if the remaining signaling questions were answered with "yes".

Conclusion

Many prediction models have been published during the past decade to predict outcomes related to breast cancer treatment. Nearly all published prediction models identified were deemed as high ROB. Mainly due to a lack of adequate reporting, many prediction models could not be implemented in clinical practice as the studies did not provide sufficient data for external validation studies or an impact assessment. Future studies should focus on improving currently available models, either by identifying specific subgroups for which no model is applicable, or by performing the required steps before clinical adoption can be justified (i.e. external validation and impact assessment) rather than developing more new models.

References

1. Bray, F. *et al.* 394 CA: A Cancer Journal for Clinicians Global Cancer Statistics 2018: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA CANCER J CLIN* **68**, 394–424 (2018).
2. Wishart, G. C. *et al.* PREDICT: a new UK prognostic model that predicts survival following surgery for invasive breast cancer. *Breast Cancer Res.* **12**, R1 (2010).
3. Michaelson, J. S. *et al.* Improved web-based calculators for predicting breast carcinoma outcomes. *Breast Cancer Res. Treat.* **128**, 827–835 (2011).
4. Blamey, R. W. *et al.* Reading the prognosis of the individual with breast cancer. *Eur. J. Cancer* **43**, 1545–1547 (2007).
5. Cardoso, F. *et al.* Early breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann. Oncol.* **30**, 1194–1220 (2019).
6. Steyerberg, E. W. *et al.* Prognosis Research Strategy (PROGRESS) 3: Prognostic Model Research. *PLoS Med.* **10**, (2013).
7. Steyerberg, E. Clinical prediction models, A Practical Approach to Development, Validation, and Updating. Springer International Publishing (2019).
8. Moons, K. G. M. *et al.* Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. *heart.bmj.com* doi:10.1136/heartjnl-2011-301246
9. Moons, K. G. M. *et al.* Risk prediction models: II. External validation, model updating, and impact assessment. *heart.bmj.com* doi:10.1136/heartjnl-2011-301247
10. Collins, G. S., Reitsma, J. B., Altman, D. G. & Moons, K. G. M. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement. *Ann. Intern. Med.* **162**, 55 (2015).
11. Moher, D. *et al.* Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine* **6**, (2009).
12. Geersing, G.-J. *et al.* Search filters for finding prognostic and diagnostic prediction studies in Medline to enhance systematic reviews. *PLoS One* **7**, e32844 (2012).
13. Moons, K. G. M. *et al.* Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies: The CHARMS Checklist. *PLoS Med.* **11**, e1001744 (2014).
14. van Smeden, M. *et al.* Sample size for binary logistic prediction models: Beyond events per variable criteria. *Stat. Methods Med. Res.* **28**, 2455–2474 (2019).
15. Wolff, R. F. *et al.* PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. *Ann. Intern. Med.* **170**, 51 (2019).
16. Royston, P., Altman, D. G. & Sauerbrei, W. Dichotomizing continuous predictors in multiple regression: a bad idea. *Stat. Med.* **25**, 127–141 (2006).
17. Damen, J. A. A. G. *et al.* Prediction models for cardiovascular disease risk in the general population: systematic review. *BMJ* **353**, i2416 (2016).
18. Wynants, L. *et al.* Prediction models for diagnosis and prognosis of covid-19: Systematic review and critical appraisal. *BMJ* **369**, (2020).

19. Phung, M. T., Tin Tin, S. & Elwood, J. M. Prognostic models for breast cancer: A systematic review. *BMC Cancer* **19**, (2019).

20. Engelhardt, E. G. *et al.* Accuracy of the online prognostication tools PREDICT and Adjuvant! for early-stage breast cancer patients younger than 50 years. *Eur. J. Cancer* **78**, 37–44 (2017)

21. van Maaren, M. C. *et al.* Validation of the online prediction tool PREDICT v. 2.0 in the Dutch breast cancer population. *Eur. J. Cancer* **86**, 364–372 (2017).

22. Steyerberg, E. W. & Vergouwe, Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *Eur. Heart J.* **35**, 1925–31 (2014).

23. van Giessen, A. *et al.* Systematic Review of Health Economic Impact Evaluations of Risk Prediction Models: Stop Developing, Start Evaluating. *Value Heal.* **20**, 718–726 (2017).

24. European Commission. Medical Devices Regulation. (2017). Available at: https://eur-lex.europa.eu/eli/reg/2017/745/2017-05-05. (Accessed: 8th February 2021)

Chapter 3

External validation of 87 clinical prediction models supporting clinical treatment decisions for breast cancer patients

Tom A. Hueting, Marissa C. van Maaren, Mathijs P. Hendriks,
Hendrik Koffijberg, Sabine Siesling

Submitted for peer-review

The supplementary data to this chapter are available upon request

Abstract:

Introduction: Numerous prediction models have been developed to support treatment-related decisions for breast cancer patients. Externally validation, a prerequisite for implementation in clinical practice, has been performed for only a few models. This study aimed to externally validate published clinical prediction models using readily available population-based Dutch data.

Methods: Patient-, tumor- and treatment-related data were derived from the Netherlands Cancer Registry (NCR). Model performance was assessed using the area under the receiver operating characteristic curve (AUC), scaled Brier score, and model calibration. Net benefit across applicable risk thresholds was evaluated with decision curve analysis.

Results: After assessing 922 models, 87 (9%) were included for validation. Models were excluded due to an incomplete model description (n=262 (28%)), lack of required data (n=521 (57%)), previously validated or developed with NCR data (n=45 (5%)), or the associated NCR sample size was insufficient (n=7 (1%)). The included models predicted survival (33 (38%) overall, 27 (31%) breast cancer-specific, and 3 (3%) other cause-specific), locoregional recurrence (n=7 (8%)), disease free survival (n=7 (8%)), metastases (n=5 (6%)), lymph node involvement (n=3 (3%)), pathologic complete response (n=1 (1%)), and surgical margins (n=1 (1%)). Seven models (8%) showed poor (AUC<0.6), 39 (45%) moderate (AUC:0.6-0.7), 38 (46%) good (AUC:0.7-0.9), and 3 (3%) excellent (AUC≥0.9) discrimination. Using the scaled Brier score, worse performance than an uninformative model was found in 34 (39%) models.

Conclusion: Comprehensive registry data supports broad validation of published prediction models. Poorly performing models are not advised to be used in clinical practice. Moderate and good performing models could be clinically useful in a Dutch setting after careful impact evaluation.

Introduction

Worldwide, over 2.2 million new cases of breast cancer were diagnosed in 2020.[1] In the Netherlands, over 17,000 women and 100 men are diagnosed with breast cancer annually, making this the most commonly diagnosed cancer in women[2] Even though the survival of breast cancer has improved throughout the past decades, the prognosis of an individual breast cancer patient strongly depends on patient and tumor characteristics, and available treatment options.[3]

To support (shared) decision-making by patients and clinicians regarding breast cancer treatment, prediction models have been developed that estimate the probability of certain outcomes using available patient and tumor characteristics. An example of such a model is PREDICT[4], which is frequently used to support clinical management on the decision to initiate adjuvant therapy.

Previously, a systematic literature review was performed to identify available prediction models that may provide valuable information to support treatment decision-making[5] A total of 922 available prediction models were identified, which were developed to predict clinical outcomes such as treatment response, lymph node involvement, adverse events, recurrence, and (breast cancer specific) survival. However, the majority of the identified models were found to be at high risk of bias according to the prediction model risk of bias assessment tool (PROBAST).[6] The clinical utility of most of these models remained unclear as a substantial number of models were not reported according to established reporting guidelines, and showed methodological flaws during the development and/or the internal validation of the model.

Moreover, prior to the use of prognostic models in a clinical setting, they should be validated both internally and externally on the target population,[7] and the clinical impact of the models on clinical practice should subsequently be assessed.[8] Still, for meaningful applications of prediction models, new models are more often developed than existing models are externally validated, and impact studies are performed even less, which means that potentially valuable information on the performance of a model is lacking.[9] This refrains existing models from being implemented in daily practice to support clinical decision-making in a certain population. However, when already available prediction models perform well on external data sets, the creation of new models will become less relevant than actually implementing valuable and validated models, and keeping these up to date.[10] Therefore, this study aimed to evaluate the performance of previously identified prediction models using readily available data obtained from the Netherlands Cancer Registry (NCR).

Methods

Study population

The performance of identified clinical prediction models was evaluated using data obtained from the NCR. The NCR is a nationwide database comprising all newly diagnosed malignant tumors in the Netherlands. The data cohort consisted of patients diagnosed with breast cancer between 2003 and 2019. Invasive and non-invasive cancers were included, as well as female and male breast cancer patients. Patients were excluded if they were younger than 18 years old, or when the cancer was diagnosed during an autopsy.

Based on the patient group targeted by a prediction model, specific subgroups of patients were extracted from the full dataset to perform the model validation. To validate the different models, the definition of included variables, and the in- and exclusion criteria were applied as described in the original paper as much as possible.

Table 1. Patient characteristics of all breast cancer patients derived from the NCR

Characteristic	**Value**	**N**	**(%)**
Total		**288784**	**100%**
Gender	Male	1784	0.6%
	Female	287000	99.4%
Age	Years (Mean (SD))	61	13.7
Year of diagnosis	2003 – 2006	57539	19.9%
	2007 – 2010	64345	22.3%
	2011 – 2014	72526	25.1%
	2015 – 2019	94374	32.7%
Malignancy	Invasive carcinoma	254395	88.1%
	Carcinoma in situ	34389	11.9%
*Stage**	0	34389	11.9%
	I	113420	39.3%
	II	95496	33.1%
	III	30825	10.7%
	IV	13420	4.6%
	Missing	1234	0.4%
Differentiation grade	1	56999	19.7%
	2	113530	39.3%
	3	76891	26.7%
	Missing	41364	14.3%
ER status	Negative	40349	14.0%
	Positive	203545	70.5%
	Missing	44890	15.5%
PR status	Negative	77977	27.0%
	Positive	161881	56.1%
	Missing	48926	16.9%
HER2 status	Negative	186141	64.5%
	Positive	29917	10.4%
	Unclear	22039	7.6%
	Missing	50687	17.5%
*Follow-up data regarding recurrences completely available over**:*	5-year	62116	21.5%
	10-year	20858	7.2%

* Stage was defined as the pathologic tumor stage, supplemented by clinical tumor stage (when pathologic stage was unknown or when the patient received neoadjuvant treatment).

** The follow-up data was actively searched for certain cohorts only in the NCR and therefore does not reflect the lost to follow-up rate.

Model selection

The previously identified 922 clinical prediction models, described in 534 papers were considered to be potential candidates for external validation and were selected based on four criteria.

First, models were selected in case sufficient details were reported to recover the underlying equation allowing the calculation of risks of the outcome for individual patients. For this, the underlying variable coefficients required to calculate the result of a model should have been available (or could have been recovered from a nomogram), and all required covariates (input variables and outcome) should have been clearly defined.

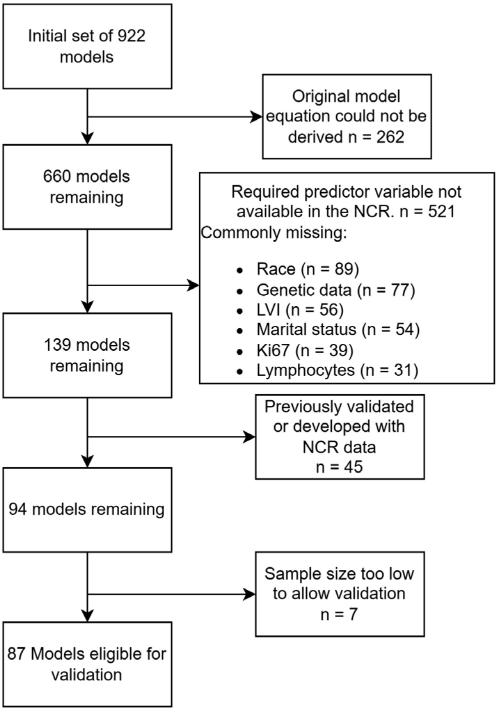
Second, the required data, including both the input and outcome data, for adequate validation of the model had to be available in the NCR.

Third, models were excluded when they were either developed by or previously validated on NCR data.

Fourth, models were excluded in case the available sample size within the NCR to validate the model was too low. For sample size considerations, the 100 events and non-events rule-of-thumb reported by Vergouwe et al. was initially used.[11] When the sample size was lower than 100 events and non-events (e.g. indicating a minimal requirement of 200 patients when the outcome occurs in 50% of the patients), additional calculations were performed according to the study by Riley et al. to determine if available data allowed validation.[12]

Several assumptions were made in the data to allow more models to be validated. As the cause of death is not recorded in the NCR, patients who died with known metastasized breast cancer were assumed to have died due to breast cancer. The breast cancer subtype definition varies in different models. When no clear definition was provided in the paper describing the development of the model, the following definition was applied for breast cancer subtype; Luminal A (HR+ & HER2-), Luminal B (HR+ & HER2-), HER2-enriched (HR- & HER2+), and triple negative (HR- & HER2-). For models predicting a time-to-event outcome that may occur more than once (e.g. metastasis or locoregional recurrence), only the first event that occurred was taken into account.

Figure 1. Flowchart of model selection.


Initial set of 922 models
Original model equation could not be derived n = 262
660 models remaining
Required predictor variable not available in the NCR. n = 521
Commonly missing:
• Race (n = 89)
• Genetic data (n = 77)
• LVI (n = 56)
• Marital status (n = 54)
• Ki67 (n = 39)
• Lymphocytes (n = 31)
139 models remaining
Previously validated or developed with NCR data n = 45
94 models remaining
Sample size too low to allow validation n = 7
87 Models eligible for validation

Abbreviations: LVI = Lymphovascular invasion, NCR = Netherlands Cancer Registry

Statistical analysis

All models were assessed on their performance in terms of discrimination, calibration, and net benefit. Discrimination concerns the ability of a model to stratify between high and low risk of the predicted outcome, and was quantified with the area under the receiving operating characteristic curve (AUC), and visualized using classification plots as proposed by Verbakel et al.[13] Discriminatory performance was considered poor (AUC<0.6), moderate (AUC:0.6-0.7), good (AUC:0.7-0.9), and excellent (AUC≥0.9). Calibration concerns the level of agreement between predicted and observed event rates and is visualized using calibration plots. Also, the Brier score and the scaled Brier score were estimated for each model. The Brier score concerns the squared differences between predicted and observed outcomes.[14] Brier scores range between 0 and 1, and a lower Brier score indicates better performance. The scaled Brier score compares the Brier score to the Brier score of an uninformative model (i.e. assuming the observed event rate is the predicted risk for all patients). A scaled Brier score <0 indicates that the model performs worse than an uninformative model. A higher scaled Brier score indicates better performance. A combination of the AUC and the scaled Brier score was used to categorize the overall performance of the models into poor (AUC<0.7 and scaled Brier≤0), moderate (either an AUC≥0.7 or a scaled Brier>0), and good (AUC≥0.7 and scaled Brier>0). Clinical usefulness was assessed by comparing the net benefit of applying the model over all feasible thresholds, and is visualized using decision curve analysis in which the added value of the model is compared to default strategies of treating all or no patients.[15]

A separate dataset was created based on the original in- and exclusion criteria reported for each of the validated models. Missing data were assessed for each separate dataset and where appropriate, missing data were handled using multiple imputation by chained equations (MICE).[16] Missing data were imputed on the complete dataset to ensure accurate estimations. The process of data imputation and model performance evaluation was repeated using 200 bootstrap samples.

Results

Patient data

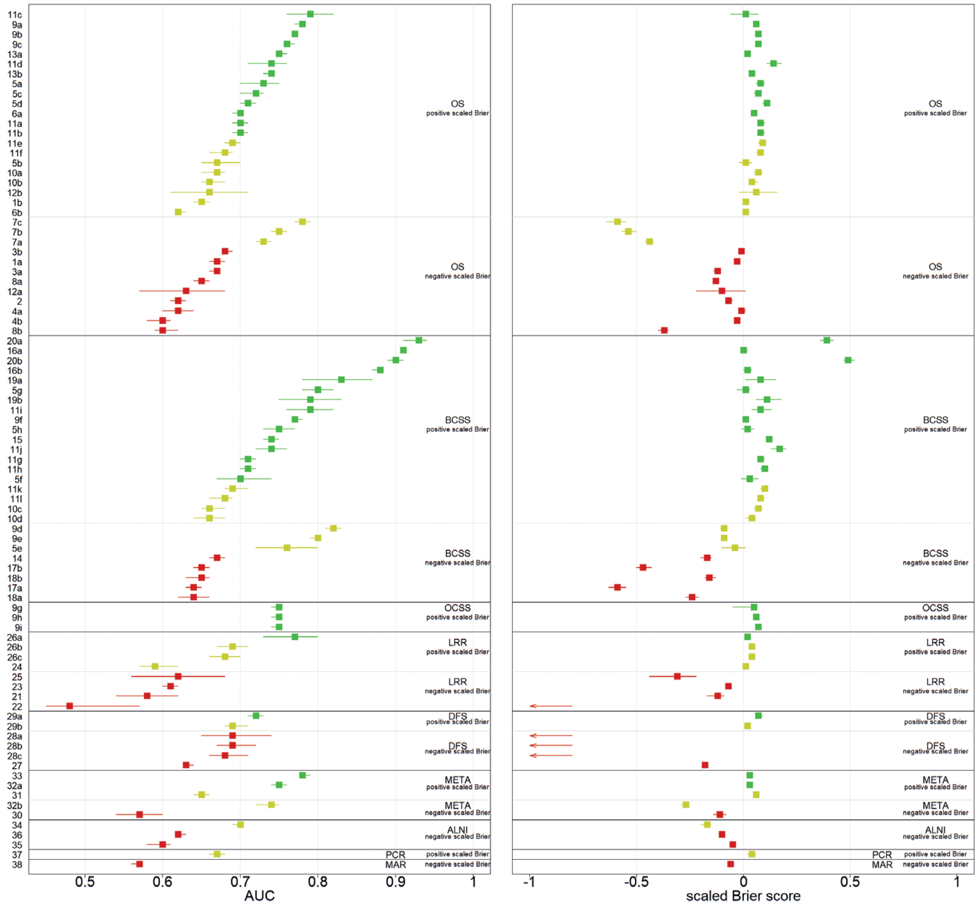
Data on 288,784 tumors diagnosed in 271,040 patients were obtained from the NCR. Patient characteristics from the data obtained from the NCR are displayed in table 1. The majority of the patients were female (n=287,000 (99.4%)). On average, patients were 61 (SD 13.7) years old when diagnosed. The number of tumors increased over the years ranging from 121,884 (42%) in 2003-2010 to 166,900 (58%) in 2011-2019. From the dataset of 288,784 breast tumors, smaller cohorts were selected according to the in- and exclusion criteria of the model being validated. For each of the validated models, detailed descriptions of the outcome, input variables, inclusion criteria, exclusion criteria, original validation, and baseline characteristics of the dataset used to validate each of the included models were summarized in the supplementary data. The sample size used to validate a model ranged between 432 and 243,930 with a median sample size of 10,368 (IQR 5,808 – 47,875).

Model selection

All 922 models were initially considered for inclusion in our study. A total of 262 (28%) models were not described with sufficient details to calculate a risk for new patients (e.g. the original model equation could not be derived due to lack of reported model coefficients) and could not be validated. Another 521 (57%) models were excluded due to the unavailability of required input or outcome data in the NCR. Data most commonly resulting in the exclusion of a model were, race (n=89), genetic data (n=77), lymphovascular invasion (LVI) (n=56), marital status (n=54), Ki67 (n=39), and lymphocytes (including tumor infiltrating tumors and indices such as monocyte-to-lymphocyte ratio) (n=31). Models developed or previously validated with NCR data (n=45 (5%)) were also excluded, and lastly, 7 (1%) models were excluded as the available sample size was too low to validate these models. Finally, a total of 38 papers reporting on a total of 87 (9%) models were included in our external validation study. The process of in- and excluding the models is visualized in the flowchart in figure 1.

An overview of the included models is provided in table 2. A total of 33 (38%) models were developed to predict overall survival (OS), 27 (31%) models predicted breast cancer-specific survival (BCSS), 3 (3%) models other cause specific survival (OCSS), 7 (8%) models disease free survival (DFS), 7 (8%) locoregional recurrence (LRR), 5 (6%) predicted metastasis, 3 (3%) models lymph node involvement (LNI), 1 (1%) model pathologic complete response (PCR), and 1 (1%) model predicted surgical margin status. Several models were developed for a specific subset of patients. For instance, the models developed by Chen et al (models 19a & 19b), were specifically aimed to provide BCSS predictions for male breast cancer patients. A short description of the specific patient subgroups per model is displayed in table 2 and more detailed descriptions can be found in the supplementary tables.

Figure 2. Visualization of the discrimination (AUC) and the scaled Brier score for each of the validated models.



The green points represent models that were considered to perform good (AUC ≥0.7 and scaled Brier score >0), yellow corresponds with a moderate performance (AUC <0.7 or scaled Brier ≤0), and red is associated with poor performance (AUC <0.7 and scaled Brier score ≤0). The model performance is presented per predicted outcome, and further divided by positive and negative scaled Brier.

Abbreviations: ALNI = Axillary Lymph Node Involvement, AUC = Area Under the Curve, BCSS = Breast Cancer Specific Survival, DFS = Disease Free Survival, MAR = Positive Surgical Margin, META = Metastasis, LRR = Locoregional Recurrence, OCSS = Other Cause Specific Survival, OS = Overall Survival, PCR = Pathologic Complete Response.

Model performance evaluation

The performance of 87 models was evaluated. For each model, the AUC, and (scaled) Brier score were calculated, and a calibration plot, classification plot, and decision curve were visualized graphically (Supplement).

The AUC, scaled Brier score, sample size used, and the event rate for each model were added to table 2. The AUC values ranged between 0.48 and 0.93. In terms of discrimination, 7 (8%) models had a poor (AUC<0.6), 39 (45%) models a moderate (AUC:0.6-0.7), 38 (44%) models a good (AUC:0.7-0.9), and 3 (3%) models an excellent (AUC≥0.9) performance on the AUC. The scaled Brier score ranged between -2.00 and 0.52 and showed an adequate performance (scaled Brier score >0) in 53 (61%) models, and a poor performance (scaled Brier score ≤0) in 34 (39%) models. Combining both measures resulted in 34 (39%) models showing a good performance (AUC ≥0.7 and scaled Brier score >0), 26 (30%) models showed a moderate performance (either an AUC<0.7 or scaled Brier score ≤0), and the remaining 27 (31%) models showed a poor performance (AUC<0.7 and scaled Brier score ≤0). The AUC and scaled Brier scores per model are described in table 2 and visualized in figure 2.

A calibration plot, classification plot, and net benefit curve were constructed for each validated model and are displayed in the supplementary data. For illustrative purposes, examples of two calibration plots, decision curves, and classification plots were displayed in figures 2, 3, and 4, respectively. For each of the figures, a model with good performance, and a model with poor performance were displayed side-to-side.

Table 2. Overview of the validated models, predictors, events, and population, grouped by outcome.

Author	ID	Specific patient sub-group	Input variables	Outcome	Original AUC*	AUC	Scaled Brier score	Sample size	Event rate
				Overall survival					
Xiong	1a	M1	Age, MFI, M, HR	1-year	0.670	0.668 (0.656 – 0.678)	-0.033 (-0.043 – -0.024)	11633	71.7%
Xiong	1b	M1	Age, MFI, M, HR	3-year	0.670	0.652 (0.642 – 0.661)	0.010 (-0.003 – 0.025)	10964	37.8%
Regierer	2	M1	MFI, HR, M	5-year	0.686	0.622 (0.614 – 0.631)	-0.065 (-0.079 – -0.050)	17608	23.9%
Fan	3a	Mast	Age, T, N, M, ER	2-year	0.800	0.665 (0.658 – 0.673)	-0.123 (-0.137 – -0.108)	86418	93.5%
Fan	3b	Mast	Age, T, N, M, ER	5-year	0.800	0.683 (0.678 – 0.687)	-0.013 (-0.024 – -0.004)	73465	79.1%
Luo	4a	M0 & HER2+	Age, ER, T, N, Tras	3-year	0.780 & 0.740	0.619 (0.598 – 0.636)	-0.005 (-0.013 – 0.005)	15107	93.5%
Luo	4b	M0 & HER2+	Age, ER, T, N, Tras	5-year	0.780 & 0.740	0.597 (0.583 – 0.610)	-0.029 (-0.038 – -0.018)	13599	87.6%
Zhang	5a	Adj Rad	Age, Gr, T, N, ER, PR	5-year	0.687 & 0.672	0.726 (0.703 – 0.747)	0.078 (0.059 – 0.096)	3208	84.1%
Zhang	5b	Adj Rad	Age, Gr, T, N, ER, PR	10-year	0.687 & 0.672	0.672 (0.650 – 0.699)	0.008 (-0.023 – 0.043)	2072	60.0%
Zhang	5c	No Rad	Age, Gr, T, N, ER, PR	5-year	0.700 & 0.696	0.715 (0.702 – 0.731)	0.067 (0.054 – 0.080)	10423	87.4%
Zhang	5d	No Rad	Age, Gr, T, N, ER, PR	10-year	0.700 & 0.696	0.711 (0.700 – 0.723)	0.106 (0.092 – 0.118)	9254	61.5%
Chen	6a	M0	Age, Gr, T, N, HR	5-year	0.822 & 0.780	0.696 (0.692 – 0.700)	0.047 (0.041 – 0.052)	170643	85.9%
Chen	6b	M0	Age, Gr, T, N, HR	5-year	0.792 & 0.800	0.622 (0.618 – 0.626)	0.007 (0.003 – 0.010)	170643	85.9%
Zhao	7a	Advanced	TNM, MS, DFS, TB, BM	1-year	0.770 & 0.710	0.731 (0.720 – 0.741)	-0.437 (-0.463 – -0.415)	8745	63.7%
Zhao	7b	Advanced	TNM, MS, DFS, TB, BM	2-year	0.770 & 0.710	0.750 (0.740 – 0.760)	-0.541 (-0.574 – -0.503)	8743	45.6%
Zhao	7c	Advanced	TNM, MS, DFS, TB, BM	3-year	0.770 & 0.710	0.776 (0.765 – 0.787)	-0.593 (-0.635 – -0.547)	8740	33.7%
Tang	8a	T1-2N1M0	Age, Topo, T, N, ER, PR, HER2, Tras	5-year	0.700	0.650 (0.638 – 0.663)	-0.129 (-0.138 – -0.116)	8774	82.2%
Tang	8b	T1-2N1M0	Age, Topo, T, N, ER, PR, HER2, Tras	10-year	0.700	0.604 (0.591 – 0.618)	-0.369 (-0.396 – -0.346)	7238	62.1%
Xu	9a	Stage I-II	Age, Gr, T, MS, SRG	3-year	0.802	0.775 (0.770 – 0.779)	0.060 (0.057 – 0.063)	175927	94.0%
Xu	9b	Stage I-II	Age, Gr, T, MS, SRG	4-year	0.795	0.769 (0.766 – 0.774)	0.067 (0.064 – 0.071)	161550	90.8%
Xu	9c	Stage I-II	Age, Gr, T, MS, SRG	5-year	0.787	0.763 (0.760 – 0.767)	0.067 (0.063 – 0.070)	147892	87.2%
Wang	10a	Bone M1	Gr, Morf, T, SRG, Chem, M, MS	3-year	0.705 & 0.678	0.665 (0.650 – 0.677)	0.070 (0.049 – 0.086)	5834	46.0%
Wang	10b	Bone M1	Gr, Morf, T, SRG, Chem, M, MS	5-year	0.705 & 0.678	0.663 (0.646 – 0.682)	0.044 (0.025 – 0.071)	5375	23.3%
Zheng	11a	M1 pre-op	Age, Gr, T, M, ER, PR, HER2, Rad, Chem	1-year	0.721	0.701 (0.689 – 0.714)	0.084 (0.074 – 0.095)	8409	75.3%
Zheng	11b	M1 pre-op	Age, Gr, T, M, ER, PR, HER2, Rad, Chem	3-year	0.721	0.703 (0.694 – 0.714)	0.081 (0.066 – 0.099)	7577	40.0%
Zheng	11c	M1 SRG	Age, Gr, T, M, ER, PR, HER2, Rad, Chem	1-year	0.713	0.786 (0.757 – 0.818)	0.011 (-0.062 – 0.074)	1994	90.5%

Table 2. Continued.

Author	ID	Specific patient sub-group	Input variables	Outcome	Original AUC*	AUC	Scaled Brier score	Sample size	Event rate
Zheng	11d	M1 SRG	Age, Gr, T, M, ER, PR, HER2, Rad, Chem	3-year	0.713	0.735 (0.714 – 0.759)	0.143 (0.110 – 0.181)	1769	59.2%
Zheng	11e	M1 no-SRG	Age, Gr, T, M, ER, PR, HER2, Rad, Chem	1-year	0.664	0.691 (0.675 – 0.704)	0.087 (0.067 – 0.104)	6415	70.5%
Zheng	11f	M1 no-SRG	Age, Gr, T, M, ER, PR, HER2, Rad, Chem	3-year	0.664	0.678 (0.661 – 0.691)	0.076 (0.056 – 0.091)	5808	34.1%
Janssen	12a	Bone M1	ECOG, M	1-year	NA	0.630 (0.573 – 0.678)	-0.095 (-0.224 – 0.008)	520	81.5%
Janssen	12b	Bone M1	ECOG, M	2-year	NA	0.657 (0.611 – 0.706)	0.058 (-0.019 – 0.159)	432	37.3%
Wang	13a	M0, mast, no neo-adj	Age, T, N, Gr, ER, PR	3-year	0.740	0.750 (0.745 – 0.756)	0.024 (0.021 – 0.027)	71758	90.0%
Wang	13b	M0, mast, no neo-adj	Age, T, N, Gr, ER, PR	5-year	0.720	0.737 (0.731 – 0.742)	0.043 (0.038 – 0.048)	65171	80.9%
Breast Cancer Specific Survival									
Abdel Rahman	14	M1 BC	M, ER, PR, HER2, Gr	4-year	0.665	0.666 (0.657 - 0.675)	-0.174 (-0.199 – -0.149)	10651	27.8%
Elwood	15	NA	HER2, Morf, Age, Etn, M, T, HR, Gr, N	10-year	0.840	0.740 (0.733 – 0.745)	0.116 (0.105 – 0.125)	48661	13.7%
Paredes Aracil	16a	NA	Age, TNM, Gr, PBC, MF	5-year	0.830	0.911 (0.908 – 0.913)	0.002 (-0.011 – 0.017)	195349	6.5%
Paredes Aracil	16b	NA	Age, TNM, Gr, PBC, MF	10-year	0.830	0.877 (0.874 – 0.881)	0.019 (0.004 – 0.036)	113615	13.8%
Wen	17a	M0, IDC or ILC	Men, T, N, ER, HER2	5-year	0.747 & 0.789	0.641 (0.628 – 0.653)	-0.591 (-0.634 – -0.548)	45517	95.1%
Wen	17b	M0, IDC or ILC	Men, T, N, ER, HER2	10-year	0.747 & 0.789	0.650 (0.639 – 0.660)	-0.465 (-0.501 – -0.433)	35270	90.6%
Wen	18a	M0, IDC or ILC	ER, HER2, T, N, Men	5-year	0.745 & 0.796	0.642 (0.623 – 0.656)	-0.239 (-0.267 – -0.209)	41122	95.8%
Wen	18b	M0, IDC or ILC	ER, HER2, T, N, Men	10-year	0.745 & 0.796	0.647 (0.634 – 0.658)	-0.157 (-0.182 – -0.134)	26164	90.4%
Zhang	5e	Adj radio	Age, Gr, T, N, ER, PR	5-year	0.699 & 0.656	0.758 (0.716 – 0.799)	-0.039 (-0.102 – 0.005)	2822	95.6%
Zhang	5f	Adj radio	Age, Gr, T, N, ER, PR	10-year	0.699 & 0.656	0.702 (0.667 – 0.735)	0.033 (-0.011 – 0.069)	1433	86.8%
Zhang	5g	Adj radio	Age, Gr, T, N, ER, PR	5-year	0.716 & 0.671	0.801 (0.780 – 0.820)	0.006 (-0.027 – 0.032)	9483	96.0%
Zhang	5h	Adj radio	Age, Gr, T, N, ER, PR	10-year	0.716 & 0.671	0.751 (0.731 – 0.772)	0.023 (-0.007 – 0.051)	7258	78.5%
Chen	19a	Male	Age, T, ER, PR, SRG	3-year	0.788	0.827 (0.782 – 0.867)	0.078 (0.010 – 0.150)	1330	94.3%
Chen	19b	Male	Age, T, ER, PR, SRG	5-year	0.825	0.789 (0.752 – 0.832)	0.112 (0.055 – 0.182)	991	89.6%
Fu	20a	ILC, stage II-IV	Age, Topo, Gr, TNM, SRG, Chem, MS	3-year	0.793 & 0.830	0.926 (0.911 – 0.936)	0.389 (0.358 – 0.419)	12246	94.1%
Fu	20b	ILC, stage II-IV	Age, Topo, Gr, TNM, SRG, Chem, MS	5-year	0.772 & 0.824	0.900 (0.889 – 0.912)	0.491 (0.466 – 0.518)	9849	89.0%
Xu	9d	Stage I-II	Age, Gr, T, MS, SRG	3-year	0.830	0.818 (0.808 – 0.828)	-0.092 (-0.105 – -0.079)	168847	99.1%
Xu	9e	Stage I-II	Age, Gr, T, MS, SRG	4-year	0.817	0.796 (0.788 – 0.803)	-0.086 (-0.099 – -0.077)	151702	98.5%

Table 2. Continued.

Author	ID	Specific patient sub-group	Input variables	Outcome	Original AUC*	AUC	Scaled Brier score	Sample size	Event rate
Xu	9f	Stage I-II	Age, Gr, T, MS, SRG	5-year	0.803	0.774 (0.766 – 0.781)	0.014 (0.011 – 0.018)	135451	97.8%
Wang	10c	Bone M1	Gr, Morf, T, SRG, Chem, M, MS	3-year	0.710 & 0.684	0.663 (0.652 – 0.677)	0.066 (0.050 – 0.085)	5834	46.0%
Wang	10d	Bone M1	Gr, Morf, T, SRG, Chem, M, MS	5-year	0.710 & 0.684	0.661 (0.642 – 0.677)	0.036 (0.009 – 0.059)	5375	23.3%
Zheng	11g	M1 pre-op	Age, Gr, T, M, ER, PR, HER2, Rad, Chem	1-year	0.722	0.708 (0.695 – 0.723)	0.083 (0.070 – 0.098)	8409	75.3%
Zheng	11h	M1 pre-op	Age, Gr, T, M, ER, PR, HER2, Rad, Chem	3-year	0.722	0.711 (0.697 – 0.722)	0.097 (0.076 – 0.114)	7577	40.0%
Zheng	11i	M1 SRG	Age, Gr, T, M, ER, PR, HER2, Rad, Chem	1-year	0.715	0.791 (0.758 – 0.822)	0.083 (0.038 – 0.132)	1994	90.5%
Zheng	11j	M1 SRG	Age, Gr, T, M, ER, PR, HER2, Rad, Chem	3-year	0.715	0.742 (0.718 – 0.764)	0.169 (0.131 – 0.203)	1769	59.2%
Zheng	11k	M1 no-SRG	Age, Gr, T, M, ER, PR, HER2, Rad, Chem	1-year	0.666	0.694 (0.680 – 0.707)	0.098 (0.084 – 0.112)	6415	70.5%
Zheng	11l	M1 no-SRG	Age, Gr, T, M, ER, PR, HER2, Rad, Chem	3-year	0.666	0.680 (0.664 – 0.693)	0.084 (0.067 – 0.101)	5808	34.1%
Other Cause Specific Survival									
Xu	9g	Stage I-II	Age, Gr, T, MS, SRG	3-year	0.813	0.749 (0.744 – 0.754)	0.049 (0.046 -0.052)	168847	98.8%
Xu	9h	Stage I-II	Age, Gr, T, MS, SRG	4-year	0.808	0.747 (0.743 – 0.752)	0.059 (0.056 – 0.062)	151702	98.2%
Xu	9i	Stage I-II	Age, Gr, T, MS, SRG	5-year	0.817	0.747 (0.743 – 0.751)	0.067 (0.063 – 0.069)	135451	97.4%
Locoregional recurrence									
Herrero-Vicent	21	Neo-adj chem	HER2, DCIS, PCR	6-year	NA	0.583 (0.544 – 0.617)	-0.124 (-0.170 – -0.088)	739	23.4%
Wobb	22	BCS, adj rad	Age, Men, Mar, ER, Gr	5-year	0.641	0.478 (0.448 - 0.565)	-1.996 (-2.235 – -1.767)	11822	2.3%
Sanghani	23	BCS	Age, LVI, Mar, T, Gr, Chem, Horm, Rad	10-year	0.660	0.592 (0.566 – 0.617)	0.006 (-0.003 – 0.016)	7343	6.8%
Li	24	T1-2N1-3M0	Age, Topo, N, T, MS	5-year	0.735 & 0.703	0.619 (0.558 – 0.682)	-0.309 (-0.435 – -0.216)	2886	3.0%
Corso	25a	Mast, no neo-adj	Age, Morf, T, N, MS, Horm, chem, rad	1-year (local)	0.700	0.765 (0.728 – 0.801)	0.016 (0.007 – 0.024)	22882	0.7%
Corso	25b	Mast, no neo-adj	Age, Morf, T, N, MS, Horm, chem, rad	5-year (local)	0.700	0.689 (0.668 – 0.708)	0.037 (0.029 – 0.045)	18498	4.0%
Corso	25c	Mast, no neo-adj	Age, Morf, T, N, MS, Horm, chem, rad	10-year (local)	0.700	0.679 (0.661 – 0.697)	0.038 (0.028 – 0.048)	15173	5.6%
Disease free survival									
Li	26	BCS	MS, Gr, N	5-year	0.700	0.610 (0.604 – 0.616)	-0.066 (-0.073 – -0.060)	44176	23.1%
Tokatli	27	M0	N, HER2, ER	5-year	0.700 & 0.715	0.633 (0.626 – 0.638)	-0.180 (-0.193 – -0.166)	58568	87.5%
Lin	28a	Age ≤ 40	N, MS	1-year	NA	0.692 (0.647 – 0.738)	-1.488 (-1.839 – -1.218)	5127	97.6%

Lin	28b	Age ≤ 40	N, MS	2-year	NA	0.693 (0.665 – 0.722)	-1.406 (-1.633 – -1.225)	4919	91.7%
Lin	28c	Age ≤ 40	N, MS	3-year	NA	0.684 (0.660 – 0.710)	-1.608 (-1.811 – -1.442)	4759	87.5%
Paredes Aracil	29a	M0	Age, TNM, MF, Gr	5-year	0.750	0.718 (0.707 – 0.729)	0.073 (0.063 – 0.085)	21653	12.0%
Paredes Aracil	29b	M0	Age, TNM, MF, Gr	10-year	0.750	0.692 (0.678 – 0.705)	0.018 (0.001 – 0.034)	7750	25.5%
Metastasized disease									
Dowsett	30	Postmenopausal, HR+	T, N, Age, Gr	5 to 10 year	0.678	0.574 (0.540 - 0.604)	-0.107 (-0.143 - -0.080)	5716	5.5%
Lin	31	M1 BC	Sex, Age, Morf, N, Gr, ER, PR, HER2	Liver metastasis	0.660 & 0.650	0.652 (0.641 – 0.663)	0.056 (0.048 – 0.066)	10312	24.7%
Lim	32a	Adj rad	Age, MS, T, N	5-year	0.812	0.748 (0.738 – 0.759)	0.026 (0.019 – 0.035)	24464	90.8%
Lim	32b	Adj rad	Age, MS, T, N	10-year	0.812	0.735 (0.722 – 0.746)	-0.273 (-0.291 – -0.256)	8601	68.6%
Boutros	33	Invasive BC	T, N, ER, PR	M1	0.861 & 0.638	0.783 (0.780 – 0.788)	0.028 (0.025 – 0.032)	243930	4.7%
Axillary lymph node involvement									
Zhang	34	T1-T3	Age, Top, N, T, Morf, MS	ALNI	0.716 & 0.701	0.696 (0.687 – 0.704)	-0.168 (-0.196 – -0.147)	12873	77.9%
Meretoja	35	Micro or ITC SLN	MF, T	ALNI	0.682	0.596 (0.581 – 0.614)	-0.052 (-0.062 – -0.039)	5601	16.0%
Houvanaeghel	36	cN-	Age, T, Morf, Gr, MS	ALNI	0.682 & 0.686	0.622 (0.619 – 0.625)	-0.101 (-0.106 – -0.095)	164213	24.3%
Pathologic complete response									
Schipper	37	cN+	T, Morf, ER, PR, HER2, Tras, Chem	PCR	0.770	0.674 (0.662 - 0.684)	0.039 (0.023 - 0.056)	13422	29.0%
Positive surgical margin									
Pan	38	BCS	HR, HER2, T, N, MF	Surgical margin	0.720 & 0.690	0.566 (0.562 – 0.570)	-0.064 (-0.068 – -0.060)	113499	17.5%

** Two values for the original AUC were displayed when the original model validation was assessed in multiple cohorts, using e.g. split sample or internal and external datasets.*
Abbreviations: Adj = Adjuvant, ALNI = Axillary Lymph Node Invasion, BM = Brain Metastasis, Chem = Chemotherapy, DFS = Disease Free Survival, ER = Estrogen Receptor status, Etn = Ethnicity, Gr = Grade, HER2 = HER2 status, Horm = hormonal therapy, HR = Hormone Receptor status, Mar = Surgical Margin, Mast = Mastectomy, Men = Menopausal status, MF = Multifocality, MFI = Metastasis Free Interval, Morf = Morfology, MS = Molecular Subtype, M = Metastasis, N = Nodal stage, PBC = Previous Breast cancer, PCR = Pathologic Complete Response, PR = Progesterone Receptor status, Rad = Radiotherapy, SRG = Surgery, T = Tumor size/stage, TB = Tumor burden, TNM = Stage, Top = Tumor Topography, Tras = Trastuzumab

Discussion

In this study, a total of 87 prediction models were externally validated using data from the nationwide NCR and 34 (39%) models showed a good discriminative performance and calibration. On AUC alone, 41 (47%) models showed good performance (AUC ≥0.7), and on the scaled Brier score, 53 (61%) models showed a better performance than an uninformative model. The net benefit of the validated models was assessed using decision curve analysis. It is difficult to provide summary measures of the net benefit for the validated models as the relevant threshold probabilities are necessary to interpret the curve and the thresholds differ between models. Additionally, the threshold probabilities should not be selected based upon the results displayed in a decision curve, but should rather be selected based on a clinically reasonable range.[17] Assessing these ranges was not the aim of the current study, but the provided decision curves can be used as input for future studies elaborating more on the clinical usefulness and impact of implementing one or more of the included models in clinical practice.

To validate the included models, several assumptions had to be made due to the lack of a complete and transparent description of the model in the underlying paper. For instance, the models 18a & 18b developed by Wen et al. predict 5- and 10-year BCSS, respectively, using the log odds of positive lymph nodes as a predictor.[18] The paper provided a definition of this predictor, but did not provide a base value for the logarithmic transformation. Also, Wen et al.[18] presented their model in a nomogram in which the log odds has to be entered as a value between 1 and 4, but no transformation of the predictor was provided. The poor performance of the model may be caused by this lack of transparency and a potentially useful model cannot be applied in clinical practice yet. Similar difficulties were identified for the validation of the models 7a – 7c provided by Zhao et al.[19] where there were some ambiguous definitions regarding both the predictors and the outcome. Zhao et al. for instance mention both overall and BCSS as the outcome, no proper definitions were provided for oligo-metastasis, breast cancer subtype, or advanced breast cancer. As the cause of death is not available in the NCR, disease specific mortality was assumed to occur when the patient died while being diagnosed with metastasized disease. The adequate performance found in multiple models predicting BCSS indicates that this assumption was appropriate. Several papers described multiple models that predicted OS and BCSS for metastasized breast cancer patients, such as the models 10a – 10d and 11a – 11l. Due to our definition of BCSS, the dataset used to validate these models was exactly the same (including the OS and BCSS outcomes). Still, differences found in model performance were small and insignificant so we do not expect that this assumption has negatively impacted our results.

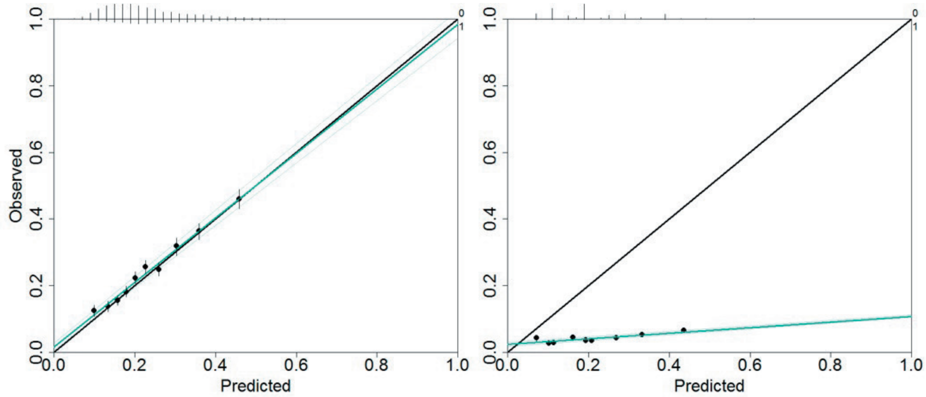
The design of the validated models affected the performance measures. For instance, model 23 incorporated LVI as a predictor, where missingness of the predictor was dealt with by modelling "unknown" as a possible input option. However, the coefficient for "unknown" was lower than the other possible input options for the predictor (i.e. LVI or no LVI). As a result, predicted probabilities were lower for all patients compared to a situation in which the predictor values would not be missing, due to the fact that LVI was missing entirely in the NCR. Also, the predictor had no discriminative value this way, as it was equivalent in all patients. Another remarkable finding concerns the models 9d – 9f predicting BCSS over 3, 4, and, 5-year, respectively, where the predicted probability can be higher after 5-years than after 3 or 4 years. It becomes difficult to explain and interpret these results well when applying these models for patient care, regardless of their performance.

The inclusion and exclusion criteria of the original models were applied as much as possible, but some discrepancies were found between the described criteria in the papers describing the development of the models and the group of patients for which the models could be applied. For instance, the models 20a and 20b described by Fu et al.[20] include the location of the tumor in the breast as a predictor (e.g. axillary tail, central, lower inner, lower outer, upper inner, or upper outer), but the data in the NCR also include patients with a tumor in an overlapping region. As it was unclear how Fu et al. dealt with these patients, these patients were excluded from the subgroup used for validation of this model, making the results only valid for a smaller group of patients.[20]

A strength of the current study concerns the large data set used to validate the models. In addition, due to the inclusion of as many identified prognostic models as possible, a total of 87 models could be validated. Given that a total of 922 models were initially considered for external validation, the number of 87 models seems to be low. The majority of the models could not be validated with NCR data due to the unavailability of several required variables such as race, genetic data, LVI, Marital status, KI67, and lymphocytes. As these data were incorporated in many different models, it is likely to assume that they provide relevant prognostic information and may become valuable additions for future data collection in the NCR or other registries. On the other hand, successful adoption of clinical prediction models relies on both performance and applicability. A model that performs very well, but requires input data that is not routinely collected may be less likely to be widely adopted in clinical practice. The NCR provided a large database with many relevant data items, but some of the commonly missing variables were missing for various reasons. For instance, due to a lack of consistency in definitions of cutoffs and methods to estimate Ki67, the variable is not routinely collected.[21] Alternative modelling methods may be applied to improve the applicability of prediction models without losing too much of its predictive performance by e.g. creating sub models in which the users of the models are enabled to still use the model when one or more of the predictors are not available, although estimates will become a little less accurate (reflected in larger confidence intervals).[22]

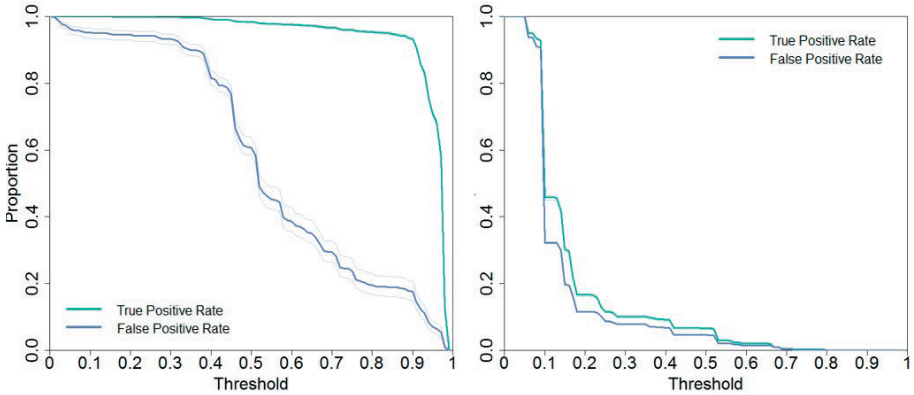
Multiple models showed a good performance in Dutch breast cancer patients. However, before these models can be used in clinical practice, further analyses may be valuable. A potentially useful next step concerns the update and re-calibration of likely valuable models. Subsequent impact studies could further define the value of incorporating some of the validated models in clinical practice. Cost-effectiveness analyses are often omitted, but are perfectly capable of estimating the actual benefits to patients and to the healthcare system when models are used in practice.[9] As highlighted by Vickers et al. a model with good performance does not necessarily indicate a valuable model.[17] Additionally, in the European Union, the use of web-apps to calculate patient-tailored predictions to inform clinical management requires the certification of the software incorporating the model under the medical devices regulation.[23] Developers should take into account the different steps needed to get valuable decision support into clinical practice even before models are developed to improve the efficiency and impact of prediction model development.

Figure 3. Examples of calibration plots to visualize the calibration.



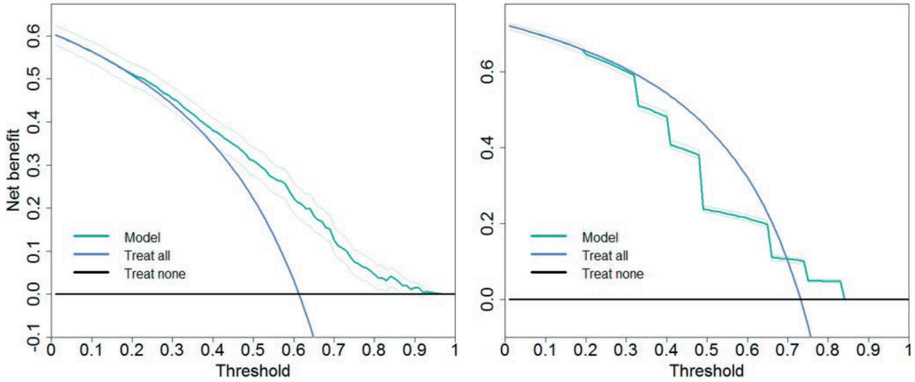
The black 45 degree line is the reference line and indicates perfect calibration. The green line is the fitted regression line. The small bars on top of the plot display a histogram of predicted risks. A taller bar represents more frequently predicted risks. The bars are stratified by 0 (non-events, displayed above the line) and 1 (events, displayed below the line). Depicted examples show good calibration (Left: model 31) and poor calibration (Right: model 22).

Figure 4. Examples of classification plots to visualize discrimination.



The green line is the true positive rate (sensitivity) and the purple line represents the false positive rate (1- specificity). The left plot concerns a model with high discrimination (model 20a with AUC = 0.926) and the right is an example of a model with barely any discriminatory power (model 38 with AUC = 0.566)

Figure 5. Examples of decision curves visualizing the net benefit.



Green line = model, purple line = treat all, black line = treat nobody. The Left curve is an example of a model with mostly higher net benefit than default strategies (model 11k) and the figure on the right shows a model with barely any added net benefit compared to default strategies (model 14)

Conclusion

The external validity of 87 prediction models to support treatment decisions of breast cancer patients was assessed. On a large Dutch registry dataset, 34 (39%) models showed a good performance, 26 (30%) models showed a moderate performance, and 27 (31%) models showed a poor performance, according to our predefined definitions. From the models showing good performance, 14 (41%) predicted BCSS, 13 (38%) predicted OS, 3 (9%) predicted OCSS, 2 (6%) predicted metastasis, 1 (3%) predicted DFS, and 1 (1%) predicted LRR. These results allow the next step towards clinical use. After careful evaluation to assess the impact of incorporating the models with a clear intended use in a usable tool, clinical adoption in the Dutch health care setting can be justified.

References

1. Cancer Today. Available at: https://gco.iarc.fr/today/online-analysis-sunburst?v=2020&mode=cancer&mode_population=continents&population=900&populations=900&key=asr&sex=2&cancer=20&type=0&statistic=5&prevalence=0&population_group=0&ages_group%5B%5D=0&ages_group%5B%5D=17&group_cancer=1&include_nmsc=1&include_nmsc_other=1. (Accessed: 9th February 2022)
2. NKR Cijfers. Available at: https://iknl.nl/nkr-cijfers?fs%7Cepidemiologie_id=526&fs%7Ctumor_id=292%2C295%2C297&fs%7Cregio_id=550&fs%7Cperiode_id=564%2C565%2C566%2C567%2C568%2C569%2C570%2C571%2C572%2C573%2C574%2C575%2C576%2C577%2C578%2C579%2C580%2C581%2C582%2C583%2C584%2C585%2C586%2C587%2C588%2C589%2C590%2C591%2C592%2C593%2C563%2C562%2C561&fs%7Cgeslacht_id=644&fs%7Cleeftijdsgroep_id=677&fs%7Cjaren_na_diagnose_id=687&fs%7Ceenheid_id=703&cs%7Ctype=line&cs%7CxAxis=periode_id&cs%7Cseries=tumor_id&ts%7CrowDimensions=periode_id&ts%7CcolumnDimensions=tumor_id&lang%7Clanguage=nl. (Accessed: 9th February 2022)
3. Howlader, N., Cronin, K. A., Kurian, A. W. & Andridge, R. Differences in breast cancer survival by molecular subtypes in the United States. *Cancer Epidemiol. Biomarkers Prev.* **27**, 619–626 (2018).
4. Candido dos Reis, F. J. *et al.* An updated PREDICT breast cancer prognostication and treatment benefit prediction model with independent validation. *Breast Cancer Res.* **19**, (2017).
5. Hueting, T. A., van Maaren, M. C., Hendriks, M. P., Koffijberg, H. & Siesling, S. Clinical prediction models to support treatment decisions in breast cancer patients: a systematic review.
6. Wolff, R. F. *et al.* PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. *Ann. Intern. Med.* **170**, 51 (2019).
7. Ramspek, C. L., Jager, K. J., Dekker, F. W., Zoccali, C. & van Diepen, M. External validation of prognostic models: what, why, how, when and where? *Clin. Kidney J.* **14**, 49–58 (2021).
8. Kappen, T. H. *et al.* Evaluating the impact of prediction models: lessons learned, challenges, and recommendations. *Diagnostic Progn. Res.* **2**, 11 (2018).
9. van Giessen, A. *et al.* Systematic Review of Health Economic Impact Evaluations of Risk Prediction Models: Stop Developing, Start Evaluating. *Value Heal.* **20**, 718–726 (2017).
10. Moons, K. G. M. *et al.* Risk prediction models: II. External validation, model updating, and impact assessment. *Heart* **98**, 691–8 (2012).
11. Vergouwe, Y., Steyerberg, E. W., Eijkemans, M. J. C. & Habbema, J. D. F. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. *J. Clin. Epidemiol.* **58**, 475–483 (2005).
12. Riley, R. D. *et al.* Minimum sample size for external validation of a clinical prediction model with a binary outcome. *Stat. Med.* **40**, 4230–4251 (2021).
13. Verbakel, J. Y. *et al.* ROC curves for clinical prediction models part 1. ROC plots showed no added value above the AUC when evaluating the performance of clinical prediction models. *J. Clin. Epidemiol.* **126**, 207–216 (2020).

14. Steyerberg, E. W. *et al.* Assessing the performance of prediction models: A framework for traditional and novel measures. *Epidemiology* **21**, 128–138 (2010).
15. Vickers, A. J. & Elkin, E. B. Decision Curve Analysis: A Novel Method for Evaluating Prediction Models. *Med. Decis. Mak.* **26**, 565–574 (2006).
16. White, I. R., Royston, P. & Wood, A. M. Multiple imputation using chained equations: Issues and guidance for practice. *Stat. Med.* **30**, 377–399 (2011).
17. Vickers, A. J. & Cronin, A. M. Everything you always wanted to know about evaluating prediction models (but were too afraid to ask). *Urology* **76**, 1298–1301 (2010).
18. Wen, J. *et al.* Development and validation of a prognostic nomogram based on the log odds of positive lymph nodes (LODDS) for breast cancer. *Oncotarget* **7**, 21046–53 (2016).
19. Zhao, J. *et al.* Development and validation of a nomogram in survival prediction among advanced breast cancer patients. *ncbi.nlm.nih.gov*
20. Fu, R. *et al.* A nomogram for determining the disease-specific survival in invasive lobular carcinoma of the breast: a population study. *ncbi.nlm.nih.gov*
21. Dowsett, M. *et al.* Assessment of Ki67 in breast cancer: recommendations from the International Ki67 in Breast Cancer working group. *J. Natl. Cancer Inst.* **103**, 1656–1664 (2011).
22. Hoogland, J. *et al.* Handling missing predictor values when validating and applying a prediction model to new patients. *Wiley Online Libr.* **39**, 3591–3607 (2020).
23. Medical Devices Regulation. (2017). Available at: https://eur-lex.europa.eu/eli/reg/2017/745/2017-05-05. (Accessed: 8th February 2021)

Chapter 4

Improved risk estimation of locoregional recurrence, secondary contralateral tumors and distant metastases in early breast cancer: the INFLUENCE 2.0 model

Vinzenz Völkel, Tom A. Hueting, Teresa Draeger, Marissa C. van Maaren, Linda de Munck, Luc J.A. Strobbe, Gabe S. Sonke, Marjanka K. Schmidt, Marjan van Hezewijk, Catharina G.M. Groothuis-Oudshoorn, Sabine Siesling

Adapted from:
Breast Cancer Research and Treatment 189, 817 – 826 (2021)
DOI: 10.1007/s10549-021-06335-z

Abstract

Purpose: To extend the functionality of the existing INFLUENCE-nomogram for locoregional recurrence (LRR) of breast cancer towards the prediction of secondary primary tumors (SP) and distant metastases (DM) using updated follow-up data and the best suitable statistical approaches.

Methods: Data on women diagnosed with non-metastatic invasive breast cancer were derived from the Netherlands Cancer Registry (n=13,494). To provide flexible time-dependent individual risk predictions for LRR, SP, and DM, three statistical approaches were assessed; a Cox proportional hazard approach (COX), a parametric spline approach (PAR), and a random survival forest (RSF). These approaches were evaluated on their discrimination using the Area Under the Curve (AUC) statistic and on calibration using the Integrated Calibration Index (ICI). To correct for optimism, the performance measures were assessed by drawing 200 bootstrap samples.

Results: Age, tumor grade, pT, pN, multifocality, type of surgery, hormonal receptor status, HER2-status, and adjuvant therapy were included as predictors. While all three approaches showed adequate calibration, the RSF-approach offers the best optimism-corrected 5-year AUC for LRR (0.75, 95%CI: 0.74 - 0.76) and SP (0.67, 95%CI: 0.65 – 0.68). For the prediction of DM, all three approaches showed equivalent discrimination (5-year AUC: 0.77 – 0.78), while COX seems to have an advantage concerning calibration (ICI<0.01). Finally, an online calculator of INFLUENCE 2.0 was created.

Conclusions: INFLUENCE 2.0 is a flexible model to predict time-dependent individual risks of LRR, SP and DM at a 5-year scale; it can support clinical decision-making regarding personalized follow-up strategies for curatively treated non-metastatic breast cancer patients.

Introduction

In the Netherlands, more than 14,000 women per year are diagnosed with invasive breast cancer,[1] rendering it the most frequently diagnosed malignancy among women.[2] Early detection and advanced treatment strategies have led to improved survival during the last decade.[3-5] The current average 5-year survival rate of women diagnosed with breast cancer (all stages) is 88% in the Netherlands.[6] The Dutch breast cancer guideline recommends annual mammograms and physical examinations during the first five years following curative treatment, unless bilateral mastectomy was performed.[7] This follow-up program is uniform for all patients and does not take individual risk profiles into account. To avoid unnecessary follow-up visits and examinations possibly inflicting psychological harm[8-10] and causing additional societal costs, the creation of personalized follow-up patterns based on individual risk-estimations would be reasonable.

In 2015, Witteveen et al.[11] developed the "INFLUENCE nomogram", which estimates an individual breast cancer patient's five-year recurrence risk as well as conditional annual risks of developing a local or regional recurrence based on different patient, tumor and treatment characteristics. Thus, it can be used to support clinical decision making; nevertheless, it neglects some relevant factors: Breast cancer follow-up aims not only at the detection of locoregional recurrences (LRR) but also of secondary primary contralateral breast tumors (SP).[7] Additionally, an estimate of the risk for developing metachronous distant metastasis (DM) is relevant for understanding a patient's prognosis and might influence decision-making regarding an optimal follow-up strategy. Besides, the HER2 status is not among the predictors of the current INFLUENCE nomogram although it has a considerable influence on therapy decisions.[12-13] From a statistical point of view, the INFLUENCE nomogram is based on five logistic regression models yielding risk estimations for the subsequent year at five arbitrary fixed time points. Other statistical approaches may contribute to improve its performance. For routine implementation in clinical practice, more detailed risk estimations for periods with flexible length are required. Thus, the patients' (changing) need for customized information could be better served and it would be possible to tailor follow-up schemes exactly to the development of individual risk profiles over time.

Aiming to incorporate all these factors, it was our aim to update the existing INFLUENCE nomogram towards an advanced INFLUENCE 2.0 model based on a large Dutch nationwide cohort. This new model is supposed to predict flexible time-dependent individual risks of LRR, SP, and DM in curatively treated non-metastatic breast cancer patients. To make the development of INFLUENCE 2.0 transparent, this paper describes the selection process among three candidates for the optimal statistical approach and describes the performance of the final model, which will be made available online in a user friendly calculator.

Methods

Study population and variables

Data for the development of the INFLUENCE 2.0 model were derived from the Netherlands Cancer Registry (NCR), a nationwide database collecting records of all newly diagnosed malignant tumors in the country hosted by the Netherlands Comprehensive Cancer Organisation (IKNL) since 1989. After notification through the nationwide pathology archive (PALGA), information on each patient is collected by specially trained registration clerks directly from patient files. The data include patient demographics, tumor-, and treatment characteristics. Vital status and date of death are regularly retrieved through linkage with the national municipality registry. Using the NCR database, we selected all women with non-metastatic (pT1-3, any pN) primary invasive adenocarcinoma of the breast, diagnosed in 2007, 2008 or the first quarter of 2012. For this cohort, active follow-up for the first five years following successful removal of the primary tumor was conducted and information on recurrences occurring within five years from diagnosis was collected. Patients were excluded in case of positive resection margins of the primary tumor, if a neoadjuvant therapy was conducted, or if surgery took place later than 180 days after diagnosis. Missing data was assumed to be missing at random. Therefore, only patients without missing data concerning potential predictor variables were included.

The INFLUENCE 2.0 model aims to estimate individual time-dependent risks for three types of events, defined according to consensus-based definitions[14]:

- Locoregional recurrence, LRR, defined as reappearance of the tumor in the ipsilateral breast, chest wall or regional lymph nodes
- Second Primary breast cancer, SP, defined as secondary primary tumor of the contralateral breast
- Distant metastasis, DM, defined as pathologically or radiologically confirmed re-appearance of tumor tissue at any location in the body

An individual is regarded to be at risk for any of these events starting the day following radical surgical removal of the primary tumor. In case of multiple events, only the first event was considered.

The following variables were selected as predictors for the named events based on previous studies and clinical expertise: age, pT-stage, pN-stage, multifocality, grading, hormone receptor status (estrogen receptor (ER)- and progesterone receptor (PR)-status), antihormonal therapy, human epidermal growth factor receptor 2 (HER2-status), type of surgery, adjuvant chemotherapy, adjuvant radiation therapy and antibody therapy. Since hormone receptor status and antihormonal therapy are highly dependent on each other (e.g. patients with negative ER-status do obviously not receive antihormonal therapy), the predictors were merged. A similar linkage exists between HER2-status and antibody therapy.

Model development

The INFLUENCE 2.0 model was designed to enable its users to choose a prediction period of variable length within five years after successful primary surgery. To optimize model performance, three statistical approaches were tested to find the best-performing model-algorithm: A Cox proportional hazards approach (COX), a parametric spline approach (PAR), and a random survival forest (RSF):

- The Cox proportional hazards approach[15] is regarded a semiparametric model since it does not assume any particular baseline survival distribution. However, it takes for granted that the predictors have a fixed effect on the underlying hazard function.
- If a changing effect of one or more predictor variables over time is assumed, the parametric spline approach might be a better choice. Basically, it consists of several piecewise defined spline functions which are joined in so-called "knots". In every piece, the influence of a predictor on the hazard function can be different.
- The Random Survival Forest[16, 17] is an extension of the classical Random Forest concept for binary outcomes[18] to analyze right censored time-to-event data. A forest of survival trees is grown using a log-rank splitting rule to select the optimal predictor variables. Survival estimates are constructed with a Kaplan-Meier estimator[19] within each terminal node, at each time.

Model performance

The three potential statistical approaches were validated and compared on their predictive ability using performance measures for calibration and discrimination.[20]

Calibration concerns the congruence between observed and predicted events. To provide quantified summary measures of model calibration, the Integrated Calibration Index (ICI, weighted average), E50 (median) and E90 (90th percentile) were calculated at t = 1, 2, 3, 4, and 5 years. These measures denote the absolute difference between observed and predicted probabilities.[21]

Discrimination was quantified using the area under the receiver operating characteristic curve (AUC). The AUC reflects the probability of a random sample of individuals with an event having a higher predicted risk than a random sample of individuals without an event. An AUC of 1.0 indicates perfect discrimination, whereas 0.5 is equal to chance. The AUC was measured based on a quarterly time frame over the whole five-year prediction period to assess the three approaches' difference in AUC over time using a cumulative/dynamic approach as described by Kamarudin et al.[22]

The performance measures were obtained as apparent and adjusted values. The apparent results reflect the performance of the tested approaches in the same data used to train them. Additionally, adjusted performance measures were estimated in 200 bootstrap samples. They reflect the performance of an approach trained in a bootstrap sample applied on the entire dataset. The difference between apparent and adjusted performance denotes the level of optimism. A low level of optimism indicates a more robust performance. The adjusted results represent the optimism-corrected performance and were used to decide upon the optimal statistical approach for the final model.[23]

Ultimately, INFLUENCE 2.0 is meant to support the tailoring of optimal individual follow-up strategies aiming at the detection of LRR and SP as potentially curable events. Therefore, discrimination was selected as key measure in the comparison of the three tested statistical approaches predicting these events. In contrast to this, knowing the risk of DM can only serve an informative purpose; predicting DM means predicting the risk of a palliative situation in which classical follow-up would not make sense, anymore. Consequently, calibration was considered the central indicator in selecting the most appropriate statistical approach to predict this event.

Software and online model

For the analyses, R version 3.6.3 (R Foundation for Statistical Computing, Vienna, Austria; http://www.R-project.org/) was used. To develop the RSF, COX and PAR algorithms, the packages "randomForestSRC[24], "survival"[25, 26], and "rstpm2"[27, 28] were used . For the performance analyses, we employed the packages "timeROC"[29], and "boot".[30] Based on the best performing statistical approach, an online calculator of the INFLUENCE 2.0 model was developed and made available on www.evidencio.org , an online platform for medical prediction models.

Results

Descriptive statistics

In total, 17,014 patients with an invasive adenocarcinoma of the breast diagnosed in 2007, 2008, or the first quarter of 2012 were identified from the NCR. Of those, 13,494 met all eligibility criteria. Supplementary figure 1 gives a detailed overview of the exclusion process.

All relevant characteristics of the patient cohort are shown in table 1. The majority of the patients (98%) had a pT1 or pT2 tumor, no lymph node involvement (65%), low tumor grade (70% grade 1 or 2), and a unifocal tumor (85%). In about 60% of the cases, breast-conserving surgery was performed. Adjuvant radiation therapy and adjuvant chemotherapy were administered in 67% and 38% of the patients, respectively. Over 80% of the patients were ER and/or PR positive and about 40% of them received antihormonal therapy. Of all patients, less than 15% were Her2-positive, of whom 60% received antibody-treatment. Within five years, 385 (2.8%), 411 (3.0%), and 848 (6.3%) patients developed a LRR, SP, or DM, respectively, as their first event. A total of 11,839 (87.7%) remained free of recurrence.

Table 1. Patient characteristics

Variable	Option	N (%) total = 13494
Inclusion year	Inclusion year = 2007	5508 (41.1%)
	Inclusion year = 2008	5621 (41.7%)
	Inclusion year = 2012	2365 (17.5%)
Age-group	50 – 59	3091 (22.9%)
	60 – 69	3635 (26.9%)
	70 – 79	3531 (26.2%)
	≥80	3237 (24.0%)
Grading	1	3409 (25.3%)
	2	6047 (44.8%)
	3	4038 (29.9%)
pT	pT1	8692 (64.4%)
	pT2	4514 (33.5%)
	pT3	288 (2.1%)
pN	pN0	8782 (65.1%)
	pN1	3493 (25.9%)
	pN2	790 (5.9%)
	pN3	429 (3.2%)
Multifocality	No	11425 (84.7%)
	Yes	2069 (15.3%)
Surgery	Breast conserving surgery	7942 (58.9%)
	Mastectomy	5552 (41.1%)
Chemotherapy	No	8366 (62%)
	Yes	5128 (38%)
Radiotherapy	No	4403 (32.6%)
	Yes	9091 (67.4%)
Hormonal therapy	HR+ & no therapy	6560 (48.6%)
	HR+ & therapy	4881 (36.2%)
	HR-	2053 (15.2%)
Targeted therapy	HER2+ & no therapy	678 (5.0%)
	HER2+ & therapy	1015 (7.5%)
	HER2-	11801 (87.5%)
First event	LRR	385 (2.8%)
	SP	411 (3.0%)
	DM	848 (6.3%)
	None	11839 (87.7%)

Abbreviations: N = number of patients, pT = pathological tumor stage, pN = pathological nodal stage, LRR = Locoregional Recurrence, SP = Secondary Primary, DM = Distant metastasis

Table 2. Calibration results

Outcome	Time (years)	COX			PAR			RSF		
		ICI	E50	E90	ICI	E50	E90	ICI	E50	E90
Locore-gional	1	0.0005	0.0003	0.0007	0.0019	0.0014	0.0016	**0.0009**	**0.0006**	**0.0015**
	2	0.0011	0.0008	0.0016	0.0030	0.0021	0.0025	**0.0023**	**0.0018**	**0.0033**
	3	0.0017	0.0013	0.0027	0.0028	0.0022	0.0032	**0.0044**	**0.0023**	**0.0039**
	4	0.0023	0.0018	0.0037	0.0023	0.0019	0.0036	**0.0064**	**0.0035**	**0.0055**
	5	0.0027	0.0022	0.0044	0.0028	0.0021	0.0042	**0.0094**	**0.0057**	**0.0096**
Second primary	1	0.0010	0.0009	0.0017	0.0030	0.0024	0.0060	**0.0010**	**0.0008**	**0.0017**
	2	0.0013	0.0011	0.0023	0.0020	0.0016	0.0036	**0.0018**	**0.0014**	**0.0032**
	3	0.0016	0.0015	0.003	0.0022	0.0021	0.0033	**0.0037**	**0.0028**	**0.0063**
	4	0.0021	0.0018	0.0038	0.0021	0.0019	0.0036	**0.0049**	**0.004**	**0.0088**
	5	0.0025	0.0022	0.0044	0.0026	0.0022	0.0047	**0.0073**	**0.0059**	**0.0135**
Distant metastasis	1	**0.0007**	**0.0004**	**0.0012**	0.0030	0.0024	0.0034	0.0014	0.0008	0.0017
	2	**0.0017**	**0.0009**	**0.0024**	0.0055	0.0035	0.0062	0.0060	0.0036	0.0058
	3	**0.0026**	**0.0015**	**0.0038**	0.0054	0.0035	0.0075	0.0119	0.0074	0.0123
	4	**0.0032**	**0.0020**	**0.0045**	0.0043	0.0028	0.0075	0.0166	0.0106	0.0201
	5	**0.0037**	**0.0024**	**0.0054**	0.0033	0.0024	0.0054	0.0216	0.0143	0.0299

The displayed results represent the optimism-corrected values. The ICI is the integrated calibration index, E50 is the median absolute difference between observed and expected, and E90 is the 90th percentile of the absolute difference. Bold results display the results for the models that were selected as the best performing model. For LRR, and SP, the decision was primarily based on the discrimination.

Internal validation and comparison of modelling approach

For the prediction of LRR, the optimism corrected discrimination is displayed graphically in figure 1a. The RSF approach shows significantly higher AUC values after the first year until the end of year five compared to the COX and PAR approaches. The AUCs of the COX, PAR, and RSF models at year five were 0.73 (95%CI: 0.72 - 0.73), 0.73 (95%CI: 0.72 – 0.73), and 0.75 (95%CI: 0.74 – 0.76), respectively. On average, the optimism in the AUCs was higher for the RSF approach (optimism = 0.04) than for the PAR approach (optimism = 0.02) and the Cox approach (optimism = 0.01). Calibration is displayed in table 2; it shows that all three modelling approaches show adequate calibration at all tested time points, reflected by an ICI, E50, and E90 below 0.01. Based on these outcomes, the RSF was selected for the final INFLUENCE 2.0 model as optimal approach to predict the risk for LRR.

For the prediction of SP, the RSF approach shows superior performance concerning discrimination compared to the other approaches at all time points (figure 1b). The optimism-corrected AUCs at year five for the COX, PAR, and RSF approaches were 0.62 (95%CI: 0.60 – 0.62), 0.62 (95%CI: 0.60 – 0.62), 0.67 (95%CI: 0.65 – 0.68), respectively. On average, the optimism in the AUCs for the RSF approach was higher (optimism = 0.08) than for the PAR approach (optimism = 0.03) and the Cox approach (optimism = 0.02). Calibration is displayed in table 2 and shows that all three modelling approaches show adequate calibration at all tested time points, reflected by an ICI, E50, and E90 below 0.01, with an exception for the RSF approach at year 5 (E90 = 0.0135). Finally, the RSF was selected as best performing approach to predict the risk for SP.

For the prediction of DM, calibration is displayed in table 2; it shows that the COX approach proofed to be best calibrated. Generally, all three statistical approaches showed mostly adequate calibration at each of the tested annual time points, reflected by an ICI below 0.01. However, the RSF approach seems to be associated with a lower level of accuracy at some time points. For the years 3, 4, and 5 its ICI is 0.012, 0.017, and 0.022, respectively. The performance of the approaches concerning discrimination is displayed in figure 1c. With exception of the first year, all three modelling approaches showed similar performance on discrimination. The optimism-adjusted AUCs at year five for the COX, PAR, and RSF approaches were 0.77 (95%CI: 0.77 – 0.78), 0.77 (95%CI: 0.77 – 0.78), and 0.78 (95%CI: 0.77 – 0.78), respectively. On average, the optimism in the AUCs for the RSF approach was higher (optimism = 0.02) than for the PAR approach (optimism = 0.01) and the Cox approach (optimism = 0.004). Based on these results, the COX approach was selected for the final INFLUENCE 2.0 model to predict the risk for DM. Table 3 gives an overview of the underlying coefficients.

Online calculator

The final INFLUENCE 2.0 model returns risk predictions for LRR, SP, and DM based on the selected statistical approaches; an easy-to use online risk calculator is available via: https://www.evidencio.com/models/show/2238. The online calculator estimates the risks and the 95% confidence intervals based on the 200 bootstrapped models.

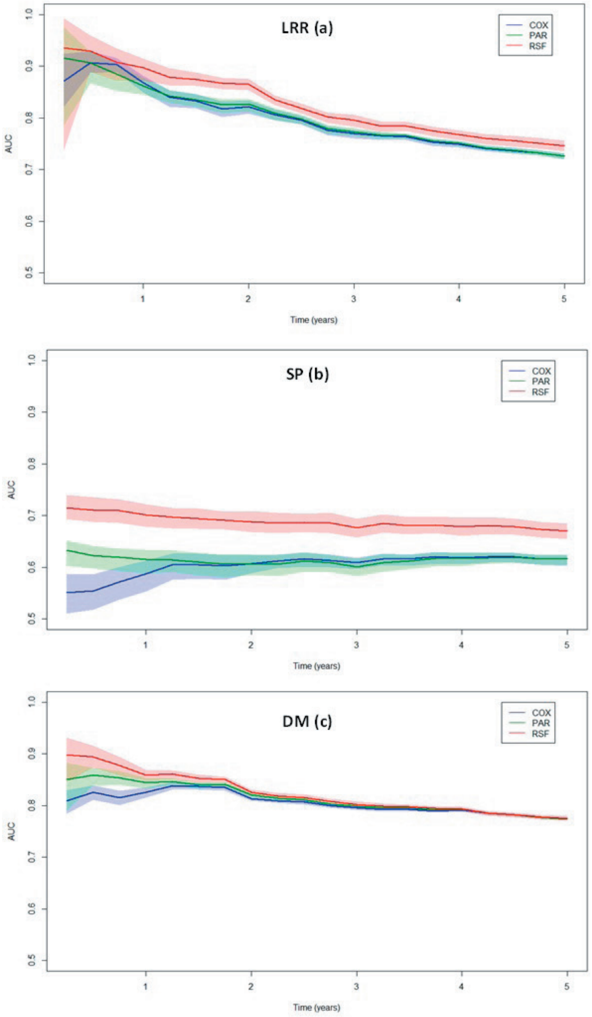
Table 3. Coefficients of the Cox regression model selected to predict distant metastasis

Variable	Option	Hazard ratio	95% CI
Age	<60	Reference	
	60-70	0.916	0.7530 - 1.1133
	70-80	0.841	0.6775 - 1.0443
	≥80	1.240	0.9693 - 1.5857
Grade	I	Reference	
	II	2.359	1.7941 - 3.1004
	III	4.081	3.0610 - 5.4404
Tumor stage	pT1	Reference	
	pT2	2.280	1.9338 - 2.688
	pT3	2.499	1.7857 - 3.4976
Nodal stage	pN0	Reference	
	pN1	1.879	1.5714 - 2.2469
	pN2	4.109	3.2221 - 5.2395
	pN3	7.503	5.8160 - 9.6785
Multifocality	No	Reference	
	Yes	1.242	1.0426 - 1.4783
Surgery	Breast conserving surgery	Reference	
	Mastectomy	0.915	0.7405 - 1.1302
Chemotherapy	No	Reference	
	Yes	0.676	0.5436 - 0.8413
Radiotherapy	No	Reference	
	Yes	0.913	0.7332 - 1.1371

Table 3. Continued

Variable	Option	Hazard ratio	95% CI
Hormone receptor status & treatment	Negative	Reference	
	Positive with treatment	0.506	0.4258 - 0.6014
	Positive without treatment	0.751	0.5868 - 0.9603
HER2-status & treatment	Negative	Reference	
	Positive with treatment	0.704	0.5301 - 0.9360
	Positive without treatment	1.188	0.9634 - 1.4642
	Time		
Baseline hazard	Year 1	0.002417	
	Year 2	0.007263	
	Year 3	0.012676	
	Year 4	0.017466	
	Year 5	0.021168	

Figures 1a, 1b, and 1c Area under the Receiver operating characteristic curve (AUC) per quarter year



Discussion

In this study, we developed a model predicting the risks for LRR, SP, and DM within 5-years after primary surgery for patients with curatively treated non-metastatic breast cancer. For this purpose, three different statistical approaches were compared concerning discrimination and calibration. Based on discrimination, the RSF approach showed superior performance in the prediction of LRR and SP as compared to COX and PAR. However, the COX approach showed a higher level of agreement between the predicted and observed risks, which was decisive for the selection of the best performing approach for the prediction of DM, as the discriminatory performance concerning this event was similar between the modelling approaches.

Comparison to the original INFLUENCE nomogram and other related prediction models
Compared to the original INFLUENCE nomogram, the INFLUENCE 2.0 model comes with a variety of updates leading to improved flexibility and a broader application range regarding predictable events. Concerning clinical decision making, discrimination is arguably the most relevant indicator for model performance. The AUC of the five annual prediction models of the original INFLUENCE-nomogram which is exclusively concentrating on the endpoint LRR starts with 0.84 for the first year and decreases to 0.62 in the fifth year.[11] A direct comparison to the AUC values of the INFLUENCE2.0 model is not possible due to differences in outcome definition (i.e. the original INFLUENCE nomogram predicted the risk of LRR in a given year, assuming the patient was event-free at the start of that year). While discerning high- and low-risk patients for SP seems to be difficult reflected by an AUC between 0.6 and 0.7 for all tested approaches, all other adjusted AUC values reported in this paper were found to be higher than 0.7, indicating a fairly good discriminative ability of the new INFLUENCE 2.0 model. Notwithstanding this, our study shows the importance of finding the optimal statistical approach and model architecture:

In contrast to the logistic regression model of the original INFLUENCE nomogram, a Cox regression-based approach offers the advantage that it can deal with censoring and make time-dependent predictions for periods of variable length in just one model. However, it is based on the proportional hazards assumption and, therefore, cannot incorporate changing event rates over time as easily as the other modelling approaches, which might be the reason that most of the time it showed the lowest discriminative ability. Technically, this problem should be solved by the parametric spline function. However, when concentrating on the time-dependent performance it is evident that both semiparametric models suffer from a lower level of discriminative ability throughout the whole prediction time compared to the more flexible non-parametric RSF-approach, which requires more computational power but is not subjected to any preliminary assumptions. Still, the predictor-outcome relation is unknown for the RSF-approach, making it difficult to assess the impact of specific characteristics on the estimated risks. The creation of an online calculator improves the transparency of the RSF approach but is explicitly not meant to be used for what-if scenarios. Concerning calibration, the Cox model showed the highest level of agreement between predicted and observed risks, reflected by an average ICI close to 0. To assess the adequacy of calibration, the ICI values should be compared with the observed absolute event rates which were 2.8%, 3.0%, and 6.3% at year 5 for LRR, SP, and DM, respectively; therefore, an ICI below 0.01 was regarded as adequate. In view of this threshold, the RSF approach's calibration with for example an ICI up to 0.022 for the prediction of DM at year 5 has to be regarded as suboptimal.

Apart from INFLUENCE, several other interesting prediction tools on breast cancer recurrence have been developed. For instance, Corso et al.[31] also came up with a time-dependent prediction model for LRR. At five years after surgery, the cumulative AUC in their validation cohort was 0.77 for patients with breast-conserving surgery and 0.69 with mastectomy. The RSF-based INFLUENCE 2.0 model is characterized by a 5-year AUC of 0.77 regardless of the primary surgical procedure, indicating a tendency towards better discrimination. Giardielo et al.[32] compared three models predicting the risk of contralateral breast cancer (CBC): the Manchester formula, CBCrisk, and PredictCBC in patients with invasive breast cancer (BC). They used data of 132,756 patients (4682 CBC) from 20 international studies with a median follow-up of 8.8 years. The AUCs at five years were: 0.59 (95% Prediction interval (PI): 0.54 to 0.64) for CBCrisk, 0.61 (95% PI: 0.59 to 0.63) for the Manchester formula, 0.63 (95% PI: 0.52 to 0.74) and 0.59 (95% PI: 0.46 to 0.71) for PredictCBC-1A (for settings where BRCA1/2 mutation status is available) and PredictCBC-1B (for the general population), respectively. They concluded that the current CBC risk prediction models provide only moderate discrimination, and the Manchester formula was poorly calibrated. Therefore, the RSF-based INFLUENCE 2.0 model on SP with its adjusted AUC of 0.67 represents an important step towards better risk estimations on this endpoint in the general population, even without information on genetics.

Limitations and strengths in clinical use

Although the bootstrapped model validation showed adequate performance, some limitations of the INFLUENCE 2.0 model should be taken into account. As stated initially, no patients with neo-adjuvant treatment or not invasive in-situ-tumors were included. Second, the set of predictors was obviously limited to the items collected by the NCR. Other potential predictors such as e.g. Ki67 were not registered due to comparability-issues caused by differing determination-methods of different pathology labs.[33] Moreover, information concerning family history, genetic markers or gene signatures such as Mammaprint® or Oncotype were only available for a small number of patients and could therefore not be included in the analyses. Even though pT3 patients were included in the analysis, the majority (98%) of data used to develop the model comprised pT1 and pT2 patients. The use of imputation techniques to deal with missing data could have resulted in the inclusion of more pT3 patients. However, only 42 pT3 patients, which is equivalent to a 0.3% share of all patients (data not shown), were excluded due to missing data. No other subgroup was misrepresented in our dataset, and the sample size was deemed sufficient to perform a complete-case analysis. Future research is required to broaden the applicability of the INFLUENCE 2.0 model and to improve its performance, e.g. by including some of the above mentioned additional predictors. Further external validation studies and potential model updates should aim to enable model use for patients who received neoadjuvant treatment or to extend the risk prediction period towards 10 years after primary surgery.

Despite these limitations, INFLUENCE 2.0 in its current state can provide substantial added value for patients, health professionals and the health care system as a whole if it is used to tailor follow-up for patients with curatively treated non-metastatic breast cancer. Using individual risk predictions could effectively contribute to decrease the number of potentially unnecessary follow-up visits for patients at a low risk of recurrence. Thus, the overall sensitivity of the breast cancer follow-up program would increase and psychological stress and costs caused by unnecessary examinations in low-risk patients could be avoided.[34] However, the successful implementation of risk-based follow-up requires a truly shared decision process which currently is often not reflected by clinical reality. A review of 42 studies revealed that patients were insufficiently involved in the decision-making process that affected their follow-up, indicating a need for further improvement.[35] With its easy-to-use online interface, the INFLUENCE 2.0 model might be an important step towards more direct patient participation, as recommended by the 2019 guideline on diagnosis, treatment and follow-up for early breast cancer[36] provided by the European society for medical oncology (ESMO): "The interval of [follow-up] visits should be adapted to the risk of relapse and patients' needs".[7,36] Following this recommendation, the risk estimations provided by INFLUENCE 2.0 do not necessarily have to be used together with strict thresholds to discern between high and low risk patients who should or should not receive follow up, but can serve as a reliable source of information to find the optimal follow-up strategy, which also has to account for other important factors like the optimal quality of life or patient preference. Further studies are ongoing to assess the impact of implementing the model in the shared decision-making process between clinicians and patients.

Conclusion

INFLUENCE 2.0 is a flexible risk prediction model for breast cancer recurrence and secondary primary tumors that might be a valuable aid for health care professionals. Together with an appropriate strategy to use its individual, event-specific, time-dependent risk predictions it can support the establishment of a personalized breast cancer follow-up scheme in daily practice.

References

1. IKNL - integraal kankercentrum Nederland Cijfers over kanker: Incidentie, Aantal Borstkanker. https://iknl.nl/nkr-cijfers?fs%7Cepidemiologie_id=506&fs%7Ctumor_id=1%2C280&fs%7Cregio_id=530&fs%7Cperiode_id=545%2C546%2C547%2C548%2C549%2C550%2C551%2C552%2C553%2C554%2C555%2C556%2C557%2C558%2C559%2C560%2C561%2C562%2C563%2C564%2C565%2C566%2C567%2C568%-2C569%2C570%2C571%2C572%2C544%2C543%2C542%2C541&fs%7Cgeslacht_id=622&fs%7Cleeftijdsgroep_id=655&fs%7Cjaren_na_diagnose_id=665&fs%7Ceenheid_id=681&cs%7Ctype=line&cs%7CxAxis=periode_id&cs%7Cseries=tumor_id&ts%7CrowDimensions=periode_id&ts%7CcolumnDimensions=tumor_id&lang%7Clanguage=nl
2. Stewart BW (ed) (2014) World Cancer Report 2014. IARC Press, Lyon
3. Holleczek B, Arndt V, Stegmaier C et al. (2011) Trends in breast cancer survival in Germany from 1976 to 2008--a period analysis by age and stage. Cancer Epidemiol 35:399–406. https://doi.org/10.1016/j.canep.2011.01.008
4. Hübner J, Katalinic A, Waldmann A et al. (2020) Long-term Incidence and Mortality Trends for Breast Cancer in Germany. Geburtshilfe Frauenheilkd 80:611–618. https://doi.org/10.1055/a-1160-5569
5. Yoshimura A, Ito H, Nishino Y et al. (2018) Recent Improvement in the Long-term Survival of Breast Cancer Patients by Age and Stage in Japan. J Epidemiol 28:420–427. https://doi.org/10.2188/jea.JE20170103
6. IKNL - integraal kankercentrum Nederland Cijfers over kanker: Overleving borst. https://iknl.nl/nkr-cijfers?fs%7Cepidemiologie_id=507&fs%7Ctumor_id=282&fs%7Coverlevingssoort_id=512&fs%7Cperiode_van_diagnose_id=580%2C579%2C578%2C577%2C576%2C575&fs%7Cjaren_na_diagnose_id=665%2C666%2C667%2C668%2C669%2C670%2C671%2C672%2C673%2C674%2C675%2C676&cs%7Ctype=line&cs%7CxAxis=jaren_na_diagnose_id&cs%7Cseries=periode_van_diagnose_id&ts%7CrowDimensions=periode_van_diagnose_id&ts%7CcolumnDimensions=jaren_na_diagnose_id&lang%7Clanguage=en
7. NABON (2012) Breast Cancer, Dutch Guideline, Version 2.0. http://www.oncoline.nl/mammacarcinoom
8. Pennery E, Mallet J (2000) A preliminary study of patients' perceptions of routine follow-up after treatment for breast cancer. Eur J Oncol Nurs 4:138-45; discussion 146-7. https://doi.org/10.1054/ejon.2000.0092
9. Kiebert GM, Welvaart K, Kievit J (1993) Psychological effects of routine follow up on cancer patients after surgery. Eur J Surg 159:601–607
10. Loprinzi CL (1995) Follow-up Testing for Curatively Treated Cancer Survivors. JAMA 273:1877. https://doi.org/10.1001/jama.1995.03520470085038
11. Witteveen A, Vliegen IMH, Sonke GS et al. (2015) Personalisation of breast cancer follow-up: a time-dependent prognostic nomogram for the estimation of annual risk of locoregional

recurrence in early breast cancer patients. Breast Cancer Res Treat 152:627–636. https://doi.org/10.1007/s10549-015-3490-4

12. T. Gamucci, A. Vaccaro, F. Ciancola et al. (2013) Recurrence risk in small, node-negative, early breast cancer: a multicenter retrospective analysis. J Cancer Res Clin Oncol 139:853–860. https://doi.org/10.1007/s00432-013-1388-2
13. McGuire S (2016) World Cancer Report 2014. Geneva, Switzerland: World Health Organization, International Agency for Research on Cancer, WHO Press, 2015. Adv Nutr 7:418–419. https://doi.org/10.3945/an.116.012211
14. Moossdorff M, van Roozendaal LM, Strobbe LJA et al. (2014) Maastricht Delphi consensus on event definitions for classification of recurrence in breast cancer research. J Natl Cancer Inst 106. https://doi.org/10.1093/jnci/dju288
15. George B, Seals S, Aban I (2014) Survival analysis and regression models. J Nucl Cardiol 21:686–694. https://doi.org/10.1007/s12350-014-9908-2
16. Ishwaran H (2007) Variable importance in binary regression trees and forests. Electron J Statist 1:519–537. https://doi.org/10.1214/07-EJS039
17. Ishwaran H, Kogalu UB (2007) Random survival forests for R. R News 7(2), 25--31.
18. Breiman L (2001) Random Forests. Machine Learning 45:5–32. https://doi.org/10.1023/A:1010933404324
19. Schemper M, Smith TL (1996) A note on quantifying follow-up in studies of failure time. Controlled clinical trials 17. https://doi.org/10.1016/0197-2456(96)00075-x
20. Moons KGM, Kengne AP, Grobbee DE et al. (2012) Risk prediction models: II. External validation, model updating, and impact assessment. Heart 98:691–698. https://doi.org/10.1136/heartjnl-2011-301247
21. Austin PC, Steyerberg EW (2019) The Integrated Calibration Index (ICI) and related metrics for quantifying the calibration of logistic regression models. Stat Med 38:4051–4065. https://doi.org/10.1002/sim.8281
22. Kamarudin AN, Cox T, Kolamunnage-Dona R (2017) Time-dependent ROC curve analysis in medical research: current methods and applications. BMC Medical Research Methodology 17:53. https://doi.org/10.1186/s12874-017-0332-6
23. Steyerberg EW (2009) Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. Statistics for Biology and Health. Springer-Verlag New York, New York, NY
24. Ishwaran H, Kogalur UB Package ‘randomForestSRC’: Fast Unified Random Forests for Survival, Regression, and Classification (RF-SRC)
25. Therneau TM, Lumley T, Atkinson, E., Crowson, C. Package ‘survival’: Survival Analysis
26. Therneau TM, Grambsch PM (2001) Modeling survival data: Extending the Cox model, 2. ed. Statistics for Biology and Health. Springer, New York
27. Clements Mea rstpm2: Smooth Survival Models, Including Generalized Survival Models
28. Liu X-R, Pawitan Y, Clements M (2018) Parametric and penalized generalized survival models. Stat Methods Med Res 27:1531–1546. https://doi.org/10.1177/0962280216664760

29. Blanche P, Dartigues J-F, Jacqmin-Gadda H (2013) Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. Stat Med 32:5381–5397. https://doi.org/10.1002/sim.5958

30. Davison AC, Hinkley DV (2009) Bootstrap methods and their application, 11. print. Cambridge series on statistical and probabilistic mathematics. Cambridge Univ. Press, Cambridge, NY

31. Corso G, Maisonneuve P, Massari G et al. (2020) Validation of a Novel Nomogram for Prediction of Local Relapse after Surgery for Invasive Breast Carcinoma. Ann Surg Oncol 27:1864–1874. https://doi.org/10.1245/s10434-019-08160-7

32. Giardiello D, Steyerberg EW, Hauptmann M et al. (2019) Prediction and clinical utility of a contralateral breast cancer risk model. Breast Cancer Res 21:144. https://doi.org/10.1186/s13058-019-1221-1

33. Penault-Llorca F, Radosevic-Robin N (2017) Ki67 assessment in breast cancer: an update. Pathology 49. https://doi.org/10.1016/j.pathol.2016.11.006

34. Geurts SME, Vegt F de, Siesling S et al. (2012) Pattern of follow-up care and early relapse detection in breast cancer patients. Breast Cancer Res Treat 136:859–868. https://doi.org/10.1007/s10549-012-2297-9

35. Ligt KM de, van Egdom LSE, Koppert LB et al. (2019) Opportunities for personalised follow-up care among patients with breast cancer: A scoping review to identify preference-sensitive decisions. Eur J Cancer Care (Engl) 28:e13092. https://doi.org/10.1111/ecc.13092

36. Cardoso F, Kyriakides S, Ohno S et al. (2019). Early breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Annals of Oncology, 30(8), 1194-1220. https://doi.org/10.1093/annonc/mdz173

PART II

Prostate Cancer

Chapter 5

External validation of prediction models predicting the probability of lymph node involvement in prostate cancer patients

Tom A. Hueting, Erik B. Cornel, Diederik M. Somford, Hanneke Jansen, Jean-Paul A. van Basten, Rick G. Pleijhuis, Ruben A. Korthorst, Job A.M. van der Palen, Hendrik Koffijberg

European Urology Oncology 5 411 – 417 (2018)
DOI: 10.1016/j.euo.2018.04.016

The supplementary data of this chapter are available online at https://dx.doi.org/10.1016/j.euo.2018.04.016

Abstract

Background: Multiple statistical models predicting lymph node involvement (LNI) in prostate cancer (PCa) patients exist to support clinical decision-making regarding extended pelvic lymph node dissection (ePLND). In this study, we aimed to validate models predicting LNI in Dutch PCa patients.

Methods: Sixteen prediction models were validated using a patient cohort of 1,001 men who underwent ePLND. Patient characteristics included serum prostate specific antigen (PSA), clinical tumor (cT) stage, primary and secondary Gleason scores, number of biopsy cores taken, and number of positive biopsy cores. Model performance was assessed using the area under the curve (AUC) of the receiving operator characteristic (ROC) curve. Calibration plots were used to visualize over- or underestimation of the models.

Results: Lymph node involvement was identified in 276 (28%) patients. Patients with LNI had a higher PSA, higher primary Gleason pattern, higher Gleason score, higher number of harvested nodes, higher number of positive biopsy cores, and higher cT stage, compared to patients without LNI. Predictions generated by the 2012 Briganti nomogram (AUC = 0.76) and the MSKCC web-calculator (AUC = 0.75) were found most accurate. Calibration had a decisive role in the selection of the most accurate models due to overlapping confidence intervals for the AUCs. Underestimation of LNI probability was present in patients with a predicted probability <20%. The omission of model updating was a limitation of this study.

Conclusion: Models predicting LNI in PCa patients were externally validated in a Dutch patient cohort. The 2012 Briganti and the MSKCC nomograms were the most accurate prediction models available.

Introduction

Prostate carcinoma (PCa) is the second most frequently diagnosed cancer among males worldwide, with the highest incidence rates in the United States (168.3/100,000 cases), followed by France (132.1/100,000 cases), and Australia (111.1/100,000 cases).[1] The incidence rate in the Netherlands is 92.4/100,000 cases.[2] Incidence rates vary highly between countries due to increased use of prostate specific antigen (PSA) testing since the late 1980s.[3] The treatment options and prognosis of patients with PCa strongly depend on the presence of metastasis. Metastases are predominantly located in the bone and in the lymph nodes. The risk of lymph node involvement (LNI) depends on tumor aggressiveness and tumor volume, which is estimated by, digital rectal examination (DRE) of the prostate (clinical tumor stage (cT)), serum PSA, and tissue patterns (i.e. Gleason score) determined on prostate biopsies.[4] Most lymphogenic metastasis occur in the pelvic lymph nodes, which are readily accessible for surgical removal.[5]

An extended pelvic lymph node dissection (ePLND) is the most accurate method to detect LNI. However, ePLND is invasive and has a risk of complications such as lymph leakage, lymph edema, and trombo-embolic events[6,7], and is therefore offered in selected cases such as patients scheduled for radical prostatectomy (RP) or precluding external beam radiotherapy with curative intent.

To limit the impact of the potential morbidity of ePLND for all localized PCa patients, selection of candidates has been suggested by using cut-off values for risk of LNI. Current European PCa guideline states that the indication for ePLND is based on a risk estimation of lymph node metastasis over 5% by using a prediction model.[4] Whereas the Dutch guideline recommends a 10% threshold[8], and the American guideline recommends a 2% threshold.[9]

Several prediction models have been developed to predict the probability of LNI in PCa patients. Predicting LNI is possible using, for example, artificial neural networks, logistic regression, classification and regression trees (CART), and simple linear formulas. Most of the prediction models have been updated to reflect recent clinical practice and new insights regarding ePLND.

Predictive models such as the nomograms by Briganti, Memorial Sloan Kettering Cancer Center (MSKCC), Partin tables, and Roach formula are recommended in several guidelines.[4,8,9] However, no external validation has yet been performed in the Netherlands. This study therefore aimed to externally validate existing models predicting LNI in a Dutch PCa patient cohort.

Materials and methods

Patient data was collected retrospectively in the Canisius Wilhelmina hospital (CWZ) in Nijmegen, The Netherlands, including patients who underwent ePLND and concomitant RP between October 2008 and December 2016, and Ziekenhuisgroep Twente hospital (ZGT) in Hengelo, The Netherlands, including patients who underwent ePLND (either with or without concomitant RP) between December 2014 and May 2017. In addition, data from the ProZIB initiative was used. The main goal of the ProZIB initiative is to get insight in the clinical practice concerning prostate cancer care in the Netherlands and to evaluate quality of care. Patients diagnosed with prostate cancer between Oct 2015 - Apr 2016, were identified through the population-based Netherlands Cancer Registry (NCR). ProZIB data contained patients who underwent ePLND either with or without concomitant RP. Patients treated in the CWZ or ZGT were removed from the ProZIB database to avoid duplicates.

Patient pseudonymity was guaranteed. Patients were included for validation if they underwent ePLND and had histopathological results available (i.e. PSA, cT-stage, Gleason score, biopsy cores, harvested lymph nodes, and positive lymph nodes). Patients with less than ten harvested lymph nodes were excluded to make sure the performed PLND was adequate. The applied ePLND template comprised the removal of nodes overlying the external iliac vessels, internal iliac artery, and the nodes located within the obturator fossa. Optionally, the areas of the common iliac artery and the pre-sacral area can be included. Patient data with missing information regarding biopsy cores taken with corresponding positivity were included for validation, but could not be used for validation in certain models using biopsy core information as predictor.

Every validated model used at least pre-operative PSA (ng/ml), cT-stage, and Gleason score as predictors. The cT-stage was defined according to the International Union Against Cancer (UICC) TNM classification, edition 7.0.[10] With regard to the Gleason score, either the Gleason sum score was used as a predictor or the primary and secondary patterns were used as separate predictors. Some models also used measures based on biopsy cores taken, such as percentage of positive cores or total amount of positive and negative cores. One model used the total amount of excised lymph nodes as a predictor. The CWZ, ZGT, and ProZIB databases contained information on 270, 109, and 622 patients, respectively. Patient data regarding biopsy cores was missing in 18, 1, and 57 (total 76, 7.6%) patients. After merging the three databases, a total of 1,001 patients were eligible for validation of models without biopsy core predictors and 925 patients for models including biopsy core predictors.

Descriptive statistics were reported using frequencies for categorical variables, means with standard deviations for normal distributed continuous variables, and medians with interquartile ranges for non-normal distributed continuous variables. Characteristics of patients with and without histologically proven LNI were reported separately. Significant differences ($p < 0.05$) between both groups were assessed using Fishers exact test for categorical variables, independent sample t-tests for normally distributed continuous variables, and Mann-Whitney U test for non-normal distributed variables.

A total of 16 models were validated, methods for the selection of the models were added as supplementary data. Model coefficients were derived and made available on www.evidencio.com for validation purposes. Evidencio is an online platform that allows researchers to translate prediction models into user-friendly online calculators, facilitating the application and (external) validation of prediction models. The Area under the curve (AUC) of the receiving operator characteristic (ROC) curve was used to quantify model accuracy. Model over-and underestimation was assessed using calibration plots. Calibration plots show the agreement between the predicted and the observed LNI. Characteristics of the calibration were described in terms of calibration slope and intercept. The slope reflects how well the predictions fit with the observed outcome over the range of the predicted risks and is ideally equal to 1. The intercept (i.e. calibration-in-the-large) quantifies if the average predicted risk corresponds to the average observed outcome, and is preferably equal to 0.[11] Given the extent of the validation, only the four best performing models were reported complete with ROC curves and calibration plots. These four models were assessed more thoroughly by looking at the calibration in a subset of patients with a predicted low probability of LNI (<20%), as in these patients the question whether or not to perform ePLND is particularly relevant.[4,8] Validation was based on the intercepts and coefficients of the original models, i.e. no model update was performed for the current validation cohort.

Table 1. Baseline characteristics of the validation cohort.

	Positive lymph nodes	Negative lymph nodes	Total	p-value
No (%)	276 (27.6%)	725 (72.4%)	1001 (100%)	
Treatment (%)				
RP	169 (61.2%)	621 (85.7%)	790 (78.9%)	
No RP	107 (38.8%)	104 (14.3%)	211 (21.1%)	
Age, yr				
Mean (SD)	66.5 (6.2)	66.5 (5.8)	66.5 (5.9)	0.95
PSA, ng/ml				
Median (IQR)	14.7 (7.6 – 28.0)	9.9 (6.7 – 16.4)	10.6 (7.0 – 19.6)	<0.0001
Biopsy cores (total)				
Mean (SD)	10.0 (2.2)	10.2 (2.6)	10.2 (2.5)	0.18
Biopsy cores (positive)				
Mean (SD)	7.2 (2.9)	5.2 (2.8)	5.7 (3.0)	<0.0001
Harvested lymph nodes				
Median (IQR)	17 (13 – 22)	15 (12 – 20)	16 (12 – 21)	0.0005
Clinical T stage				
cT1	32 (11.6%)	229 (31.6%)	261 (26.1%)	0.0005
cT2	128 (46.4%)	335 (46.2%)	463 (46.3%)	
cT3	109 (39.5%)	156 (21.5 %)	265 (26.5%)	
cT4	7 (2.5%)	5 (0.7%)	12 (1.2%)	
Primary Gleason pattern				
≤ 3	92 (33.3%)	380 (52.4%)	472 (47.2%)	<0.0001
≥ 4	184 (66.7%)	345 (47.6%)	529 (52.8%)	
Secondary Gleason pattern				
≤ 3	73 (26.4 %)	229 (31.6%)	302 (30.2%)	0.12
≥ 4	203 (73.6 %)	496 (68.4%)	699 (69.8%)	
Gleason Score				
≤6	11 (4.0 %)	103 (14.2%)	114 (11.4%)	0.0005
7 (3 + 4)	70 (25.4%)	237 (32.7%)	307 (30.7%)	
7 (4 + 3)	56 (20.3%)	117 (16.1%)	173 (17.3%)	
8	62 (22.5%)	163 (22.5%)	225 (22.5%)	
9	68 (24.6%)	95 (13.1%)	163 (16.3%)	
10	9 (3.3%)	10 (1.4%)	19 (1.9%)	

Differences in between group of patients with positive and negative lymph nodes were compared. IQR = Interquartile Range, PSA = Prostate specific antigen, RP = Radical Prostatectomy, SD = Standard deviations

Results

Baseline characteristics of the validation cohort are displayed in table 1. Significant differences were found between patients with and without LNI in PSA, positive biopsy cores, primary Gleason score, cT stage, and Gleason sum in the dataset.

An overview of all validated models, including predictors and accuracy estimates achieved on the validation cohort, is displayed in table 2. The more recent updated models performed better than the corresponding original models. The model by Briganti et al. published in 2012 performed better than the Briganti models from 2006-2007. The MSKCC model including biopsy core information as predictors performed better than the Godoy nomogram as well as the MSKCC model without biopsy cores. The most recent update of the Partin tables by Tosoian et al. performed slightly better than the Makarov

Partin tables, and the Eifler Partin tables. Out of the three formulas by Roach, Nguyen, and Yu (Yale formula), the most recent formula (Yale formula) performed best.

The ROC plots showing the AUC for the best-performing model of each of the four types of models are displayed in figure 1. Figure 2 displays the calibration plots for these four models, including separate plots for a subgroup of patients with predicted risks below 20%.

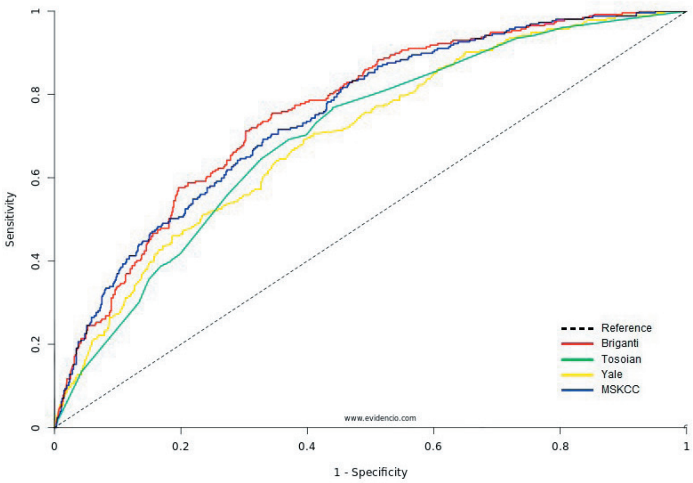
Briganti and MSKCC showed comparable calibration, both underestimating the risk of LNI in patient groups with an observed probability of LNI below 25-40% and overestimating the risk of LNI in patient groups with a higher observed probability of LNI. Tosoian and Yale showed an overall underestimation of the risk of LNI. All four models showed an underestimation of the predicted probabilities in patient groups with an observed probability below 20%.

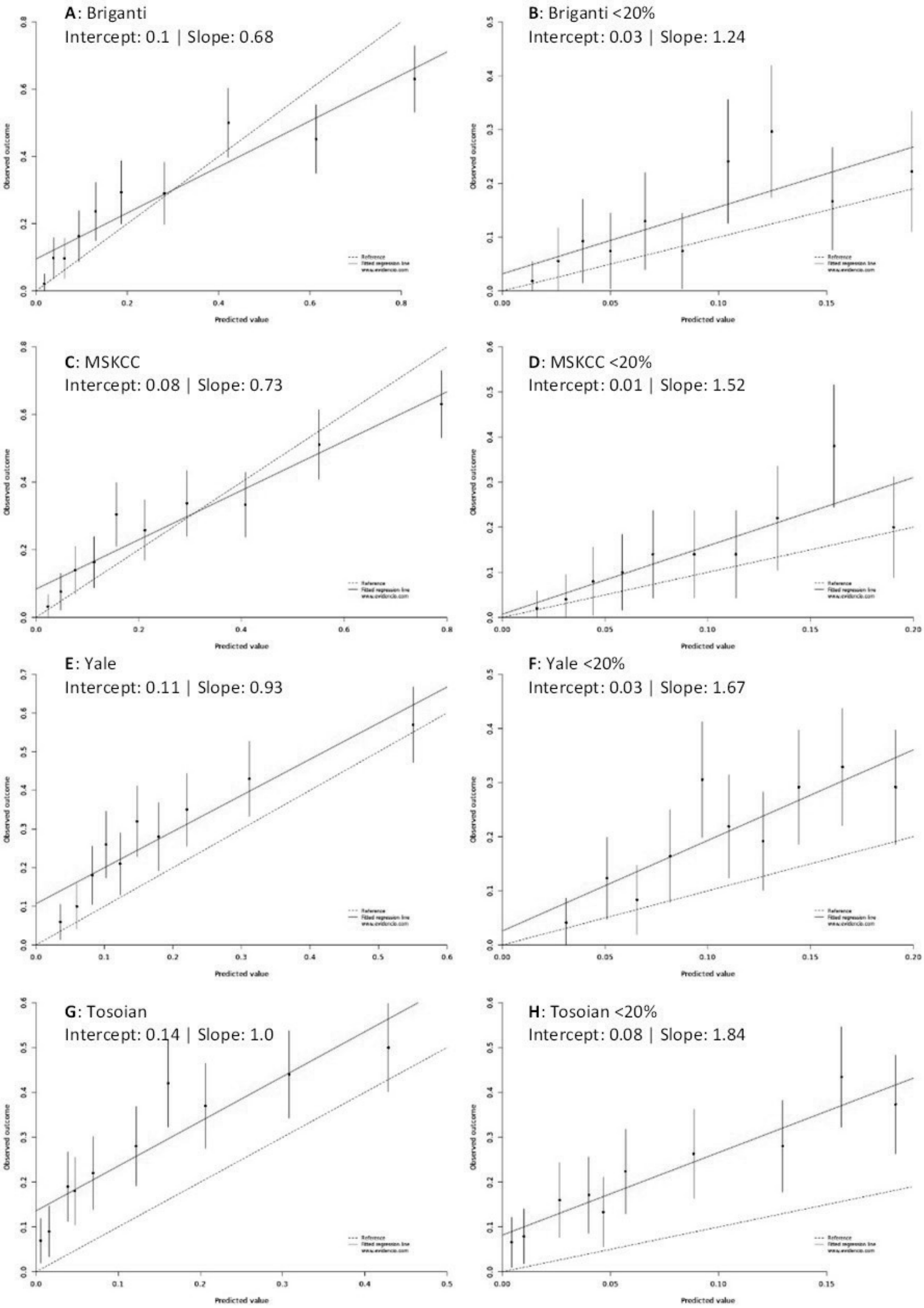
Table 2. model performances

Update - line	Model	Predictors	AUC (95% CI)
Briganti's nomograms:	Briganti 2006[14]	PSA, cT stage, Gleason sum	0.69 (0.65 – 0.72)
	Briganti 2006[15]	PSA, cT stage, Gleason sum, no. of lymph nodes removed	0.70 (0.66 – 0.73)
	Briganti 2007 (% cores)[16]	PSA, cT stage, Gleason sum, percentage of biopsy cores positive	0.72 (0.69 – 0.76)
	Briganti 2007 (n cores)[16]	PSA, cT stage, Gleason sum, no. of biopsy cores positive	0.71 (0.67 – 0.74)
	Briganti 2012[17]	PSA, cT stage, primary gleason, secondary gleason, percentage of biopsy cores positive	0.76 (0.73 – 0.79)
Formulas	Roach Formula[18]	PSA, Gleason sum	0.66 (0.63 – 0.70)
	Nguyen Formula[19]	PSA, cT stage, Gleason sum	0.68 (0.64 – 0.71)
	Yale Formula[20]	PSA, cT stage, Gleason sum	0.70 (0.66 – 0.74)
Partin tables	Makarov[21]	PSA, cT stage, Gleason sum	0.69 (0.66 – 0.73)
	Eifler[22]	PSA, cT stage, Gleason sum	0.69 (0.66 – 0.73)
	Tosoian[23]	PSA, cT stage, Gleason sum	0.70 (0.67 – 0.74)
MSKCC models	Godoy (MSKCC)[24]	PSA, cT stage, Gleason sum	0.70 (0.66 – 0.74)
	MSKCC web calculator (excl. biopsy cores)[25]	PSA, cT stage, Primary gleason, secondary gleason	0.71 (0.67 – 0.74)
	MSKCC web calculator (incl. biopsy cores)[25]	PSA, cT stage, primary gleason, secondary gleason, no. of biopsy cores positive, no. of biopsy cores negative	0.75 (0.72 – 0.78)
Asian population based model	Yonsei Nomogram[26]	PSA, cT stage, Gleason sum	0.69 (0.65 – 0.72)
Sentinel lymph node dissection based model	Winter[27]	PSA, cT stage, Gleason sum	0.69 (0.66 – 0.73)

Models were grouped in the different lines of updates throughout the years except for the Yonsei nomogram and the nomogram by Winter et al. which stand by themselves. AUC = Area under the curve, MSKCC = Memorial Sloan Kettering Cancer Center, PSA = Prostate specific antigen.

Figure 1. Receiving Operator Characteristic (ROC) curve of the four best performing prediction models in their line of updated models




A: Briganti
Intercept: 0.1 | Slope: 0.68
B: Briganti <20%
Intercept: 0.03 | Slope: 1.24
C: MSKCC
Intercept: 0.08 | Slope: 0.73
D: MSKCC <20%
Intercept: 0.01 | Slope: 1.52
E: Yale
Intercept: 0.11 | Slope: 0.93
F: Yale <20%
Intercept: 0.03 | Slope: 1.67
G: Tosoian
Intercept: 0.14 | Slope: 1.0
H: Tosoian <20%
Intercept: 0.08 | Slope: 1.84
Observed outcome
Predicted value
Reference
Fitted regression line
www.evidencio.com

Discussion

The purpose of this study was to validate available predictions models for predicting LNI in Dutch PCa patients. Three databases were combined to validate sixteen models predicting LNI. The most recent updated prediction models achieved higher AUCs than the older versions. The most recent update of the Briganti model showed the highest AUC (0.76) and the MSKCC nomogram including biopsy cores achieved a comparable AUC (0.75). Still, the 95% confidence intervals of several validated models overlap with the confidence intervals of the models by Briganti and the MSKCC. Thus it remains uncertain if either of these two models truly predict LNI better than the other validated models.

The validation cohort included data from the ProZIB initiative (62%). The ProZIB data contained patient information collected from all Dutch hospitals treating prostate cancer patients. Therefore, the outcome of this study is likely to be representative for the Dutch population. In the period of data collection, it was already known that PLND could be omitted in PCa patients with low risk of LNI.[8] Therefore, one might assume that quite a large proportion of, predominantly low-risk, patients did not undergo PLND and thus were not included for validation. This may be reflected by the fact that LNI was present in 27.6% of our patients, while this was 8% for Briganti's 2012 nomogram. The Dutch guideline advises to omit PLND in PCa patients with a calculated probability of LNI below 10%. However, it is known that several Dutch hospitals (among which CWZ) follow the European guideline recommending a 5% threshold to perform PLND. The ZGT applied a 10% threshold for patients undergoing RP, and a 15% threshold for patients undergoing radiation therapy. Out of 925 patients with a risk lower than 10% according to Briganti's 2012 nomogram, 338 (33.7%) patients still underwent PLND. If the Dutch guideline had been followed in all cases, PLND would have been incorrectly omitted in 27 patients with positive LNI (false-negatives). In addition, morbidity and costs of the procedure could have been prevented in 311 patients with negative LNI in who PLND could be safely omitted (true-negatives). In patients with a predicted probability below 5%, 189 (18.9%) patients still received PLND, of which 12 patients with and 177 patients without positive LNI. It should be noted that treating physicians might have had alternative reasons that led to a well-considered decision to perform PLND in patients that were not recommended to receive the procedure. Since the data was collected retrospectively, however, it cannot be determined what the basis of this decision was.

To our knowledge, this is the first validation study on LNI in PCa patients in the Netherlands. Yet, several other external validations have been performed. An overview of previous validation studies is provided as supplement. Updating validated models might have improved outcomes of the validation. To do a proper update, all validated models should be updated and then validated again. Although this could be of interest for future research, it was considered not feasible for all validated models and fell outside the scope of the current study. Notably, the AUCs in our validations were lower than the AUCs reported in previous external validations. The highest AUC in our validation was 0.76 on the Briganti nomogram.

Differences found between the current validation and other validation studies could also be caused by several other factors. Most external validation cohorts exist of patients receiving PLND with concomitant RP. The current cohort contains both patients undergoing RP as well as PLND alone. Robot-assisted surgery is increasingly used to perform RP and PLND throughout the past years. As our cohort included patients from 2008 to 2016, methods to dissect the lymph nodes may have been different (i.e. Robot-assisted, laparoscopic, or open), however, the surgical method used was not available in the data. Moreover, the locations of dissected lymph nodes were not registered in the data, making it unclear if there may have been differences in applied ePLND templates. It is also possible that there are differences in the methods performed to take biopsy cores. The article presented by Briganti does not state a certain method of performing biopsies influences risk predictions nor is this indicated for the web calculator of the MSKCC. For instance, the amount and percentage of positive biopsy cores can differ if biopsies were guided by MRI. Another explanation for the underestimation can be that physicians based the choice to perform PLND on other factors than the predicted risk alone, such as enlarged lymph nodes at staging MRI or PET/CT. Overestimation of the predictions were often found in the higher risk groups and may partly be explained by the inclusion of patients with a high serum PSA value (i.e. >50 ng/ml), which were often excluded in the development cohorts.[20] The combination of underestimation in the lower predicted risks and overestimation in the higher predicted risks may also explain why the found AUCs were notably lower than in most external validation studies.

There has not been a consensus on the suggested thresholds described by the different developers of the models, validations, and international guidelines. It seems that advised thresholds were based on expert-opinions on the clinically acceptable sensitivity and specificity without thorough quantitative analysis of the impact of using different risk threshold values.[12] Therefore, the outcomes of the current validation were used as input for a new study applying cost-effectiveness analysis to identify the optimal risk threshold to perform or omit PLND.

Recently, Gandaglia et al. published a novel model predicting LNI risk.[13] This model uses the amount of biopsy cores containing the high grade PCa and the amount of biopsy cores containing the lower grade PCa. These data were not present in the current cohort and therefore this model could not be validated. The new model seems to be an update of the 2012 Briganti nomogram, which showed best performance in this validation. Collecting how many biopsies contained primary and secondary Gleason scores may be useful for future validations, and potentially more accurate predictions regarding LNI.

Conclusion

Out of sixteen validated models, the Briganti nomogram from 2012 showed the best performance with an AUC of 0.76. The MSKCC nomogram showed comparable results with an AUC of 0.75. The confidence intervals of the AUC of these models overlap with AUCs of multiple other validated models, however the nomograms by Briganti and MSKCC showed adequate calibration. Based on these results, it is advised to either use the Briganti or the MSKCC nomogram to predict the risk of LNI in PCa patients.

References

1. Torre LA, Siegel RL, Ward EM, et al: Global cancer incidence and mortality rates and trends—an update. Cancer Epidemiology and Prevention Biomarkers 25:16-27, 2016
2. Organization NCC: Digits about cancer in the Netherlands,
3. Zhou CK, Check DP, Lortet-Tieulent J, et al: Prostate cancer incidence in 43 populations worldwide: an analysis of time trends overall and by age group. International journal of cancer 138:1388-1400, 2016
4. Mottet N, Bellmunt J, Bolla M, et al: EAU-ESTRO-SIOG guidelines on prostate cancer. Part 1: screening, diagnosis, and local treatment with curative intent. European urology 71:618-629, 2017
5. Datta K, Muders M, Zhang H, et al: Mechanism of lymph node metastasis in prostate cancer. Future oncology (London, England) 6:823-836, 2010
6. Ploussard G, Briganti A, De La Taille A, et al: Pelvic lymph node dissection during robot-assisted radical prostatectomy: efficacy, limitations, and complications—a systematic review of the literature. European urology 65:7-16, 2014
7. Fossati N, Willemse P-PM, van den Bergh RC, et al: The benefits and harms of different extents of lymph node dissection during radical prostatectomy for prostate cancer: a systematic review. European Urology, 2017
8. Reijke d, Th.M., Coenen JLLM, Gietema JA, et al: Dutch guideline prostate cancer. Dutch Association of Urology, 2016
9. Mohler JL, Armstrong AJ, Bahnson RR, et al: Prostate cancer guideline. Journal of the National Comprehensive Cancer Network, 2016
10. Sobin LH, Gospodarowicz MK, Wittekind C: TNM classification of malignant tumours, John Wiley & Sons, 2011
11. Debray TP, Vergouwe Y, Koffijberg H, et al: A new framework to enhance the interpretation of external validation studies of clinical prediction models. Journal of clinical epidemiology 68:279-289, 2015
12. Hansen J, Rink M, Bianchi M, et al: External validation of the updated Briganti nomogram to predict lymph node invasion in prostate cancer patients undergoing extended lymph node dissection. The Prostate 73:211-218, 2013
13. Gandaglia G, Fossati N, Zaffuto E, et al: Development and Internal Validation of a Novel Model to Identify the Candidates for Extended Pelvic Lymph Node Dissection in Prostate Cancer. European Urology
14. Briganti A, Chun FK-H, Salonia A, et al: Validation of a nomogram predicting the probability of lymph node invasion among patients undergoing radical prostatectomy and an extended pelvic lymphadenectomy. European urology 49:1019-1027, 2006
15. Briganti A, Chun FKH, Salonia A, et al: Validation of a nomogram predicting the probability of lymph node invasion based on the extent of pelvic lymphadenectomy in patients with clinically localized prostate cancer. BJU international 98:788-793, 2006

16. Briganti A, Karakiewicz PI, Chun FK-H, et al: Percentage of positive biopsy cores can improve the ability to predict lymph node invasion in patients undergoing radical prostatectomy and extended pelvic lymph node dissection. European urology 51:1573-1581, 2007
17. Briganti A, Larcher A, Abdollah F, et al: Updated nomogram predicting lymph node invasion in patients with prostate cancer undergoing extended pelvic lymph node dissection: the essential importance of percentage of positive cores. European urology 61:480-487, 2012
18. Roach M, Marquez C, Yuo H-S, et al: Predicting the risk of lymph node involvement using the pre-treatment prostate specific antigen and Gleason score in men with clinically localized prostate cancer. International Journal of Radiation Oncology* Biology* Physics 28:33-37, 1994
19. Nguyen PL, Chen M-H, Hoffman KE, et al: Predicting the risk of pelvic node involvement among men with prostate cancer in the contemporary era. International Journal of Radiation Oncology* Biology* Physics 74:104-109, 2009
20. Yu JB, Makarov DV, Gross C: A new formula for prostate cancer lymph node risk. International Journal of Radiation Oncology* Biology* Physics 80:69-75, 2011
21. Makarov DV, Trock BJ, Humphreys EB, et al: Updated nomogram to predict pathologic stage of prostate cancer given prostate-specific antigen level, clinical stage, and biopsy Gleason score (Partin tables) based on cases from 2000 to 2005. Urology 69:1095-1101, 2007
22. Eifler JB, Feng Z, Lin BM, et al: An updated prostate cancer staging nomogram (Partin tables) based on cases from 2006 to 2011. BJU international 111:22-29, 2013
23. Tosoian JJ, Chappidi M, Feng Z, et al: Prediction of Pathologic Stage Based on Clinical Stage, Serum PSA, and Biopsy Gleason Score: Partin Tables in the Contemporary Era. BJU international, 2016
24. Godoy G, Chong KT, Cronin A, et al: Extent of pelvic lymph node dissection and the impact of standard template dissection on nomogram prediction of lymph node involvement. European urology 60:195-201, 2011
25. MSKCC: Pre-operative tool to predict probability of lymph node involvement in prostate cancer patients. Memorial Sloan Kettering Cancer Center, Memorial Sloan Kettering Cancer Center, 2017
26. Kim KH, Lim SK, Kim HY, et al: Yonsei nomogram to predict lymph node invasion in Asian men with prostate cancer during robotic era. BJU international 113:598-604, 2014
27. Winter A, Kneib T, Henke RP, et al: Sentinel lymph node dissection in more than 1200 prostate cancer cases: rate and prediction of lymph node involvement depending on preoperative tumor characteristics. International Journal of Urology 21:58-63, 2014

Chapter 6

External validation of the Memorial Sloan Kettering Cancer Centre and Briganti nomograms for the prediction of lymph node involvement of prostate cancer using clinical stage assessed by magnetic resonance imaging

Timo F.W. Soeterik, Tom A. Hueting, Bas Israel, Harm H.E. van Melick, Lea M. Dijksman, Saskia Stomps, Douwe H. Biesma, Hendrik Koffijberg, Michiel Sedelaar, J. Alfred Witjes, Jean-Paul A. van Basten

BJU International 128-2 236 – 243 (2021)
DOI: 10.1111/bju.15376

The supplementary data to this chapter are available online at https://dx.doi.org/10.1111/bju.15376

Abstract

Objectives: To evaluate the impact of using clinical stage assessed by multi-parametric magnetic resonance imaging (mpMRI) on the performance of two established nomograms for prediction of lymph node involvement (LNI) in patients with prostate cancer.

Methods: Patients undergoing robot-assisted extended pelvic lymph node dissection (e-PLND) from 2015-2019 at three teaching hospitals were retrospectively evaluated. Risk of pelvic LNI was calculated four times for each patient, using DRE- and mpMRI T-stage, in the MSKCC (2018) and Briganti (2012) nomograms. Discrimination (area under the curve [AUC]), calibration, and net benefit of these four strategies were assessed and compared.

Results: A total of 1,062 patients were included, of whom 301 (28%) had LNI. Using DRE T-stage resulted in AUCs 0.71 (95%CI 0.70-0.72) for MSKCC and 0.73 (95%CI 0.72–0.74) for Briganti, whereas AUCs for mpMRI T were 0.72 (95%CI 0.71–0.73) for MSKCC and 0.75 (95%CI 0.74–0.76) for Briganti. MpMRI T-stage resulted in improved calibration compared with DRE T-stage. Combined use of mpMRI T-stage and Briganti 2012 was shown to be superior in terms of AUC, calibration, and net benefit. Use of mpMRI T-stage led to increased sensitivity for the detection of LNI for all risk thresholds in both models, countered by a decreased specificity, compared with DRE T-stage.

Conclusion: The mpMRI T-stage is an appropriate alternative for DRE T-stage to determine nomogram-based risk of LNI in PCa patients, and was associated with improved model performance of both the MSKCC 2018 and Briganti 2012 nomograms

Introduction

Identification of lymph node involvement (LNI) is an essential component of the general staging work-up in patients with newly-diagnosed prostate cancer, and is indicated in patients with a risk of LNI above 5%.[1] Even minimal tumour involvement of the lymphatic system is thought to have pivotal impact on disease prognosis, and should be established to identify patients with an increased risk of disease recurrence.[2]

Currently, advances within the field of clinical imaging, particularly prostate-specific membrane antigen (PSMA) positron-emission tomography/computer tomography (PET/CT), are rapidly evolving. However, since the sensitivity of PSMA PET/CT for the detection of LNI in primary prostate cancer is only moderate, it cannot yet replace extended pelvic lymph node dissection (e-PLND) to exclude LNI.[3,4] Thus, e-PLND remains the preferred option for lymph node staging in primary prostate cancer.[1]

Performing e-PLND in patients undergoing radical prostatectomy is associated with unfavourable intraoperative and perioperative outcomes, including symptomatic lymphocele development (in up to 18%), bleeding (2.7%), infections (3.6%), and ureteral damage (0.8%), whereas there is no high-level evidence for a direct therapeutic effect.[5,6] Therefore, e-PLND should be reserved for carefully selected patients.

Both the European Association of Urology (EAU) and the National Comprehensive Cancer Network (NCCN) guidelines recommend the use of nomograms to guide patient selection for e-PLND.[1,7] Several of these prediction tools have been developed over the years.[8] The Memorial Sloan Kettering Cancer Centre (MSKCC) pre-radical prostatectomy (update 2018) and Briganti 2012 nomograms are the two most established models.[9,10] In a recent validation study using a contemporary cohort of patients with Prostate cancer, the 2012 Briganti and the 2018 MSKCC nomograms were identified as the most accurate prediction tools available, with a reported area under the curve (AUC) of 0.76 and 0.75, respectively.[8]

Both the MSKCC 2018 and Briganti 2012 include clinical T-stage assessed by digital rectal examination (DRE) as one of the input parameters.[9,10] However, recent guideline updates include the recommendation for performing multi-parametric MRI (mpMRI) prior to prostate biopsy.[1,11] As a result, MRI staging information will become increasingly available in newly diagnosed patients. In addition, mpMRI potentially enables a more accurate estimation of local tumour extent compared with DRE.[12] However, it is not clear if the use of mpMRI T-stage results in more accurate nomogram-based LNI risk prediction.

Therefore, we will evaluate if replacing DRE T-stage by mpMRI T-stage results in a more accurate LNI risk prediction by the MSKCC 2018 and Briganti 2012 nomograms.

Methods

Study population

After receiving institutional review board approval, patients diagnosed with prostate cancer undergoing e-PLND from January 2015 to September 2019 at three Dutch teaching hospitals (St. Antonius Hospital Nieuwegein/Utrecht, Hospital Group Twente Almelo/Hengelo, and Canisius Wilhelmina Hospital Nijmegen), were included.

Patients underwent e-PLND combined with radical prostatectomy or prior to radiation therapy. In general, patients with a risk of LNI >5% (based on DRE T-stage), calculated using the MSKCC web calculator,[9] were considered as candidates for e-PLND. However, deviations were allowed at the discretion of the treating urologist.

Clinical T-stage established by DRE, radiological T-stage determined using mpMRI, preoperative prostate-specific antigen (PSA), highest International Society of Urological Pathology (ISUP) grade observed on most recent biopsy, total number of biopsy cores taken on systematic biopsy and the relative number of cores containing Prostate cancer were collected. Patients were included if they underwent systematic biopsies with or without MRI-guided target biopsy and mpMRI for local staging prior to e-PLND. Patients undergoing salvage e-PLND or those who received androgen deprivation therapy prior to e-PLND were excluded.

Covariates and endpoints

PSA, clinical stage assessed by DRE (DRE T-stage), clinical stage assessed by mpMRI (mpMRI T-stage), total number and relative number of positive biopsy cores as well as pathological lymph node status were collected during standard clinical practice. Gleason scoring was done according to the ISUP 2014 consensus statement.[13]

DRE was performed during the primary diagnostic work-up by urologists with >5 years of experience with diagnosing and staging prostate cancer. DRE consisted of systematic palpation of all prostate regions including both lateral sides, the posterior region and the sulcus. DRE was performed in either dorsal lithotomy or lateral position. Findings were classified according to the clinical classification of the American Joint Committee on Cancer.[14]

During the study period, 3-Tesla MRI scanners were used at the three institutions. Radiological reporting was performed by the local dedicated uro-radiologists. Reporting was done according to the PI-RADS v2 guidelines.[15] MpMRI T-stages were defined as T2a, (unilateral suspicious lesion, involving <50% of the prostatic lobe), T2b, (unilateral suspicious lesion, involving >50% of the prostatic lobe) T2c (bilateral suspicious lesion), T3a (definite or high-degree of suspicion for extraprostatic extension [EPE]), T3b (definite or high-degree of suspicion of seminal vesicle invasion) and T4 (invades adjacent structures). Protocols of mpMRI performed at the three institutions are presented in the supplemental section (Supplemental section, Table S1.).

The e-PLND template included removal of nodes overlying the external iliac vessels, internal iliac artery, and the nodes located within the obturator fossa.[16]

All resected nodal tissue was submitted for pathologic evaluation, performed by experienced uro-pathologists. The total number of lymph nodes found in the tissue, as well as the number of nodes containing prostate cancer metastasis were assessed. Histopathological evaluation was performed in accordance with the ISUP consensus statement.[17]

Statistical analysis

The risk of LNI was estimated a total of four times per patient: using both the MSKCC and Briganti 2012 nomograms, with both DRE- and mpMRI T-stage. Other covariates used for LNI risk calculation included most recent preoperative serum PSA level, highest ISUP grade found on either systematic or target biopsy, as well as the number of positive cores and the total number of cores taken on systematic biopsy. Model discrimination was quantified using the AUC, and refers to the probability of a random patient with LNI (pN1) having a higher predicted risk than a random patient without LNI (pN0).[18] Classification plots, showing the true and false positive rates per risk threshold were used to visualize discriminatory ability.[19] Model calibration, which refers to the agreement between observed and predicted LNI, was assessed by plotting calibration curves and by determining calibration-in-the-large and calibration slopes.[18] The calibration-in-the-large indicates whether predicted probabilities are systematically too low or too high. Perfect calibration is characterized by a calibration-in-the-large of 0, and a calibration slope of 1.[18] The scaled Brier score, which is the average squared difference between the actual outcomes (i.e. LNI) and predicted probabilities, was also determined. A scaled Brier score close to 1 shows overall poor predictive ability, whereas a scaled Brier score of 0 corresponds with perfect risk prediction of the model.[18] Decision curve analysis was performed to determine net benefit of the models over multiple clinically relevant thresholds. The calculated net benefit of the models was compared to the scenarios of treating either all or no patients.[20] A systematic analysis was performed to determine the number of patients (with or without LNI) in whom e-PLND would be advised, for LNI risk thresholds between 1%-15%. Missing data were handled by using multiple imputations by chained equations.[21] A total of 10 imputed datasets were created. Model performance measures were estimated by bootstrapping each imputed dataset 500 times. To select the best performing approach, the different approaches were compared head-to-head by estimating in how many bootstrap samples a specific approach resulted in the highest pooled AUC measure. Statistical analysis was performed using R v3.6.3. (R Project for Statistical Computing, www.r-project.org).

Table 1. Baseline characteristics of the validation cohort

	Overall **N (%), median (IQR)**	**pN0** **N (%), median (IQR)**	**pN1** **N (%), median (IQR)**
No. of patients	1062 (100)	761 (72)	301 (28)
Hospital			
SAH	258 (24)	186 (24)	72 (24)
HGT	246 (23)	159 (21)	87 (29)
CWH	558 (53)	416 (55)	142 (47)
Age (years)	67 (63 - 71)	68 (63 - 71)	67 (63 - 71)
PSA (ng/ml)	10 (6.6 – 18)	9.3 (6.2 – 16)	13 (7.8 – 22)
Total cores	10 (10 - 12)	10 (9 – 12)	10 (10 – 12)
Unknown	5 (0)	4 (1)	1 (0)
Total positive cores	5 (3 – 8)	5 (3 – 7)	7.1 (3.3)
Unknown	6 (1)	5 (1)	1 (0)
Percentage of positive cores (%)	0.50 (0.33 – 0.75)	0.50 (0.25 – 0.67)	0.75 (50 - 100)
Unknown	7 (1)	6 (1)	1 (0)
Biopsy ISUP grade			
1	78 (7)	65 (9)	13 (4)
2	245 (23)	191 (25)	54 (18)
3	280 (26)	202 (27)	78 (26)
4	253 (24)	189 (25)	64 (21)
5	201 (19)	110 (14)	91 (30)
Unknown	5 (0)	4 (0)	1 (0)
DRE T-stage			
T1c	384 (36)	301 (40)	83 (28)
T2a	328 (31)	248 (33)	80 (27)
T2b	84 (8)	57 (7)	27 (9)
T2c	77 (7)	52 (7)	35 (12)
T3a	169 (16)	103 (14)	66 (22)
T3b	7 (7)	3 (0)	4 (1)
T4	3 (0)	1 (0)	2 (0)
Unknown	10 (1)	6 (1)	4 (1)
mpMRI T-stage			
T1c	40 (4)	36 (5)	4 (1)
T2a	301 (28)	250 (33)	51 (17)
T2b	41 (4)	29 (4)	12 (4)
T2c	120 (11)	103 (14)	17 (6)
T3a	376 (35)	261 (34)	115 (38)
T3b	160 (15)	69 (9)	91 (30)
T4	22 (2)	12 (2)	10 (3)
Unknown	2 (0)	1 (0)	1 (0)
Biopsy type			
TRUS-guided SB	694 (65)	479 (63)	215 (71)
TRUS-SB+ MRI-TB	368 (35)	282 (37)	86 (29)
Total nodes resected	**20 (13 - 25)**	**17 (12 - 24)**	**20 (14 - 28)**

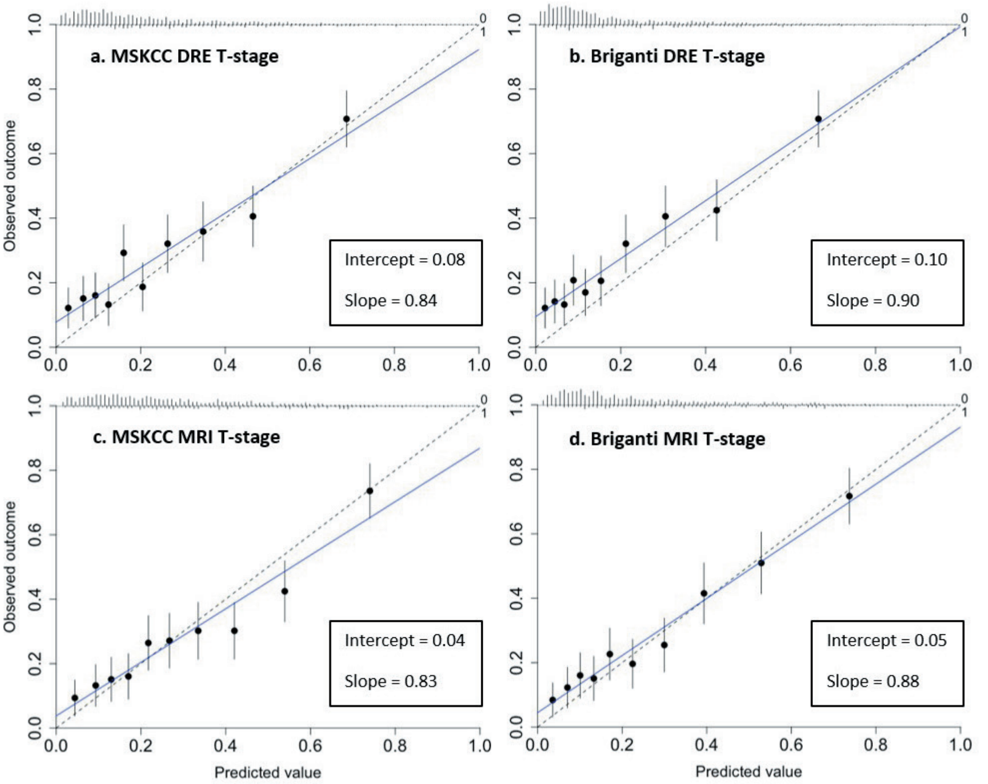
SAH: St. Antonius Hospital, HGT: Hospital Group Twente, CWH: Canisius Wilhelmina Hospital, SD: standard deviation, TRUS: transrectal ultrasonography, SB: systematic biopsy, MRI-TB: magnetic resonance image-guided target biopsy

Results

Study population

A total of 1,062 patients fulfilled the inclusion criteria. Overall, 301 (28%) patients had histologically confirmed LNI. The median number of lymph nodes removed was 20 (IQR: 13-25). A total of 21 patients (2%) had one or more covariates missing including DRE T-stage (N=11), mpMRI T-stage (N=2), and biopsy data (N=8). Baseline characteristics of the study cohort are presented in Table 1.

Figure 1. Calibration plots of both nomograms based on DRE T-stage (A & B) and mpMRI T-stage (C & D)



Each dot represents mean observed probability of 106 patients (10%). De blue line represents the linear regression line based on the mean observed probabilities per subgroup. The dashed line represents the ideal situation with 100% agreement between predicted and observed probabilities. The lines on the top of the chart reflect the density of the events and the non-events. Events are seen below the line (labelled as 1), non-events are seen above the line (labelled as 0).

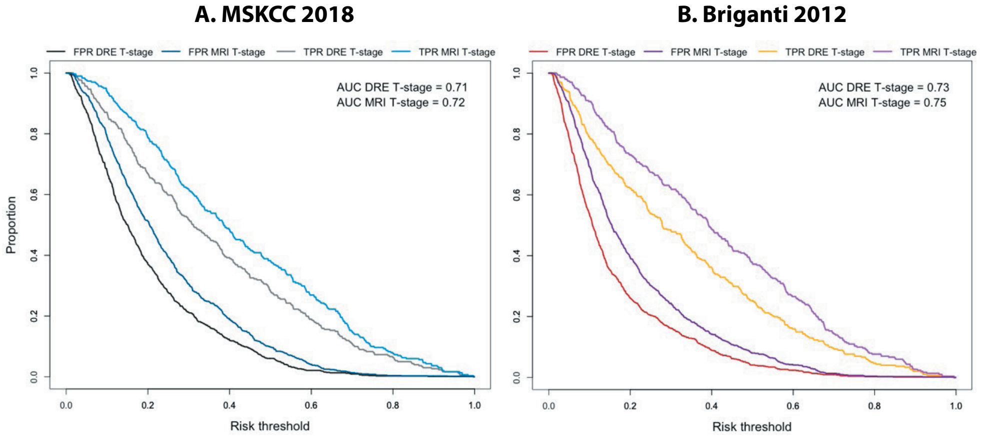
Model performance using DRE or mpMRI T-stage

Initial validation included use of DRE T-stage assessed. Discrimination in terms of AUC was respectively 0.71 (95% CI 0.70 – 0.72) for MSKCC and 0.73 (95% CI 0.72 – 0.74) for Briganti. Mean predicted probability for LNI was respectively 24% for the MSKCC and 21% for the

Briganti nomogram, where the observed LNI rate was 28% (Table 2.). On visual exploration of calibration plots, we also observed systematic underestimation of the predicted risk of both nomograms, particularly for risk thresholds between 0% and 30% (Figures 1A & 1B).

When using mpMRI T-stage, discrimination in terms of AUC, increased to 0.72 (95% CI 0.71 – 0.73) for the MSKCC and 0.75 (95% CI 0.74 – 0.76) for the Briganti model (Supplemental section, Table S2.). Mean predicted probability for LNI changed to 30% for MSKCC and 27% for Briganti, respectively. As shown in the calibration plots, the agreement between predicted and observed probabilities was comparable (both moderate calibration) for both DRE T-stage and mpMRI T-stage. The calibration-in-the-large was closer to 0 when using mpMRI instead of DRE for both MSKCC (0.04 (95% CI: 0 – 0.08) versus 0.08 (95% CI: 0.04 – 0.12)), and Briganti (0.05 (95% CI: 0 – 0.08) versus 0.10 (95% CI: 0.06 – 0.13)). (Supplemental section, Table S2.). In a head-to-head comparison, calculating the LNI risk using mpMRI T-stage with the Briganti nomogram led to higher AUCs in all bootstrap samples, compared with Briganti DRE T-stage as well as both DRE T-stage and mpMRI T-stage with the MSKCC nomogram.

Figure 2. Classification plots of both nomograms displaying sensitivity, specificity established using both DRE T-stage and mpMRI T-stage



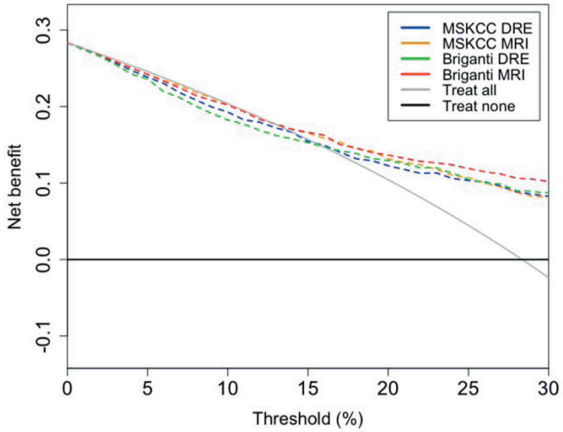
Risk threshold: value of the estimated risk that is used to classify patients as "test positive" (eligible for ePLND). True positive rate: proportion of patients with pN1 above the risk threshold, false positive rate: proportion of patients with pN0 above the risk threshold. AUC: the area under the ROC curve. PR = false positive rate, TPR = true positive rate

Clinical usefulness

Using mpMRI T-stage resulted a in higher true-positive rate and a higher false-positive rate for the detection of positive lymph nodes for all risk thresholds, compared to using DRE T-stage (Figure 2.). Use of mpMRI T-stage led to increased sensitivity for the detection of LNI for all risk thresholds in both models, countered by a relatively lower specificity, compared with DRE T-stage In Tables S3 and S4 of the Supplemental section, total number of missed LNI cases per risk thresholds are presented, combined with rates of performed ePLND and number of positive LNI cases. For all thresholds, number of missed LNI cases was lower when mp-MRI T stage was used, countered by higher rates of unnecessary ePLND (pN0).

Decision curve analysis revealed that use of mpMRI T-stage in both nomograms resulted higher net benefits, compared with DRE T-stage, for the clinically most relevant risk thresholds between 5% and 20%. Net benefits for both the MSKCC and Briganti nomogram, using mpMRI T-stage, were comparable for this range of risk thresholds. For risk thresholds ranging from 20% and 30%, the combined use of mpMRI T-stage with the Briganti nomogram would lead to the highest net benefit (Figure 3.).

Figure 3. Decision curves of the four performed validation scenario's compared to the default strategies



* The figure shows the net benefit for the threshold probabilities between 0 and 0.30. The dashed lines display the net benefit for the four models. The grey line represents the scenario in which all patients would undergo ePLND ("treat all"). The black line represents the scenario in which no patients would undergo ePLND ("treat none").

Discussion

Use of mpMRI T-stage for nomogram-based LNI risk assessment led to higher AUCs, comparable agreement between predicted and observed probabilities and higher net benefits compared with DRE T-stage, in both the MSKCC 2018 and Briganti 2012 nomograms. In our study population, use of DRE T-stage would lead to overall LNI risk underestimation in the clinically relevant range of risk thresholds (0-30%). In the head-to-head comparison, combined use of the mpMRI T-stage with the Briganti 2012 nomogram led to the most accurate LNI risk prediction.

Our study acknowledges the robustness of both the MSKCC 2018 and Briganti 2012 nomograms, since model performance was still fair to good, even when the model was applied in a patient population with substantial higher prevalence of the predicted outcome compared with the development populations. In our cohort, LNI prevalence (28%) was substantially higher compared to both MSKCC's (7% [internal communication MSKCC research team]) and Briganti's (8%) populations.[10] Thereby, our results show both models are applicable in a contemporary patient cohort. In addition, our analysis confirmed that mpMRI T-stage can be safely used as impute parameter for these nomograms, even leading to improved accuracy of the predicted LNI risk compared with DRE T-stage.

This study's main findings add up to the available body of literature supporting the additional value of mpMRI information for predicting presence of LNI in Prostate cancer. For example, Porpiglia *et al*. showed MRI has an important role in LNI risk prediction in patients with a nomogram predicted risk <5%.[22] Huang et al. showed that addition of the PI-RADS score to led to improved AUC for both nomograms, increasing from 75% to 86% for Briganti and from 79% to 88% for MSKCC, respectively.[23]

Recently, two new nomograms have been introduced, including mpMRI and target biopsy features such as maximum diameter of the index lesion and maximum percentage of tumour involvement in one core.[24,25] Of these, the 2019 Briganti nomogram was recently externally validated showing excellent characteristics including an AUC of 79% and high agreement between predicted and observed probabilities for risk thresholds below 35%.[26] In their head-to-head comparison, the 2019 Briganti nomogram outperformed the Briganti 2017 and MSKCC 2018 in terms of discrimination, calibration and net benefit. Although these new nomograms potentially enable improved LNI risk prediction due to the addition of mpMRI guided target biopsy, they both include complex features which may not be always available in clinical practice, such as maximum diameter of the index lesion on mpMRI and highest tumour length in millimetres of all biopsy cores taken.[24,25] In addition, more external validation studies are warranted to confirm the accuracy of these new nomograms in external patient populations. Data derived from external validation studies is crucial, as model transportability needs to be adequate to prevent systematic wrong decision-making.

Our results do not support the statements in a recent position paper on Prostate cancer staging by Paner *et al.*, who suggested that DRE should not be replaced by mpMRI for establishing clinical T-stage.[27] In our study, mpMRI outperformed DRE in terms of AUC for nomogram-based LNI risk prediction, as the use mpMRI T stage resulted in higher AUCs for all bootstrap samples. This is most likely the consequence of the main advantage that mpMRI has over DRE for determining local tumour extent, which is the visualisation of the prostate gland as a whole and improved detection of non-organ confined disease. Our study group has confirmed this in a recent study, as the reported sensitivity for the detection of non-organ confined disease was significantly lower for DRE compared with mpMRI (51% vs. 13%, p<0.001).[28]

Although our main study results favour the use of mpMRI T-stage for nomogram-based LNI risk prediction, there are arguments against replacing DRE with mpMRI T-stage that should be mentioned. First, disadvantages of MRI include reader inter-observer variability and quality differences regarding mpMRI reading.[27] However, a previous study by Angulo *et al*. showed interobserver inconsistency also to be an issue for DRE, resulting in a low ability to reproduce clinical staging on DRE among multiple examiners.[29]

Second, use of mpMRI compared with DRE would lead to upstaging of clinical T-stage in one-third of the patients.[29] Although mpMRI can provide valuable prognostic information for specific patients, including those with non-organ-confined disease which was missed on DRE, the high upstaging rates bear the risk of overstaging and hence overtreatment in patients with favourable risk disease.[28]

To select patients for e-PLND, it remains important to take into account patient's preferences, age and prognostic tumour parameters other than those included in the nomograms to distinguish the patients who would benefit from additional e-PLND from those in whom this intervention would potentially do more harm than good.

In addition, the trade-off between subjecting node-negative patients to the concomitant risks of e-PLND versus the potential advantages e-PLND in the specific node-positive subgroup, remains to be explored. Future studies should focus on finding the optimum risk threshold at which the benefits of e-PLND, at best, outweigh the harms.

Although this study has several strengths, such as the inclusion of a multi-centre cohort representing the real-wold prostate cancer population and large study sample with a sufficient number of events for adequate external validation, it is not exempt of limitations.

First, data used in this study was derived from routine clinical practice, and no central review of DRE, mpMRI and histopathological evaluation was performed. In addition, the majority of the data were collected retrospectively, which could have led to measurement bias. Lastly, the indication to perform an e-PLND in this patient cohort was done using nomogram-based LNI risk estimation (risk of LNI >5%). Even though this is according to current EAU guideline recommendations, and reflects contemporary clinical practice, this could have introduced bias due to the selection of patients for e-PLND with higher risk of LNI (reflected by the relatively high LNI prevalence). For instance, selecting patients with higher risk of LNI (and prevalence) could explain the counterintuitive finding on DCA, showing that a "treat all" approach would lead to higher net benefit compared with nomogram-based selection for risk thresholds between 0%-15%.

Conclusion

The MSKCC and Briganti 2012 nomograms showed to be adequate models for the prediction of LNI in patients with Prostate cancer when using either mpMRI T-stage or DRE T-stage. The use of mpMRI T-stage led to improved model discrimination, equal calibration, and lower rates of missed LNI cases. Using the mpMRI T-stage with the Briganti 2012 nomogram was shown to be the most accurate strategy for LNI risk prediction.

References

1. Mottet N, van den Bergh RCN, Briers E, et al. EAU – ESTRO – ESUR – SIOG Guidelines on Prostate Cancer 2019 2019; presented at the EAU Annual Congress Barcelona 2019. Arnhem, The Netherlands: European Association of Urology Guidelines Office.
2. Wilczak W, Wittmer C, Clauditz T, et al. Marked Prognostic Impact of Minimal Lymphatic Tumor Spread in Prostate Cancer. Eur Urol 2018;74 376-86.
3. Luiting HB, van Leeuwen PJ, Busstra MB, et al. Use of gallium-68 prostate-specific membrane antigen positron-emission tomography for detecting lymph node metastases in primary and recurrent prostate cancer and location of recurrence after radical prostatectomy: an overview of the current literature. BJU Int. 2020;125 206-14.
4. Perera M, Papa N, Roberts M et al. Gallium-68 Prostate-specific Membrane Antigen Positron Emission Tomography in Advanced Prostate Cancer-Updated Diagnostic Utility, Sensitivity, Specificity, and Distribution of Prostate-specific Membrane Antigen-avid Lesions: A Systematic Review and Meta-analysis. Eur.Urol 2020;77:403-417
5. Fossati N, Willemse PM, Van den Broeck T, et al. The Benefits and Harms of Different Extents of Lymph Node Dissection During Radical Prostatectomy for Prostate Cancer: A Systematic Review. Eur Urol 2017;72:84-109.
6. Oderda M, Diamond R, Albissini S, et al. Indications for and complications of pelvic lymph node dissection in prostate cancer: accuracy of available nomograms for the prediction of lymph node invasion. BJUI 2020; Sep 1. doi: 10.1111/bju.15220. Online ahead of print.
7. Carroll PR, Parsons JK, Andriole G, et al. NCCN Guidelines Insights: Prostate Cancer Early Detection, Version 2.2016. J Natl Compr Canc Netw 2016;14:509-19.
8. Hueting TA, Cornel EB, Somford DM, et al. External Validation of Models Predicting the Probability of Lymph Node Involvement in Prostate Cancer Patients. Eur Urol Oncol 2018;1:411-7.
9. Memorial Sloan Kettering Cancer Center. Pre-radical prostatectomy tool to predict probability of lymph node involvement in prostate cancer patients. www.mskcc.org/nomograms/prostate/pre_op.
10. Briganti A, Larcher A, Abdollah F, et al. Updated nomogram predicting lymph node invasion in patients with prostate cancer undergoing extended pelvic lymph node dissection: the essential importance of percentage of positive cores. Eur Urol 2012;61:480-7.
11. Bjurlin MA, Carroll PR, Eggener S, et al. Update of the Standard Operating Procedure on the Use of Multiparametric Magnetic Resonance Imaging for the Diagnosis, Staging and Management of Prostate Cancer. J Urol 2020;203:706-712
12. de Rooij M, Hamoen EH, Witjes JA, Barentsz JO, Rovers MM. Accuracy of Magnetic Resonance Imaging for Local Staging of Prostate Cancer: A Diagnostic Meta-analysis. Eur Urol 2016;70:233-45.
13. Epstein JI, Zelefsky MJ, Sjoberg DD, et al. A Contemporary Prostate Cancer Grading System: A Validated Alternative to the Gleason Score. Eur Urol 2016;69:428-35.

14. Edge SB, Byrd DR, Compton CC, Fritz AG, Greene FL, Trotti A, editors. AJCC cancer staging manual (7th ed). New York, NY: Springer; 2010.

15. Weinreb JC, Barentsz JO, Choyke PL, et al. PI-RADS Prostate Imaging - Reporting and Data System: 2015, Version 2. Eur Urol 2016;69:16-40.

16. Mattei A, Fuechsel FG, Bhatta Dhar N, et al. The template of the primary lymphatic landing sites of the prostate should be revisited: results of a multimodality mapping study. Eur Urol 2008;53:118-25.

17. Berney DM, Wheeler TM, Grignon DJ, et al. International Society of Urological Pathology (ISUP) Consensus Conference on Handling and Staging of Radical Prostatectomy Specimens. Working group 4: seminal vesicles and lymph nodes. Mod Pathol 2011;24:39-47.

18. Moons KG, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med 2015;162:1-73.

19. Verbakel J, Steyerberg EW, Uno H, et al. ROC curves for clinical prediction models part 1. ROC plots showed no added value above the AUC when evaluating the performance of clinical prediction models. J Clin Epidemiol 2020;126:207-216

20. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making 2006;26:565-74.

21. Van Buuren S, Groothuis-Oudshoorn K,. Multivariate Imputation by Chained Equations in R. J Stat Soft 2011;45:1-67.

22. Porpiglia F, Manfredi M, Mele F, et al. Indication to pelvic lymph nodes dissection for prostate cancer: the role of multiparametric magnetic resonance imaging when the risk of lymph nodes invasion according to Briganti updated nomogram is <5. Prostate Cancer Prostatic Dis. 2018;21:85-91.

23. Huang C, Song G, Wang H, et al. Preoperative PI-RADS Version 2 scores helps improve accuracy of clinical nomograms for predicting pelvic lymph node metastasis at radical prostatectomy. Prostate Cancer Prostatic Dis 2020;23:116-126.

24. Gandaglia G, Ploussard G, Valerio M, et al. A Novel Nomogram to Identify Candidates for Extended Pelvic Lymph Node Dissection Among Patients with Clinically Localized Prostate Cancer Diagnosed with Magnetic Resonance Imaging-targeted and Systematic Biopsies. Eur Urol 2019;75:506-14.

25. Draulans C, Everaerts W, Isebaert S, et al. Development and External Validation of a mpMRI- and ISUP-Based Add-on Prediction Tool to Identity Prostate Cancer Candidates for Pelvic Lymph Node Dissection. J Urol 2020;203:713-718.

26. Gandaglia G, Martini A, Ploussard G, et al. External Validation of the 2019 Briganti Nomogram for the Identification of Prostate Cancer Patients Who Should Be Considered for an Extended Pelvic Lymph Node Dissection. Eur Urol 2020; S0302-2838(20)30198-6. Online ahead of print.

27. Paner GP, Stadler WM, Hansel DE, Montironi R, Lin DW, Amin MB. Updates in the Eighth Edition of the Tumor-Node-Metastasis Staging Classification for Urologic Cancers. Eur Urol 2018;73:560-9.

28. Soeterik TFW, Van Melick HHE, Dijksman LM, et al. Multi-parametric magnetic resonance imaging should be preferred over digital rectal examination for prostate cancer local staging and disease risk classification. Urology 2020 Oct 28; S0090-4295(20)31296-6. Online ahead of print

29. Angulo JC, Montie JE, Bukowsky T et al. Interobserver consistency of digital rectal examination in clinical staging of localized prostatic carcinoma. Urol Oncol 1995;1:199–205.

Chapter 7

Optimizing the risk threshold of lymph node involvement for performing extended pelvic lymph node dissection in prostate cancer patients: a cost-effectiveness analysis

Tom A. Hueting, Erik B. Cornel, Ruben A. Korthorst, Rick G. Pleijhuis, Diederik M. Somford, Jean-Paul A. van Basten, Job A.M. van der Palen, Hendrik Koffijberg

Urologic Oncology: Seminars and Original Investigations 39: 72.e7 – 72.e14 (2021)
DOI: 10.1016/j.urolonc.2020.09.014

The supplementary data to this chapter are available online at https://dx.doi.org/10.1016/j.urolonc.2020.09.014

Abstract

Background: Extended pelvic lymph node dissection (ePLND) may be omitted in prostate cancer (PCa) patients with a low predicted risk of lymph node involvement (LNI). The aim of the current study was to quantify the cost-effectiveness of using different risk thresholds for predicted LNI in PCa patients to inform decision making on omitting ePLND.

Methods: Five different thresholds (2%, 5%, 10%, 20%, and 100%) used in practice for performing ePLND were compared using a decision analytic cohort model with the 100% threshold (i.e. no ePLND) as reference. Compared outcomes consisted of quality-adjusted life years (QALYs) and costs. Baseline characteristics for the hypothetical cohort were based on an actual Dutch patient cohort containing 925 patients who underwent ePLND with risks of LNI predicted by the MSKCC web-calculator. The best strategy was selected based on the incremental cost effectiveness ratio (ICER) when applying a willingness to pay (WTP) threshold of €20,000 per QALY gained. Probabilistic sensitivity analysis was performed with Monte Carlo simulation to assess the robustness of the results.

Results: Costs and health outcomes were lowest (€4,858 and 6.04 QALYs) for the 100% threshold, and highest (€10,939 and 6.21 QALYs) for the 2% threshold, respectively. The ICER for the 2%, 5%, 10%, and 20% threshold compared with the first threshold above (i.e. 5%, 10%, 20%, and 100%) were €189,222/QALY, €130,689/QALY, €51,920/QALY, and €23,187/QALY respectively. Applying a WTP threshold of €20.000 the probabilities for the 2%, 5%, 10%, 20%, and 100% threshold strategies being cost-effective were 0.0%, 0.3%, 4.9%, 30.3%, and 64.5% respectively.

Conclusion: Applying a WTP threshold of €20.000, completely omitting ePLND in PCa patients is cost-effective compared to other risk-based strategies. However, applying a 20% threshold for probable LNI to the Briganti 2012 nomogram or the MSKCC web-calculator, may be a feasible alternative, in particular when higher WTP values are considered.

Introduction

Extended Pelvic lymph node dissection (ePLND) in patients with prostate carcinoma (PCa) is still the most accurate staging method for lymph node involvement (LNI).[1,2] However, the value of ePLND in the treatment of pelvic lymph node metastasis is an ongoing topic of debate for several years.[3] A recent systematic review suggested that there is no evidence for any beneficial therapeutic effect of the procedure.[4] Still, prospective randomized trials on the potential benefit of ePLND on PCa outcomes is lacking. Therefore, omitting ePLND in PCa patients with low predicted risk of LNI, that is, below a certain risk threshold based on prediction models, may be advised. Applying such a risk threshold can prevent unnecessary complications in node negative patients, and reduce health care expenditure.[5] The ePLND is generally performed as part of a radical prostatectomy (RP), or is performed as a stand-alone procedure prior to radiotherapy. Having one or more positive lymph nodes worsens the prognosis of the disease.[6] Selecting those patients expected to benefit most from ePLND is the crux regarding its controversy, and the key to its efficient and beneficial use.

Several tools have been developed to predict the risk of LNI in PCa patients, supporting urologists in the decision to perform or omit ePLND. Predictions are made based on prostate specific antigen (PSA), primary and secondary Gleason scores, clinical T-stage, and either percentage of positive biopsy cores or amount of positive and negative biopsy cores taken.[7–9] Several guidelines recommend to base the decision to perform ePLND on the predicted risk of LNI. However, these guidelines recommend usage of different prediction tools: either the Briganti nomogram, Memorial Sloan Kettering Cancer Center (MSKCC) nomogram, Partin Tables or the Roach formula.[7–10] These four tools predict a different risk for the same patient, and consequently their recommended risk threshold to perform ePLND also varies between 2% (NCCN guideline), 5% (EAU guideline), and-10% (Dutch guideline) risk of LNI. As a result, it remains difficult for urologists to assess whether a patient might benefit from ePLND or not. This may result in differences in patient management across hospitals and urological practices, and thus differences in both quality and costs of care.[11]

Although the recommended risk thresholds for the four prediction models are different, they are all derived based on a (perceived) optimal balance between the chance of false positive and of false negative classifications of patients. However, such thresholds do not account for the consequences of such false positive and of false negative classifications, based on subsequent patient management decisions, in terms of health outcomes and health care costs. In a cost-effectiveness analysis the optimal risk threshold for ePLND can be derived accounting for all relevant health and economic aspects.

A recent validation study assessed a total of 16 tools on their performance at predicting LNI in Dutch PCa patients.[12] The validation study demonstrated that the Briganti 2012 nomogram and the MSKCC web-calculator were best at predicting LNI. Currently, the cost-effectiveness of using different risk thresholds for ePLND is unknown. Therefore, the purpose of this study is to apply a cost-effectiveness analysis to identify the best risk threshold for the MSKCC web-calculator and the Briganti 2012 nomogram, from a set of five realistic threshold values, to inform decision making on performing ePLND in a Dutch healthcare context. The cost-effectiveness analysis of the MSKCC web-calculator is shown in the paper. The analysis of the Briganti 2012 nomogram is displayed in supplemental data S3 (scenario 2).

Methods

Target population

The proportion of patients with and without pathohistologically proven LNI above and below different risk thresholds applied in practice (e.g. 2%, 5%, 10%, and 20%) were derived from the recently performed validation study by Hueting et al. and used as input in the decision analytic model.[12] The derived proportions for the MSKCC web-calculator are displayed in table 1. The population used for the validation study consisted of 1001 Dutch PCa patients of which 925 were eligible for validating the MSKCC web-calculator. The number of patients with confirmed LNI that would have been missed when applying a 2%, 5%, 10%, 20% or 100% risk threshold to perform ePLND were 1 (0.1%), 12 (1.3%), 27 (2.9%), 72 (7.8%), and 276 (29.8%), respectively. On the other hand, unnecessary ePLND could have been spared in patients with confirmation of having no LNI. Applying a 2%, 5%, 10%, 20% or 100% risk threshold resulted in the safe omission of ePLND in 53 (5.7%), 177 (19.1%), 311 (33.6), 458 (49.5%), and 649 (70.2%) patients, respectively. The applied ePLND template included removal of the nodes overlying the internal and external iliac artery, nodes located within the obturator fossa, and optionally within the common iliac artery and presacral areas.

Model development

A decision tree was constructed to evaluate the cost-effectiveness of different risk thresholds for performing ePLND (figure 1). The development of the tree was based on published clinical guidelines.[1,2,13] Five strategies were compared; applying a 2%, 5%, 10%, 20% and a 100% threshold to the predicted risk of LNI to guide application of ePLND. The 100% threshold represents the strategy in which no ePLND is performed in any patient (i.e. all patients will have a risk of LNI less than 100%) and was used as a reference strategy in the analysis. All patients in the decision tree underwent RP and based on their characteristics and predicted risk, and the selected risk threshold, they did or did not receive ePLND. The patients who received ePLND could experience complications from the procedure. Patients with histopathologically proven LNI received either observation, Adjuvant hormonal deprivation therapy (ADT) or a combination of ADT and adjuvant radiotherapy (ART) as indicated in the EAU guideline.[1]

Probabilities of ePLND related complications, adjuvant treatment, quality of life values (utilities) for the health outcomes following with or without concomitant ePLND, and costs were derived from available literature. An overview of the probabilities, utilities, and costs used in the analysis, with respective evidence sources, is shown in tables 1 and 2.

In the decision analytic model, patients receiving either ADT or a combination of ADT+ART when having proven LNI have a survival benefit compared to patients who do not receive adjuvant treatment. However, there is a lack of substantial evidence for any treatment benefit in patients receiving ePLND compared to patients who did not receive ePLND. For this reason, we also analyzed a scenario in which the 10-year survival outcomes are similar for patients with positive LNI regardless of whether ePLND was performed. This scenario analysis was added in the supplementary data S3 (Scenarios 3 & 4). Due to lack of evidence, several assumptions were necessary to develop the decision analytic model. The assumptions made were outlined in supplementary data S4

Outcomes

The strategies were compared in terms of health outcomes (Quality-Adjusted Life Years (QALYs)) and costs. One QALY equals one year in perfect health.[14] In the model, each strategy results in one of the end nodes, representing the consequences of (not) performing ePLND, experiencing complications, and receiving subsequent treatment. Expected health outcomes in QALYs were calculated using post RP survival data from available literature and is added as supplementary data S1.[6,15–19] Survival outcomes were reported as progression free survival, biochemical recurrence, metastasized disease, and overall mortality. Reported outcomes were different between patients with and without LNI, and different in patients with LNI who received adjuvant treatment compared to patients with LNI who did not receive without adjuvant treatment. QALYs were calculated by multiplying the probability of these outcomes by its corresponding utility value and summing these values over the total time span of ten years following RP. As available survival data from published papers were mostly limited to ten year survival rates, a ten year time horizon was applied to avoid data extrapolation. QALYs were discounted with 1.5% and costs with 4.0% each year according to the Dutch guideline to perform Health Economic evaluations.[20] The derived health outcomes and an example of the calculation of QALYs are presented in the supplementary data S1. In the calculation of QALYs expected over a 10-year period, health outcomes were allowed to change over time. For instance it was found that the health state utility of RP was 0.67 for the first year following treatment, and increased to 0.90 for the second year following treatment.[21,22] Comparable utilities were found for ART of which the utility value for the first year following treatment was 0.73,[21] and 0.89 for the second year[23]. Disease burden for thromboembolic events was taken into account in patients experiencing this complication for the first 18 months following treatment, by then, either the patient died from the event or would be completely cured. According to Versteegh et al.[24] the age specific utility of healthy individuals aged 60-69 years in the Netherlands is 0.84, and 0.85 for patients aged 70-79. These utilities were used as a ceiling value so that patients with PCa could not have a higher utility value than the average utility observed in healthy individuals of the same age.

The mean costs and corresponding standard errors of ePLND, ART & ADT were derived from pricelists (passantenprijslijsten) published by Dutch hospitals.[32] Annual management costs of biochemical recurrent disease and metastasized disease originate from a U.S. population described in 2012 and were converted from Dollars to Euros (conversion rate 1 USD = 0.765 Euro per December 2012) and adjusted to 2019 using the Dutch consumer price indices.

The five strategies were compared, amongst each other, using the incremental cost effectiveness ratio (ICER) in which the difference in mean costs is divided by the difference in mean QALYs achieved.

Analysis

To reflect uncertainty in the evidence used in the model all parameters were described with parametric distributions. Beta distributions were used for all utilities and probabilities. Gamma distributions were used for costs. Uncertainty in outcomes was then assessed by performing a probabilistic sensitivity analysis generating 5,000 samples. Results were visualized in the incremental cost-effectiveness plane using the 100% risk threshold strategy as a reference. ICERs were assessed by decreasing the threshold step-by-step to assess the additional costs of improving health outcomes by performing more and more ePLND procedures (20% vs 100%, 10% vs 20%, 5% vs 10%, and 2% vs 5%). The probabilities of strategies being cost-effective were visualized in a cost-effectiveness acceptability curve. For decision making, a willingness to pay (WTP) of €20,000/QALY was applied, which is the lower bound of the WTP range applied in the Netherlands as advised by the national healthcare institute.[25] To inform decision makers from other countries with different WTP thresholds, a cost-effectiveness acceptability curve (CEAC) was displayed with thresholds ranging between €0/QALY and €100,000/QALY. The costs used in the analyses were derived from a health care perspective, using only direct and indirect medical costs. All analyses were performed using Microsoft Excel 2016.

Table 1. Model input per threshold based on a validation study of 925 Dutch patients in the MSKCC web-calculator.

Threshold	Proportion of patients with predicted LNI risk exceeding the threshold	Proportions of patients with predicted LNI risk below the threshold	Proportion of patients with positive LNI below threshold*	Proportion of patients with positive LNI above threshold*
2%	0.96	0.04	0.001	0.30
5%	0.84	0.162	0.01	0.29
10%	0.66	0.345	0.03	0.27
20%	0.37	0.635	0.08	0.21
100%	0	1	0.30	0

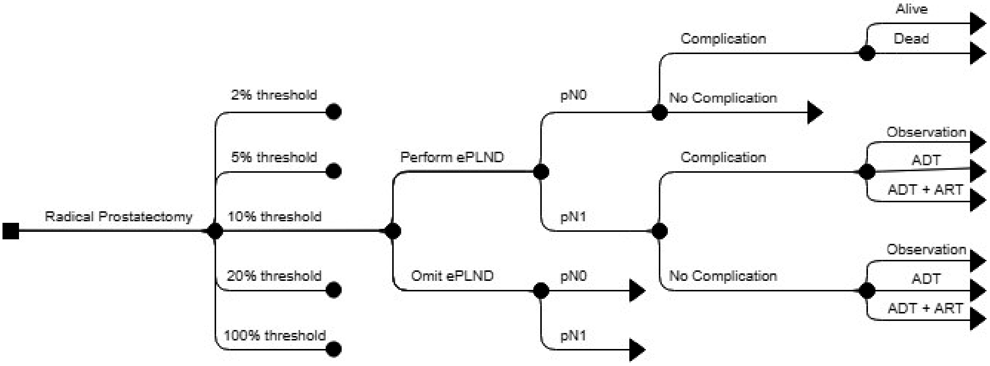
Uncertainty for each probability was assessed using beta distribution.
*In the decision tree, for all five options, there are two branches. The branch not showing in this table is the complement of the shown probability for that branch.

Table 2. Input parameters of the model

Parameter	Mean	SD	Distribution	Source:
Utilities				
Biochemical recurrence	0.67	0.24	Beta	[21]
Metastatic disease	0.25	0.11	Beta	[21]
Orchiectomy	0.87	0.16	Beta	[21]
Hormonal injection	0.83	0.19	Beta	[21]
Adjuvant Radiotherapy*	0.73	0.3	Beta	[21]
Radiotherapy post 1st year**	Disutility -0.11		N.A.	[23]
DVT	0.84	0.09	Beta	[27]
PE	0.63	0.13	Beta	[27]
Age specific 60-69 years	0.84	0.18	Beta	[24]
Age specific 70-79 years	0.85	0.15	Beta	[24]
Probabilities				
Lymphocele	0.067	0.01	Beta	[5]
DVT***	0.019	0.004	Beta	[5,28]
PE***	0.015	0.006	Beta	[5,28]
Lymphocele & DVT***	0.082	0.032	Beta	[5]
Lymphocele & PE***	0.028	0.019	Beta	[5]
DVT Death	0.021	0.002	Beta	[29]
PE Death	0.020	0.002	Beta	[29]
Observation	0.28	0.012	Beta	[16]
Adjuvant Hormonal Therapy	0.49	0.013	Beta	[16]
Adjuvant Radiotherapy	0.23	0.011	Beta	[16]
Costs (€)				
PLND	5912	1066	Gamma	[30]
Orchiectomy	4342	269	Gamma	[18]
Hormonal injection	633	51	Gamma	[30]
Adjuvant radiotherapy	2133	166	Gamma	[18]
Yearly management of biochemical recurrent disease	1992	490	Gamma	[31]
Yearly management of metastasized disease	2394	611	Gamma	[31]
DVT	1187	259	Gamma	[27]
PE	4221	922	Gamma	[27]

* Utility values for radical prostatectomy and adjuvant radiotherapy only accounted for the first year following treatment
** Utility values accounted for the second year following treatment
*** Utility values for DVT and PE accounted for the first 18 months following treatment.
Abbreviations: DVT: Deep venous thrombosis, PE: Pulmonary embolism, ePLND: pelvic lymph node dissection.

Figure 1. Decision analytic model



Results

The decision analytic model is displayed in figure 1. The branches behind the first probability node (i.e. 5% threshold) were defined identical for all risk thresholds, but not shown to improve visual clarity. Complications in the decision tree consist of lymphoceles (mean incidence: 6.7% ± 1.0%), DVT (1.9% ± 0.4%) and PE (1.5% ± 0.6%), with an increased probability of DVT (8.2% ± 3.2%) and PE (2.8% ± 1.9%) once lymphoceles occurred. Estimated outcomes in the model are displayed in table 3, showing that the mean QALYs range from 5.0 to 6.8, and the mean costs range from €3,823 to €17.697. Differences in outcomes are caused by LNI and treatment received (ePLND, ADT, and ART), see supplementary data S1 for calculation. For all five strategies analyzed, utilities and costs assigned to health outcomes were identical, however, the probabilities of receiving ePLND, and proportion of patients with LNI receiving ePLND was different between strategies. The cost of management of biochemical recurrent and metastasized disease, and ADT injections were the only costs induced annually. Costs for these outcomes were multiplied by the probability of the outcome for each year and summed over ten years. The treating physician has fewer options to personalize further treatment options in patients who did not receive ePLND, causing an increased risk of disease progression (i.e. biochemical recurrence, metastasized disease, and death) in patients with undetected LNI. In the supplementary data, costs have been converted to US Dollars to calculate the results.

Figure 2 displays the results of the probabilistic sensitivity analysis for all five strategies. Displayed are incremental QALYs and incremental costs for the 2%, 5%, 10%, and 20% risk thresholds compared to the 100% risk threshold (reference). The majority of simulated samples are found in the northeast quadrant meaning that both costs and QALYs are higher for the 2%, 5%, 10%, and 20% thresholds compared to the 100% risk threshold.

The cost-effectiveness acceptability curve (CEAC) shows the probabilities of the five analyzed strategies being cost-effective for WTP thresholds between €0/QALY and €100,000/QALY (figure 3). The CEAC shows that the 100% strategy has the highest probability of being cost-effective when applying a €20,000/QALY WTP threshold. Probabilities for the 2%, 5%, 10%, 20%, and 100% strategies being cost-effective at this WTP threshold were 0.0%, 0.3%, 4.9%, 30.3%, and 64.5%, respectively.

The results of the analysis performed on the Briganti 2012 nomogram instead of the MSKCC web-calculator are presented in supplementary data S3: Scenario 2. The alternative scenario in which patients with confirmed LNI did not have any treatment benefit over patients with unidentified LNI are displayed in supplementary data S3: Scenario 3 and 4. The alternative scenario shows that the 100% threshold strategy is dominant over the other strategy thresholds.

Table 3. Calculated QALYs and Costs used for the outcomes in the decision tree

Health states	Average calculated QALYs	Average calculated costs
Positive LNI without AT without ePLND	5.04	€ 8,640
Positive LNI without AT with ePLND	5.03	€ 14,653
Positive LNI with ADT*	5.85	€ 16,192
Positive LNI with ADT and ART*	5.77	€ 17,798
Negative LNI without ePLND	6.49	€ 3,823
Negative LNI with ePLND	6.48	€ 9,836

Calculations were added as supplementary data. Note: Costs can increase and QALYs can decrease based on the probability of DVT or PE occurring, these probabilities differ per threshold caused by different input probabilities. Abbreviations: ADT = Adjuvant hormonal therapy, ART = Adjuvant radiotherapy, AT = Adjuvant therapy, LNI = Lymph node involvement, ePLND = Pelvic lymph node dissection, QALY = Quality adjusted life year. *PLND included

Table 4. Results of the five thresholds analyzed in the decision tree using a time horizon of 10 years

Threshold	Average QALYs after 10 year	Average costs after 10 year		ICER*	QALY differences*	Cost differences*
100%	6.05	€4,867				
20%	6.17	€ 7,357	20% vs 100%:	€ 20,631	0.12	€ 2,490
10%	6.20	€ 9,178	10% vs 20%:	€ 60,607	0.03	€ 1,821
5%	6.21	€ 10,300	5% vs 10%:	€ 116,960	0.01	€ 1,122
2%	6.21	€ 11,050	2% vs 5%:	€ 682,469	0	€ 750

ICERS were calculated from top to bottom, displaying the ICER of each step taken. The 100% threshold regards the scenario in which no ePLND would be performed. Abbreviations: ICER = Incremental cost-effectiveness ratio, QALY = Quality adjusted life years. * Compared with the row above

Figure 2. Probablistic sensitivity analysis using the 100% threshold as reference for comparison.

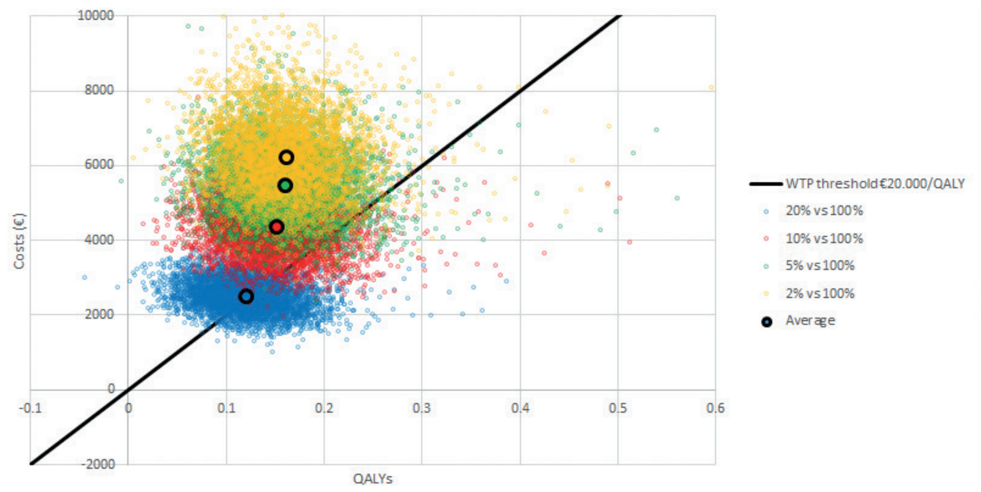
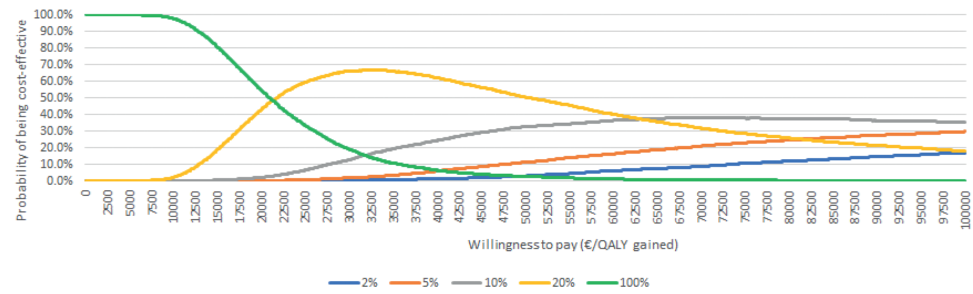


Figure 3. Cost-effectiveness acceptability curve



Discussion

The purpose of this study was to identify the best threshold value for the Briganti 2012 nomogram and MSKCC web-calculator, from a set of five realistic threshold values, to perform or omit ePLND in prostate cancer patients using a cost-effectiveness analysis. When the risk threshold decreases from 100% to 2% health outcomes consistently improve and costs consistently increase. Applying a WTP of €20,000/QALY gained, decreasing the risk threshold from 100% to lower values would not be cost-effective, that is, would result in too limited health benefits to outweigh the additional costs. This implies that, from a health economic perspective, for this WTP value, and using these prediction models, ePLND should not be performed in this patient group. However, for higher WTP threshold values, for example, €30,000/QALY gained and higher, use of a 20% or 10% risk threshold has the highest probability of being cost-effective. Such threshold values may appear to be high compared to previous recommendations. This makes sense as evaluations only focusing on health outcomes will, in this case, always prefer low threshold values. The cost-effectiveness analysis enables to estimate the optimal threshold to perform or omit ePLND. However, the optimal threshold may be different from the best strategy identified here, as we choose to evaluate five plausible threshold values (based on current guideline recommendations) rather than evaluate all possible threshold values.

Our study had several limitations. In this evaluation it was assumed that patients receiving ePLND with histopathological proven LNI may consequently receive ADT or a combination of ADT and ART. However, currently there is no consensus on the most effective timing and treatment modality for administering ADT, which may lead to variation in healthcare outcomes and costs in practice. In addition, the reported outcomes in available literature were, with exception of two randomized controlled trials,[15,26] solely based on retrospective data. Consequently, high quality evidence was not available for all input parameters of the decision analytic model. The decision analytic model was based on treatment recommendations from the EAU guideline[1] in which three postoperative treatment strategies were discussed; observation, ADT, or a combination of ADT and ART. The strategies were substantiated by the long term survival data reported bij Touijer et al.[18] However, in current clinical practice alternative treatment options may also be applied. In addition, certain urological methods may not support adjuvant treatment based on lymph node status, but are followed by postoperative procedures based on the presence of residual disease (i.e. reflected by (in) measurable PSA levels). The discrepancy between guideline recommendations and clinical practice may be partly explained by the fact that the outcomes of interest are often reported over 10-years after treatment. For instance, the fairly recent paper by Touijer et al. in 2018 reflects clinical decision making in patients who received treatment between 1988 and 2010.[18]

Certain complications caused by ePLND such as neurological, vascular, and ureteral damage could not be taken into account in the analysis since evidence was lacking regarding their impact on quality of life and costs. However, it is unlikely that these complications would have had a large impact on the outcomes, because their risk is lower than 1%.[5] In addition, anxiety or reassurance for (not) knowing whether cancer had spread to the pelvic lymph nodes may support the decision to perform ePLND and may also influence outcomes following RP with or without concomitant ePLND. Yet, anxiety and reassurance were not incorporated into the current analysis as evidence regarding effect size and duration is lacking.

The current analysis was performed using a hypothetic cohort for which the baseline characteristics were based on a cohort of Dutch prostate cancer patients who underwent ePLND. As the analysis focused on the Dutch health care setting, generalizability of the results to other health care settings may be limited, especially for settings in which the patient characteristics vary highly from Dutch prostate cancer patients (i.e. with more high risk prostate cancers). In addition, the applied WTP threshold of €20,000/QALY was used as recommended by the Dutch government. Other WTP thresholds may be used by other countries. Figure 3 displays a range of feasible WTP thresholds to inform decision making in different health care settings.

The scenario in which the Briganti 2012 nomogram was assessed showed similar results as the MSKCC web-calculator (Supplement S3: Scenario 2). The analysis showed that applying a lower threshold (i.e. performing more ePLNDs) resulted in better health outcomes (e.g. higher QALYs). However, high-quality evidence to substantiate a beneficial therapeutic effect of ePLND is still lacking. Therefore, an alternative scenario was assessed in which patients with confirmed LNI did not experience any treatment benefit compared to patients with unidentified LNI (Supplement S3: Scenario 3&4). The results of this scenario showed that performing no more ePLND (i.e. applying a 100% threshold) is the dominant strategy compared to performing ePLND in patients with a risk above a 2%, 5%, 10%, or a 20% threshold, even for WTP thresholds up to €100,000 per QALY gained. Even in the absence of evidence supporting direct therapeutic value of ePLND a cost-effectiveness analysis may be valuable, for instance, to assess potential cost savings from ePLND, to identify the optimal risk threshold for providing ePLND, to inform policy makers on value-based aspects and trade-offs related to ePLND, or to guide future research on this topic. Until evidence on the true therapeutic value of ePLND becomes available, it remains unclear whether performing ePLND is cost-effective at all.

Conclusion

The current results suggest very limited value of ePLND in patients with risk of LNI less than 10%. Which risk would be 'high enough' to consider ePLND is likely to be topic of further discussion, and part of the shared decision making process between clinicians and patients. However, Finally, when new evidence on the actual therapeutic value of ePLND would become available, the presented analysis should be updated.

References

1. Mottet, N. *et al.* EAU-ESTRO-SIOG Guidelines on Prostate Cancer. Part 1: Screening, Diagnosis, and Local Treatment with Curative Intent. *Eur. Urol.* 71, 618–629 (2017).
2. Reijke, d., Th.M., Coenen, J. L. L. M., Gietema, J. A., Moorselaar, v., R.J.A., Vogel, W. V. Dutch guideline prostate cancer. *Dutch Assoc. Urol.* (2016).
3. Briganti, A. *et al.* Pelvic Lymph Node Dissection in Prostate Cancer. *Eur. Urol.* **55**, 1251–1265 (2009).
4. Fossati, N. *et al.* The Benefits and Harms of Different Extents of Lymph Node Dissection During Radical Prostatectomy for Prostate Cancer: A Systematic Review. *Eur. Urol.* 72, 84–109 (2017).
5. Loeb, S., Partin, A. W. & Schaeffer, E. M. Complications of pelvic lymphadenectomy: do the risks outweigh the benefits? *Rev. Urol.* 12, 20–4 (2010).
6. Boorjian, S. A. *et al.* Long-Term Outcome After Radical Prostatectomy for Patients With Lymph Node Positive Prostate Cancer in the Prostate Specific Antigen Era. *J. Urol.* 178, 864–871 (2007).
7. Briganti, A. *et al.* Updated Nomogram Predicting Lymph Node Invasion in Patients with Prostate Cancer Undergoing Extended Pelvic Lymph Node Dissection: The Essential Importance of Percentage of Positive Cores. *Eur. Urol.* 61, 480–487 (2012).
8. MSKCC. Pre-operative tool to predict probability of lymph node involvement in prostate cancer patients. (2017). Available at: https://www.mskcc.org/nomograms/prostate/pre-op.
9. Tosoian, J. J. *et al.* Prediction of pathological stage based on clinical stage, serum prostate-specific antigen, and biopsy Gleason score: Partin Tables in the contemporary era. *BJU Int.* 119, 676–683 (2017).
10. Roach, M., Marquez, C., Yuo, H., … P. N.-I. J. of & 1994,. Predicting the risk of lymph node involvement using the pre-treatment prostate specific antigen and Gleason score in men with clinically localized prostate cancer. *Elsevier*
11. Porter, M. E. What Is Value in Health Care? *N. Engl. J. Med.* 363, 2477–2481 (2010).
12. Hueting, T. A. *et al.* External Validation of Models Predicting the Probability of Lymph Node Involvement in Prostate Cancer Patients. *Eur. Urol. Oncol.* 1, 411–417 (2018).
13. Leyh-Bannurah, S.-R. *et al.* North American Population-Based Validation of the National Comprehensive Cancer Network Practice Guideline Recommendation of Pelvic Lymphadenectomy in Contemporary Prostate Cancer. *Prostate* 77, 542–548 (2017).
14. Whitehead, S. J. & Ali, S. Health outcomes in economic evaluation: the QALY and utilities. *Br. Med. Bull.* 96, 5–21 (2010).
15. Messing, E. M. *et al.* Immediate versus deferred androgen deprivation treatment in patients with node-positive prostate cancer after radical prostatectomy and pelvic lymphadenectomy. *Lancet Oncol.* 7, 472–479 (2006).
16. Touijer, K. A. *et al.* Survival Outcomes of Men with Lymph Node-positive Prostate Cancer After Radical Prostatectomy: A Comparative Analysis of Different Postoperative Management Strategies. *Eur. Urol.* 73, 890–896 (2018).

17. Touijer, K. A., Mazzola, C. R., Sjoberg, D. D., Scardino, P. T. & Eastham, J. A. Long-term outcomes of patients with lymph node metastasis treated with radical prostatectomy without adjuvant androgen-deprivation therapy. *Eur. Urol.* 65, 20–5 (2014).
18. Da Pozzo, L. F. *et al.* Long-Term Follow-up of Patients with Prostate Cancer and Nodal Metastases Treated by Pelvic Lymphadenectomy and Radical Prostatectomy: The Positive Impact of Adjuvant Radiotherapy. *Eur. Urol.* 55, 1003–1011 (2009).
19. Boehm, K., et al. "No impact of blood transfusion on oncological outcome after radical prostatectomy in patients with prostate cancer." World journal of urology 33.6 (2015): 801-806.
20. IJzerman, M. J. Richtlijn voor het uitvoeren van economische evaluaties in de gezondheidszorg. *Zorginstituut Ned.* (2016).
21. Stewart, Susan T., et al. "Utilities for prostate cancer health states in men aged 60 and older." Medical care (2005): 347-355.
22. Koerber, F., Waidelich, R., Stollenwerk, B. & Rogowski, W. The cost-utility of open prostatectomy compared with active surveillance in early localised prostate cancer. *BMC Health Serv. Res.* 14, 163 (2014).
23. Reed, S. D., Stewart, S. B., Scales, C. D. & Moul, J. W. A Framework to Evaluate the Cost-Effectiveness of the NADiA ProsVue Slope to Guide Adjuvant Radiotherapy among Men with High-Risk Characteristics Following Prostatectomy for Prostate Cancer. *Value Heal.* 17, 545–554 (2014).
24. M. Versteegh, M. *et al.* Dutch Tariff for the Five-Level Version of EQ-5D. *Value Heal.* 19, 343–352 (2016).
25. Zwaap, J., Knies, S., van der Meijden, C., Staal, P. & van der Heiden, L. Cost-effectiveness in practice. *National Healthcare Institute.* (2015). Available at: https://english.zorginstituutnederland.nl/publications/reports/2015/06/16/cost-effectiveness-in-practice.
26. Messing, Edward M., et al. "Immediate hormonal therapy compared with observation after radical prostatectomy and pelvic lymphadenectomy in men with node-positive prostate cancer." New England Journal of Medicine 341.24 (1999): 1781-1788.
27. Bamber, L. *et al.* Cost-effectiveness analysis of treatment of venous thromboembolism with rivaroxaban compared with combined low molecular weight heparin/vitamin K antagonist. *Thromb. J.* 13, 20 (2015).
28. Allaf, M. E., Partin, A. W. & Carter, H. B. The importance of pelvic lymph node dissection in men with clinically localized prostate cancer. *Rev. Urol.* 8, 112–9 (2006).
29. Flinterman, L. E., van Hylckama Vlieg, A., Cannegieter, S. C. & Rosendaal, F. R. Long-Term Survival in a Large Cohort of Patients with Venous Thrombosis: Incidence and Predictors. *PLoS Med.* 9, e1001155 (2012).
30. Pricelists. Catharina Wilhelmina Ziekenhuis, Ziekenhuis Groep Twente, St. Antonius, Amsterdam Medisch Centrum, Viecurie, Treant, Medisch Centrum Leeuwarden, Tjongerschans, Elkerliek, Anthoni van Leeuwenhoek. (2019).
31. Cooperberg, M. R. *et al.* Primary treatments for clinically localised prostate cancer: a comprehensive lifetime cost-utility analysis. *BJU Int.* 111, 437–450 (2013).

Chapter 8

Summary & Samenvatting

General summary

Clinical prediction models are statistical tools that can be used to estimate the probability of a patient to either have a specific outcome or to develop an outcome in time. This probability is estimated based on patient or disease-specific input variables. It provides insights into the diagnosis (e.g. disease status) or prognosis (e.g. 5-year survival probability) of a patient, and can subsequently be used to support (shared) decision-making regarding the optimal management of the disease. Prediction models are developed and evaluated using data from patients that can be classified in similar patient groups (e.g. diagnosed with estrogen receptor positive breast cancer), but with varying disease characteristics (e.g. tumor stage, treatment received, nodal involvement etc.).

Before the available models are used to support in routine healthcare decision-making some challenges on the identification of currently existing models (accessibility), review of the quality of the models (transparency), assessment how well they perform on external validation (generalizability), and investigation of the potential benefit of recalibrating the validated models (updating). Subsequently, models showing adequate performance will be ready for implementation in clinical practice after clearly defined intended model use is described (interpretation), and the intended model use is substantiated by evidence regarding added value (impact assessment).

In this thesis, multiple studies aiming to overcome the challenges are described using examples on breast and prostate cancer. Since breast and prostate cancer are among the top three most commonly diagnosed cancers in women and men, respectively, there is a large amount of data available to establish clinical prediction models for patients diagnosed with breast or prostate cancer. Currently available models for breast and prostate cancer are required to be critically assessed to demonstrate which models are valuable and which information is still lacking when used in Dutch care.

Chapter 2 describes the systematic literature review that was performed to identify all clinical prediction models that were developed between 2010 and 2020 for patients diagnosed with breast cancer to predict outcomes related to treatment decision-making. A total of 922 prediction models were described in 534 articles. A large majority of the identified models were found to be at high risk of bias according to the prediction model risk of bias tool (PROBAST). The outcomes predicted with the different models concerned mortality, disease recurrence, lymph node involvement, adverse events, treatment response, menopausal status, quality of life, surgical margins, receiving treatment, cosmetic outcome, or nipple-areola complex invasion. Commonly used predictors included age, tumor size, and lymph node involvement. A substantial number of models demonstrated flaws in the reporting or execution of development and/or validation of the model, making their clinical utility uncertain.

Chapter 3 describes the external validation of the models that were identified in the systematic review described in Chapter 2. The models were validated when sufficient data was provided to apply the described model to new patients, when the required data was sufficiently available in the Netherlands Cancer Registry (NCR), and when the models were not already developed or validated using the NCR data. Finally, 87 models could be externally validated. A total of 34 (39%) models showed a good performance on the NCR data, defined as a scaled Brier score >0 and C-index ≥0.7. Another 26 (30%) models showed a moderate performance (scaled Brier score >0 and C-index <0.7 or a scaled Brier score ≤0 with C-index ≥0.7). The remaining 27 (31%) models showed poor performance (Scaled Brier score ≤0 and C-index <0.7)

Chapter 4 provides the steps that were taken to develop and internally validate an updated INFLUENCE model. The first model was developed using logistic regression, older patient data, did not incorporate all desired predictor variables, and lacked the prediction of contralateral breast cancer. The newly updated model includes the desired improvements, and also incorporates the prediction of distant metastasis. Three modelling techniques (cox regression, parametric spline, and random survival forest) were compared to predict three outcomes (locoregional recurrence, second primary contralateral breast cancer, and distant metastasis). The best performing models were selected based on discrimination and calibration of the outcomes. The random survival forest model was found to be the best performing model for the prediction of locoregional recurrence (AUC: 0.75) and second primary contralateral breast cancer (AUC: 0.67), and the Cox regression model most accurately predicted distant metastasis (AUC 0.77). An online calculator was constructed to use the newly developed models for patient care.

Chapter 5 describes the external validation of models predicting pelvic lymph node involvement (LNI) in prostate cancer patients. International guidelines currently recommend the performance of pelvic lymph node dissection for prostate cancer patients, but the recommendations have changed over the years, making the use of clinical prediction models to assess which patient benefits from a dissection more important. Using data from two hospitals, supplemented by registry data, the models were externally validated. Based on discrimination and calibration, the models developed by Briganti et al. (AUC: 0.76) and Memorial Sloan Kettering Cancer Center (MSKCC) (AUC: 0.75) performed best and are recommended for the estimation of LNI risk in Dutch prostate cancer patients.

Chapter 6 further assesses the models by Briganti et al. and MSKCC by comparing their performance using different methods to estimate the input variables. Data such as clinical tumor stage can be estimated using digital rectal examination or multiparametric magnetic resonance imaging. The two methods to define the variable may result in significantly different tumor staging and may affect the performance of prediction models incorporating this variable. The models developed by Briganti et al. and MSKCC were compared head-to-head by assessing both models on discrimination, calibration, and net benefit, with T-stage determined using mpMRI and DRE. For both models, the mpMRI T stage is an appropriate alternative instead of DRE T stage when using the models to predict LNI. The AUCs with DRE T-stage were 0.71 and 0.73, and 0.72 and 0.75 with mpMRI T-stage, for the MSKCC and Briganti models, respectively. The head-to-head comparison showed that the model developed by Briganti et al. using mpMRI T-stage performed best.

Chapter 7 presents a cost-effectiveness analysis to compare the impact of applying either the Briganti or MSKCC model in clinical practice. International guidelines recommend different models to predict the risk of LNI in prostate cancer patients and apply different thresholds to assess which patient is at sufficiently low risk to safely omit extensive pelvic lymph node dissection. The analysis was performed using a decision analytic model where the assumption was made that ePLND may indirectly improve the prognosis in pN1 patients by improved indication of patients for adjuvant treatment. The different thresholds recommended in international guidelines resulted in varying incremental costs and incremental quality adjusted life years (QALYs) over 10 years. Results were similar for both the Briganti and MSKCC models. Applying a willingness-to-pay (WTP) threshold of €20.000,- per QALY gained, performing no ePLND is cost-effective compared to other feasible risk-based strategies. However, with a higher WTP threshold, a 20% risk threshold for probable LNI may be a cost-effective alternative.

Nederlandse samenvatting

Klinische predictiemodellen zijn statistische instrumenten die kunnen worden gebruikt om de waarschijnlijkheid te schatten dat een patiënt een bepaald resultaat zal hebben of dat een bepaald resultaat zich in de tijd zal ontwikkelen. Deze waarschijnlijkheid wordt geschat op basis van patiënt- of ziekte specifieke inputvariabelen. De waarschijnlijkheid geeft inzicht in de diagnose (bv. ziektestatus) of prognose (bv. 5-jaars overlevingskans) van een patiënt, en dit kan vervolgens worden gebruikt ter ondersteuning van de (gedeelde) besluitvorming omtrent de optimale zorg van de ziekte. Predictiemodellen worden ontwikkeld en geëvalueerd aan de hand van gegevens van patiënten die in soortgelijke patiëntengroepen kunnen worden ingedeeld (bv. gediagnosticeerd met oestrogeenreceptor positieve borstkanker), maar met verschillende ziektekenmerken (bv. tumorstadium, behandeling, lymfeklierbetrokkenheid, enz.)

Voordat de beschikbare modellen worden gebruikt ter ondersteuning van de routinematige besluitvorming in de gezondheidszorg, zijn er een aantal uitdagingen: identificatie van de momenteel bestaande modellen (toegankelijkheid), beoordeling van de kwaliteit van de modellen (transparantie), beoordeling hoe goed zij presteren bij externe validatie (generaliseerbaarheid), en onderzoek naar het potentiële voordeel van het kalibreren van de gevalideerde modellen (updaten). Vervolgens zullen modellen met adequate prestaties klaar zijn voor implementatie in de klinische praktijk nadat het beoogde modelgebruik duidelijk is omschreven (interpretatie), en het beoogde modelgebruik is onderbouwd met bewijs omtrent de toegevoegde waarde (impact beoordeling).

In dit proefschrift worden verschillende studies beschreven die een antwoord willen vinden op deze uitdagingen, aan de hand van voorbeelden van borst- en prostaatkanker. Aangezien borst- en prostaatkanker respectievelijk tot de top drie van meest gediagnosticeerde kankersoorten bij vrouwen en mannen behoren, is er een grote hoeveelheid gegevens beschikbaar om klinische predictiemodellen te ontwikkelen voor patiënten bij wie borst- of prostaatkanker wordt gediagnosticeerd. Momenteel beschikbare modellen voor borst- en prostaatkanker dienen kritisch te worden beoordeeld om aan te tonen welke modellen waardevol zijn en welke informatie nog ontbreekt bij gebruik in de Nederlandse zorg.

Hoofdstuk 2 beschrijft het systematische literatuuronderzoek dat is uitgevoerd om alle klinische predictiemodellen te identificeren die tussen 2010 en 2020 zijn ontwikkeld voor patiënten met de diagnose borstkanker om uitkomsten te voorspellen met betrekking tot de besluitvorming over de behandeling. In totaal werden 922 predictiemodellen beschreven in 534 artikelen. Een grote meerderheid van de geïdentificeerde modellen bleek een hoog risico op bias te hebben volgens de predictiemodel risk of bias tool

(PROBAST). De uitkomsten die met de verschillende modellen werden voorspeld betroffen mortaliteit, terugkeer van de ziekte, lymfeklierbetrokkenheid, ongewenste gevolgen, behandelingsrespons, menopauzestatus, kwaliteit van leven, chirurgische snijvlakken, behandeling, cosmetisch resultaat, of tepel-areola complex. Vaak gebruikte voorspellers waren leeftijd, tumorgrootte en lymfeklierbetrokkenheid. Een aanzienlijk aantal modellen vertoonde gebreken in de rapportage of uitvoering van de ontwikkeling en/of validatie van het model, waardoor hun klinische bruikbaarheid onzeker was.

Hoofdstuk 3 beschrijft de externe validatie van de modellen die werden geïdentificeerd in de systematische review beschreven in Hoofdstuk 2. De modellen werden gevalideerd wanneer er voldoende data beschikbaar was om het beschreven model toe te passen op nieuwe patiënten, wanneer de benodigde data in voldoende mate beschikbaar was in de Nederlandse Kankerregistratie (NKR), en wanneer de modellen niet reeds ontwikkeld of gevalideerd waren met behulp van de NKR data. Uiteindelijk konden 87 modellen extern gevalideerd worden. In totaal lieten 34 (39%) modellen een goede prestatie zien op de NKR data, gedefinieerd als een geschaalde Brier score >0 en AUC ≥0.7. Nog eens 26 (30%) modellen presteerden matig (geschaalde Brier-score >0 en AUC <0,7 of een geschaalde Brier-score ≤0 met AUC ≥0,7). De overige 27 (31%) modellen presteerden slecht (Scaled Brier score ≤0 en AUC <0.7).

Hoofdstuk 4 beschrijft de stappen die werden ondernomen om een geüpdatete INFLUENCE model te ontwikkelen en intern te valideren. Het eerste model was ontwikkeld met behulp van logistische regressie, oudere patiëntgegevens, bevatte niet alle gewenste voorspellende variabelen, en miste de voorspelling van contralaterale borstkanker. Het nieuwe geüpdatete model bevat de gewenste verbeteringen en bevat ook de voorspelling van verre metastase. Drie modelleertechnieken (cox regressie, parametrische spline, en random survival forest) werden vergeleken om drie uitkomsten te voorspellen (locoregionaal recidief, tweede primaire contralaterale borstkanker, en afstandsmetastase). De best presterende modellen werden geselecteerd op basis van discriminatie en kalibratie van de uitkomsten. Het random survival forest model bleek het best presterende model te zijn voor de voorspelling van locoregionaal recidief (AUC: 0.75) en tweede primaire contralaterale borstkanker (AUC: 0.67), en het Cox regressiemodel voorspelde afstandsmetastase het meest accuraat (AUC 0.77). Er werd een online calculator gemaakt om de nieuw ontwikkelde modellen te gebruiken voor patiëntenzorg.

Hoofdstuk 5 beschrijft de externe validatie van modellen die de uitzaaiing van lymfeklieren in de bekkenregio bij prostaatkankerpatiënten voorspellen. Internationale richtlijnen bevelen momenteel de uitvoering van een lymfeklierdissectie aan voor prostaatkankerpatiënten. De aanbevelingen zijn in de loop der jaren veranderd, waardoor het gebruik van klinische predictiemodellen om te beoordelen welke patiënt baat heeft bij een dissectie belangrijker is geworden. Met behulp van gegevens van twee ziekenhuizen, aangevuld met registergegevens uit de NKR, werden de modellen extern gevalideerd. Op basis van discriminatie en kalibratie presteerden de modellen ontwikkeld door Briganti et al. (AUC: 0.76) en het Memorial Sloan Kettering Cancer Center (MSKCC) (AUC: 0.75) het beste en worden aanbevolen voor de schatting van het risico op lymfeklieruitzaaiing bij Nederlandse prostaatkankerpatiënten.

Hoofdstuk 6 beoordeelt verder de modellen van Briganti et al. en MSKCC door hun prestaties te vergelijken met behulp van verschillende methoden om de input variabelen te schatten. Gegevens zoals klinisch tumorstadium kunnen worden geschat met behulp van digitaal rectaal onderzoek (DRE) of multiparametrische magnetic resonance imaging (mpMRI). De twee methoden om de variabele te definiëren kunnen resulteren in een significant verschillende tumorstadiëring en kunnen de prestatie beïnvloeden van predictiemodellen waarin deze variabele is opgenomen. De modellen ontwikkeld door Briganti et al. en MSKCC werden met elkaar vergeleken door beide modellen te beoordelen op discriminatie, kalibratie, en netto winst, waarbij het T-stadium werd bepaald met mpMRI en met DRE. Voor beide modellen is het mpMRI T-stadium een geschikt alternatief in plaats van het DRE T-stadium wanneer de modellen worden gebruikt om lymfeklieruitzaaiing te voorspellen. De AUCs met DRE T-stadium waren 0.71 en 0.73, en 0.72 en 0.75 met het mpMRI T-stadium, voor respectievelijk het MSKCC- en het Briganti-model. De head-to-head vergelijking toonde aan dat het model ontwikkeld door Briganti et al. met mpMRI T-stadium het beste presteerde.

Hoofdstuk 7 presenteert een kosteneffectiviteitsanalyse om de impact te vergelijken van het toepassen van het Briganti of MSKCC model in de klinische praktijk. Internationale richtlijnen bevelen verschillende modellen aan om het risico op lymfeklieruitzaaiing bij prostaatkankerpatiënten te voorspellen en hanteren verschillende drempels om te beoordelen welke patiënt een voldoende laag risico heeft om een uitgebreide lymfeklierdissectie (ePLND) veilig achterwege te laten. De analyse werd uitgevoerd met gebruikmaking van een beslissingsanalytisch model waarbij de veronderstelling werd gemaakt dat ePLND indirect de prognose bij pN1 patiënten kan verbeteren door een betere indicatie van patiënten voor adjuvante behandeling. De verschillende drempels die in internationale richtlijnen worden aanbevolen resulteerden in verschillende incrementele kosten en incrementele voor kwaliteit gecorrigeerde levensjaren (QALY's) over 10 jaar. De resultaten waren vergelijkbaar voor zowel het Briganti- als het MSKCC-model. Bij toepassing van een willingness-to-pay (WTP)-drempel van € 20.000,- per gewonnen QALY is het uitvoeren van geen ePLND kosteneffectief in vergelijking met andere haalbare risicogeoriënteerde strategieën. Bij een hogere WTP-drempel kan een risicodrempel van 20% voor waarschijnlijke lymfeklieruitzaaiing echter een kosteneffectief alternatief zijn.

Chapter 9

General discussion

General discussion

The use of prediction models to support clinical decisions in the diagnosis and treatment of patients with breast or prostate cancer has become an integral part of everyday clinical practice.[1,2] The number of clinical prediction models that are being developed is increasing exponentially. Using the search strategy provided by Ramspek et al.[3], over 17,000 prediction model studies were published in 2020, while that number used to be about 7,300 in 2015, and about 3,800 in 2010. Even though the number of publications keeps increasing rapidly, the number of models being recommended in clinical guidelines does not seem to increase as fast. The Dutch breast cancer guideline recommends a small set of clinical prediction models that may be applied in clinical practice. For breast cancer, multiple international guidelines only recommend a few models to be used for treatment decision-making. Recommended models include PREDICT and genetic profiles such as Mammaprint or Oncotype DX.[2,4–6] Also for prostate cancer, the number of recommended models in international guidelines seems to be limited. Multiple of the recommended models are intended to assess the risk of lymph node involvement (LNI) (i.e. the Briganti nomogram,[7] Roach formula,[8] Partin tables,[9] MSKCC nomogram,[10] or the Gandaglia nomogram).[11] According to the European Association of Urology (EAU) prostate cancer guideline,[12] patients with an LNI risk over 5% (for the Briganti nomogram[7]) or 7% (for the Gandaglia nomogram[11]) are recommended to undergo an extended pelvic lymph node dissection (ePLND) during radical prostatectomy. However, these recommended thresholds seem to be based on consensus regarding acceptable sensitivity and specificity rather than the impact of the applied thresholds on health outcomes or cost-effectiveness of care. The fact that from such a large number of available models, only few are recommended for clinical use, indicates that there is still much to be done to demonstrate the (added) value of current models in clinical oncology. It is crucial to identify currently existing models (accessibility), review the quality of the models (transparency), assess how well they perform on external validation (generalizability), and investigate the potential benefit of recalibrating the validated models (updating). Subsequently, models showing adequate performance will be ready for implementation in clinical practice after clearly defined intended model use is described (interpretation), and the intended model use is substantiated by evidence regarding added value (impact assessment). In this thesis, multiple studies aiming to address these challenges are described using examples on breast and prostate cancer.

The funnel from the large number of published studies on prediction models to the small number of models recommended in clinical guidelines is also reflected by the studies reported in this thesis. All of the 922 identified models by the review reported in **chapter** 2 were considered for external validation in the study described in **chapter 3**. However, only 87 models could be externally validated with data from the Netherlands Cancer

Registry (NCR), and a minority of 34 models showed good performance. In order to optimize available models for the prediction of locoregional recurrence (LRR) and distant metastases (DM), and due to a lack of viable alternatives for the prediction of second primary contralateral breast cancer (SP), in **chapter 4**, a new set of models was developed to predict LRR, DM and SP. The three models were incorporated in an online tool with an intended use to support risk based follow-up strategies for patients treated for breast cancer. Additionally, **chapter 5** identified 16 different models developed to predict LNI in prostate cancer, but only two models performed well enough to be considered for another validation study outlined in **chapter 6** and the subsequent cost-effectiveness analysis reported in **chapter 7**.

The main conclusions of this thesis are:

- An overwhelming majority of clinical prediction models to support treatment decisions for breast cancer patients have been developed and internally validated using deficient methods and were reported incompletely.
- Due to the lack of transparency and available patient data, only a limited number of models for breast cancer patients could be validated. Of the models that could be validated, 34 (39%) performed well, 26 (30%) performed moderately, and 27 (31%) performed poorly and cannot be advised to be used in clinical practice.
- A newly updated INFLUENCE 2.0 model was developed using random survival forest and Cox regression models and showed robust performance on internal validation. An online web-calculator was established to predict the risk of LRR, SP, and DM, allowing physicians to personalize follow-up care for individual patients.
- Out of 16 models the Briganti 2012 nomogram and the Memorial Sloan Kettering Cancer Center online calculator showed the best performance and are advised to be used for LNI risk prediction in Dutch prostate cancer patients.
- The Briganti 2012 nomogram, using tumor stage determined by multiparametric magnetic resonance imaging (mpMRI), is advised to be used to predict LNI in prostate cancer patients.
- Application of lower risk thresholds on the Briganti 2012 nomogram and the MSKCC online calculator result in improved health outcomes and higher costs. Applying a willingness-to-pay threshold of €20,000, the complete omission of an ePLND in prostate cancer patients is cost-effective compared to other risk-based strategies. However, when higher willingness to pay thresholds are used, a 20% threshold for probable lymph node involvement may be a viable alternative.

With the conclusions and results of the studies in this thesis, important steps are taken towards the deployment of valuable prediction models in daily clinical practice for breast and prostate cancer in the Netherlands. However, several shortcomings have still been identified around the key challenges (i.e. accessibility, transparency, generalizability, updating, impact assessment, and interpretation) in the development, validation, and evaluation of prediction models in clinical oncology.

Accessibility

Depending on the goal of applying a developed clinical prediction model, different levels of accessibility are sufficient. For validation purposes, a transparent description of the model and the underlying equation makes the model sufficiently accessible. However, for clinical adoption of a prediction model, an interface is required to enable access to the model for clinical users lacking technical and statistical knowledge. To improve accessibility and allow clinical interpretation of a prediction model, a nomogram can be presented in the publication.[13] However, the calculation and presentation of a model is preferably available as a digital tool to enable integration in the physician's workflow (e.g. using online calculators).[14] But even when online calculators are created, accessibility may change over time. For instance, Adjuvant! Online was recommended in clinical guidelines, but the website (www.newadjuvant.com) is no longer available.[15] Another online tool, CancerMath (www.lifemath.net) is still online available. However, CancerMath is no longer usable due to the dependency of Adobe Flash Player, which is not supported anymore.[16] Even though CancerMath was found to provide accurate predictions in an external validation study using Dutch patients,[17] the accessibility of the tool currently relies on the update of the web application. The creation of online calculators has become easier in the past years with the use of software provided by e.g. Shiny (https://shiny.rstudio.com/) or Evidencio (www.evidencio.com). Both software platforms provide tools that allow users to freely create interactive tools and calculators. Nevertheless, the continuous availability of the online calculators will depend on the software party with which it was developed, and costs should be covered regarding maintenance and hosting of an online calculator. The best way to keep models accessible, is to describe the model and underlying equations in a transparent manner, in open-access publications or reports, preferable in combination with clearly annotated programming codes including example calculations.

Transparency

A complete and transparent description of a developed prediction model is arguably the most important step for carrying out the relevant activities required to ultimately adopt a model in clinical practice. Transparency is required to perform external validation studies, impact assessments, and to create accessible tools such as online calculators. Issues regarding the transparent reporting and the conduct of clinical prediction model development and validation studies have been known for years.[18,19] Multiple initiatives were taken to improve the quality of studies regarding the development and validation of prediction models. The transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) checklist has been developed to provide guidance regarding all required topics to ensure a complete and transparent description.

Also, multiple articles and books were published providing step-by-step guidance on appropriate methods to develop, validate, and update clinical prediction models.[20–23] In addition, the prediction model risk of bias assessment tool (PROBAST) was developed for critical appraisal of prediction model studies. A high risk of bias according to the PROBAST tool was associated with discriminative performance at validation.[24] Regardless of these efforts, the majority of developed prediction models still lack proper description, were developed using inadequate methods, and were found to be at high risk of bias as illustrated by various systematic reviews, including the review reported in **chapter 2**. Another systematic review by Yang et al. observed only limited improvements in the conduct and reporting of prediction models over the past years.[25] One of the key issues concerns the lack of a complete presentation (i.e. description of all model coefficients, including the intercept) of the final model, which was also the reason for exclusion of 262 (28%) of the models considered for external validation in **chapter 3**. In addition, black box machine learning models such as neural networks are increasingly being used. A complete and transparent description of black box models is difficult to report, but not impossible.[26] Authors should consider feasible options to enable external validation studies by independent researchers when developing clinical prediction models.[27] For instance, the INFLUENCE 2.0 model described in **chapter 4** predicts two of the three outcomes with a random survival forest model. To present the final model in a transparent manner, the model has been made available as an online calculator. Generally, to improve transparency specifically of machine learning-based prediction models, TRIPOD and PROBAST updates with extensions for the use of artificial intelligence in prediction model studies are underway.[28]

Generalizability

The generalizability of clinical prediction models is assessed in external validation studies using data collected in a different, but comparable setting than in the development of the model.[29,30] Ideally, external validation is performed by independent researchers.[31] Even though external validation is a crucial step that must be taken before models can be used in a setting other than the one in which they were deployed, external validation is carried out much less frequently than development of prediction models.[3,32–34] And even when external validation is performed, the methodological quality and reporting was mostly found to be inadequate.[35,36] In this thesis, only 87 identified models were externally validated in the study outlined in **chapter 3,** while a total of 922 models identified in the review reported in **chapter 2** were considered for external validation using data from the NCR. Reasons for not having been able to carry out the external validation included the lack of accessible, transparent, and a completely described final model, but also concerned the lack of sufficiently available patient data required to validate the models. This regards the availability of large enough sample sizes for external validation studies, and the availability of the required predictor and outcome information.[37,38] For instance, in the study reported in **chapter 5**, a recent and potentially valuable prediction model could not be externally validated due to the unavailability of any information regarding one of the incorporated predictors.[11] To ensure that newly developed prediction models can be validated in different settings, the creation of sub-models based on observed covariates, or an imputation approach using fixed chained equations are feasible solutions.[39] The proposed solutions are especially helpful when the association between predictor and outcome is weak and would improve the possibilities of applying models more easily in practice when predictor data is not routinely collected. This was the case, for example, for models considered for external validation in **chapter 3** in which outcomes partly were predicted by race (used in 89 models), or marital status (used in 54 models). However, it is debatable to which extent these variables may reflect e.g. underlying cultural differences and racial disparities in the health care system that create an unwanted bias when applying such models in practice.[40,41]

In prediction model validation studies, the performance of a model is evaluated using various measures and visualization options. Recommended aspects concern discrimination, calibration, and clinical usefulness.[22,42] The term validation suggests that a final conclusion can be drawn about whether or not a model is valid for the target population. Elaborate guidance has been published regarding the interpretation of discrimination, calibration, and clinical usefulness, but the use of predefined thresholds to conclude poor or good performance is arbitrary, and should generally be avoided.[43–45] Rather, the performance should be assessed in the context in which the model is applied, and which alternatives are available (e.g. other models or no model). For instance, a model showing modest discrimination with a C-statistic of e.g. 0.63 can still be useful if the related current clinical decision involves a "toss-up", and alternative models showed a worse performance.[42,44] In addition, general performance measures may be lower if a model performs well in only certain patient subgroups. This has been demonstrated, for example, in external validation studies of PREDICT.[46,47] On the other hand, external validation studies can also show that a model is more widely applicable than intended during the development of the model. This is the case for the validation study reported in **chapter 6** in which it was concluded that the existing models to predict LNI in prostate cancer patients could also be used for patients whose tumor stage was measured with an MRI instead of a DRE.

To enable independent investigators to perform external validation studies of online calculators more easily, the online platform Evidencio has implemented a validation module on the online platform (www.evidencio.com). The module semi-automatically provides performance measures including the C-statistic, (scaled) Brier score, and displays insightful plots including a classification plot, a calibration plot, and a decision curve.[48]

Updates

To improve predictions for future patients in a new setting, models can be updated. Methods to update a model range from a re-estimation of the model intercept, model recalibration (re-estimation of intercept and slope), model extension or reduction by adding or removing one or multiple predictors, to a complete revision of the model.[38,49] The updated INFLUENCE model of which the development and internal validation is described in **chapter 4** concerns a complete revision of the model. The new INFLUENCE 2.0 model can be considered an update as both models have an equivalent intended use, predict the risk of LRR, have been developed using data from the same registry (NCR), and all variables included in the previous model are also included in the updated model. However, the performance of parsimonious updating methods was not feasible as the initial INFLUENCE model was developed using logistic regression, and the updated model was developed with random survival forest and Cox regression. Also, the INFLUENCE 2.0 model was extended with predictions of the risk for SP and DM. The previous INFLUENCE model showed adequate performance on external validation in German patients.[50] Although the updated INFLUENCE model was internally validated, it cannot be assumed that the updated INFLUENCE model is also directly generalizable in the German population until this updated model is also externally validated.[38]

After models have proven to be valuable when implemented in practice, their performance should repeatedly be assessed over time as the context in which the model is used may change, or the way predictor variables are measured may change. Collection of data necessary to continuously validate (and subsequently update) implemented prediction models over time can be challenging. For instance, the models predicting LNI in prostate cancer patients highlighted in **part II** of this thesis are intended to be used to omit an ePLND in patients with a low risk (e.g. <5%). However, the ePLND procedure is still the standard method to assess LNI in prostate cancer patients. Omitting ePLND in patients with the lowest risk will result in a lack of available outcome data in this particular subgroup of patients. This effect was also partly reflected in the data used in **chapter 5**, where the observed event rate was 28%, compared to the event rate of 8% in one of the best performing models.[7] Patients with a risk <5% were still sufficiently present in the data in order to appropriately validate the models predicting LNI (n = 175), but the models were not updated due to the risk of introducing unwarranted bias in the newly derived models. Prior to the implementation of prediction models in clinical practice, appropriate steps should be determined on the data collection required to ensure that the model remains valid over time.

Updating a model is generally recommended when the predictive performance of a model is shown to be inadequate in external validation.[51] However, updating models already showing adequate performance can also be sensible, for instance when overall performance is adequate, but not in all patient subgroups. Eventually, slightly different versions of robust models, optimized in different local or regional settings would become available. The decision to update a model for a specific setting also relies on factors other than model performance alone. For instance, the results of the external validation reported in **chapter 5** showed a miscalibration of the Briganti 2012 model[7] for patients with an observed risk above 40%. However, clinical decision making is performed for patients with an LNI risk between 0% and 20%. Updating the model may improve overall performance, but not necessarily for the relevant risk thresholds. Therefore, it is recommended to perform decision curve analysis to assess the added net benefit of the updated model compared to the original model, at relevant risk thresholds.[52]

Impact assessment

Once models have been shown to provide accurate predictions, they may be ready for clinical use. Still, a strong case can be made for the need of further impact assessment prior to the integration of prediction models for decision support, as a model with good performance on discrimination and calibration can still be clinically irrelevant.[53] For instance, a model predicting lymph node involvement can be perfectly calibrated with a c-statistic of 0.8. However, such a model would still be useless if all predicted risks are between 60% and 90%, and the threshold relevant for decision making is set at 5% - 7%. All patients would be treated similarly when this well performing model is applied in clinical practice. An indication of clinical usefulness of a model is performed using decision curve analysis, in which the net benefit of applying a model over all feasible thresholds is compared to default strategies where either all or no patients are treated.[52] The real impact on health outcomes and costs of the use of a model for actual decision support is preferably assessed using (cluster-)randomized clinical trials.[38] However, randomized trials are often not feasible, especially when models predict long term outcomes (i.e. over 5 to 10 years). Also, most randomized trials are bound by strict protocols and therefore do not necessarily reflect real-world use of prediction models. Alternatively, impact assessment can be performed using e.g. health economic modelling where the use of the model is compared to usual care (i.e. no prediction model used, or the currently used prediction model).[54] Performance of a health economic evaluation is outlined in **chapter 7**. To assess the cost-effectiveness of applying prediction models to decide which prostate cancer patients require an ePLND, several assumptions have been made regarding the effectiveness of an ePLND. For instance, limited evidence was available for the effect of ePLND on relevant patient outcomes (i.e. recurrence and mortality), leaving the utility of prediction models for these clinical decisions uncertain.

Adequate clinical evaluations of prediction model implementations, such as cost-effectiveness analyses, are conducted far too infrequent, as illustrated by van Giessen et al.[55] Performing impact studies of course requires time and financial resources. However, such studies can directly contribute to further insight into the added value of prediction models and thereby their implementation in clinical guidelines and clinical practice.

Interpretation

After all, the important steps have been carefully carried out regarding the development, (external) validation, and impact assessment, a model is ready to be implemented in the clinical workflow. Clinical prediction models aim to accurately estimate the probability of an outcome with a set of predictors. The identification of causal factors is performed in etiological studies with a different objective than prediction studies.[56] However the aims, methodology, and the interpretation of the results for both studies are often conflated.[57] If there is confusion about the objectives of a study, for example if predictors are interpreted as modifiable risk factors while there is a risk of bias due to "confounding by severity",[58] undesired effects may be found in the eventual use of the model.[59] For proper clinical adoption, it is of paramount importance that the intended use of the model is clearly defined to ensure the correct interpretation by clinicians.[60,61] For example, the updated INFLUENCE model presented in **chapter 4** predicts the risk of breast cancer recurrence using predictors such as chemotherapy, radiotherapy, hormonal therapy, and targeted therapy. Even though it may seem evident, it is a pitfall to think that the model can be used to support the decision to administer these therapies to a patient based on their risk-reducing effect. The model was developed using observational data from patients for whom the decision to administer treatment had already been taken prior to the intended use of the model.[62] INFLUENCE is therefore not suitable to support the decision to initiate adjuvant treatment, whereas the PREDICT model has been developed to support this decision. For the development of PREDICT, patient data were obtained from randomized controlled trials to predict treatment benefit properly.[63] The correct interpretation of the model can well be supported by properly designed online calculators. Potential end-users should be taken into consideration for the interface design of the online calculator. For instance, online calculators such as PREDICT (https://breast.predict.nhs.uk) and INFLUENCE (www.evidencio.com/models/show/2238) are also accessible for patients. To ensure proper interpretation of the tools, effective risk communication strategies were investigated, which were subsequently used for the development of the designed interfaces.[64,65]

Overall, the correct interpretation of clinical prediction models starts with complete and transparently reported studies on the development and validation of a model, using the TRIPOD checklist. But even if the studies conducted are transparently reported, the scientific information can be challenging to understand for clinicians.[61] With the increasing use of black box machine learning models, the interpretation of e.g. the predictor-outcome association is more difficult than in regression models.[66]

Future perspectives

Throughout the past decades, more data have become available for more researchers. Data availability may be improved even further by adopting a different and more systematic approach to the collection, processing, and storage of patient data. Various initiatives have emerged over the years to make efficient use of patient data available in electronic health records. An example of such an initiative is the personal health train, which aims to provide a distributed learning infrastructure that enables the re-use of health data where the data remains in control of the data owner.[67,68] Success of such initiatives all heavily rely on the application of FAIR (Findable, Accessible, Interoperable, and Reusable) data principles and standard frameworks such as SMART on FHIR.[69,70] And even when best practices are applied, it remains challenging to accurately utilize unstructured data that could potentially be valuable for predictive analytics.[71,72] Also, the observational and retrospective nature of these data shall always remain to be a limitation, and clinical trials may be required to answer certain research questions (e.g. for the incorporation of treatment benefit in clinical prediction models).[73]

Wider availability of data allows researchers to validate future prediction models which consequently can be updated and/or implemented. This seems plausible since several studies have already demonstrated the prognostic value of certain variables that are not yet included in the current recommended models. For instance, the National Institute for Health and Care Excellence (NICE) performed a systematic review into the effect of lifestyle on breast cancer specific outcomes for the update of the NICE guidance. Associations were found between a healthy lifestyle and lower risk of recurrence. A healthy lifestyle was defined as achieving and maintaining healthy weight, limiting alcohol intake below five units per week, and regular physical activity. Also, evidence was found that both dietary changes and physical activity increase survival in patients with invasive breast cancer.[74] Yet, the identified lifestyle factors were not yet incorporated in the models predicting these outcomes. To reliably incorporate lifestyle factors into prediction models, patient-generated health data (PGHD) is needed.[75] For instance, physical activity can be determined with the use of wearables, such as wrist-worn activity trackers.[76]

Another example of PGHD used to support clinical decisions concerns patient-reported outcome measures (PROMs).[77] PROMs provide insight into disease-related symptomatic adverse events, physical function, and quality of life.[78] With these insights, clinical prediction models can be complemented to promote shared decision-making between patients and physicians.[79,80] For the treatment of localized prostate cancer, PROMs highlighted differences between treatment options' effect on bowel, sexual, and urinary function, and associated quality of life.[81] The differences can subsequently be used to inform patients of the risks and benefit trade-offs associated with the different treatment options. Ideally, expected effects of treatment options on both PROMS and clinical outcomes (such as survival benefit) would be tailored to the patient using prediction models. However, the development of models accurately predicting PROMs is still challenging, due to the lack of PROM collection at baseline[82] and the low response rates (e.g. 48.3%[83] and 39.3%[84]) to PROM surveys.

As more data are becoming available for researchers, the use of machine learning, a subset of artificial intelligence, is also becoming more popular.[85] The application of machine learning for clinical decision support has been a topic of debate a lot for the past years.[86] Although the use and benefits of machine learning seem valuable, caution must be exercised since some promising applications ultimately failed to reach their potential. An example of such an application concerns IBM Watson for Oncology, which initially promised to disrupt health care by developing an artificially intelligent doctor, but could not live up to its promises.[87] Current applications of artificial intelligence in clinical oncology mostly aim to provide relevant insights using imaging data.[88] Studies show conflicting results when machine learning methods are compared on performance to more traditional methods. Examples are available of improved performance in machine learning type models,[89] which was also found for two of the three outcomes in the updated INFLUENCE 2.0 model (**chapter 4)**. However, other studies showed that the use of machine learning methods did not outperform more traditional methods for the development of clinical prediction models.[90,91] In the end, it seems to be pointless to argue whether the methods used can be classified as machine learning and whether they actually perform better than more traditional statistical techniques. The ultimate goal of a model is to provide valuable insights into an outcome probability of an individual with which relevant clinical decisions can be supported that will eventually improve patient and societal outcomes.

The use of clinical prediction models and the ability to continuously evaluate and improve existing models over time rely on the availability of the models where they can benefit health care decision making. Integration in the clinician's workflow is preferred, but different health care organizations use a variety of software. More generic, centralized solutions may solve the challenges that arise from depending on Electronic Health Record (EHR) vendors. An example of a software platform that offers this solution is Evidencio (www.evidencio.com). Evidencio developed an online validation module that facilitates the performance of semi-automated external validations of prediction models hosted on Evidencio by uploading anonymous patient data.[48] Besides, Evidencio enables to integrate web-based calculators into third party software such as the EHR, allowing for a continuous exchange of patient data relevant to the implemented prediction models. By combining the features provided on Evidencio's software platform, successful implementations can eventually enable continuously automated updating of prediction models. Automated creation of updated model versions requires validation of the employed system rather than the validation of the updated model.[92]

The use of software interfaces incorporating clinical prediction models intended to support clinical decision-making has not gone unnoticed by regulatory bodies. The European Union (EU) published the medical devices regulation in 2017 that requires software intended to provide information which is used to take decisions with diagnostic or therapeutic purposes to be classified as a medical device.[93] The legislation is another barrier that has to be dealt with before prediction models can be implemented in clinical practice, but ensures that the software has proven its value when available as a certified medical device due to the requirements for clinical evaluation.[94] As the MDR replaced the previous legislation in May 2021, and one of the requirements for the clinical evaluation concerns the mandatory performance of a systematic review, it is likely to assume that many more systematic reviews on clinical decision support tools (i.e. online calculators) are going to be published. In addition, the clinical evaluation of a medical device requires evidence regarding the acceptable benefit-risk ratio based on the state of the art in medicine. Even though the use of health economic modelling as a means to perform the clinical evaluation is not made explicit in the MDR legislation nor the guidance documents, a cost-effectiveness analysis is arguably suitable to assess the benefit-risk ratio (i.e. incremental cost-effectiveness ratio) compared to the state of the art in medicine (i.e. usual care). However, manufacturers usually seem to focus on clinical evidence prior to the conformity assessment and on economic endpoints after medical devices have been placed on the market.[95] Currently, guidance on the performance of health technology assessment (HTA) for medical devices in the EU is lacking.[96] Only several member states dedicated a chapter to medical devices in official HTA guidelines, including the Dutch guideline for economic evaluations.[97] As the current Dutch guideline dates from before the MDR legislation was published, this would now require an update. Fortunately, the broadening of the guidelines on HTA in medical devices, among others, has not gone unnoticed in the Netherlands.[98] In addition, the EU adopted new regulation on HTA that is going to be applied from 2025.[99] Nonetheless, the recent implementation of the MDR legislation is likely going to cause an increase in the studies assessing the impact of clinical prediction models applied in practice in the European Union.

All these efforts together will help solve the problems of accessibility, transparency, generalizability, updating, impact assessment, and interpretation of prediction models in breast and prostate cancer, leading to improved clinical decision-making and ultimately benefiting patients.

References

1. Vickers, A. J. Prediction models in cancer care. *CA. Cancer J. Clin.* **61**, (2011).
2. Cardoso, F. *et al.* Early breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann. Oncol.* **30**, 1194–1220 (2019).
3. Ramspek, C. L., Jager, K. J., Dekker, F. W., Zoccali, C. & van Diepen, M. External validation of prognostic models: what, why, how, when and where? *Clin. Kidney J.* **14**, 49–58 (2021).
4. Borstkanker - Risicoprofilering - Richtlijn - Richtlijnendatabase. Available at: https://richtlijnendatabase.nl/richtlijn/borstkanker/risicoprofilering.html. (Accessed: 18th February 2022)
5. Glinsky, G. V., Glinskii, A. B., Stephenson, A. J., Hoffman, R. M. & Gerald, W. L. Gene expression profiling predicts clinical outcome of prostate cancer. *J. Clin. Invest.* **113**, 913–923 (2004).
6. Paik, S. *et al.* A Multigene Assay to Predict Recurrence of Tamoxifen-Treated, Node-Negative Breast Cancer. *N. Engl. J. Med.* **351**, 2817–2826 (2004).
7. Briganti, A. *et al.* Updated Nomogram Predicting Lymph Node Invasion in Patients with Prostate Cancer Undergoing Extended Pelvic Lymph Node Dissection: The Essential Importance of Percentage of Positive Cores. *Eur. Urol.* **61**, 480–487 (2012).
8. Roach, M. *et al.* Predicting the risk of lymph node involvement using the pre-treatment prostate specific antigen and gleason score in men with clinically localized prostate cancer. *Int. J. Radiat. Oncol. Biol. Phys.* **28**, 33–37 (1994).
9. Tosoian, J. J. *et al.* Prediction of pathological stage based on clinical stage, serum prostate-specific antigen, and biopsy Gleason score: Partin Tables in the contemporary era. *BJU Int.* **119**, 676–683 (2017).
10. Prostate Cancer Nomograms | Memorial Sloan Kettering Cancer Center. Available at: https://www.mskcc.org/nomograms/prostate. (Accessed: 18th February 2022)
11. Gandaglia, G. *et al.* A Novel Nomogram to Identify Candidates for Extended Pelvic Lymph Node Dissection Among Patients with Clinically Localized Prostate Cancer Diagnosed with Magnetic Resonance Imaging-targeted and Systematic Biopsies. *Eur. Urol.* **75**, 506–514 (2019).
12. Mottet, N. *et al.* EAU-ESTRO-SIOG Guidelines on Prostate Cancer. Part 1: Screening, Diagnosis, and Local Treatment with Curative Intent. *Eur. Urol.* **71**, 618–629 (2017).
13. Iasonos, A., Schrag, D., Raj, G. V. & Panageas, K. S. How to build and interpret a nomogram for cancer prognosis. *J. Clin. Oncol.* **26**, 1364–1370 (2008).
14. Kappen, T. H. *et al.* Evaluating the impact of prediction models: lessons learned, challenges, and recommendations. *Diagnostic Progn. Res.* **2**, 11 (2018).
15. Ravdin, P. M. *et al.* Computer Program to Assist in Making Decisions About Adjuvant Therapy for Women With Early Breast Cancer. *J. Clin. Oncol.* **19**, 980–991 (2001).
16. Michaelson, J. S. *et al.* Improved web-based calculators for predicting breast carcinoma outcomes. *Breast Cancer Res. Treat.* **128**, 827–835 (2011).
17. Hoveling, L. A. *et al.* Validation of the online prediction model CancerMath in the Dutch breast cancer population. *Breast Cancer Res. Treat.* **178**, 665–681 (2019).

18. Bouwmeester, W. *et al.* Reporting and Methods in Clinical Prediction Research: A Systematic Review. *PLoS Med.* **9**, e1001221 (2012).
19. Mallett, S., Royston, P., Dutton, S., Waters, R. & Altman, D. G. Reporting methods in studies developing prognostic models in cancer: a review. *BMC Med.* **8**, 20 (2010).
20. Steyerberg, E. Clinical prediction models. (2019).
21. Riley, R., Windt, D. van der, Croft, P. & Moons, K. Prognosis research in healthcare: concepts, methods, and impact. (2019).
22. Steyerberg, E. W. & Vergouwe, Y. Towards better clinical prediction models: Seven steps for development and an ABCD for validation. *Eur. Heart J.* **35**, 1925–1931 (2014).
23. Steyerberg, E. W. *et al.* Prognosis Research Strategy (PROGRESS) 3: Prognostic Model Research. *PLoS Med.* **10**, (2013).
24. Venema, E. *et al.* Large-scale validation of the prediction model risk of bias assessment Tool (PROBAST) using a short form: high risk of bias models show poorer discrimination. *J. Clin. Epidemiol.* **138**, 32–39 (2021).
25. Yang, C., Kors, J., Ioannou, S. & … L. J.-J. of the. Trends in the conduct and reporting of clinical prediction model development and validation: a systematic review. *academic.oup.com*
26. de Hond, A. A. H. *et al.* Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. *npj Digital Medicine* **5**, (2022).
27. Van Calster, B., Wynants, L., Timmerman, D., Steyerberg, E. W. & Collins, G. S. Predictive analytics in health care: how can we know it works? *J. Am. Med. Informatics Assoc.* **26**, 1651–1654 (2019).
28. Collins, G. S. *et al.* Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open* **11**, (2021).
29. Altman, D. G., Vergouwe, Y., Royston, P. & Moons, K. G. M. Prognosis and prognostic research: Validating a prognostic model. *BMJ* **338**, 1432–1435 (2009).
30. Justice, A. C., Covinsky, K. E. & Berlin, J. A. Assessing the generalizability of prognostic information. *Ann. Intern. Med.* **130**, 515–524 (1999).
31. Altman, D. G. & Royston, P. What do we mean by validating a prognostic model? *Stat. Med.* **19**, 453–473 (2000).
32. Siontis, G. C. M., Tzoulaki, I., Castaldi, P. J. & Ioannidis, J. P. A. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. *J. Clin. Epidemiol.* **68**, 25–34 (2015).
33. Phung, M. T., Tin Tin, S. & Elwood, J. M. *Prognostic models for breast cancer: A systematic review*. *BMC Cancer* **19**, (BioMed Central Ltd., 2019).
34. Strijker, M. *et al.* Systematic review of clinical prediction models for survival after surgery for resectable pancreatic cancer. *Br. J. Surg.* **106**, 342–354 (2019).
35. Collins, G. S. *et al.* External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. *BMC Med. Res. Methodol.* **14**, 40 (2014).

36. Van Den Boorn, H. G. *et al.* Prediction models for patients with esophageal or gastric cancer: A systematic review and meta-analysis. *PLoS One* **13**, (2018).

37. Riley, R. D. *et al.* Minimum sample size for external validation of a clinical prediction model with a binary outcome. *Stat. Med.* **40**, 4230–4251 (2021).

38. Moons, K. G. M. *et al.* Risk prediction models: II. External validation, model updating, and impact assessment. *heart.bmj.com* doi:10.1136/heartjnl-2011-301247

39. Hoogland, J. *et al.* Handling missing predictor values when validating and applying a prediction model to new patients. *Wiley Online Libr.* **39**, 3591–3607 (2020).

40. O'Keefe, E. B., Meltzer, J. P. & Bethea, T. N. Health Disparities and Cancer: Racial Disparities in Cancer Mortality in the United States, 2000–2010. *Front. Public Heal.* **3**, (2015).

41. Ward, E. *et al.* Cancer Disparities by Race/Ethnicity and Socioeconomic Status. *CA. Cancer J. Clin.* **54**, 78–93 (2004).

42. Steyerberg, E. W. *et al.* Assessing the performance of prediction models: A framework for traditional and novel measures. *Epidemiology* **21**, 128–138 (2010).

43. Van Calster, B. *et al.* Calibration: The Achilles heel of predictive analytics. *BMC Med.* **17**, (2019).

44. Vickers, A. J., van Calster, B. & Steyerberg, E. W. A simple, step-by-step guide to interpreting decision curve analysis. *Diagnostic Progn. Res.* **3**, (2019).

45. Royston, P. & Altman, D. G. Visualizing and assessing discrimination in the logistic regression model. *Stat. Med.* **29**, 2508–2520 (2010).

46. van Maaren, M. C. C. *et al.* Validation of the online prediction tool PREDICT v. 2.0 in the Dutch breast cancer population. *Eur. J. Cancer* **86**, 364–372 (2017).

47. Engelhardt, E. G. *et al.* Accuracy of the online prognostication tools PREDICT and Adjuvant! for early-stage breast cancer patients younger than 50 years. *Eur. J. Cancer* **78**, 37–44 (2017).

48. van Steenbeek, C. D. *et al.* Facilitating validation of prediction models: a comparison of manual and semi-automated validation using registry-based data of breast cancer patients in the Netherlands. *BMC Med. Res. Methodol.* **19**, 117 (2019).

49. Vergouwe, Y. *et al.* A closed testing procedure to select an appropriate method for updating prediction models. *Stat. Med.* **36**, 4529–4539 (2017).

50. Voelkel, V. *et al.* Predicting the risk of locoregional recurrence after early breast cancer: an external validation of the Dutch INFLUENCE-nomogram with clinical cancer registry data from Germany. *J. Cancer Res. Clin. Oncol.* (2019). doi:10.1007/S00432-019-02904-4

51. Janssen, K. J. M., Vergouwe, Y., Kalkman, C. J., Grobbee, D. E. & Moons, K. G. M. A simple method to adjust clinical prediction models to local circumstances. *Can. J. Anesth.* **56**, 194–201 (2009).

52. Vickers, A. J., Cronin, A. M., Elkin, E. B. & Gonen, M. Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers. *BMC Med. Inform. Decis. Mak.* **8**, (2008).

53. Vickers, A. J. & Cronin, A. M. Everything you always wanted to know about evaluating prediction models (but were too afraid to ask). *Urology* **76**, 1298–1301 (2010).

54. Jenniskens, K. *et al.* Decision analytic modeling was useful to assess the impact of a prediction model on health outcomes before a randomized trial. *J. Clin. Epidemiol.* **115**, 106–115 (2019).
55. van Giessen, A. *et al.* Systematic Review of Health Economic Impact Evaluations of Risk Prediction Models: Stop Developing, Start Evaluating. *Value Heal.* **20**, 718–726 (2017).
56. Shmueli, G. To explain or to predict? *Stat. Sci.* **25**, 289–310 (2010).
57. Ramspek, C. L. *et al.* Prediction or causality? A scoping review of their conflation within current observational research. *Eur. J. Epidemiol.* **36**, 889–898 (2021).
58. Salas, M., Hofman, A. & Stricker, B. H. C. Confounding by Indication: An Example of Variation in the Use of Epidemiologic Terminology. *Am. J. Epidemiol.* **149**, 981–983 (1999).
59. Bosco, J. L. F. *et al.* A most stubborn bias: no adjustment method fully resolves confounding by indication in observational studies. *J. Clin. Epidemiol.* **63**, 64–74 (2010).
60. van Geloven, N. *et al.* Prediction meets causal inference: the role of treatment in clinical prediction models. *Eur. J. Epidemiol.* **35**, 619–630 (2020).
61. Kappen, T. H. & Peelen, L. M. Prediction models: The right tool for the right problem. *Curr. Opin. Anaesthesiol.* **29**, 717–726 (2016).
62. Groenwold, R. H. H. *et al.* Explicit inclusion of treatment in prognostic modeling was recommended in observational and randomized settings. *J. Clin. Epidemiol.* **78**, 90–100 (2016).
63. Wishart, G. C. *et al.* PREDICT: a new UK prognostic model that predicts survival following surgery for invasive breast cancer. *Breast Cancer Res.* **12**, R1 (2010).
64. Farmer, G. D. *et al.* Redevelopment of the Predict: Breast Cancer website and recommendations for developing interfaces to support decision-making. *Wiley Online Libr.* **10**, 5141–5153 (2021).
65. Kramer, I., Cardozo, J. & … D. G.-D. Preferences for graphical presentation of probabilities in a contralateral breast cancer risk prediction model: an exploratory interview study among breast cancer. *scholarlypublications …*
66. Stiglic, G. *et al.* Interpretability of machine learning-based prediction models in healthcare. *zitniklab.hms.harvard.edu* **10**, (2020).
67. Beyan, O. *et al.* Distributed analytics on sensitive medical data: The personal health train. *Data Intell.* **2**, 96–107 (2020).
68. Geleijnse, G. *et al.* Prognostic factors analysis for oral cavity cancer survival in the Netherlands and Taiwan using a privacy-preserving federated infrastructure. *Sci. Rep.* **10**, (2020).
69. Mandel, J. C., Kreda, D. A., Mandl, K. D., Kohane, I. S. & Ramoni, R. B. SMART on FHIR: A standards-based, interoperable apps platform for electronic health records. *J. Am. Med. Informatics Assoc.* **23**, 899–908 (2016).
70. Choudhury, A., van Soest, J., Nayak, S. & Dekker, A. Personal Health Train on FHIR: A Privacy Preserving Federated Approach for Analyzing FAIR Data in Healthcare. *Commun. Comput. Inf. Sci.* **1240 CCIS**, 85–95 (2020).
71. Sun, W. *et al.* Data processing and text mining technologies on electronic medical records: A review. *J. Healthc. Eng.* **2018**, (2018).
72. Tayefi, M. *et al.* Challenges and opportunities beyond structured data in analysis of electronic health records. *Wiley Online Libr.* **13**, (2021).

73. Kent, D. M. *et al.* The Predictive Approaches to Treatment effect Heterogeneity (PATH) statement. *Ann. Intern. Med.* **172**, 35–45 (2020).

74. Recommendations | Early and locally advanced breast cancer: diagnosis and management | Guidance | NICE. Available at: https://www.nice.org.uk/guidance/ng101/chapter/Recommendations#lifestyle. (Accessed: 18th February 2022)

75. Jim, H. S. L. *et al.* Innovations in research and clinical care using patient-generated health data. *CA. Cancer J. Clin.* **70**, 182–199 (2020).

76. Breteler, M. J. M. *et al.* Measuring free-living physical activity with three commercially available activity monitors for telemonitoring purposes: Validation study. *JMIR Form. Res.* **3**, (2019).

77. Greenhalgh, J. *et al.* How do patient reported outcome measures (PROMs) support clinician-patient communication and patient care? a realist synthesis. *J. Patient-Reported Outcomes* **2**, (2018).

78. Kluetz, P. G. *et al.* Focusing on core patient-reported outcomes in cancer clinical trials: Symptomatic adverse events, physical function, and disease-related symptoms. *Clin. Cancer Res.* **22**, 1553–1558 (2016).

79. Damman, O. C. *et al.* The use of PROMs and shared decision-making in medical encounters with patients: An opportunity to deliver value-based health care to patients. *J. Eval. Clin. Pract.* **26**, 524–540 (2020).

80. de Ligt, K. M., van Egdom, L. S. E., Koppert, L. B., Siesling, S. & van Til, J. A. Opportunities for personalised follow-up care among patients with breast cancer: A scoping review to identify preference-sensitive decisions. *Eur. J. Cancer Care (Engl).* **28**, (2019).

81. Donovan, J. L. *et al.* Patient-Reported Outcomes after Monitoring, Surgery, or Radiotherapy for Prostate Cancer. *N. Engl. J. Med.* **375**, 1425–1437 (2016).

82. van Egdom, L. S. E., Pusic, A., Verhoef, C., Hazelzet, J. A. & Koppert, L. B. Machine learning with PROs in breast cancer surgery; caution: Collecting PROs at baseline is crucial. *Breast J.* **26**, 1213–1215 (2020).

83. Rose, M. *et al.* Patient-reported outcome after oncoplastic breast surgery compared with conventional breast-conserving surgery in breast cancer. *Breast Cancer Res. Treat.* **180**, 247–256 (2020).

84. Kowalski, C. *et al.* A multicenter paper-based and web-based system for collecting patient-reported outcome measures in patients undergoing local treatment for prostate cancer: first experiences. *J. Patient-Reported Outcomes* **4**, 1–7 (2020).

85. Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H. & Aerts, H. J. W. L. Artificial intelligence in radiology. *Nat. Rev. Cancer* **18**, 500–510 (2018).

86. Nagendran, M. *et al.* Artificial intelligence versus clinicians: Systematic review of design, reporting standards, and claims of deep learning studies in medical imaging. *BMJ* **368**, (2020).

87. Strickland, E. IBM Watson, heal thyself: How IBM overpromised and underdelivered on AI health care. *IEEE Spectr.* **56**, 24–31 (2019).

88. Kann, B. H., Hosny, A. & Aerts, H. J. W. L. Artificial intelligence for clinical oncology. *Cancer Cell* **39**, 916–927 (2021).

89. Moncada-Torres, A., van Maaren, M. C., Hendriks, M. P., Siesling, S. & Geleijnse, G. Explainable machine learning can outperform Cox regression predictions and provide insights in breast cancer survival. *Sci. Rep.* **11**, (2021).

90. Christodoulou, E. *et al.* A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J. Clin. Epidemiol.* **110**, 12–22 (2019).

91. Witteveen, A., Nane, G. F., Vliegen, I. M. H., Siesling, S. & IJzerman, M. J. Comparison of Logistic Regression and Bayesian Networks for Risk Prediction of Breast Cancer Recurrence. *Med. Decis. Mak.* **38**, 822–833 (2018).

92. Jenkins, D. A. *et al.* Continual updating and monitoring of clinical prediction models: time for dynamic prediction systems? *Diagnostic Progn. Res.* **5**, (2021).

93. Regulation (EU) 2017/745 on medical devices. (2017). Available at: https://eur-lex.europa.eu/eli/reg/2017/745/2017-05-05. (Accessed: 8th February 2021)

94. Group, M. D. C. *MDCG 2020-1 Guidance on Clinical Evaluation (MDR) / Performance Evaluation (IVDR) of Medical Device Software*.

95. Blankart, C. R. *et al.* Regulatory and HTA early dialogues in medical devices. *Health Policy (New. York).* **125**, 1322–1329 (2021).

96. Blüher, M. *et al.* Critical Review of European Health-Economic Guidelines for the Health Technology Assessment of Medical Devices. *Front. Med.* **6**, (2019).

97. Zwaap, J., Knies, S., van der Meijden, C., Staal, P. & van der Heiden, L. Cost-effectiveness in practice. *National Healthcare Institute.* (2015). Available at: https://english.zorginstituutnederland.nl/publications/reports/2015/06/16/cost-effectiveness-in-practice.

98. Enzing, J. J., Knies, S., Boer, B. & Brouwer, W. B. F. Broadening the application of health technology assessment in the Netherlands: A worthwhile destination but not an easy ride? *Heal. Econ. Policy Law* **16**, 440–456 (2021).

99. Regulation (EU) 2021/2282 on health technology assessment, EUR-Lex - 32021R2282 - EN - EUR-Lex. Available at: https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:32021R2282. (Accessed: 18th February 2022)

Dankwoord, Curriculum Vitae, List of publications

Dankwoord

Tot slot, misschien wel het belangrijkste stuk van het hele proefschrift, het dankwoord. Het schijnt het meest gelezen hoofdstuk in een proefschrift te zijn. Misschien wel omdat vrienden en familie vaak een boekje krijgen en benieuwd zijn of zij nog benoemd worden in het dankwoord, want de rest van het proefschrift is realistisch gezien niet voor iedereen even interessant. Daarom begin alvast ik met een bedankje aan jou. Ik weet natuurlijk niet wie je bent of waarom je dit leest, maar hoe dan ook bedankt dat je de moeite neemt om mijn proefschrift te bekijken! Nu kan in ieder geval niemand mij vertellen dat ik ze ben vergeten te vermelden in het dankwoord. Toch zijn er een aantal mensen die ik in het bijzonder wil bedanken. Niet omdat dit nou eenmaal hoort, maar omdat ze het zo ontzettend verdienen, want zonder jullie had dit proefschrift er niet gelegen.

Als eerst wil ik natuurlijk mijn fantastische promotieteam bedanken, Prof. Dr. Sabine Siesling, Prof. Dr. Ir. Erik Koffijberg en dr. Marissa van Maaren. Mijn dank aan jullie is oneindig groot voor het vertrouwen, jullie support en de waardevolle, maar ook altijd gezellige bijeenkomsten.

Ik voel me ontzettend bevoorrecht om jou als promotor te hebben gehad, Sabine. Ik heb enorm veel respect en bewondering voor de manier waarop jij de grote hoeveelheid werk die je verzet weet te combineren met zoveel humor en daarbij ook nog oog hebt voor de persoonlijke omstandigheden van betrokkenen. Je hebt me altijd scherp weten te houden op het einddoel. En als het even iets anders liep dan gepland, heb je mij vooral geleerd om te kijken naar de mogelijkheden en deze dan ook te pakken. Als ik terugdenk aan onze samenwerking, herinner ik me vooral alle leuke momenten buiten het werken aan het proefschrift om. Eén van deze momenten was een letterlijk hoogtepunt, hiken door de bergen van Innsbruck na een symposium in Hall in Tirol, waar de voortzetting van mijn promotie in gevaar zou zijn gekomen bij een kleine misstap over de sneeuw, maar bedankt dat je heelhuids beneden bent gekomen! Ik hoop dan ook de komende jaren nog veel met je te mogen blijven samenwerken. Zonder jouw inzet, vertrouwen en visie was dit proefschrift er niet geweest. Heel erg bedankt Sabine!

Bij jou is het eigenlijk allemaal begonnen, Erik. Eind 2016 ging op zoek naar een opdracht voor mijn masterscriptie. Ik weet nog dat ik altijd onder de indruk was van de kennis en kunde die je over wist te dragen tijdens de colleges. Ik besloot te reageren op een opdracht over de kosteneffectiviteitsanalyse van een nieuwe technologie genaamd 'DiagnOSAS'. Een medestudent, Floris, had ook interesse in deze opdracht. Jij bracht Floris en mij in contact met Rick en Rob van DiagnOSAS. Rick en Rob vertelden over DiagnOSAS en wat ze daarnaast deden met Evidencio. Na enkele gesprekken konden Floris en ik beide aan de slag met onze scripties. Floris bij DiagnOSAS en ik bij Evidencio. Niet alleen tijdens het schrijven van de masterscriptie, maar gedurende de afgelopen jaren heb ik ontzettend veel van jou geleerd over het uitvoeren van onderzoek en het schrijven van de bijbehorende artikelen. Daarnaast heb je me laten zien hoe je op een zeer correcte en integere manier om kan gaan met verschillende belangen die er kunnen spelen rondom de uitvoer en de resultaten van de onderzoeken. Ik vind het dan ook geweldig om jou sinds kort als professor aan te mogen spreken, dit heb je zo dik verdiend! Ik hoop ook met jou in de toekomst nog veel samen te mogen blijven werken. Heel erg bedankt voor de altijd fijne samenwerking Erik!

Als laatste kwam jij bij het promotieteam, Marissa. Ik vond het bijzonder fijn om regelmatig samen met jou van gedachte te kunnen wisselen over allerlei zaken rondom de promotie. Of het nou inhoudelijk ging over gebruik van de beste methoden, het promotieproces en de bijbehorende planning, of over de invulling van het feest na de promotieverdediging, je kon me altijd van goed advies voorzien. Inhoudelijk ben je altijd erg sterk, het kwam daarmee natuurlijk erg goed uit dat je vanuit je eigen promotieonderzoek al brede ervaring had met de ontwikkeling en validatie van predictiemodellen bij borstkankerpatiënten en je had ook al eerder gewerkt met de software van Evidencio. Ik heb onze samenwerking daarom ook altijd als erg prettig en heel waardevol ervaren en we zullen ongetwijfeld veel samen blijven werken. Heel erg bedankt Marissa!

Alle leden van de beoordeling- en promotiecommissie, hartelijk dank voor het lezen en beoordelen van het proefschrift. Ik kijk erg uit naar de vragen die jullie hebben tijdens de verdediging.

Ook zou ik graag Dr. Ir. Rob Mentink in het bijzonder willen bedanken. Als directeur van Evidencio ben je zeker veel betrokken geweest bij mijn promotieproces. Ik ben je dan ook uitzonderlijk dankbaar voor alle kansen die je me hebt gegeven, het vertrouwen dat je me blijft geven in mijn werk en voor alle mooie momenten die we samen hebben beleefd de afgelopen jaren. Ook heb ik altijd veel gehad aan de motiverende werking van je nuchterheid, visie en enthousiasme waar je mee op kantoor komt. Dank je wel Rob!

Onmisbaar in mij dankwoord is dr. Rick Pleijhuis. Het is uitermate inspirerend om te zien hoe jij als arts en ondernemer al je verschillende werkzaamheden weet te combineren en tegelijkertijd zo nuchter, toegankelijk en realistisch blijft. Ik heb dan ook vaak dankbaar gebruik gemaakt van je enorme hoeveelheid aan parate kennis die je hebt over van alles en nog wat. Heel erg bedankt Rick!

Alle collega's bij ons op kantoor in Haaksbergen van Evidencio, DiagnOSAS en Orange-M-Health, bedankt voor het creëren van de altijd fijne en gezellige werksfeer. Na een tijd van thuiswerken en afstand houden kunnen we hopelijk weer steeds meer genieten van de borrels op de vrijdag en bijbehorende feestjes.

Verder vind ik het belangrijk om te benoemen dat dit proefschrift alleen maar tot stand heeft kunnen komen door de brede beschikbaarheid van patiëntdata. Hiervoor spreek ik dan ook graag mij dank uit aan eenieder die betrokken is geweest bij het verzamelen van deze data en de data vervolgens beschikbaar hebben gemaakt voor onderzoek.

Ook spreek ik graag mijn dank uit aan alle co-auteurs die betrokken zijn geweest bij de verschillende hoofdstukken in het proefschrift. Een aantal co-auteurs wil ik in het bijzonder bedanken: Dr. Jean-Paul van Basten, Dr. Rik Somford, Dr. Erik Cornel, Drs. Ruben Korthorst, Drs. Mathijs Hendriks, Dr. Vinzenz Völkel en Dr. Timo Soeterik. Jean-Paul, Rik, Erik en Ruben, dank voor jullie uitzonderlijk waardevolle input en feedback op de prostaatkankerstukken waar ik destijds als masterstudent mee aan de slag ben gegaan, mede dankzij jullie toewijding ligt dit proefschrift er nu. Mathijs, dank voor de prettige samenwerking en jouw waardevolle klinische inzichten op de predictiemodellen voor borstkanker. Vinzenz, thank you for our fruitful collaboration and your trust in me to finalize work for the INFLUENCE model during the often hectic times of the pandemic, danke schön! Timo, dank voor de altijd gezellige en waardevolle meetings die we hebben gehad, jouw skills over zowel prostaatkanker als predictiemodellen zorgden voor een vlot verloop van de studie.

Floris, we hebben veel samen opgetrokken bij de uitvoer van onze kosteneffectiviteit studies op kantoor in Haaksbergen. Het was altijd prettig om van gedachte te kunnen wisselen met jou over onze vergelijkbare onderzoeken, dank je wel!

Paranimfen Stein en Erik, bedankt dat jullie deze bijzondere dag samen met mij van dichtbij mee willen gaan maken.

Vrienden en familie, heel erg bedankt voor alle support, betrokkenheid, maar ook zeker voor alle leuke en mooie momenten die we samen meegemaakt hebben. Bram, Stein, Joris, Justin, Sander, Niels, Wessel en Frank, mooi dat we na al die jaren nog altijd een hechte vriendengroep zijn. Ik ben dan ook blij en dankbaar om deze mijlpaal weer samen met jullie te kunnen vieren!

Wouter, Anne, Lotte en Maurice, mijn broer en zusje en aanhangers, heel fijn dat jullie er altijd zijn om samen wat leuks te doen. Ik ben heel erg dankbaar voor de goede band die we met z’n allen hebben!

Papa en mama, dank jullie wel dat jullie er altijd voor mij zijn. Ik kan en heb altijd op jullie kunnen bouwen en voel me altijd erg door jullie gesteund. Jullie hebben nooit veel van mij gevraagd, mama is al trots als ik alleen nog maar mijn veters kan strikken en papa heeft altijd gezegd om iets te gaan doen wat ik leuk vind en mij energie geeft. Toch waren jullie al zo trots toen ik een paar jaar geleden aangaf dat ik ging promoveren. Jullie zijn al die tijd ook heel betrokken geweest. Nu is het dan eindelijk zo ver dat het proefschrift er ligt en ben ik zo blij en trots dat ik dit verheugde moment samen met jullie mag vieren!

Sylvia, Fred, Luuk, Charlotte, Sepp en Niene, mijn schoonfamilie. Jullie zijn eigenlijk een bonusfamilie die ik er bij Maartje bij heb gekregen. Jullie hebben altijd voor de broodnodige afleiding gezorgd met de vele gezellige momentjes, lekkere etentjes en leuke uitjes. Ook stonden jullie altijd klaar om te helpen en zorgden jullie ervoor dat ik tijd en ruimte had om o.a. aan het proefschrift te werken. Dank jullie wel dat jullie er zijn! In het bijzonder wil ik graag Sepp bedanken voor jouw inhoudelijke bijdrage. Ze zeggen dat wijsheid met de jaren komt, maar de briljante inzichten die jij nu al deelt, hebben een onuitwisbare indruk op mij achtergelaten.

Tot slot, het laatste en belangrijkste dankwoord is natuurlijk voor mijn vrouw, Maartje! Zonder jou zou ik nu niet staan waar ik sta en zou ik niet zijn wie ik ben. Jij hebt mij na de afronding van mijn studie fysiotherapie geïnspireerd om eens verder te kijken naar de mogelijkheden om een master te doen op de universiteit. Dit bleek uiteindelijk een schot in de roos te zijn. Je hebt me al die jaren ontzettend goed gesteund en gaf mij alle vrijheid, ruimte en tijd die ik nodig had. Ik ben zo enorm blij en dankbaar dat je in mijn leven bent. Het gelukkigst ben ik als ik jou en onze dochter Mette samen zie lachen. Ik kijk enorm uit naar alle mooie herinneringen die we samen nog gaan maken!

Curriculum Vitae

Tom Hueting was born on June 6th 1992 in Hengelo, Overijssel, the Netherlands. Together with his brother and sister he grew up in Haaksbergen. After graduating from the Assink Lyceum in Haaksbergen in 2010, he started with the study physical therapy at the Saxion university of applied sciences in Enschede. In 2015, he received his bachelor degree and started working as a physical therapist at a nursing home in Groenlo and a practice for physical therapy in Haaksbergen. He remained to work as a physical therapist in the private practice at the same time as performing the master Health Sciences at the University of Twente in Enschede. In 2017, he graduated cum laude, after which he started to work at two start-up companies, Evidencio and DiagnOSAS, where he also performed his master's thesis. Since 2019 he officially started as a PhD candidate at the department of Health Technology and Services Research at the University of Twente. His PhD project concerned the development, validation, and evaluation of clinical prediction models in breast and prostate cancer. The results of his project are described in this thesis. After completing his PhD, he remains employed at Evidencio and DiagnOSAS in Haaksbergen. Tom is happily married to Maartje, and is a proud father of their daughter Mette.

List of publications

Published

1. Hueting TA, Cornel EB, Somford DM, Jansen H, van Basten JA, Pleijhuis RG, Korthorst RA, van der Palen JAM, Koffijberg H. External Validation of Models Predicting the Probability of Lymph Node Involvement in Prostate Cancer Patients. Eur Urol Oncol. 2018 Oct;1(5):411-417. doi: 10.1016/j.euo.2018.04.016. Epub 2018 Jun 28. PMID: 31158080.
2. Hoveling LA, van Maaren MC, Hueting T, Strobbe LJA, Hendriks MP, Sonke GS, Siesling S. Validation of the online prediction model CancerMath in the Dutch breast cancer population. Breast Cancer Res Treat. 2019 Dec;178(3):665-681. doi: 10.1007/s10549-019-05399-2. Epub 2019 Aug 30. PMID: 31471837.
3. Voelkel V, Draeger T, Groothuis-Oudshoorn CGM, de Munck L, Hueting T, Gerken M, Klinkhammer-Schalke M, Lavric M, Siesling S. Predicting the risk of locoregional recurrence after early breast cancer: an external validation of the Dutch INFLUENCE-nomogram with clinical cancer registry data from Germany. J Cancer Res Clin Oncol. 2019 Jul;145(7):1823-1833. doi: 10.1007/s00432-019-02904-4. Epub 2019 Mar 29. PMID: 30927074; PMCID: PMC6571079.
4. Hueting TA, Cornel EB, Korthorst RA, Pleijhuis RG, Somford DM, van Basten JA, van der Palen JAM, Koffijberg H. Optimizing the risk threshold of lymph node involvement for performing extended pelvic lymph node dissection in prostate cancer patients: a cost-effectiveness analysis. Urol Oncol. 2021 Jan;39(1):72.e7-72.e14. doi: 10.1016/j.urolonc.2020.09.014. Epub 2020 Oct 26. PMID: 33121913.
5. Soeterik TFW, Hueting TA, Israel B, van Melick HHE, Dijksman LM, Stomps S, Biesma DH, Koffijberg H, Sedelaar M, Witjes JA, van Basten JA. External validation of the Memorial Sloan Kettering Cancer Centre and Briganti nomograms for the prediction of lymph node involvement of prostate cancer using clinical stage assessed by magnetic resonance imaging. BJU Int. 2021 Aug;128(2):236-243. doi: 10.1111/bju.15376. Epub 2021 May 16. PMID: 33630398.
6. Völkel V, Hueting TA, Draeger T, van Maaren MC, de Munck L, Strobbe LJA, Sonke GS, Schmidt MK, van Hezewijk M, Groothuis-Oudshoorn CGM, Siesling S. Improved risk estimation of locoregional recurrence, secondary contralateral tumors and distant metastases in early breast cancer: the INFLUENCE 2.0 model. Breast Cancer Res Treat. 2021 Oct;189(3):817-826. doi: 10.1007/s10549-021-06335-z. Epub 2021 Aug 2. PMID: 34338943; PMCID: PMC8505302.

Submitted for Peer Review:

1. Hueting TA, van Maaren MC, Hendriks MP, Koffijberg H, Siesling S. Clinical prediction models to support treatment decisions in breast cancer patients: a systematic review.
2. Hueting TA, van Maaren MC, Hendriks MP, Koffijberg H, Siesling S. External validation of 87 clinical prediction models supporting clinical treatment decisions for breast cancer patients.